MEMORIES FOR THE

INTELLIGENT INTERNET OF THINGS

BETTY PRINCE | DAVID PRINCE

WILEY

Memories for the Intelligent Internet of Things

Memories for the Intelligent Internet of Things

Betty Prince and David Prince
Memory Strategies International
Texas, USA

WILEY

This edition first published 2018
© 2018 John Wiley & Sons Ltd

All rights reserved. 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 or otherwise, except as permitted by law. Advice on how to obtain permission to reuse material from this title is available at http://www.wiley.com/go/permissions.

The right of Dr. Betty Prince and Mr. David Prince to be identified as the authors of this work has been asserted in accordance with law.

Registered Office(s)
John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, USA
John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK

Editorial Office
The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK

For details of our global editorial offices, customer services, and more information about Wiley products visit us at www.wiley.com.

Wiley also publishes its books in a variety of electronic formats and by print-on-demand. Some content that appears in standard print versions of this book may not be available in other formats.

Limit of Liability/Disclaimer of Warranty
While the publisher and authors have used their best efforts in preparing this work, they make no representations or warranties with respect to the accuracy or completeness of the contents of this work and specifically disclaim all warranties, including without limitation any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives, written sales materials or promotional statements for this work. The fact that an organization, website, or product is referred to in this work as a citation and/or potential source of further information does not mean that the publisher and authors endorse the information or services the organization, website, or product may provide or recommendations it may make. This work is sold with the understanding that the publisher is not engaged in rendering professional services. The advice and strategies contained herein may not be suitable for your situation. You should consult with a specialist where appropriate. Further, readers should be aware that websites listed in this work may have changed or disappeared between when this work was written and when it is read. Neither the publisher nor authors shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

Library of Congress Cataloging-in-Publication Data

Names: Prince, Betty, author. | Prince, David, 1966– author.
Title: Memories for the intelligent internet of things / by Dr. Betty Prince, David Prince.
Description: Hoboken, NJ : John Wiley & Sons, 2018. | Includes bibliographical references and index. | Descr
Identifiers: LCCN 2018001009 (print) | LCCN 2018012072 (ebook) | ISBN 9781119296409 (pdf) | ISBN 9781119298953 (epub) | ISBN 9781119296355 (cloth)
Subjects: LCSH: Computer storage devices. | Internet of things.
Classification: LCC TK7895.M4 (ebook) | LCC TK7895.M4 P735 2018 (print) | DDC 004.67/8–dc23
LC record available at https://lccn.loc.gov/2018001009

Cover Design: Wiley
Cover Image: Images courtesy of Betty and David Prince
background: © flowerstock/Shutterstock

Set in 10/12pt Warnock by SPi Global, Pondicherry, India

10 9 8 7 6 5 4 3 2 1

Contents

Introduction to the Intelligent Internet of Things *xi*

1 Smart Cities as the Prototype of the Intelligent Internet of Things *1*
1.1 Overview *1*
1.2 Smart Cities *1*
1.3 Smart Commerce as an Element of the Smart City *1*
1.3.1 Smart Inventory Control *1*
1.3.2 Smart Delivery *3*
1.3.3 Smart Marketing Using Artificial Intelligence *3*
1.4 Smart Residences *4*
1.4.1 A City of Smart Connected Homes *4*
1.5 People as Center of Smart Connected Homes *5*
1.5.1 Wearable Electronics *5*
1.5.2 Control Electronics *6*
1.6 Smart Individual Transportation *6*
1.6.1 Overview of Smart Automobiles *6*
1.6.2 Driving Aids *7*
1.6.3 Engine Processors *8*
1.6.4 Auto Body Processors *8*
1.6.5 Infotainment Processors *8*
1.6.6 Autonomous Cars *8*
1.7 Smart Transportation Networks *9*
1.7.1 Smart Public Conveyance Networks *9*
1.7.2 Individual Automotive Traffic Control *9*
1.7.3 Smart Highways *10*
1.8 Smart Energy Networks *10*
1.8.1 Smart Electrical Meters *10*
1.8.2 Smart Electrical Grids *12*
1.9 Smart Connected Buildings *12*
1.9.1 Smart Office Buildings *12*
1.9.2 Smart Factories *13*
1.9.3 Intelligent Hospitals *13*
1.9.4 Smart Public Buildings *14*
1.10 Thoughts *15*
References *15*

2 Memory Applications for the Intelligent Internet of Things *17*
2.1 Introduction *17*
2.2 Comparisons of the Various Nonvolatile Embedded Memories Characteristics *18*
2.2.1 Embedded EEPROM, Flash, and Fuse Devices *18*
2.2.2 Embedded Emerging Memory Devices in MCU *19*
2.2.3 Required Properties of Embedded Nonvolatile Memories in Various Applications *21*
2.3 Circuits Using Ultralow Power MCU with Embedded Memory for Energy Harvesting *23*
2.3.1 Introduction to Ultralow Power MCU Using Energy Harvesting *23*
2.3.2 Ultralow Power MCU with Embedded Flash Memory for Energy Harvesting *24*
2.3.3 Ultralow Power MCU with Embedded FeRAM Memory for Energy Harvesting *24*
2.3.4 Ultralow Power MCU with Embedded RRAM Memory for Energy Harvesting *26*
2.3.5 Ultralow Power MCU for Energy Harvesting Power Management *26*
2.4 Ultralow Power Battery Operated Flash MCU *27*
2.4.1 Introduction to Ultralow Power Battery Operated Flash MCU *27*
2.4.2 Ultralow Power Battery Operated Flash MCU with Embedded Flash Memory *28*
2.4.3 Ultralow Power Battery Operated MCU with Embedded RRAM *29*
2.4.4 Ultralow Power Battery Operated MCU with Embedded FeRAM *30*
2.5 Nonvolatile MCUs Using Emerging Memory for Nonvolatile Logic *32*
2.5.1 Nonvolatile Logic Arrays Using FeRAM *32*
2.5.2 Nonvolatile Logic Arrays Using MTJ MRAM *35*
2.5.3 Processors with RRAM for Nonvolatile Logic Arrays *37*
2.6 Communication Protocols for Memory Sensor Tags *41*
2.6.1 Radio Frequency Identification (RFID) Tags *41*
2.6.2 Near Field Communications (NFC) *42*
2.6.3 Bluetooth-Based Beacons and Sensor Nodes *43*
2.6.4 IoT Devices with Wi-Fi *46*
2.6.5 IoT Devices with USB Connectivity *47*
2.6.6 Single Wire Connectivity *48*
2.6.7 Zigbee Interface *48*
2.6.8 ANT Interface *48*
2.7 Wearable Medical Devices *49*
2.7.1 Overview of Wearable Medical Devices *49*
2.7.2 Miniature Hearing Aids Using FeRAM Memory *50*
2.7.3 Body Sensor Node Platforms Using CB-RAM Memory *50*
2.7.4 “Store Mostly” Healthcare Systems Using MRAM *50*
2.7.5 Wearable Biomonitoring with NFC and eFeRAM Memory *51*
2.7.6 Wearable Healthcare System with ECG Processor Using FeRAM *52*
2.8 Low Power Battery Operated Medical Devices and Systems *55*
2.8.1 Overview of Low Power Battery Operated Medical Devices *55*
2.8.2 Low Power Battery Operated Medical Devices Using eFlash *55*

2.8.3 LP Battery Operated Medical Devices Using Embedded Emerging Memories *59*
2.8.4 Security for Medical Systems *60*
2.9 Automotive Network Applications *61*
2.9.1 Overview of the Automotive Application *61*
2.9.2 Early Advanced Automotive Driver Assistance Systems *64*
2.9.3 More Recent Advanced Driver Assistance Systems (ADAS) *65*
2.9.4 Automotive Navigation and Positioning *66*
2.9.5 Under-the-Hood Applications *66*
2.9.6 MONOS Memory for Under-the-Hood Applications *68*
2.9.7 Automotive Infotainment *69*
2.9.8 Secure Automotive *70*
2.9.9 Automotive Body Processors *70*
2.10 Smart Electrical Grid and Digital Utility Smart Meters *71*
2.10.1 Overview of the Smart Meter Market *71*
2.10.2 Smart Meter Chips with Embedded Flash Memory *71*
2.10.3 Smart Meter Chips with Large Embedded Flash Memory *71*
2.11 Consumer Home Systems and Networks *74*
2.11.1 Remote Controls *74*
2.11.2 Environmental Sensors *75*
2.11.3 Home Network Systems *75*
2.12 Motor Control Chips with Embedded Memory *76*
2.12.1 Small System Motor Control Using Embedded Memory *76*
2.12.2 Motor Control for Multiple Motors Using Embedded MONOS Memory *76*
2.12.3 Motor Control with Embedded NV FeRAM *77*
2.13 Smart Chip Cards in Advanced Applications *77*
2.14 Analysis of Big Data Server Memory Hierarchy for Storing IoT *78*
References *80*

3 Embedded Flash and EEPROM for Smart IoT *89*
3.1 Introduction to eFlash and eEEPROM for Smart IoT *89*
3.1.1 Overview of eFlash and eEEPROM for Smart IoT *89*
3.1.2 Summary of Application Requirements for Embedded Flash in IoT *90*
3.2 Single Poly Floating Gate eFlash/EEPROM Cells for IoT *91*
3.2.1 Overview of Single Poly Floating Gate eFlash/EEPROM for IoT *91*
3.2.2 Early Single Polysilicon Floating Gate EEPROMS *91*
3.2.3 Single Poly EEPROM Cells for Specialty Applications *96*
3.2.4 Multitime-Programmable Single Poly Embedded Nonvolatile eMemories *99*
3.2.5 Recent Single Poly Fully CMOS Embedded EEPROM Devices *103*
3.2.6 Single Polysilicon eNVM in High Voltage CMOS *106*
3.3 eFlash Cells Using Multiple Single Polysilicon CMOS Logic Transistors *107*
3.4 Split Gate Technology for Floating Gate Embedded Flash *112*
3.4.1 Early Split Gate Embedded Flash Floating Gate Technology *112*
3.4.2 Issues, Peripherals, and Applications-Specific FG Split Gate Memory *116*
3.4.3 Advanced Split Gate Floating Gate Technology below 50 nm *124*
3.5 Stacked Flash and Processor TSV Integration *127*

3.6 OTP/MTP Embedded Flash Cells and Fuses *127*
3.7 Stacked Gate Double Poly Flash *130*
3.8 Charge Trapping eFlash *133*
3.8.1 Overview of Early Embedded Charge Trapping Memory *133*
3.8.2 Embedded 40 nm Charge Trapping (MONOS) Flash MCU *136*
3.8.3 Embedded 28 nm Charge Trapping (MONOS) Flash MCU *139*
3.8.4 Embedded Application-Specific 1T-MONOS Flash Macro *141*
3.8.5 FinFET SG-MONOS *142*
3.8.6 Embedded Charge Trapping (SONOS) NOR Flash *144*
3.8.7 Embedded 2T SONOS NVM in HV CMOS *147*
3.8.8 Self-Aligned Nitride Logic NVM *148*
3.8.9 p-Channel SONOS Embedded Flash *149*
3.8.10 Charge Trap eFlash for Low Energy Applications *150*
3.8.11 Blocking and Tunnel Oxide of DT BE-SONOS Performance *151*
3.8.12 Novel Embedded Charge Trap Memories *152*
3.9 Split Gate CT eFlash Nanocrystal Storage *158*
3.10 Novel Embedded Flash Memory *160*
References *161*

4 Thin Film Polymer and Flexible Memories *169*
4.1 Overview *169*
4.2 Organic Ferroelectric Memories *169*
4.2.1 Characteristics and Features of Organic Ferroelectric Memories *169*
4.2.2 Printable Ferroelectric Embedded Memories *174*
4.2.3 IoT Applications of Thin Film Ferroelectric Memory *179*
4.3 Polymer Ferroelectric Tunnel Junctions *181*
4.4 Types and Characteristics of Polymer Resistive RAMs with Flexible Substrate *181*
4.4.1 Overview of Polymer Resistive RAMs with Flexible Substrate *181*
4.4.2 Parylene-C-Based Resistive RAM *182*
4.4.3 Cu Atom Switches *184*
4.4.4 Inorganic Thin Film Resistive RAMs on Flexible Substrates *187*
4.4.5 IZO and IGZO Resistive RAM Memories *189*
4.4.6 Other Polymer Resistive RAMS with Flexible Substrates *192*
4.5 Charge Trapping Nanoparticle (NP) Memory on Flexible Substrates *199*
4.5.1 Overview of Charge Trapping NP Memory on Flexible Substrates *199*
4.5.2 Carbon Nanotube Charge Trapping Memory with Flexible Substrates *200*
4.5.3 Inkjet Printed Nanoparticle Memory *201*
4.5.4 Other Nanoparticle Charge Trapping Memories on Flexible Substrates *202*
4.6 Transfer of Conventional Memory Chips on to Flexible Substrates *206*
4.6.1 Transfer of Silicon Chips Using SOI Base Wafers *206*
4.6.2 Creating Thin Chips Using an Underlying Cavity *208*
4.6.3 Fan-Out Wafer Level Packaging for Assembling Silicon Chips on Flexible Substrate *210*
References *215*

5 Neuromorphic Computing Using Emerging NV Memory Devices *221*
5.1 Overview of Resistive RAMs and Ferroelectric RAMs in Neuromorphic Systems *221*
5.2 Various Resistive RAMs for use as Synapses in Neuromorphic Systems *221*
5.2.1 Metal Oxide Resistive RAM (MO-RRAMs) as Synapses *221*
5.2.2 Conductive Bridge RRAM (CB-RRAM) as Synapses *224*
5.2.3 Phase Change Memory (PCM) as Synapses *225*
5.2.4 PCMO RRAM as Synapses *226*
5.2.5 RRAM with Simultaneous Potentiation and Depression *228*
5.2.6 Other Nonvolatile Memories with Analog Properties *229*
5.3 3D Neuromorphic Memories *230*
5.3.1 Neuromorphic Architectures as Dense TSV 3D Structures *230*
5.3.2 3D Vertical RRAMs as Synapses Connecting Neurons *231*
5.4 Modeling and Characterization of RRAMs as Synaptic Devices *236*
5.5 Spiking Neural Nets, STDP, Potentiation, and Depression *239*
5.5.1 Introduction to Spiking Neural Networks *239*
5.5.2 Hybrid RRAM/CMOS STDP Neuromorphic Systems *239*
5.5.3 Memory Synapse and Neuron Systems *244*
5.5.4 Novel RRAM Synapse Applications *247*
5.6 Neural Network Systems Using Ferroelectric RAM Technology *250*
5.6.1 Neural Network Circuits Using Ferroelectric Memory (FeMEM) Synapses *250*
5.6.2 Using the FeMEM in Neural Network Circuits *251*
5.6.3 Ferroelectric Tunnel Junctions in Neuromorphic Circuits *252*
5.7 Early Neuromorphic Computers Using Phase Change Memory *254*
5.8 Resistive RAMs in Neuromorphic System Design and Application *257*
5.8.1 Design for Synaptic Devices for Neuromorphic Computing *257*
5.8.2 Using RRAMs in Various Neuromorphic Computing Applications *259*
5.8.3 Large RRAM Array Design for Neuromorphic Computing *260*
5.8.4 Advantages of RRAM over SRAM Crossbar Arrays in Matrix Multiplication *262*
5.9 Neuromorphic Memories Using Polymer and Flexible Memories *262*
References *266*

6 Big Data Search Engines and Deep Computers *271*
6.1 Overview of Big Data Search Engines and Deep Computers *271*
6.2 Content Addressable Memories Made Using Various Emerging Nonvolatile Memories *271*
6.2.1 Ternary CAMs Using Resistive RAMS *272*
6.2.2 CAMs Made Using Magnetic Memory *273*
6.2.3 CAMs Using Other Emerging Memories *276*
6.3 Components of Large Search Engines and Artificial Neural Networks *276*
6.3.1 Using RRAMs in Look-Up Tables in Large Search Engines *276*
6.3.2 Using STT MRAM in Large Artificial Neural Networks *278*
6.4 Memory Issues in Deep Learning Systems *281*
6.4.1 Issues with Partitioning SRAM and RRAM Synaptic Arrays *281*

6.4.2 Issues of RRAM Variability for Extreme Learning Machine Architectures *283*
6.4.3 Issues with RRAM Memories in Restricted Boltzman Machines *284*
6.4.4 Large Neural Networks Using Memory Synapses *287*
6.5 Deep Neural Nets for IoT *289*
6.5.1 Types of Deep Neural Nets for IoT *289*
6.5.2 Deep Neural Nets for Noisy Data *291*
6.5.3 Deep Neural Nets for Speech and Vision Recognition *293*
6.5.4 Deep Neural Nets for Other Applications *298*
References *299*

7 Memory in Security Issues for IoT *303*
7.1 Introduction to Memory in Security Issues for IoT *303*
7.2 Memories Used as Physical Unclonable Functions (PUFs) *303*
7.2.1 Using RRAM for a Physical Unclonable Function *304*
7.2.2 Using MRAMs as Physical Unclonable Functions *311*
7.2.3 Using Flash Memory as a Physical Unclonable Function *315*
7.2.4 Other Memories used as Physical Unclonable Functions *316*
7.3 On-Chip Memory-Based Security Systems *316*
7.3.1 Introduction to On-Chip Security Systems *316*
7.3.2 Physically Secure Key and TAG Storage *316*
7.3.3 Face and Feature Detection in Security Systems *319*
7.3.4 Security in Embedded Systems *320*
References *321*

Index *323*

Introduction to the Intelligent Internet of Things

The Internet of Things (IoT) has evolved from an older concept of Machine to Machine communications composed of specialized networks of things sending and receiving data obtained from the environment without the necessity of human intervention. The "Internet of Things" (IoT) is a concept of everyday objects that have network connectivity and can send, receive, and analyze data. An "Intelligent Internet of Things" will send, receive, and analyze data as well as have the capability to act with an intervention resulting from the analysis. Individual specialized networks are expected to communicate their knowledge to other networks to improve the function of a web of networks extending over the span of human existence.

These networks are everywhere. In retail stores, early bar codes provided scanned knowledge of inventory, its price, and some recorded indication of its origin. Bar codes have been replaced in high end retail stores with RFID tags, which can provide information on the state of the merchandise. For example, temperature monitors on wine bottles can detect temperature and record information on the maximum temperature experienced. This information can be radioed from the tag to a nearby receiver, which can analyze the state of the inventory and make recommendations on pricing or returns. These can then be recorded on the original RFID tag.

There are other examples of intelligent networks of systems communicating with each other. Wearable medical devices can take data and transmit it regularly to the medical provider node, which aids in medical monitoring and evaluation of patients. It also can implement relevant medical practices, in response to the original input, by feeding the recommended treatment back to the wearable devices. The outcome of the treatment could be collected with similar data to rapidly evolve successful new treatments.

Networks in smart homes are another example. Sensors can detect motion, fire, smoke, state of door locks; control cameras or audio devices record this information and turn on/off household equipment under the guidance of a smart network controller. An intelligent house network can turn on the installed sprinkler system as a result of rising temperature or fumes, which indicate a fire, along with setting off the fire alarm, notifying the fire department, and alerting first responders to the presence and location of people in the structure. Networks in traffic management systems in "Smart Cities" use embedded intelligence to control traffic lights to improve the flow of traffic, which could be detected from sensors along the curb set to register automobile location and detect speed. These sensors could also flag traffic monitors if excess speed is detected. The Smart City as a web of interacting networks supporting human existence is discussed in Chapter 1.

The MCU requirements for the various IoT applications differ significantly and affect the type of embedded nonvolatile memory that can be used in the MCU for that application. Embedded memory considerations include: standby power, active power, endurance, program and erase voltage, read and write speed, and data retention. These depend on the particular application the MCU will be used in. There are also options for the types of embedded memory technology to use.

Applications of interest include: ultralow power MCU used with energy harvesting and those used with battery operated applications, processors with nonvolatile arrays and power gating, processors used in intermittent operations, and communications processors. Processors used in automotive network applications have a different set of requirements for embedded memory. The varying characteristics of different IoT applications for processors and their embedded memories are discussed in Chapter 2.

For IoT processors to be made at low cost and in high volume, the memory in these processors needs to be manufacturable in high volume on standard CMOS logic processing lines. Currently most wafer fabrication areas have simple conventional embedded Flash memory and EEPROM macros available to be used as IP in MCUs run on their processing lines. As the demand for IoT processors has risen, new configurations of these well understood logic compatible memories have been developed. Chapter 3 discusses the status and development of embedded floating gate Flash and EEPROM memory as well as charge trapping memories that are CMOS logic compatible.

The requirements for both memories and processors in battery powered or energy harvesting sensors are very low power and very low cost. For wearable devices, the circuitry must be flexible. Circuits that can be made in large volume without the expense of semiconductor processing might lower the cost so a significant amount of effort has gone into ferroelectric, charge trapping, and resistance memories that can be made by inkjet printing or screen printing and can also form logic circuits. Chapter 4 discusses the evolution of these efforts to produce low power, low cost, flexible memories for the Internet of Things. It also discusses flexible circuits with higher performance made from thinned silicon chips mounted on a flexible substrate.

Local area networks (LANs) on the edge of the Internet can potentially use the memory-based neuromorphic computers that are currently reaching an early level of capability. The use of these local intelligent nodes mean that local data can be analyzed and the results sent to the Cloud. This can provide an extra level of data security. The development of these memory-based neuromorphic computers is discussed in Chapter 5.

The significant amount of data that is collected by the many sensor nodes and the identification data of these nodes must be stored where it is widely available, generally in the banks of servers that make up the Internet Cloud. The memory hierarchies in these servers are critical for the efficient functioning of the system. Sophisticated search engines associated with the Cloud servers need a new level of artificial intelligence that is currently under development. A few such artificial intelligence systems are discussed in Chapter 6.

Chapter 7 discusses various aspects of Internet Security involving memory devices. The use of physical unclonable functions (PUF) based in emerging memory devices such as MRAMs and RRAMs is also covered here.

1

Smart Cities as the Prototype of the Intelligent Internet of Things

1.1 Overview

The Intelligent Internet of Things (IoT) is in the process of development. Many advances in this smart interactive human environment are already established. Many others are still to come. Significant visionary concepts have been floated, resulting in a general weariness with the hyperbole inherent in such futuristic concepts. This should not obscure the serious advances that have been made and are still to come. This chapter will attempt to envision the future while concretely treating current developments in smart IoT and near term future trends.

1.2 Smart Cities

The Intelligent Internet of Things is most developed in the urbanized areas of the developed countries where efficient networks for commerce, transportation, public utilities, residences, and life essentials are well developed as a matter of necessity.

An overview of these smart networks as currently envisioned can be framed by the concept of the smart city, as indicated in Figure 1.1.

Many of the networks used by people can be thought of as elements of the smart city. These include: networks of modes of transportation, commerce, services, residences, and public utilities.

1.3 Smart Commerce as an Element of the Smart City

1.3.1 Smart Inventory Control

The original inventory control tag was the bar code, which offers basic inventory data in a machine readable form and works by varying the widths and spacing of parallel lines. Following after the bar code was the QR Code optical label, which includes information about the item to which it is attached. The QR Code offers fast readability, more information than a bar code, and can be read with a smart phone. It often leads to a web page with product information.

Memories for the Intelligent Internet of Things, First Edition. Betty Prince and David Prince.
© 2018 John Wiley & Sons Ltd. Published 2018 by John Wiley & Sons Ltd.

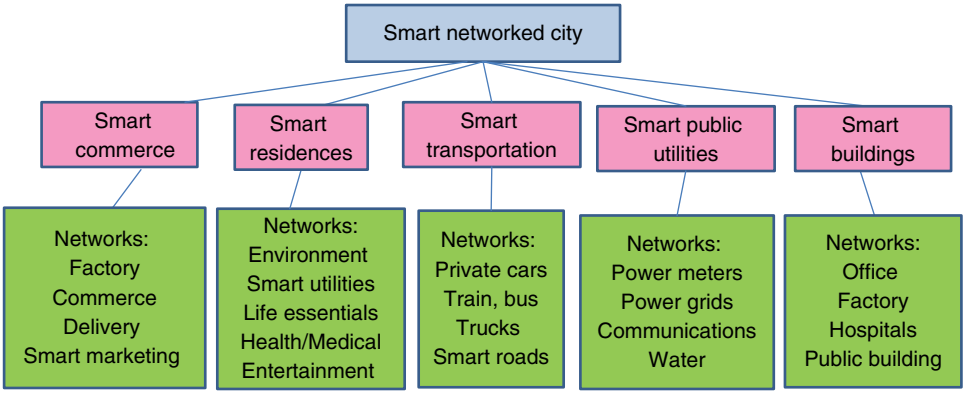


Figure 1.1 Elements of the smart city include connected networks of transportation, commerce, public services, connected residences, and public utilities.

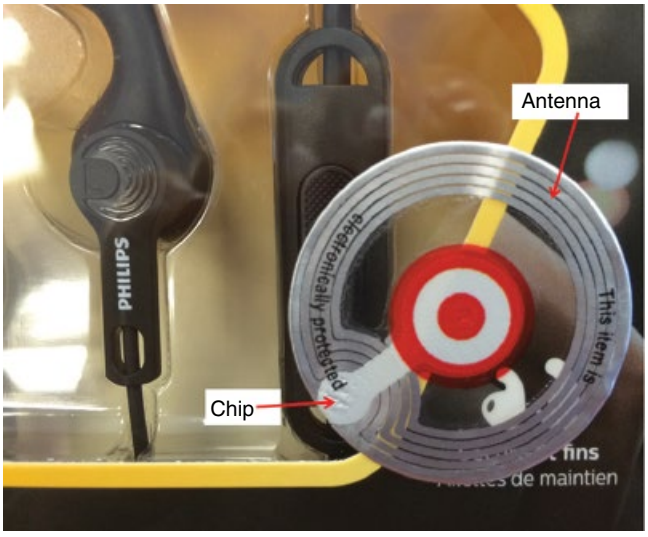


Figure 1.2 RFID tag on retail merchandise showing antenna and chip. Photo by B. Prince.

Smart commerce began with the early machine-to-machine (M-to-M) communication devices. Early radio frequency identification (RFID) tags were used in factory and retail inventory, as shown in Figure 1.2. Unique ID codes could be stored locally and later in the cloud. The origin of the devices along with their manufacturing and transit history and even environmental influences could be saved and transmitted when needed. Prices could be fixed, stored, and changed locally.

While RFID was an early implementation of a communications protocol, other remote communication protocol standards have been developed such as: near field communications (NFC) and Bluetooth. Communications protocols such as USB, SPI, and I2C are used within a connected system. WiFi is also used to connect electronic devices to a wireless local area network (WLAN) using specific frequencies.

Within a retail environment, tagged inventory as small as an individual item can be tracked, priced, and repriced and environmental information measured and stored. For example, sensors with thermal detectors can track the environment of an expensive wine or a medicine to be sure that it has been held within permitted temperature ranges.

1.3.2 Smart Delivery

Scheduled delivery has become an everyday reality, with on-line stores routinely making timely deliveries with networks of agents using interconnected reporting systems. Delivery drones are not yet an everyday occurrence but pilot studies are being conducted. Amazon, in December of 2016, indicated it had delivered its first package by drone to a customer in Cambridge, England, using a drone with a range of 10 miles and a maximum load of 5 pounds [1]. An illustration of a commercially available drone carrying a camera and a control unit is shown in Figure 1.3.

1.3.3 Smart Marketing Using Artificial Intelligence

Artificial Intelligence can survey individual historical interests and purchases, and can predict consumer preferences. Smarter tags with transmit capability, called "Beacons," can be used to lure potential buyers with targeted advertisements to their cell phones. These smart tags can also be used for tracking mobile objects in a controlled environment with the use of stationary tracking beacons.

Figure 1.3 Drone carrying camera with control unit. Photo by B. Prince.

1.4 Smart Residences

1.4.1 A City of Smart Connected Homes

IoT for Intelligent homes potentially includes: environmental control, external access and safety control, vision systems, light control, comfort control (auto adjusting beds and chairs), information control, entertainment control, health and medical alert, food and water control, safety alerts (tornado, fire, flood), animal access and care, and child monitoring and care. Initially remote control of these systems can be provided to the owner. Eventually such systems will be automated and robotic with owner access on demand. An illustration of the concept of a smart connected home is shown in Figure 1.4.

In the intelligent home, a controller device can be the owner's mobile handheld computer (smartphone), which can be used to monitor the smart devices and provide control as modifications to the settings are desired. Normally, intelligent home sensors would be expected to function automatically without the need for human intervention.

An automated home could be expected to use the GPS locator in the owner's phone to anticipate his/her arrival and biological inputs to secure entry. In the owner's absence, thermostats would be expected to maintain temperatures at timed preset levels. The thermostat setting could be adjusted remotely from the resident's smart phone. When the house is occupied, wearable monitors on the residents could communicate with the thermostat to adjust the light and comfort level for the residents automatically. Minithermostats that control a network of air conditioners could be used to

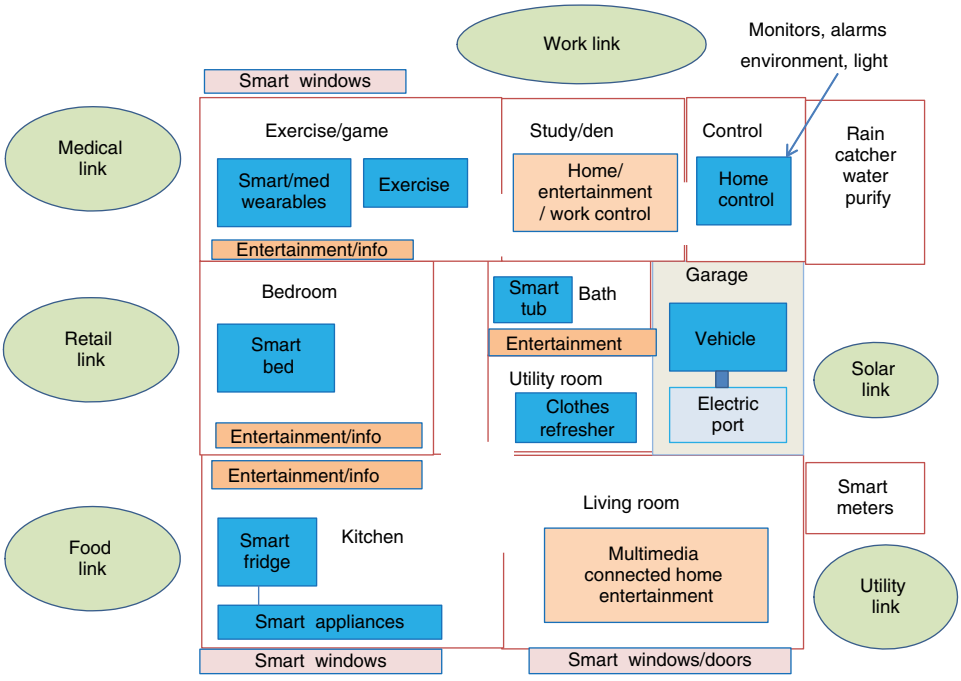


Figure 1.4 Concept of an intelligent connected home.

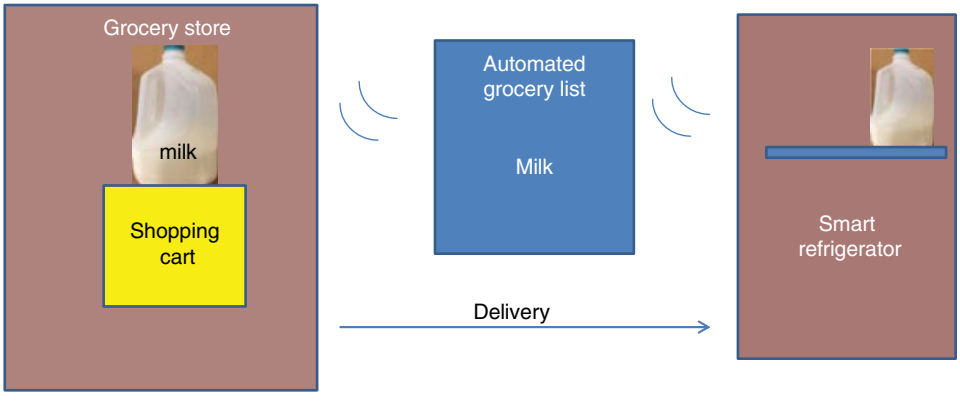


Figure 1.5 Example of transmitted autofill of grocery list using sensors for a low milk level, NFC for transmission of milk to a grocery list, WiFi for transmission of purchase request to grocery store, followed by delivery of milk to the home.

individualize the temperature to the location and requirements of individual users. Light sensors could open and close shades at appropriate times and turn on outside lights as required and these could also be controlled and programmed on demand. Domestic appliances can be turned on and off remotely by either the owner or by a preset controller in the home network.

A smart refrigerator network, illustrated in Figure 1.5, could detect the condition and amount of each item from smart tags on the items and transmit an automated grocery list to the store [2]. The store in turn could find the requested items and have them either ready for pick-up or delivered to the home.

Autonomous systems using smart meters can monitor electrical usage and adjust flexible electrical equipment run times to correspond to lower cost energy times. Motion sensors can maintain security. Appropriate devices for implementation of these features are available today. Further out in time, the health and diet of the residents could be monitored, along with the quantity of nutritional elements in the house and grocery shopping lists could be automatically refilled [2].

1.5 People as Center of Smart Connected Homes

1.5.1 Wearable Electronics

The people in the smart homes can become another center for an Intelligent Internet of Things. Wearable exercise and health monitoring electronics can potentially form such a network. At its simplest, wearable electronics are machine-to-machine networks. Sensors worn on the body record a parameter such as heart rate or temperature and these measurement data are recorded and transmitted to a device that monitors and records the reading, such as a smart phone. The information could also be transmitted on to health professionals, with automated advice or prescription delivery being returned.

An example of intelligent connected wearable electronics is a system that senses, records, transmits, analyzes, instructs, and enables a response that could be by either a person or a device. An example is a wearable blood glucose detector associated with a

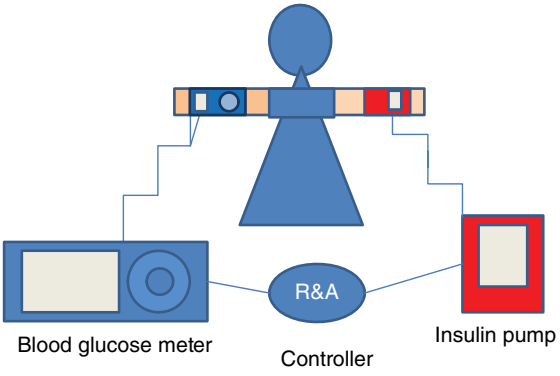


Figure 1.6 Illustration of a possible future wearable system – a wearable blood glucose detector network with an intelligent monitor and connected insulin pump.

wearable insulin pump and an analysis device. This glucose detector could measure and record the user's blood glucose level and report it to the recording and analysis (R&A) device, which could be an application in a smart phone. The R&A device could note the glucose level and compare it to an acceptable level. If it is above a certain level, the R&A device could communicate with an insulin pump, which would inject a determined amount of insulin into the blood. Such a wearable glucose detector is available today as well as a wearable insulin pump that can be combined with the glucose detector [3–5]. Both can be worn as armbands, as indicated in Figure 1.6. Other smart wearable devices might include enhancements of normal bodily functions such as radar added to vision or color enhancement of sound.

Wearable electronics are an example of systems that require very low standby power since the wearable devices are either powered by small coin cell batteries that provide a low current and can last a significant amount of time or are powered by some form of energy harvesting. These systems require ultralow power microcontrollers with embedded memory that is both nonvolatile for low standby power and has low operating power requirements.

1.5.2 Control Electronics

The home resident can be expected to have available simple automated control for all the functions of the home, whether they are in the home or away. This control is already available for many features such as cameras, lights, curtains, safety monitors, and alarms. Control of appliances and utilities is not far behind. All of this will be exercised from a handheld or worn device, currently a smartphone.

1.6 Smart Individual Transportation

1.6.1 Overview of Smart Automobiles

Automobiles are another example of an Intelligent Network of Things application. There are over 100 microcontrollers in the modern high end automobile [6]. These microcontrollers tend to fall into the classifications of: body processors, infotainment processors, driving aids, and engine processors, as shown in Figure 1.7 [6].

1.6.2 Driving Aids

An early example of a smart application driving aid is "anti-lock" braking in a car in which wheel speed sensors can detect if one or more wheels are beginning to lock up during braking. If a wheel tries to lock up, a series of hydraulic valves limit braking on that wheel and initiate a pumping action, which helps stop the car safely without skidding in a minimum amount of time.

Another example of a smart IoT application that is a driving aid might be a back-up camera or radar that detects an obstacle in the path of the car and notifies the braking system to apply the brakes, a lane change warning device, or a blind spot warning. Such systems exist today along with traction control, cruise control, and collision avoidance systems. Some systems currently readily available are shown in Figure 1.8.

Driver-assisted applications are being developed that can increase the knowledge of the driver of hazards, road conditions, and unseen traffic and also increase the driver's skill at dealing with these traffic environment conditions. Examples include cameras and radar mounted on the car, which expand the driver's knowledge of the immediate surroundings of the car on all sides at any given time. They also include automatic parking systems and automatic trailer back-up systems.

The feedback to the driver from a driving aid must be fast, instantly understandable, unambiguous, and the response required obvious or autonomous. This requires fast sensors along with audiovisual aids, which would normally require fast processing and

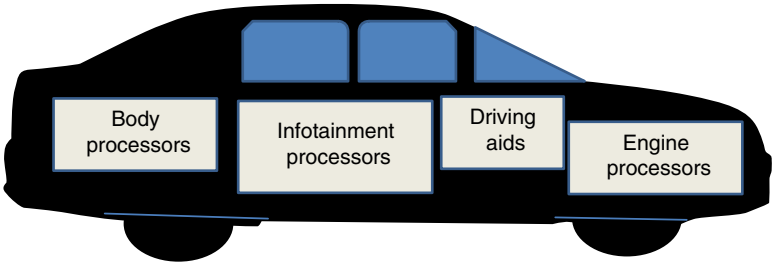


Figure 1.7 Automotive IoT networks of processors including: body processors, infotainment processors, driving aids, and engine processors. Based on T. Kono *et al.* (Renesas), *IEEE Journal of Solid-Sate Circuits*, January 2014 [6].

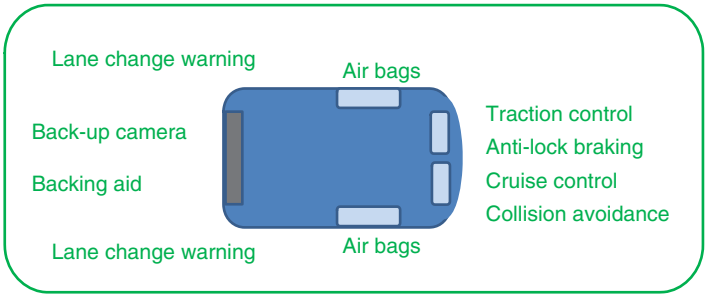


Figure 1.8 Automotive safety assistance devices currently available.

significant amounts of memory. High performance graphics processors could be used to implement these driving aids.

1.6.3 Engine Processors

Engine Control Units control an electrical system for a driving function in an automobile. These can include systems that control: power train, power steering, transmission, or engine timing. The inputs to the engine processors come from various sensors, which are primarily for engine management and performance. The Engine Processors tend to function autonomously without the requirement for driver intervention, providing primarily driver notification. For example, the failure of one of the engine sensors may result in the "check engine" light coming on as a driver notification aid. Primary criteria include performance at high temperature and high reliability. Since the engine of a car becomes hot, electronic components that will be expected to function properly in the engine compartment have temperature requirements for functionality that is generally –40 °C to +150 °C, but could be as high as 170 °C.

1.6.4 Auto Body Processors

Body processors can include: seat control units, automatic lock controllers, door modules, seat modules, central body, vehicle body, smart junction box, mirror adjust, air conditioning, lighting, seat belt sensors, and air bag controllers. These processors would be expected to include nonvolatile memories in some cases up to MBs in capacity. Air bag controllers would require fast response times. Security features are being added on these intelligent body processors.

1.6.5 Infotainment Processors

Visual information processors are critical for automobiles to reduce accidents at current high highway speeds by aiding drivers and enhancing their awareness of the external and internal environment of the car. Automotive projection systems with very wide fields of view that enhance driver vision both during the day and at night are in development. Such an optical system could include speed, navigation, obstacle awareness and avoidance, nighttime radar visual enhancement, and field of view enhancement for the driver of the car.

Parking assistance is another driver support system being developed. Such a system can use multiple cameras with 360 degree views and audible proximity sensors to aid in the parking process, or parking the car can be fully automated. Parking assistance of trailers is even available.

High performance GPS multisatellite navigation systems including complex destination guidance with both audio and visual displays are also in development. The current ease of integration of the smartphone into the automotive infotainment system offers connection between the vehicle and the Internet.

1.6.6 Autonomous Cars

Autonomous cars are in various stages of development, which would take the environmental conditions notification and act on these automatically. Fully autonomous driving would require considerations of priorities, which could be that value judgment cars

would require artificial intelligence to analyze and act on. Criteria for an application such as smart brakes are fast sensing and speed of processing, since the brakes need to be applied before the accident occurs. Inputs to an automated braking system could be either input from cameras mounted on the car or from radar, both of which sense obstacles. In driverless applications the automobile is handled by a computer, which intakes information about the road conditions and drives the car to avoid obstacles and arrive at the intended destination.

The automobile network is an example of a system that does not necessarily have an ultralow power standby requirement since the various processors are on the car battery. High performance MCUs are required, however, for speed, both of processing and of read and write functions.

It is possible that traffic would be safer if autonomously driven cars and driver controlled cars are not driven in the same lanes or on the same highways. Perhaps initially internal city areas could be reserved for autonomous cars, which are generally available for hailing upon entering these areas. Alternatively, conventional public transportation networks could be expanded for city centers.

1.7 Smart Transportation Networks

1.7.1 Smart Public Conveyance Networks

Networks of public transportation such as trains and buses are also examples of smart transportation that can be run by M-to-M. Trains can be run automatically to preset schedules. Trains can have sensors either on the train or on the track that detect other trains and adjust speed to avoid collisions. Smart networks of trains would have programmed priorities and adjust travel speeds and routes to achieve the priorities of the smart network. Smart controllers could run an entire network with only local human monitoring.

1.7.2 Individual Automotive Traffic Control

Smart cards for ticketing have been used for years and toll tags mounted on the windscreen of a car are also in widespread usage. These tags can be polled remotely. The EZ tag system used in Texas is an example [7]. In this case users receive a small radio frequency (RFID) transponder which is affixed to the inside of the windshield of their car, as shown in Figure 1.9. When passing through a toll reader, specially equipped sensors can read EZ TAG transmitters and the amount stored on the card can be decremented after each use and incremented when recharged with money at a terminal.

The tags can also be used to generate real time traffic information for use by the general public. The individual customer information is not exposed since the data taken is general traffic density and motion.

It is conceivable that such a system could be used to locate all residents of an area monitored by compatible sensors at a given time since the EZ tags do ID the specific automobile. This would depend on their being a sufficient density of sensors to detect all vehicles on the roadways in a specific area. Such a system could also be used to determine the speed of a particular vehicle by noting the time elapsed between successive tag readings.

Figure 1.9 Photograph of a Texas EZ tag RFID sticker affixed to a windshield. Photo by B. Prince.

1.7.3 Smart Highways

Smart highways should be able to coordinate traffic to ensure the smooth flow of traffic and ensure that traffic jams are eliminated and individual vehicles can safely reach their destinations in a minimum amount of time. Traffic light coordination is not unusual, but it should be flexible to accommodate the existing traffic configuration. Highway direction can also be changed depending on the time of day so that traffic in the dominant direction has more lanes. A different amount can be charged for access to high speed lanes, ensuring that each vehicle has the option of moving at a user determined rate. Sensors along the road should be able to sense the presence of cars and determine their speed by calculating the time between successive readings. The highway could help determine traffic safety by sensing speed and location of individual vehicles and communicating with each vehicle to ensure that vehicles maintain a minimum separation from each other.

Elevated and underground roadways are used to expand the amount of lane space over that available on a flat plane and were used, for example, on the Boston Dig and in many other cities. Limited access to freeways provides a smoother flow of traffic as do metered entrance ramps on to freeways. Accident detection systems can minimize the amount of time average speed is reduced due to an accident. They could also notify authorities of the occurrence of an accident and determine the vehicles involved and the ID of possible witnesses. Toll tag systems throughout the United States detect and identify a car containing a tag when the car passes a roadmark sensor.

1.8 Smart Energy Networks

1.8.1 Smart Electrical Meters

Electrical meters can connect the user to the electric utility to receive energy and determine its cost. It can connect to the devices in a smart residence or commercial building to determine essential systems and reduce electrical usage of nonessential systems

during times of high energy cost or shift power usage by essential systems to lower the cost times of day. If electrical energy is available at the specific location, such as a rooftop solar array, the smart meter can route this additional energy into the electrical grid during peak energy usage times. Energy is produced by local solar systems during the hottest hours of the day when usage by air-conditioning systems is highest. This could help reduce the amount of higher cost energy required during these peak hours. Electrical energy in and out of a specific location can be tallied by a smart meter.

Datalogging of electricity usage by a smart meter can be done frequently. The data for conventional smart meters is usually logged to a serial EEPROM due to its low cost, low power, and standard package. The advent of "advanced metering infrastructure" (AMI) meters promotes efficient electrical power generation, transmission, and distribution. Smart e-meters can log electric power parameters as frequently as every few seconds or even in millisecond intervals. With frequent datalogging the performance of the meter is impacted by the slow nonvolatile memory write speed and limited endurance cycles. For this reason alternative memory technologies, such as ferroelectric RAMs, battery backed SRAM, and nonvolatile SRAM, have been used in AMI smart meters. These memory options tend to have higher endurance and a faster write speed [8].

Smart meters can be connected to energy consuming devices in a building with instructions to take these devices off-line during peak energy times when electricity demand and prices are highest and the system is most at risk for rolling blackouts. This both saves money for the building owner and helps the electric utility avoid blackouts. Security is a significant issue for smart electricity meters, which have two-way communications between the electric utility and the consumer. Figure 1.10 is an illustration of a smart meter connected to a solar array in a building and also connected to the central control of an electric grid containing various power sources including: hydroelectric, solar, coal burning generator, and windmill farm.

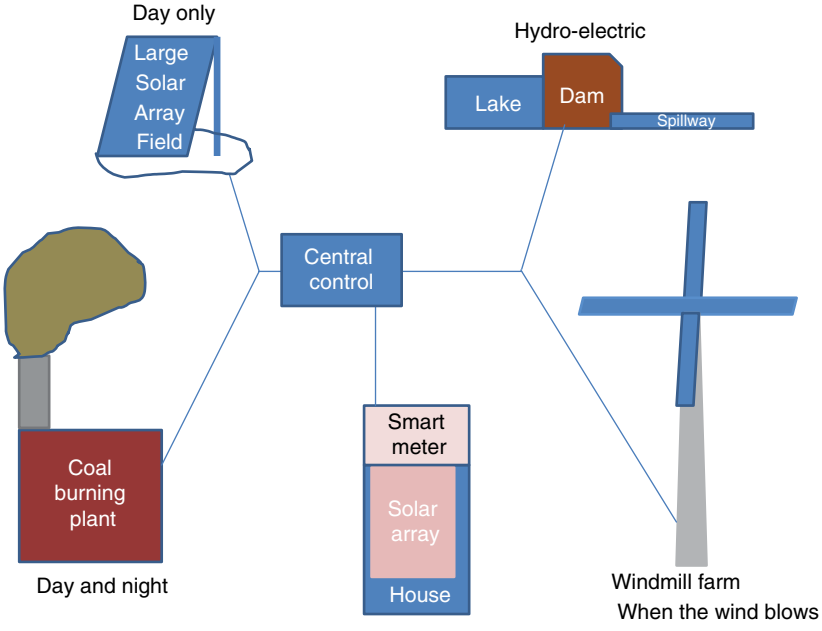


Figure 1.10 Illustration of a smart energy network controlled by smart electrical meters.

1.8.2 Smart Electrical Grids

The electrical grid is the control network for supplying energy to the city. The electrical grid can be connected to a network of energy sources including coal burning and natural gas burning plants, which are available on demand, intermittent energy plants such as solar installations, and emergency engines and hydroelectric plants, which are available on a longer term perspective.

These energy sources range from steady to intermittent to demand and can be ranked from low to high cost.

1.9 Smart Connected Buildings

1.9.1 Smart Office Buildings

A smart office building can regulate its environment to ensure the comfort and safety of the inhabitants. Temperature can be sensed and regulated at different locations. Motion sensors or heat sensors can monitor the presence of people for safety purposes and GPS monitors of ID tags can determine the locations of people within the building. Light levels can be adjusted for comfort and blinds drawn automatically. Figure 1.11 shows a picture of a thermostat for smart houses.

In the future, sensors worn by individuals could communicate with the smart building's environmental controls to permit individually temperature controlled workspaces. Individual track robots can bring requested materials to the user and individually prepared meals, for example, could be automatically delivered along the same tracks.

Air quality can be monitored on an ongoing basis to ensure high air quality to all residents. Local toxins can be sensed and isolated consistent with ensuring the safety of all

Figure 1.11 Picture of a "NEST" thermostat for smart buildings and houses. Photo by B. Prince.

inhabitants. In case of a fire, areas can be isolated automatically after checking for inhabitants. Sound can be monitored from all locations and information can be provided at each location. Sound systems can be maintained by location.

1.9.2 Smart Factories

The fully automated factory is also a reality today. Large semiconductor factories with fully automated assembly lines are operated by a few workers and maintenance staff along with sophisticated robotic machines capable of precise operations that can be programmed and the program modified remotely.

While automated factories have been around since the 1970s, fully interconnected smart factories are more recent. Both customer reliability and government reportability demand transparency and interaction during the various stages of the automated manufacturing process.

Automated monitors can generate data at every step. Smart vision monitors can replace inspectors on assembly lines. Safety can be improved by new touch interfaces with sensors to detect the presence of people. This has results in new generations of large robots with user-friendly interfaces. This ranges from pick and place for very small objects and for very large objects. Real time information can be generated by connected machines around the world as industrial robots generate real time information for monitors and regulators.

Automated inspection robots can detect faults, replacing human inspectors, and can identify and report on these faults. Robots can self-test and repair themselves. The motion and amount of force used can be adjusted automatically using vision systems during pick and place. Vision guided smart robots can pack boxes and sort random materials on a pallet in a warehouse. Materials can be transferred to smart carts for delivery and place in warehouses. These are all examples of smart machines talking with other smart machines.

An example is a phased semiconductor factory automation design developed by Hewlett-Packard that permits a semiconductor factory to incorporate progressive levels of factory control software, as well as automated material handling systems and real time tool control [9].

Distributed stacking can be designed into a warehouse with manual contingency operations. Reliability considerations can be installed using in a phased approach. Three-dimensional (3D) vision guided robots can handle random bins or mixed pallets in warehouses. Smart camera systems can inspect and report to machines staged further on in the process. Smart configurable machines can be programmed by incoming vision systems for the automated process required for multiple processes and can transmit the requirements to the appropriate machine. Varying reliability requirements can be implemented and inspected in a smart factory according to user requirements by machines programming other machines.

1.9.3 Intelligent Hospitals

Intelligent hospitals can intake physicians' instructions, order prescriptions, automate incoming patient screening and tests, and include specialists in various fields in the individual diagnosis. Reliability monitors can check the remotely monitored individuals for procedure errors and make automated corrections. Sanitation detection measures

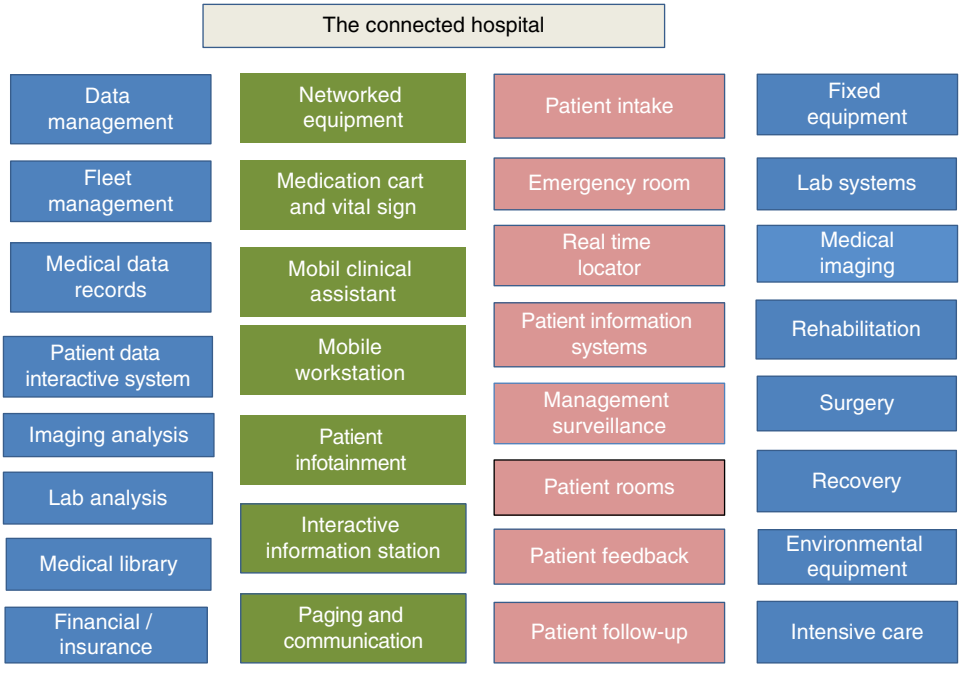


Figure 1.12 Illustration of digital connected systems in a smart hospital environment.

can be implemented to prevent hospital-generated diseases and patients can be routed to the areas best able to protect and care for them.

An illustration of possible features of a connected hospital is shown in Figure 1.12.

Advantech in Berlin, for example, configures smart hospitals in which nursing carts can deliver medications to patient rooms and deliver individualized medications [10]. Patient data management systems can automatically handle a large number of on-line patient records and ongoing treatment systems. Mobile handheld clinical assistants can provide instant records to local medical staff. Patient feedback both from automated monitors and individually entered by each patient can provide a full picture of individual care. Individual room imaging systems can aid in diagnoses and provide expert opinions and diagnosis without the expert being required to be present in the room.

Safety checks can be programmed into the system to double-check diagnoses, tests, and treatments to avoid errors. Automated background calibration systems can check instruments and automated redundant reliability systems can check for hospital errors in procedures. Digital badges can maintain a real time monitor location of each staff member to track, for example, communicable disease spread or establish medical decision chains. Digital badges for patients can monitor procedures and locations and check against prescribed procedures for that patient.

1.9.4 Smart Public Buildings

Public buildings are being networked to enhance their functions. Courthouses and Libraries are being linked to large information centers and extensive specialized sources. Museums have virtual representations that transport the viewer to distant and exotic

scenes. Extensive safety networks are being developed to keep people safe in large public gatherings. In the future, virtual museums and virtual concert halls may bring the sights and sounds to the viewer, reducing the number of large public gatherings.

1.10 Thoughts

The intelligent connected world being described here is feasible and early implementation is already taking place around us. The rate of implementation of these technologies will be sporadic, but it is clear that the more useful applications are already finding a rapid rate of adoption in urban areas where the population density lends itself to observing the advantages more quickly. Automated and intelligent use of these extensive data connections will follow.

References

1 McGoogan, C. (2016) Amazon makes first drone delivery to house in Cambridge. *The Daily Telegraph*, 14 December 2016.
2 Aitken, R. *et al.* (2014) Device and technology implications of the Internet of Things (ARM), VLSI 2014, June 2014.
3 MiniMed 530G System featuring exclusive SmartGuard technology for advanced protection against lows. Medtronic website, professional.medtronicdiabetes.com/minimed-530-g, September 1, 2016.
4 Introducing the Dexcom G5 mobile CGM system, https://www.dexcom.com/.
5 The Animas Vibe insulin pump and CGM system, https://www.animas.com/diabetes-insulin-pump-and-blood-glucose-meter.
6 Kono, T. *et al.* (2014) 40-nm embedded split-gate MONOS (SG-MONOS) Flash Macros for automotive with 160-MHz random access for code and endurance over 10 M cycles for data at the junction temperature of 170 °C (Renesas). *IEEE Journal of Solid-State Circuits*, 49 (1), 154, January 2014.
7 EZ tag. Wikipedia, 20 July 2016.
8 Singh, S. (2014) F-RAM for smart E-meters, Cypress Applications Note AN87352.
9 Gardner, D.C. (1996) Semiconductor factory automation: designing for phased automation, Hewlett-Packard, Advanced Semi Man. Conference and Workshop, ASMC Proceedings, November 12, 1996.
10 Advantech digital healthcare smart hospital (PEIL), conhIT Berlin, December 2015.

2

Memory Applications for the Intelligent Internet of Things

2.1 Introduction

The Intelligent Internet of Things (IoT) is a term used here for smart networks of processing systems communicating with each other, analyzing the data generated, and responding to the conclusions of this analysis. These networks are expected to improve smart automation in many fields. Examples of such smart networks of systems communicating with each other are: wearable medical devices that take data and transmit it regularly to the network to aid in medical monitoring and evaluation of patients; networks in smart houses that detect motion, fire, smoke, state of door locks; and control cameras or audio devices and turn on/off household equipment under the guidance of a smart network controller. Communicating tags are used everywhere on IoT sensors and they need to be very low cost and work with very low levels of intermittently harvested energy. Energy harvesting is a requirement for the many sensors required in such ubiquitous networks since replacing batteries would be prohibitive. Where batteries are used, ultralow power is required. The many devices used in an IoT network must be very low cost.

Data storage is needed in smart networks. Memories are required to collect the information and process it. Many microcontroller units (MCUs) for IoT have the required memory embedded in the processor. In addition, the processors themselves must have near zero power consumption in standby since most of their time will be spent there. They may need in some cases to power up in the state required without taking time and power to boot-up. This can be done using nonvolatile nodes in the processors, which can be implemented in some cases using high endurance memory devices such as ferroelectric RAMs (FeRAMs) or magnetic RAMs (MRAMs) on these nodes for power gating. For faster program speed for code and data storage and lower program voltage than Flash memory, resistive RAM (RRAM) could be used. For moderate performance at very low cost, phase change memory (PCM) might be used.

A few of the applications that have been studied for memories for IoT are detailed in this section. These include: ultralow power MCUs with energy harvesting sensors or compact batteries, smart communication tags, networks of wearable medical devices, smart motors, automotive networks, smart meters, and big data search engines.

Annual shipments of MCU with embedded nonvolatile memory were discussed by Renesas in January of 2014 [1]. These are illustrated in Figure 2.1 from 1980 to 2000 along with an indication of the primary embedded nonvolatile memory in each time

Memories for the Intelligent Internet of Things, First Edition. Betty Prince and David Prince.
© 2018 John Wiley & Sons Ltd. Published 2018 by John Wiley & Sons Ltd.

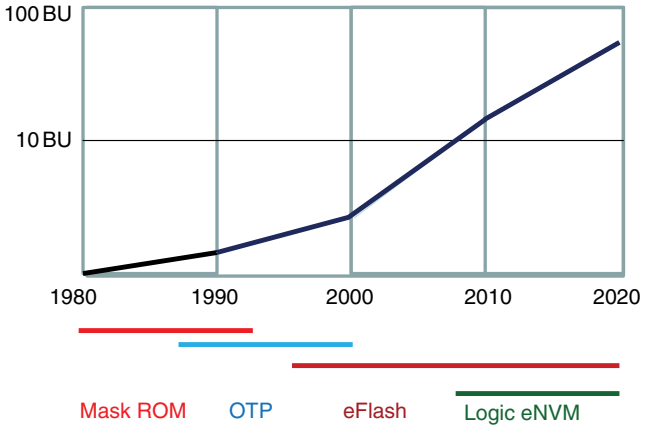


Figure 2.1 Estimate of shipments of MCU chips with embedded nonvolatile memory (BU annually). Based on T. Kono *et al.* (Renesas). *IEEE Journal of Solid-State Circuits*, January 2014 [1].

period: first generation mask ROM, second generation embedded one-time-programmable (OTP) memory and third generation embedded Flash. Logic-based eNVMs are added and a forecast is made to 2020 by extending the data trends. It was estimated by Renesas that MCUs with embedded Flash memories accounted for about 70% of all MCU shipments in 2011.

2.2 Comparisons of the Various Nonvolatile Embedded Memories Characteristics

2.2.1 Embedded EEPROM, Flash, and Fuse Devices

An early rewriteble embedded nonvolatile memory was the floating gate electrically erasable programmable read only memory (EEPROM) followed soon after by the dual polysilicon Flash memory. Both of these devices were made in the double polysilicon technology still used for standalone devices. Since conventional CMOS logic technology, in which MCUs tend to be made, does not use double polysilicon, the double polysilicon EEPROM and Flash memories added cost to the process when embedded in a processor technology.

Embedded EEPROM and Flash devices that used single polysilicon and were compatible with the conventional CMOS logic technology, were developed in the 1980s. The cost of the technology was lower than for devices made from double polysilicon but the cell size was larger. For very small amounts of embedded memory, a larger cell size was not a significant cost issue. For larger memory capacity or for off-chip memory, the double polysilicon technology continued to be used.

A comparison between the characteristics of different types of embedded Flash memories can illustrate the trade-offs between performance and array size. A comparison of various embedded Flash and EEPROMs as a function of array bit count and write cycles was done in October of 2015 by the University of Brescia [2]. Devices considered were: double polysilicon floating gate embedded Flash, single polysilicon EEPROM,

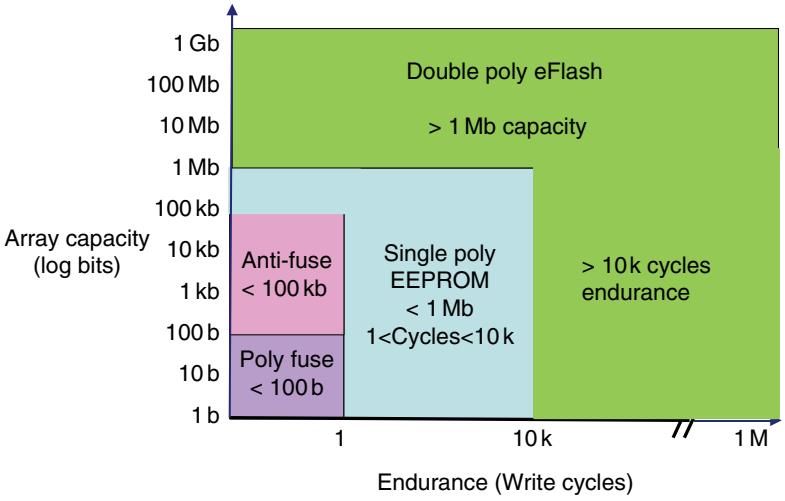


Figure 2.2 Array capacity versus endurance for various types of embedded Flash memory cells. Based on L. Milani *et al.* (University of Brescia), *IEEE Trans. on Electron Devices*, October 2015 [2].

transistor-based antifuse devices, and polysilicon fuse devices. An illustration of the bit-count versus write cycle characteristics of these embedded Flash memory technologies is shown in Figure 2.2.

The polyfuse and antifuse solutions for NVM are suitable for one-time programmable (OTP), small density (<100b) and medium density (<100Kb) applications, but multiple time programmable applications require double poly embedded Flash or single poly EEPROM. The double polysilicon eFlash is suitable for applications of more than 1 M-bit capacity where the additional cost due to the extra masks and process steps is acceptable. The single poly EEPROM is the best choice for low and medium density applications requiring up to 1 Mb of memory

2.2.2 Embedded Emerging Memory Devices in MCU

Various new memory devices have been investigated to solve the shortcomings of conventional SRAM and Flash storage. These include: magnetic RAM (MRAM), resistive RAM (RRAM), phase change memory (PCM), ferroelectric RAM (FeRAM), and charge trapping Flash memories, such as SONOS, MONOS, and nanoparticle. Many applications discussed here, particularly for IoT, will use the memory embedded in an MCU. The Flash MCU requirements for the various applications differ significantly and affect the type of embedded nonvolatile memory and RAM that can be used in the MCU for that application.

Among the differentiating characteristics of various embedded nonvolatile memories are the endurance to repeated write cycling and the write cycle time. A representation of an analysis done by Qualcomm in June of 2014 of the endurance and write cycle time of various emerging embedded nonvolatile memories is illustrated in Figure 2.3 [3]. This analysis included not only conventional embedded floating gate Flash but also the various emerging memories that are in development and production including: STT-MRAM, FeRAM, PCM, and RRAM. STT-MRAM and FeRAM tend to have the longest

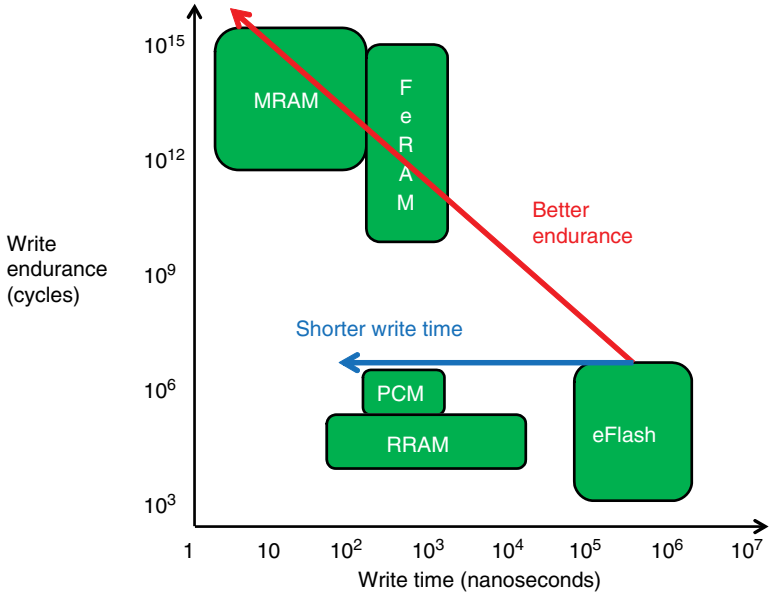


Figure 2.3 Write endurance (cycles) versus write time (ns) for different memory cell types including: SST-MRAM, FeRAM, PCM, RRAM, and embedded Flash memory. Based on S.H. Kang (Qualcomm), VLSI Technology Symposium, June 2014 [3].

write cycle endurance, between 10^{12} and 10^{15} cycles, and a shorter write cycle time. The 1T1C FeRAM, however, has a destructive read, which means that every read or write cycle affects the endurance. It is, therefore, better suited for applications with limited numbers of read cycles such as datalogging. The 1 T FeRAM does not have a destructive read but is early in its development cycle. PCM and ReRAM have lower endurances at 10^4 to 10^6 write cycles with about the same write cycle time as the FeRAM.

All of the emerging memories have superior characteristics to the conventional embedded Flash. The STT-MRAM and the FeRAM have better write endurance and shorter write times than the eFlash. The PCM and RRAM have a shorter write time but about the same endurance. The characteristics determine to a large extent the application for which the particular memory is better suited.

The growth of the Internet enabled devices has created many applications that rely on communication between multiple devices and data storage in these devices. This includes: wearable systems, systems for IoT, and search and storage for big data processing. Nonvolatile embedded memories have become a focus of attention due to the large amount of standby power required for embedded SRAM and the limitations of battery technology. Also of concern is the high program/erase voltages, slow read/write speed, and limited endurance of mainstream Flash memory.

A large number of IoT devices are expected. Gartner has forecast that 6.4 billion connected things (IoT) will be in use worldwide in 2016, up 30% from 2015 [4]. They further forecast that this will reach 20.8 billion by 2020. In January of 2013, Cisco ISBG estimated that 50 billion IoT products would be in existence by 2020 at a value in US$ of 14.4 trillion [5].

2.2.3 Required Properties of Embedded Nonvolatile Memories in Various Applications

While standalone nonvolatile memories have standardized requirements agreed on throughout the industry and defined in industry standards committees among users and vendors, embedded nonvolatile memories have reliability dependent on the specific circuit in which they are embedded and the applications for which the circuit is intended. Typical reliability requirements for embedded nonvolatile memories in general MCU applications were indicated by Samsung in May of 2014 and are shown in Figure 2.4 [6]. The special reliability requirements for smart card and automotive MCU applications were also indicated and are shown in Figure 2.5.

Further nonvolatile memory characteristics for various systems intended for IoT applications were discussed by Renesas in April of 2015 [7]. For automotive power train, auto body, and airbag applications, the embedded Flash memory was seen to be adequate for automotive system and performance requirements. As the number of connected systems and the functionality of those systems rose in 2015 for automotive systems, the performance increased for power train, body, and airbag applications. The embedded nonvolatile memory type preferred in Automotive Systems remained the eFlash macro, as shown in Figure 2.6, which indicates the automotive system and performance requirements in 2015.

Requirement	Feature
Performance (MHz)	25 – 75
Endurance (cycles)	10,000 – 100,000
Data retention (years)	10
Temperature (°C)	
General	0 to +70
Industrial	–40 to +85
Granularity	Block, byte erase

Figure 2.4 Typical reliability requirements for NV memory embedded in MCU. Based on Y.K. Lee *et al.* (Samsung), IMW, May 2014 [6].

Requirements	Smart card	Automotive
Endurance (Min.cycles)	100 k	100 k
Data retention (years)	10	20
Temperature (°C)	–40 to +85	
Body		–40 to +125
Under hood		–40 to +150
Read performance (MHz)	>20	>75

Figure 2.5 Special reliability requirements for nonvolatile embedded memory in MCU for smart card and automotive applications. Based on Y.K. Lee *et al.* (Samsung), IMW, May 2014 [6].

Requirement	Power train	Body	Air-bag
Performance (MHz)	300	200	100
Power (mA/MHz)	0.5	0.5	0.25
Temperature (°C)	–40 to +150	–40 to +125	–40 to +125
eFlash density (MB)	8	8	2
Code rewrite (cycles)	1k–10 k	1k–10 k	1k–10 k
Data rewrite (cycles)	>100 k	>100 k	>100 k

Figure 2.6 Automotive system and performance requirements in 2015. Based on T. Yamauchi (Renesas), VLSI-TSA, April 27, 2015 [7].

	Smart card	Medical sensor
Performance (MHz)	15–50	1–10
Power (mA/MHz)	0.2	0.1
Rewrite (cycles)	500 k	100 k
Memory density (KB)	512	256

Figure 2.7 Smart card and medical sensor requirements. Based on T. Yamauchi (Renesas), VLSI-TSA, April 27, 2015 [7].

These requirements are changing rapidly as the smart connected applications evolve and numbers given here are only indications at a point in time. Automotive electronics remains one of the largest of the Flash MCU sectors. New automotive applications include body and engine processors, infotainment, and driving aids up to completely autonomous cars. Automotive requirements are stringent and include a wide operational temperature range along with high performance and long endurance. This is an application that does not require ultralow power.

Smart card and medical sensors had specification requirements that could be satisfied by embedded Flash memory. For normally off sensor systems it was projected that embedded RRAM or embedded MRAM might also need to be used [7]. Smart cards are another application requiring ultralow power. This includes the rapidly growing bank card application and requires high reliability and secure communications for the card and banking network. Medical devices also required high reliability and secure communications. Smart card and medical device requirements are very different from those of the automotive applications. Power and performance requirements for these systems currently are shown in Figure 2.7. Medical sensors require lower power than smart cards, but do not require the performance at this time, as indicated in the figure.

Other IoT applications include: networks of digital utility meters, secure home networks, and networks of portable and wearable medical devices. The multiple sensors required in these connected applications require either ultralow power for infrequent changes of batteries or energy harvesting. These criteria are changing constantly and many other applications with differing requirements exist for Flash MCU. The smart

electric grid needs network controllers for digital utility meters. Home networks require security and reliable functionality. Secure wireless network chips need smart crypto-processors while high performance SSD requires integrated storage class memories.

The requirement for the handling of massive data generated by the billions of IoT devices requires large search engines made of nonvolatile memory Ternary CAMs (TCAMs) for high volume data storage in the cloud and at the edge of the Internet perhaps smart computers for processing local data for efficient use. The development of neuromorphic computers with learning capability can perhaps serve this application primarily in vision and hearing analysis. Neuromorphic computers in development are using the analog characteristics of emerging memory chips such as FeRAM, PCM, RRAM, and MRAM.

The predominant types of memory embedded in MCU today are SRAM and Flash NVM with some emerging memories such as FeRAM and RRAM beginning to be used for ultralow power applications. General capabilities of embedded Flash in MCU were discussed by Freescale (now NXP) in May of 2015 and ranged from low end 8-bit MCU in a 350 nm process, which may have 256 to 512 bytes of Flash, to a high end 32-bit automotive MCU in a 55 nm process with up to 8.5 MB of Flash memory [8]. Flash memory is used to store boot and application codes, device configuration parameters or personalization information. Requirements are application dependent.

The silicon costs associated with large amounts of embedded Flash in advanced logic processes have to compete with system-in-package processes. These systems interconnect discrete memories with an MCU often in stacked chip solution using copper pillar interconnects and through silicon vias (TSV). These packaging costs will come down with time and compete with true embedded Flash. IoT connectivity is causing situations where it is not possible to embed sufficient Flash in the MCU to satisfy the storage requirements for the required amount of data for the application. In this case, integrated MCU and nonvolatile memory system-in-package processes might be used. Wafer level integration of high performance memory and processor chips on flexible substrates is also an advancing technology.

Circuits used in IoT applications can be very specialized and require other specifications for the processor and embedded memory. Examples are ultralow power MCU, high speed MCU, and flexible high and low performance devices. In the next section the characteristics of MCU being introduced for some of these specialty applications that occur in the Internet of Things area are discussed.

2.3 Circuits Using Ultralow Power MCU with Embedded Memory for Energy Harvesting

2.3.1 Introduction to Ultralow Power MCU Using Energy Harvesting

Ultralow power MCUs are used today in systems powered by coin batteries or energy harvesters where electrical system energy is not feasible. These MCU serve a variety of IoT applications like wireless home automation, security systems, and wearable medical electronics. In future IoT networks, these applications are more likely to be powered by energy harvesting or coin cell batteries and, as a result, require electronics that consumes very low energy.

The Internet of Things (IoT) uses sensors and MCUs on various devices that are connected through the Internet. With IoT, trillions of sensors are used on devices for collecting information such as temperature, humidity, motion, strain, magnetic field strength, etc. In some cases it is feasible to harvest solar, electromagnetic, or mechanical energy and store it in batteries or capacitors to power these devices.

2.3.2 Ultralow Power MCU with Embedded Flash Memory for Energy Harvesting

Many wireless sensor applications make use of energy harvesting. Ultralow power sensor nodes in wireless sensor networks are being developed for applications such as wearable devices.

A Flash MCU intended for energy harvesting applications was discussed by ST Microelectronics in January of 2014 [9]. This MCU was intended for cost-sensitive applications and used a low end RISC processor core. Target applications included IoT and smart energy devices. This Flash MCU had 16KBytes to 128KBytes of embedded Flash memory and was expected to bridge the 8-bit and 16-bit MCU world. The MCU had crystal-less USB2.0 and CAN interfaces with a self-calibrated clock system. It also had a battery charger detection system, controlled power management, and separate power for analog and digital I/Os to support low voltage digital and high voltage analog operation.

2.3.3 Ultralow Power MCU with Embedded FeRAM Memory for Energy Harvesting

It was expected that the energy harvesting applications would require the use of emerging memories with microampere operating and nanoampere standby capabilities. A Texas Instrument's MCU with an embedded two transistor and two ferroelectric capacitor (2T2C) FeRAM cell was used as an ultralow power processor for this application [10]. This MCU ran up to 25 MIPS. Clock frequencies could be relatively slow ranging, in this case up to 16 MHz. The power supply was specified for battery operation from 1.8 to 3.6 V. The device was optimized for ultralow power mode operation including a 1.7 μA standby current and a 15 nA shutdown current. Active power was typically 126 μA/MHz. The embedded FeRAM memory offered 10^{15} cycle read and write endurance. The electrical characteristics of the MCU with embedded FeRAM for energy harvesting are summarized in Figure 2.8.

For computing devices that are transiently powered, such as in energy harvesting systems, the logic state needs to be retained on power down. Such a computing device depends on an intermittent power supply that has bursts of energy that are very short, often less than 100 ms. Purdue University in January of 2014 discussed such transiently powered computers (TPCs) [11]. The primary concern when using conventional embedded Flash memory with higher power processors is that frequent system reboots can result. An SRAM or Flash memory could be used with a battery, but the battery would tend to run down due to the long erase/write time of the Flash memory. A method was followed called Quick Recall, which used ferroelectric memory to enable long running computations in TPCs. This method saved and restored a checkpoint in only 12.6 μs, which is a significantly shorter time than would be required with Flash

Device	Feature	Characteristic
MCU	RISC	16-bit
	Clock speed (MHz)	16
	Throughput (MIPS)	25
	Power supply (V)	1.8–3.6
	Active power	
	Typical (uA/MHz)	126
	Standby power (μA)	1.7
	Shutdown current (nA)	15
eMemory	Ferroelectric RAM (FeRAM)	2T2C
	Read & write endurance (cycles)	10^{15}

Figure 2.8 Characteristics of an embedded FeRAM MCU for energy harvesting sensor applications. Based on Texas Instruments Press Release, March 24, 2015 [10].

memory. A Texas Instruments MCU with embedded FeRAM was used. The system showed a significant improvement in program execution time over using an MCU with embedded Flash memory.

Checkpointing performed at previously determined points in the program stored a snapshot of system state in the FeRAM. In case of a fault, the system went back to the most recent check point. To avoid interfering with the normal computations, checkpointing with FeRAM is done only when a drop in the supply voltage occurs. The checkpointing was completed before the power was lost by choosing an appropriate trigger voltage for the FeRAM to begin the checkpointing operation. The program state retained was that of the global variables in use by the program. This study showed that transiently powered computers could do computations even when they receive power for periods as short as 5 ms.

An energy harvesting device can also be made using a low power single ferroelectric transistor (FeFET) memory. An example of an energy harvesting device using an FeFET was given by Panasonic in June of 2014 [12]. The advantage of the FeFET over the 1T1C FeRAM is that the endurance is not affected by a read cycle since the device does not require rewrite after read.

The FeFET energy scavenger used a low cost cantilever-type 28 μm piezoelectric polyvinylidene fluoride (PVDF) film with a screen printed silver (Ag) ink electrode laminated on to a 0.125 mm polyester substrate. The PVDF cantilever converted energy from mechanical vibration into electric power. A reusable battery-less impact logging device using a vibration energy scavenger and an FeFET was shown to be functional and suitable for portable use. Figure 2.9 shows the PVDF cantilever in the circuit with the FeFET memory.

The FeFET used a ferroelectric material as a dielectric layer in a MOS FET structure, as shown in Figure 2.10. The FeFET consisted of a stack of ferroelectric Pb(Zr, Ti)O_2 (PZT) perovskite oxide film on a 30 nm thick $SrRuO_2$ bottom gate deposited on a sliced stacked trench oxide (STO) substrate by pulsed laser deposition. This combination enabled oriented growth of ZnO on top of the PZT layer with a sharp ZnO/PZT interface.

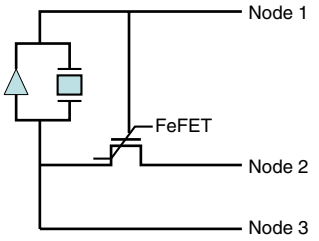


Figure 2.9 Circuit schematic of PVDF piezoelectric vibration energy scavenger cantilever in circuit with FeFET memory. Based on Y. Kaneko *et al.* (Panasonic), DRC, June 22, 2014 [12] (permission of IEEE).

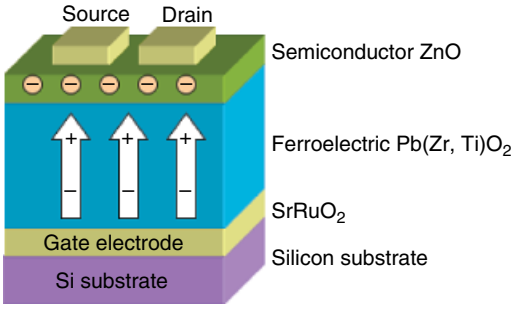


Figure 2.10 Schematic illustration of an FeFET with stacked ZnO/PZT/$SrRuO_2$ structure. Based on Y. Kaneko *et al.* (Panasonic), DRC, June 22, 2014 [12] (permission of IEEE).

2.3.4 Ultralow Power MCU with Embedded RRAM Memory for Energy Harvesting

When solar energy is harvested to power an IoT application, a boost converter can be used to increase the solar battery voltage so a nonvolatile memory can be programmed and the sensor data stored. In November of 2015, Chuo University discussed a boost converter for RRAM programming in low power IoT applications [13]. An embedded RRAM was used for the code and data storage since it has faster program speed and smaller program voltage at 3 V than NAND Flash, which uses about 20 V for programming. A RRAM cell was assumed to use 3.0 V SET voltage and 25 μA program current. A representation of the 1T1C structure of the RRAM cell array is shown in Figure 2.11.

2.3.5 Ultralow Power MCU for Energy Harvesting Power Management

Another circuit using solar energy harvesting was discussed in August of 2015 by Cypress Semiconductor [14]. This family of energy harvesting power management integrated circuits (PMIC) consisted of single chip solar cells as small as 1 cm^2 in area. These devices were fully integrated circuits making them a potential for use in smart homes, commercial buildings, and factories. The circuits were able to monitor physical and environmental conditions for these smart buildings.

The solar cells were naturally dependent on placement and available light for energy harvesting applications. Startup power was 1.2 μW and current consumption was as low as 250 mA, which optimized the power available for sensing, processing, and communications depending on the application. A complete, battery-free energy harvesting system could pair the PMIC with a Bluetooth module for low energy connectivity.

Another energy harvesting device was introduced by TI in 2016 [15]. It was designed to efficiently extract microwatts to milliwatts of power generated from a variety of high

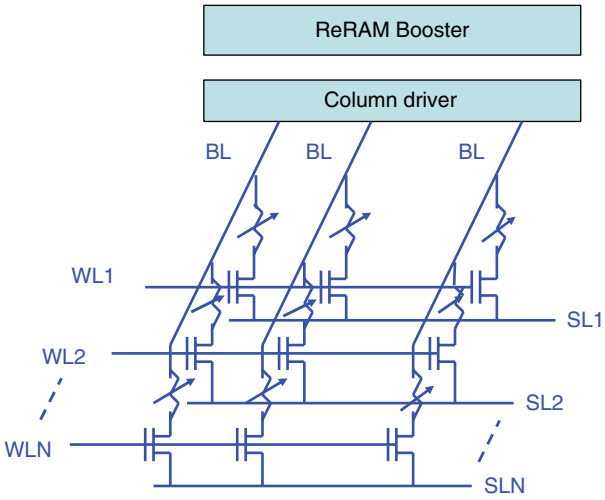


Figure 2.11 Schematic circuit diagram of the 1T1C resistive RAM (RRAM) cell array structure. Based on T. Ishii *et al.* (Chuo University), IEEE A-SSCC, November 2015 [13].

output impedance DC sources, such as solar generators or thermal electric generators, without collapsing these sources. Battery management features were used to ensure that the associated rechargeable battery was not overcharged by the extracted power or depleted beyond safe limits by a system load. The cold start voltage was 330 mV. The device allowed continuous energy harvesting from a source as low as 100 mV. Full operating quiescent current was typically 488 nA and the maximum charge current was 100 mA.

2.4 Ultralow Power Battery Operated Flash MCU

2.4.1 Introduction to Ultralow Power Battery Operated Flash MCU

Ultralow power MCUs serve a variety of IoT applications like energy metering, home automation, security systems, industrial control, and portable medical electronics, as indicated in Figure 2.12. These applications are likely to be either powered by energy harvesting or by low current batteries and require very low energy operation along with zero power standby and rapid wakeup from standby.

Some emerging memory devices, such as RRAM and FeRAM, switch faster than Flash memory, operate at lower voltages, and have a lower operating current. Like Flash memory, they are nonvolatile. MCU with these emerging memories embedded in the processor chip could provide extended battery operating life for portable devices if they are also capable of being manufactured in conventional CMOS processes at low cost.

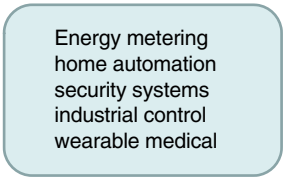


Figure 2.12 Ultra-low power IoT battery operated applications used with emerging memory technologies.

Ultralow power battery operated Flash MCU require low current consumption to prolong battery life. This low current can be obtained by various current detection and

-Current detection circuits
-Power management control circuits
-Separate power for analog and digital circuits
-Low current EEPROM macros
-Fast wake-up from standby feature
-Longer times in low power states with fast switching

Figure 2.13 Power lowering techniques for ultralow power battery operated Flash MCU systems.

power management control circuits. Separating power for analog and digital circuits can help. Using lower current EEPROMs can also reduce current consumption. Operational modes such as fast wake-up from standby and longer times in low power states combined with fast switching can also reduce power. Various of these power lowering options that have been used are shown in Figure 2.13. These include: current detection circuits, power management control circuits, separate power for analog and digital circuits, low current EEPROM macros, fast wake-up from standby and longer times in low power states with fast switching.

2.4.2 Ultralow Power Battery Operated Flash MCU with Embedded Flash Memory

An MCU with embedded Flash memory, targeted at cost-sensitive battery powered applications, was announced in January of 2014 by ST Microelectronics [16]. This MCU used 16 KB to 128 KB of embedded Flash memory and a low end RISC processor core. Target applications included IoT and smart energy devices. This series was expected to bridge the 8-bit and 16-bit world. The MCU had crystal-less USB 2.0 and CAN communication interfaces with a self-calibrated clock system. To help control power consumption, it had battery charger detection and system controlled power management. Separate power for analog and digital I/Os helped to support low voltage digital and high voltage analog operation.

In order to target applications operating on smaller batteries such as wearable devices and industrial sensor networks, lower power consumption was required. In February of 2014, STMicroelectronics announced an ultralow power Flash MCU featuring a 32 MHz RISC processor core, a 100 k sps ADC consuming 40 μA, and crystal-less USB FS2.0 [17]. Up to 64 KB of Flash memory could be integrated along with 2 KB of embedded EEPROM in the CMOS logic technology. Target applications included fitting smaller coin batteries on portable medical equipment such as fitness trackers and glucose meters. The analog-to-digital converter (ADC) consumed 40 μA at 100 k sps and up to 200 μA at 1.14 M sps, so a coin battery could operate a fitness tracker for two years. The CMOS technology had low variation over the –25 to 125 °C temperature range.

An energy efficient MCU with RISC core and embedded Flash memory intended for use in ultralow power IoT applications was introduced by Cypress in November of 2015 [18]. The device was optimized for advanced wearables along with a range of ultralow power battery-powered products. The MCUs offered up to 560 KB of Flash memory and 64 KB of SRAM. The devices had fast wake-up time from standby, which minimized power consumption by being able to remain in low power states for longer times before switching briefly to active mode. Active current was 40 μA. The device offered the

ability to move data between peripherals and RAM without CPU intervention. Communication interfaces offered advanced encryption standard (AES) hardware encryption. AES is a symmetric block cipher used to protect classified information and to encrypt sensitive data. The communication interfaces included: multifunction serial interfaces, USB, a smart card interface, and I2C for digital audio. Some analog peripherals were included.

2.4.3 Ultralow Power Battery Operated MCU with Embedded RRAM

Ultralow power MCUs that are operated on battery optimize the current available by storing data at a low current level during the write operation so the optimum data storage is obtained. An RRAM discussed by Panasonic in July of 2013 was embedded in an 8-bit MCU and was intended to replace Flash memory [19]. It had stable low current operation as a result of the electric current being kept constant during the RRAM forming process so that the resulting filaments were uniform. A metal oxide RAM (OxRAM) technology was used for the RRAM. This RRAM used Ta_2O_5 to control oxide density and defects during the deposition of a tantalum layer. The intended battery-backed applications require low power operation and include devices such as sensors, security systems, and portable healthcare devices. Electrical and reliability characteristics of 8-bit MCU with embedded RRAM intended for use in battery powered circuits are shown in Figure 2.14.

The MCU with the embedded RRAM operated at voltages as low as 0.9 V, which permitted low power consumption. Standby current was 60 nA with a 3 V power supply and an operating current of 200 nA during real time clock operation. Compared to a comparable MCU with embedded Flash (eFlash), the power consumption was about half. In a comparable eFlash MCU, the write voltage was given as 2.7 V and the write speed as 30 μs/byte, while in the MCU with eRRAM, the write voltage was 1.8 V and the write speed was 3 μs/byte.

Process technology	
CMOS Logic	130 nm
RRAM	Ta2O5 OxRAM
Electrical characteristics	
Read voltage	0.9 V
Write voltage	1.8 V
Standby current	60 nA @3 V Power
Operating current	200 nA (with real time clock)
Write speed	3 μs/byte
Power	4 uW (low speed active mode)
Reliability characteristics	
Endurance (rewrite)	100,000 cycles

Figure 2.14 Electrical and reliability characteristics of 8-bit MCU with embedded RRAM for use in battery powered circuits. An oxide RAM (OxRAM) technology was used for the RRAM. Based on Panasonic Press Release, July 30, 2013 [19].

With the embedded RRAM, the device is turned on for a much shorter time and at lower voltage so less current is consumed. The 'ewrite cycle endurance for the eRRAM was 100 000 cycles, which is comparable to the eFlash. The RRAM still needed to be installed in the back-end process, after which it was expected to be usable at various process nodes. The initial production was in Panasonic's 180 nm process. In February of 2017 Panasonic and UMC in Taiwan began joint development of a 40 nm RRAM platform [20].

2.4.4 Ultralow Power Battery Operated MCU with Embedded FeRAM

An embedded FeRAM in a low power MCU was discussed by Texas Instruments in June of 2013 [21]. Standby current consumption was reduced to 350 nA by using critical blocks and other architectural tailoring. Fast wake-up within 6.5 μs from standby mode or sleep mode was achieved using analog precharge current sources. For the lowest system energy, power management was divided into blocks.

A power management block diagram of the MCU is shown in Figure 2.15. This block diagram illustrates the details of the dedicated power subsystem. The system is partitioned into different power domains in order to power-off circuitry that is currently unused. A low dropout (LDO) linear regulator is used as a low cost, accurate power supply. The real time clock (RTC) domain is separated so only 350 nA of current is consumed while the RTC is running. The memory FeRAM LDO is also separated. The core domains are power gated. The precharge current sources directly source the power gated domains from the 3 V VCC supply for fast startup of the power gated domains.

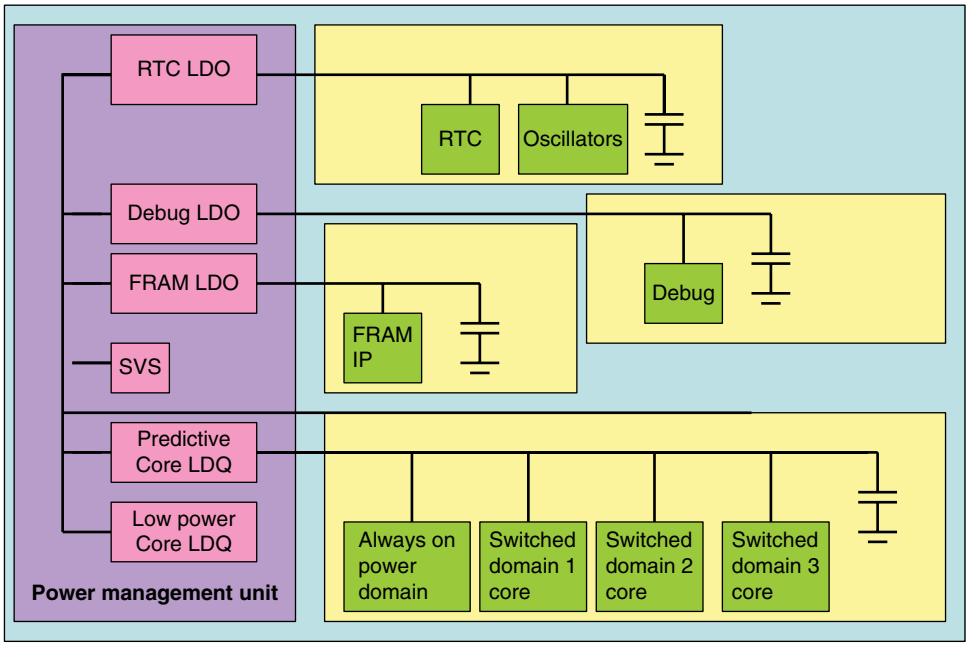
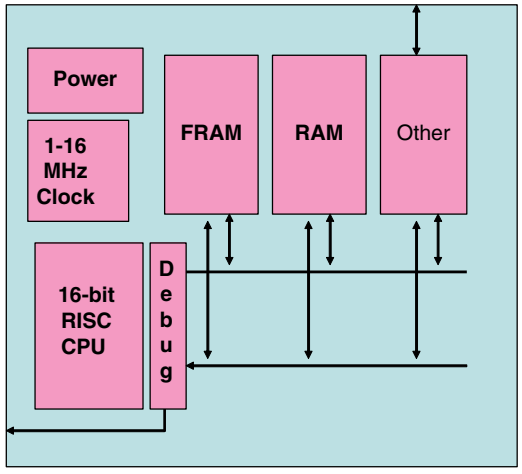


Figure 2.15 Power management block diagram of the low power FeRAM MCU. Based on A. Baumann *et al.* (Texas Instruments), VLSI Circuits Symposium, June 12, 2013 [21].

Figure 2.16 Features and target applications of 16-bit MCU with embedded FeRAM intended for use in battery powered circuits. Based on A. Baumann *et al.* (Texas Instruments), VLSI Circuits Symposium, June 12, 2013 [21])

Product	
Type	16-bit MCU + FeRAM
Chip size	0.12 mm^2
Process technology	
CMOS Logic	130 nm
FeRAM Capacity	64 K Byte
Electrical characteristics	
MCU Speed	16 MHz
Power supply (VCC)	3 V
Standby current	350 nA
Wake-up from standby	6.5 μs
Target applications	
Medical health, fitness and building automation	

Figure 2.17 Block diagram of the MCU with eFeRAM. Based on A. Baumann *et al.* (Texas Instruments), VLSI Circuits Symposium, June 12, 2013 [21].



The major blocks consist of the 16-bit MCU core and digital logic with dedicated power domains for debug and the real time clock (RTC), power management module (PMM), and analog and memories including FeRAM, RAM, and ROM. Features and target applications of the 16-bit MCU with embedded FeRAM intended for use in battery powered circuits are shown in Figure 2.16.

The 130 nm CMOS technology used 5-layer copper (Cu) metal interconnects. Chip size was 0.12 mm^2. The MCU ran at 16 MHz and the FeRAM capacity was 64 KB. A block diagram of the MCU with eFeRAM is shown in Figure 2.17. Blocks shown are: the clock, nonvolatile FeRAM, RAM, and 16b RISC processor.

Many IoT sensing applications require extended battery life for sensing and measurement applications. In March of 2016, TI announced an MCU with embedded FeRAM that offered a configurable low leakage transimpedance amplifier with 20 times lower leakage than previous solutions and therefore potentially offered an extended battery life [22]. The MCU included a configurable analog, which reduced printed circuit board (PCB) space. The analog integration permitted connecting to a wide range of sensors. It integrated a ferroelectric RAM nonvolatile memory. The fast read and write of the FeRAM permitted configuring the amount of memory allocated for application code or

data by eliminating the boundaries between Flash and RAM. The FeRAM MCU offered a higher performance than the previous generation MCU, which used Flash. Applications targeted included: medical health and fitness and building automation. TI proposed adding an energy harvesting IC to extend the battery life even further.

2.5 Nonvolatile MCUs Using Emerging Memory for Nonvolatile Logic

For MCUs in battery powered or energy harvesting applications, it is essential that standby power be ultralow. This can be accomplished in the logic by power gating, which is turning off the power in a block of circuits. If a nonvolatile memory cell is associated with a logic node then the logic state of that node is stored in the event of power-down and can be restored on startup. This latter technique essentially eliminates standby power. In this case, the MCU does not need to be supplied with power during standby and comes up in the previously stored state when the power is restored.

2.5.1 Nonvolatile Logic Arrays Using FeRAM

The state of a nonvolatile data Flip-Flop (nvDFF) can be stored in an embedded FeRAM capacitor. In February of 2013, MIT and TI discussed the design of an nvDFF with embedded ferroelectric capacitors that sense data [23]. The nvDFF was integrated into the ASIC design flow with a power management unit and a simple interface to the brown-out detection circuitry. This design managed power by operating normally "off". The process technology included embedded ferroelectric (FE) capacitors that store data. Prior to sensing, the FE capacitors are programmed to opposite data states. Identical charging currents integrate the difference in remnant charge between the two FE capacitors, as shown in the circuit schematic in Figure 2.18. The first node to cross the diode voltage drop plus a PMOS threshold pulls the internal node of the sense latch high.

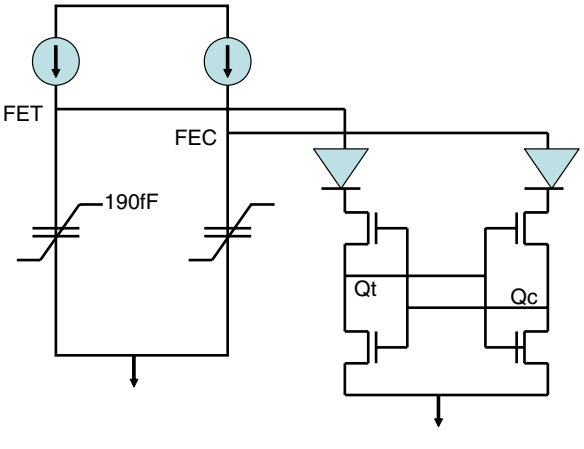


Figure 2.18 Illustration of ferroelectric (FE) capacitors connected to a nonvolatile data flip-flop (nvDFF). Based on M. Qazi *et al.* (MIT, TI), ISSCC, February 2013 [23].

The FIR filter was designed with 96 nvDFFs and about 500 gates. For the target application, which was an MCU, about 5000 nvDFFs were retained. The nvDFF has a split supply rail so sensing current ramps can be perturbed. With zero skew, all nvDFFs in the five measured chips, a total of about 21 000 nvDFFs, operated without failure and the distribution of skew indicated a below 1 ppm failure for an unskewed nvDFF. The additional energy cost from nodes glitching in the FIR filter, cycle overhead, and programmable memory unit (PMU) energy is about 1.780 pJ out of 3.439 pJ.

Commercial ultralow power (ULP) MCUs need to balance different requirements by providing high performance and very low standby power along with fast wake-up time for various real time applications. In November of 2013, TI and TSMC discussed a full high voltage threshold (HVT) 8 MHz 75 μA/MHz nonvolatile logic (NVL)-based MCU with zero standby power and fast 384 ns wake-up time [24]. In this processor, nonvolatile miniarrays are distributed though the logic domain of the MCU, which take a snapshot of the state of all sequential elements before the MCU goes into a power-gated standby mode. Upon wake-up, no boot-up is required.

These miniarrays use ferroelectric capacitors for fast low power nonvolatile state retention. The ferroelectric capacitors are programmable capacitors that retain their programmed state without a power source. The read plus write endurance of the FeRAM is $>10^{15}$.

A high bandwidth parallel connection between the flip-flops and miniarrays helped achieve a fast MCU wake-up. Nonvolatile logic (NVL) had no impact on MCU active mode performance and power. It added only 3.6% to the chip area. Since the NVL eliminates leakage in standby mode, it permits use of high performance leaky processes in MCU design.

The system architecture of the MCU is shown for a second generation full standard voltage threshold (SVT) 32 MHz nonvolatile logic MCU in the block diagram in Figure 2.19. The SVT SoC has four times higher active mode performance or 30% lower active energy than the HVT SoC, but can still achieve zero leakage using the nonvolatile logic. This created a zero leakage fast wake-up standby mode and also provided benefits in active power and performance. In particular, it is expected to be useful for energy harvesting applications.

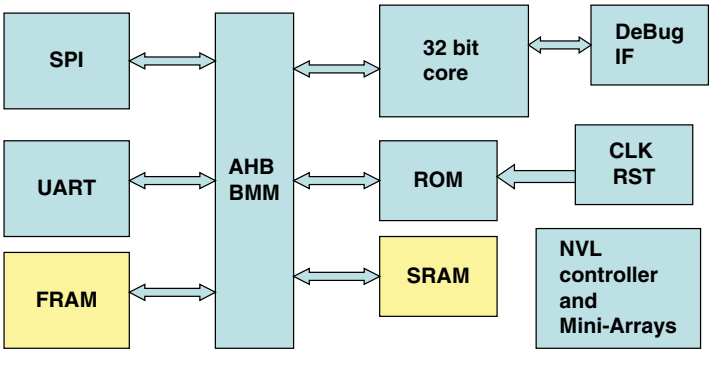


Figure 2.19 System architecture of the MCU showing 32b core and nonvolatile logic (NVL) controller. Based on S. Khanna *et al.* (TI, TSMC), A-SSCC, November 11, 2013 [24].

This 130 nm nonvolatile logic (NVL) based 32b MCU SoC backs up its working state, which includes all flip-flops, upon receiving a power interrupt. TI reported in January of 2014 that this MCU has zero leakage in sleep mode and requires less than 400 ns to restore the system state on power-up [25]. A low power technology is used with a 1.5 V power supply. Nonvolatile ferroelectric capacitor-based miniarrays back up the machine state and permit the chip to wake up instantly after a power cycle. Without the NVL, a chip would either need to keep all flip-flops powered, which would result in high standby power, or the system would need to spend energy and time rebooting after each power-up.

NVL permits systems to use leakier processes to achieve higher performance and lower dynamic power while retaining zero leakage in sleep mode. In this case, the addition of the NVL only adds 3.6% to the SoC area. Process cost adders are minimized since only two additional masks are required for the ferroelectric NVL implementation. A functional diagram of the NVL SoC is shown in Figure 2.20.

The voltage regulation, supply detection and PM state machines are off-chip. The MCU SoC discussed has zero sleep mode leakage with a 400 ns wake-up time. The chip has 100% digital state retention. System wake-up time is reduced from 100 s of microseconds to a few microseconds since no boot-up is required after a power cycle. Before entering the sleep mode, data from all flip-flops and latches in the system are transferred to the ferroelectric capacitor-based NVL arrays.

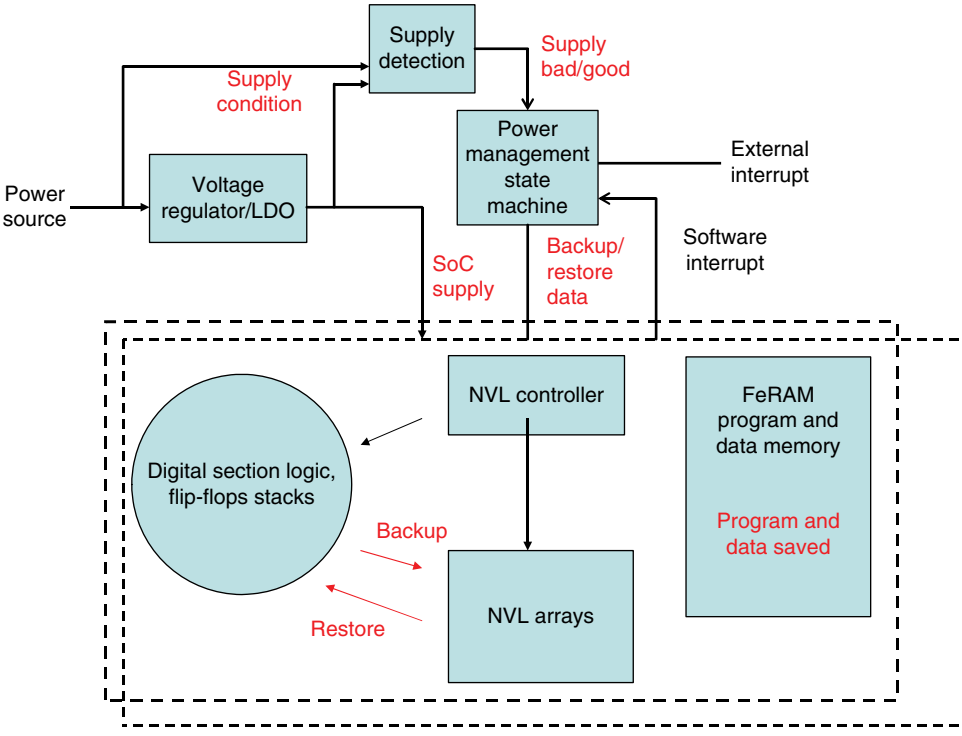


Figure 2.20 Functional diagram of NVL 32-bit SoC with FeRAM program and data memory. Based on S. Khanna *et al.* (Texas Instruments), IEEE JSSC, January 2014 [25].

The NVL can be used in any digital system with two system requirements: the system must be edge triggered synchronous and the clocks for all flip-flops must be stopped in the inactive state while going into the sleep mode. The SoC has conventional FeRAM memory. The NV elements added to the SoC are tested and the reliability ascertained. NVL bit-cell read signal margins are measured and there is a test and debug capability.

2.5.2 Nonvolatile Logic Arrays Using MTJ MRAM

Another processor using power gating and a nonvolatile array was described in November of 2013 by Tohoku University and NEC [26]. This power-gated MPU uses a nonvolatile flip-flop (nvFF) with a magnetic tunnel junction (MTJ) MRAM type of memory. The nvFF was used to store the MPU's internal state. The power-gating operation for this MPU had a 3 μs entry/exit delay penalty during power on and power off. This was one order of magnitude faster than a conventional MPU in deep power-down mode. An nvFF circuit was developed to obtain this small entry/exit delay with stable fast store and recall. The power-gated mode was easy to control. This processor core was based on a 32-bit pipeline architecture.

In the power-gated MPU, all necessary processor internal states are stored in NV memory. This includes executing programs in the instruction memory and some temporary calculation results in the register file. This means the MPU can immediately return from the suspended state and restart program execution. This move into and from the power-off mode does not require obtaining or storing data from external memories, which makes the entry/exit delay short. A nonvolatile latch circuit diagram using MTJ MRAM elements in CMOS technology is shown in Figure 2.21.

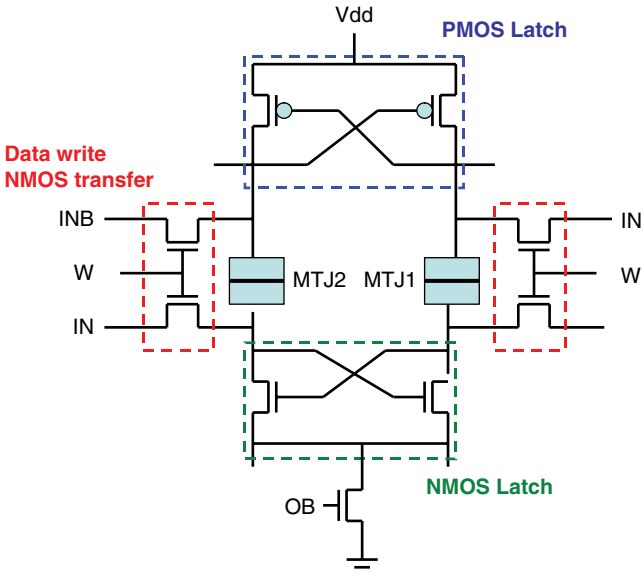


Figure 2.21 Basic NV latch circuit diagram using magnetic tunnel junction (MTJ) nonvolatile memory elements in CMOS technology. Based on H. Koike *et al.* (Tohoku University, NEC), A-SSCC, November 11, 2013 [26]. (Permission of IEEE).

A detailed SPICE simulation analysis of a CMOS/MTJ hybrid circuit was executed and the design and fabrication of a prototype MPU chip was done. Results showed that power gating had a 3 μs penalty for entry/exit delay. This was an order of magnitude faster than for a conventional MPU deep power-down.

Processors with MTJ-MRAM nonvolatile arrays can also be used for intermittent operations.

A fully nonvolatile MCU with an embedded emerging NV-RAM memory was discussed by NEC and Tohoku University in February of 2014 [27]. This 90 nm CMOS 16-bit MCU used a three transistor SpinRAM memory consisting of two transistors and one MTJ MRAM (2 T + 1MTJ) and was intended for use in applications such as battery-powered sensor nodes. The MCU operated at 20 MHz with 145 μW operation with a 1 V power supply in the active state and 4.5 μW intermittent operation with a 120 ns wake-up time and 0.1% active ratio without needing to forward reboot code from memory. Power consumption was sufficiently low that battery powered sensor nodes were effectively free of maintenance. A block diagram of the MCU with embedded SpinRAM memory is shown in Figure 2.22.

The MCU had a 16b RISC CPU core, a 64 KB RAM/ROM unified SpinRAM macro and other logic circuits including timers, a 32-bit multiplier, a power management module (PMM), universal serial interfaces (USCI), and a direct memory access (DMA) module. It had some redundant words and columns to replace defective cells and an error correction code (ECC) circuit for write failures.

To eliminate backup and restore overhead through the memory bus, 4096 nonvolatile magnetic flip-flops (MFF) were used to capture the context of the CPU. Two instructions permitted software to flexibly back up and restore data to the MFFs. V_{cc} was 1.8 V to 3.3 V with a 1 V internal power supply provided by a DC/DC converter. The MCU supported three low power modes: standby mode, power-gating mode and sleep mode. In standby mode, all cores remained on and the main clock was gated, resulting in a 117 μW leakage current. In power-gating mode, 1.6 μW was consumed with all cores turned off. In sleep mode the DC/DC converter was turned off and static power was zero.

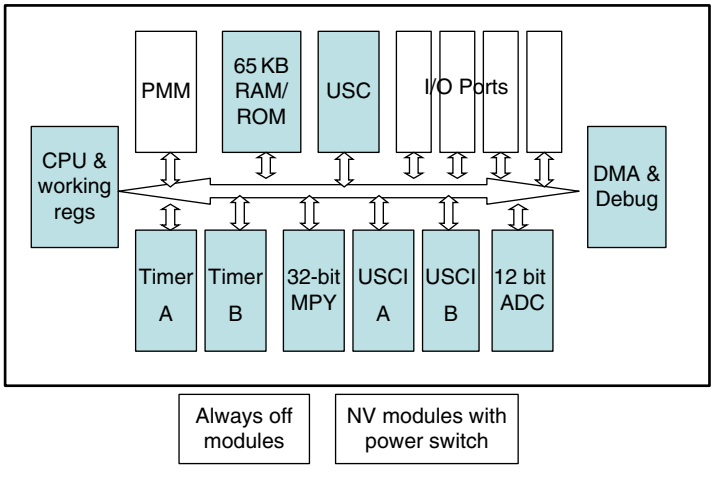


Figure 2.22 Block diagram of nonvolatile 65 KB RAM/ROM unified (2T+1MTJ) SpinRAM MCU. Based on N. Sakimura *et al.* (NEC, Tohoku University), IEEE ISSCC, February 2014 [27].

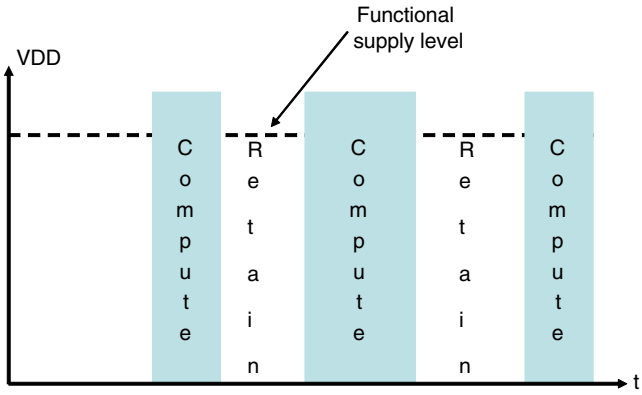


Figure 2.23 VDD versus time of computations in energy harvesting type system with NVDFF showing computations interspersed with times of retained data. Based on M. Qazi *et al.* (Cypress, MIT, TI), *IEEE Journal of Solid-State Circuits*, January 2014 [28].

A similar digital system where there is no distinction between ON and OFF and where the computational state is stored in NV Logic elements was discussed in January of 2014 by Cypress, TI, and MIT [28]. This system was intended to operate from harvested energy without requiring reboot. A nonvolatile D-Flip-Flop (nvDFF) was designed in 130 nm CMOS technology. The nvDFF was integrated into an ASIC design flow.

A test case nonvolatile FIR filter is discussed with a power management unit that automatically saved and restored the state of a one-bit indicator of energy availability. Correct operation was verified over power cycle intervals from 4.8 μs to 1 day. The round trip save–restore energy is 3.4 pJ per nvDFF. Measurements were made to validate the capability to make a 10 parts per minute (ppm) failure rate for embedded system applications.

An illustration of the designed operation of nonvolatile processing is shown in the graph in Figure 2.23, which shows VDD plotted against time. The system performs useful computations in between unpredictable interruptions in the availability of harvested energy. During an interruption the data are retained.

2.5.3 Processors with RRAM for Nonvolatile Logic Arrays

The power of many SoCs tends to be dominated by the leakage power of the embedded memories and the status registers. Leakage current and active power can be reduced by power supply scaling to the near threshold range or even into the subthreshold range. Low power memories can also help reduce the leakage. RRAMs are nonvolatile and therefore have low standby power dissipation and can be operated at low voltages. In November of 2014, EPFL, the Swiss Federal Institute of Technology, discussed energy and reliability trade-offs in a low voltage RRAM-based nonvolatile flip-flop (nvFF) design [29]. In this study, RRAM nvFF are optimized for operation in the low voltage and subthreshold ranges. A 180 nm CMOS process was used with a compatible voltage of 2 V and a low current compliance of 10 μA.

Three low voltage nvFF circuits were evaluated for power dissipation and reliability. Topologies are used that have two complementary programmed RRAM devices. Monte

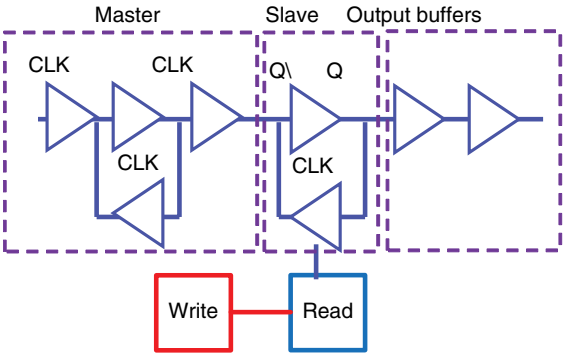


Figure 2.24 Conventional master–slave nonvolatile flip-flop. Based on I. Kazi *et al.* (EPFL), *IEEE Trans. on Circuits and Systems*, November 2014 [29].

Carlo simulations are used to determine parametric variations and confirm data restore operations from the RRAM devices in subthreshold ranges as low as 400 mV. Minimal energy is found for a near voltage threshold (V_t) read operation with a total read plus write energy of 735 fJ. A block diagram of a nonvolatile flip-flop is shown in Figure 2.24. This is a conventional master–slave flip-flop with write and read registers indicated.

RRAMS can be used in field programmable gate arrays (FPGA) as well as nvFF. In December of 2014, CEA and LETI Minatec discussed RRAMS for both FPGA and for nvFF [30]. In FPGAs, replacing the SRAM with RRAM saves area since the cell size of the RRAM is smaller than that of the SRAM and the nonvolatility suppresses standby power consumption. The leakage current through the RRAM during operation depends on the high resistance state (HRS) value. For this reason, a specific type of RRAM, a conductive bridge RAM (CBRAM) with a dual layer electrolyte stack and resistance ratio (high resistance state divided by low resistance state) greater than 10, was suggested to minimize operating leakage current. The resistance ratio is also called the ON/OFF current ratio. The CBRAM was made of a dual layer of solid electrolyte, HfO_2 and GeS_2, embedded between an Ag electrode and a W plug.

The use of RRAM in fixed logic circuits such as nvFFs permits zero standby leakage and fast power ON/OFF operation. The requirements for an RRAM in an nvFF are: endurance, low operating voltage, and fast ON/OFF operation. An HfO_2/Ti metal oxide RRAM cell, an OxRAM, with less than 10 ns switching time at 1 V and endurance up to 10^8 cycles was also suggested for this application. This OxRAM's active layer was HfO_2 between TiN and Ti/TiN electrodes. A schematic cross-section is shown in Figure 2.25. Figure 2.25(a) shows the OxRAM stack and Figure 2.25(b) shows the CBRAM stack with GeS_2 and with the HfO_2/GeS_2 electrolyte with an Ag top electrode and a W bottom electrode. Devices with a large memory window such as the HfO_2/GeS_2 CBRAM enable high density, low leakage cells for FPGA. Cells with low voltage, time switching, and high endurance such as the HfO_2/Ti OxRAM permit a fast switching fixed logic design.

Nonvolatile SRAMs (nvSRAM) are conventionally constructed using two nonvolatile memory devices and a 6T SRAM. In June of 2015, several Taiwanese labs and universities discussed an RRAM-based 7T1R nonvolatile SRAM [31]. The nvSRAM had a two times reduction in stored energy and 94 times reduction in restore energy for frequent-off and instant-on applications. The store energy is reduced by using only a single RRAM device rather than the two normally used. Also the DC short current is suppressed during the restore operation by using a pulsed overwrite method. A circuit schematic of the 7T1R nvSRAM cell is shown in Figure 2.26.

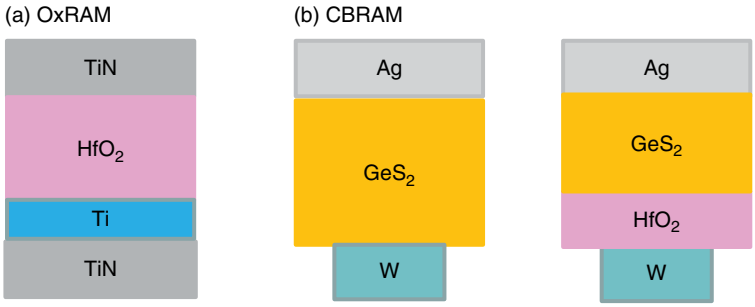


Figure 2.25 Schematic of (a) OxRAM RRAM stack and (b) CBRAM with HfO_2/GeS_2 electrolyte. Based on E. Vianello *et al.* (CEA LETI), IEDM, December 2014 [30].

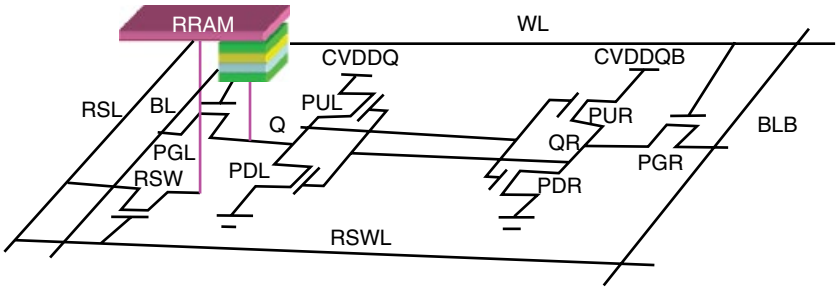


Figure 2.26 3D circuit schematic illustration of 7T1R nvSRAM. Based on A. Lee (NTHU, NDL, NCHU, EOL ITRI), Symposium on VLSI Technology, June 2015 [31].

A high restore yield was attained by using a differentially supplied initialization method. This initialization and overwrite 7T1R nvSRAM improved breakeven time by 6 times compared to previous nvSRAMS. A 16 K-bit 7T1R nvSRAM was made using an HfO_x RRAM in a 90 nm CMOS process. In SRAM mode the read operation was the same as for the 6 T SRAM. The separation of CVDD lines permitted applying different voltages to VDDQ and VDDQB to improve the write margin.

In February of 2016, Tsinghua University, NTHU, and the University of California, Los Angeles, discussed a 65 nm RRAM enabled nonvolatile processor that had a 6 times reduction in restore time and a 4 times higher clock frequency when using Adaptive Data Retention and Self-Write termination [32]. The processor used an adaptive nvFlipFlop with RRAM, a code storage RRAM macro, and an adaptive nvSRAM using RRAM. The system architecture of the RRAM enabled nvProcessor is shown in Figure 2.27.

A 65 nm fully logic compatible RRAM based nonvolatile processor (NVP) achieved time and space adaptive data retention. A one macro nvSRAM with self-write termination (SWT) was integrated to increase the clock frequency and reduce stored energy. The adaptive retention made the RRAM write endurance, which was 10^6–10^{12} cycles, sufficient for most applications. The NVP operated at 100 MHz with 20 ns per 0.45 J restore time/energy. It had a 6 times reduction in restore time, more than 6000 times reduction in restore energy, and 4 times higher clock frequency than previous designs.

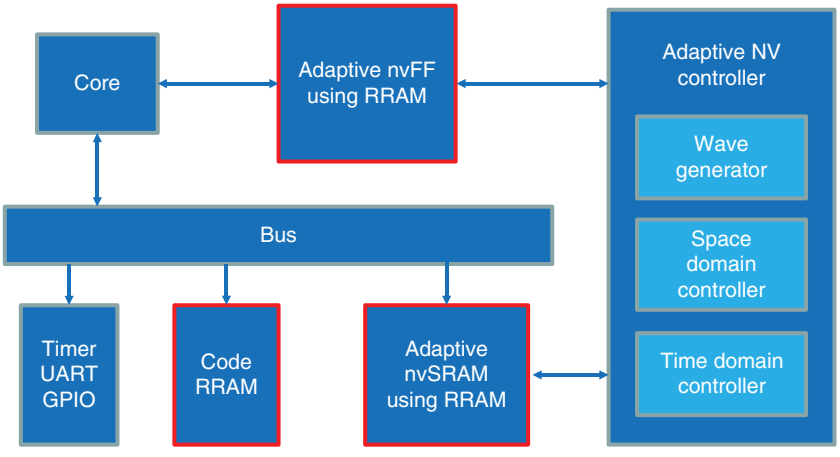


Figure 2.27 System architecture of the RRAM enabled NV-processor. Based on Y. Liu *et al.* (Tsinghua University, National Tsing Hua University, University of California, Los Angeles), ISSCC, February 2016 [32].

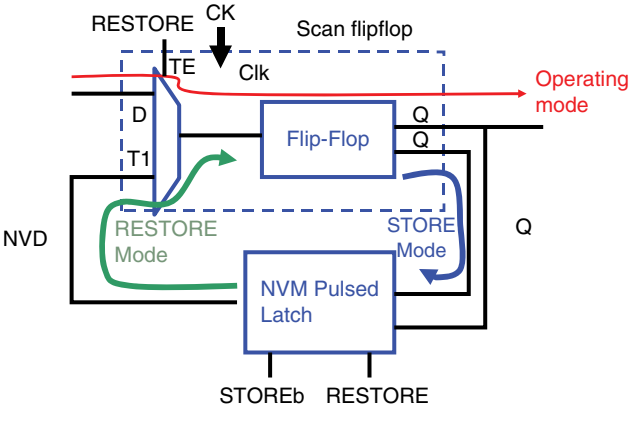


Figure 2.28 Schematic diagram of nonvolatile scan flip-flop. Based on A. Levisse *et al.* (University of Grenoble Alpes), *IEEE SOI-3D-Subthreshold Microelectronics Technology*, October 2014 [33].

In October 2014, an OxRAM-based pulsed latch was described by the University of Grenoble Alpes [33]. A schematic diagram of the nonvolatile scan flip-flop is shown in Figure 2.28. The operating modes included store and restore. This device was made in 28 nm CMOS fully depleted SOI (FDSOI) technology. It used a nonvolatile flip-flop (nvFF), thereby providing an ultralow standby power design for emerging connected devices for the Internet of Things (IoT), which operated on battery or harvested energy sources. The inclusion of nonvolatility in the flip-flop eliminated power consumption in the sleep mode while maintaining the system state. The nvFF was designed as an OxRAM based pulsed latch tied to a regular flip-flop to provide an ultrawide voltage range. The addition of nonvolatility cut off the flip-flop leakage but used a 63 pJ data store and restore. This added less than a 15% delay penalty.

2.6 Communication Protocols for Memory Sensor Tags

2.6.1 Radio Frequency Identification (RFID) Tags

Radio frequency identification (RFID) tags are used to identify and track everything from warehouse inventory to automobiles on freeways [34]. To increase usage, RFID tags must be very low power and low cost. For smart RFID tags containing SoC with embedded memory to be used, the application must require both fast access and nonvolatile reprogrammability. For example, RFID tags can reduce costs in automated manufacturing while maintaining control by recording stages of automated manufacturing. They can be used in retail tracking of expensive inventory and large shipping containers. RFID is also used in access control, payment systems, automatic meter reading systems, and many others.

RFID tags in retail are an upgrade from the ubiquitous Universal Product Code (UPC) bar code, which is on most retail products and could potentially replace them as the cost falls. While UPC is a read only memory (ROM), the RFID chip has the potential to be a read and/or write memory, that is, a ROM, a PROM (programmable read only memory), or an EEPROM (an electrically erasable programable read only memory). RFID tags can be inductively coupled consisting of metal coils and antenna powered by a magnetic field generated by the RFID reader, which induces a current in the tag. It can also be capacitively coupled using conductive ink instead of metal coils to transmit data.

RFID tags can be passive, semipassive or active. They all consist of a microchip and an antenna. Passive RFID tags rely on the reader for power and can be read up to 20 feet away.

Active and semipassive tags contain a battery. Data are stored in the RFID tag chip. When the chip's antenna receives a signal from the RFID reader's antenna, the tag uses its antenna to send a radio wave signal back to the reader. The power used by the chip to send its signal can either come from an internal battery or from energy harvesting most frequently coming from the reader's electromagnetic field. Active and semipassive tags can broadcast high frequencies from 850 to 950 MHz that can be read 100 feet or more away. Additional batteries can be used to boost the range of the tag. High frequency RFID tags such as in smart cards can have a range of 3 feet or less. Other signal enhancement devices can be used. For example, traffic monitors include automated tolls that generally use RFID tags affixed to the car's windshield (windscreen). The windshield enhances the signals, which are read by readers powered by electricity and can have a range of about 12 meters.

The RFID tag, to be useful, must have assigned a unique identifier and trillions of devices might eventually require identifiers. The electronic product code (EPC) identifier is commonly used and contains 96 bits of information. Information stored in the RFID chip is usually written in the product markup language (PML). The issue today for the universal acceptance of RFID is cost reduction in tags and readers. The processor in an RFID tag can be a conventional silicon MCU with embedded nonvolatile memory, but still needs to be cost reduced. Plastic printed roll-to-roll processing of RFID tag circuits including the MCU and memory has been developed and promises to eventually reduce the cost of an RFID tag; however, the performance of printed circuits is not yet as high as that of silicon.

2.6.2 Near Field Communications (NFC)

Near field communications (NFC) is a short range wireless communications protocol that uses RF signaling and enables communications between electronic devices that are either touched together or within a range of a few centimeters, generally less than 4 cm, permitting passing the device in front of the scanner. NFC devices combine RF signaling and an MCU with host interface circuitry and memory. The NFC protocol simplifies the pairing of Bluetooth or Wi-Fi devices. The host interface can be a 38.4 kbps universal asynchronous receiver/transmitter (UART) or a 1 Mbps serial port.

NFC tags let everyday objects interact with smart devices such as cell phones and other mobile terminals. In February of 2015, Panasonic's NFC tag platform was shown [35]. These tag ICs use low power, fast nonvolatile FeRAM along with secure data transmission with standardized encryption and authentication. They are compatible with NFC compliant mobile terminals. This NFC chip tag integrated RF, signal-conditioning, host interface circuitry, and a 4 k-bit embedded FeRAM in a surface mount package. Applications included: home appliances, health care systems, wearable medical devices, digital paper, smart keys, and smart meters. The FeRAM had 10 year data retention and a 0.33 mA current at 1 Hz compared to an EEPROM with a 2 mA current at 1 MHz. Write speed for 64 bytes was 154 μs compared to an EEPROM at 7.6 ms.

An RFID chip can have an EPC interface as well as an NFC interface. The EPC is similar to a bar code and acts as an identifier for retail inventory when using RFID tags. A fully integrated dual frequency RFID chip with both an NFC type 3 interface and EPC Gen2V2 RF interface was announced in April of 2015 by EM Microelectronics [36]. This circuit was intended for supporting EPC functionality for retail logistics and inventory control and for communicating directly with the consumer for marketing and advertising and product registration. This chip had a common rewriteable memory area, a unique ID, and security features. It also had a dual interface memory access of NFC or EPC and user data of more than 2Kb.

An example of the use of NFC is the US Passport, which has a chip with an embedded Flash processor with an NFC capability of being read when the cover is open. Figure 2.29 shows the closed cover of a US Passport with the location of the NFC chip indicated. The Passport uses an RFID chip, which is read using an NFC scanner. It can be read only when the passport is open due to a metal shield that is embedded in the cover of the passport.

Tags can either be passive or active. Passive tags collect energy from an interrogating reader interface while active tags have a local power source. An example of an NFC tag family was introduced in February of 2016 by NXP for smart home applications. These tags combined a passive NFC interface and a contact I2C interface together with a nonvolatile memory [37]. This tag offered zero power capability along with password protection, full memory access from both interfaces, and an originality signature for protection against cloning. This second generation part had four times the data transfer performance of the first generation along with backward compatibility to earlier generations. It had a pass-through mode with a 64-byte SRAM buffer, a 1912 bit EEPROM user memory, 32-bit password protection, NFC Type 2 tag, and a 7B serial number. An originality signature was based on elliptic curve cryptography for authentication. It used energy harvesting from the NFC field to power external devices. The operating temperature range was –40 to +105 °C.

Figure 2.29 Picture of US Passport with the cover closed, with the position of the NFC chip indicated. Photo by B. Prince.

2.6.3 Bluetooth-Based Beacons and Sensor Nodes

Bluetooth is a standard protocol for securely transmitting data using a 2.4 GHz wireless link. It is used for short range, low power, and low cost wireless transmission between electronic devices. Standard Bluetooth has a range of 1–100 meters with a point-to-point topology while Bluetooth Smart has a range of 10–100 meters with a point-to-point star topology. Bluetooth can be used for smart retail beacons, which can recognize a nearby shopper's mobile device and transmit information to the targeted shopper. See Figure 2.30 for an illustration of a shopper with a smart phone interacting with a retail beacon.

These beacons can be preprogrammed to respond to particular user profiles. Macy's department stores are reported to have implemented Bluetooth beacons [38]. In September of 2014, they reported expanded use of low energy beacons to include several departments in all stores. These beacons enabled a shopper, who elected to be in the system, to receive discount coupons and rewards while they were in a store based on their location in the store. Another possible application for beacons is in museum exhibits where exhibits can detect the presence of a user and transmit interesting information about nearby exhibits. Criteria for transmission might include a defined transmission range to avoid interference between individual beacons.

An illustration of a chip for a Bluetooth beacon was announced in December of 2014 by EM Microelectronics [39]. This was a coin-sized beacon consuming less than 25 μA average current that typically provided 12 months operation from a single CR2032 coin

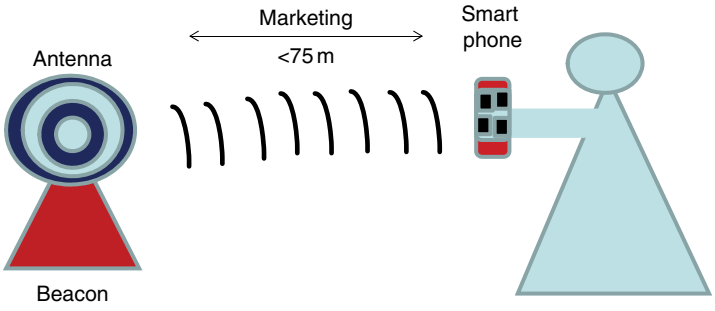


Figure 2.30 Illustration of retail beacon interacting with individual shopper's smartphone.

battery included in the beacon. The beacon used an EMM ultralow power MCU. The 0.9 V 8-bit RISC Flash MCU contained from 5 KB to 16 KB of Flash EEPROM and a miniature antenna that made the beacon detectable 75 meters away by an advanced smart phone. Its rugged package made it suitable for outdoor music festivals, sporting events, and arenas.

An early low energy Bluetooth compatible chip was introduced by Nordic Semiconductor in June of 2012 [40]. This multiprotocol Bluetooth 2.4 GHz RF Flash-based SoC was designed as a single chip solution for applications such as fitness and healthcare wearable sensors. It included 256 KB of embedded Flash and 16kB of RAM. Peripherals included: SPI, two-wire, ADC. The power supply range was 1.8 to 3.6 V. It included a DC/DC converter for 3 V coin cell batteries. A complete Bluetooth protocol stack was included along with 2.4 GHz RF protocols and 128 KB of Flash and 10kB of SRAM were available for an application code.

A Bluetooth Module Add-on for IoT Designs was introduced in August of 2015 by Silicon Labs. This Bluetooth module was intended to help developers reduce development costs and improve compliance by adding a "plug and play" Bluetooth to their designs [41]. The module provided precertified use in the markets of North America, Europe, and Asia Pacific. The modules were preloaded with the Bluetooth 4.1 compliant software stack and were field upgradable to Bluetooth 4.2 and beyond.

Two Bluetooth ICs that supported Bluetooth Low Energy (LE) ver. 4.1 communications were introduced in November of 2015 by Toshiba [42]. One tag integrated a built-in Flash ROM and the other supported both Flash ROM and NFC Forum Type 3 Tag for peer-to-peer communication. The latter device included a Flash ROM to store user programs and data along with a 64 KB user program area using an internal SRAM to execute various application programs. The chip's NFC Tag achieved easy Bluetooth peer-to-peer pairing and power ON/OFF functions by touching another NFC supported mobile device. The built-in Flash ROM eliminated any need for an external EEPROM and reduced external part counts and PCB size. The NFC tag was powered by small coin cell batteries.

It was expected that these devices would ease adoption of Bluetooth LE communications for small devices such as wearable healthcare devices, sensors, and toys. Characteristics of a typical Bluetooth LE type 3 Tag are shown in Figure 2.31. Communication was an NFC Type 3 Tag at 212 kbps/424 kbps. Security was mutual recognition by Triple DES Message Authentication code (MAC).

Peak power consumption	<6 mA at 3.3 V
Transmitter/Receiver power	<100 nA in deep sleep at 3 V
Embedded Flash ROM Capacity	192 KB
Embedded EEPROM	2 KB
Operating voltage range	1.8 V to 3.6 V
Operating current consumption	5.9 mA at 3.3 V
Current during deep sleep	<100 nA at 3.3 V
Operating temperature range	–40 °C to +85 °C
Interfaces	USTT, I2C, SPI, GPIO

Figure 2.31 Characteristics of a typical Bluetooth LE type3 tag. Based on Toshiba PR, November 17, 2015 [42].

Figure 2.32 Illustration of wireless Bluetooth speaker. Photo by B. Prince.

Bluetooth is being added to applications such as connected home, wearables, remote controls, baby monitors, beacons, electronic shelf labels, audio equipment speakers, health and fitness devices, and point of sale equipment. Figure 2.32 is a photo of a wireless Bluetooth Speaker.

In February of 2016, Silicon Labs announced an SoC family intended to support these applications [43]. These Bluetooth SoCs consume 8.7 mA in the peak receive mode and 8.8 mA at 0 dBm in the peak transmit mode and offer –94 dBm sensitivity and output power from –30 dBm to +19.5 dBm. The 40 MHz RISC processor had a powerful floating point and DSP capability. It consumed 63 μA/MHz current in the active mode. An integrated hardware cryptographic accelerator permitted autonomous encryption and decryption of Internet security protocols with minimal CPU intervention.

Another MCU intended for wearables, IoT devices, sensor hubs, building automation, personal medical devices, and smartphones was introduced in December of 2015 by STMicroelectronics [44]. This MCU included an RISC core performing at 125 DMIPS and 339 EEMBC CoreMark at 100 MHz. The accelerator helped drive dynamic power down to 89 μA/MHz with STOP current of 6 μA. Communication includes NFC, Bluetooth, or IEEE 802.15.4 radio boards. The devices have from 64 KB to 128 KB of Flash and 32 KB of SRAM. A RUN mode permits the Flash to be powered down for power savings. To save power, sensor data could be captured directly into SRAM while the CPU and Flash remained off. It had six timers including motor control and 32-bit/100 MHz timers. Communication ports included: SPI, I2C, I2S, USART, and ISO7816 interface along with a 12-bit DAC.

2.6.4 IoT Devices with Wi-Fi

A Wi-Fi platform can be designed to help add ultralow power long range connectivity to IoT designs. Wi-Fi is supported by laptops, smart phones, tablets, and other connected things, making it useful for device-to-cloud or device-to-device connectivity. In November of 2015, TI discussed an ultralow power Wi-Fi platform using a sub-1 GHz wireless MCU [45]. This wireless MCU offered 20 years of battery life for applications such as: building and factory automation, alarm and security, smart grid, and wireless sensor networks. The chip had an integrated RISC MCU sensor controller, low power modes, and 0.6 μA sleep current. The wireless sensitivity spanned from full building to citywide coverage for over 20 km on a coin cell battery with up to 128 KB of embedded Flash possible. The wireless MCU operated in various ISM bands. Communication protocols included: Bluetooth, Wi-Fi, Zigbee, and others.

A plug and play Wi-Fi module for use with Internet of Things (IoT) applications was discussed in February of 2016 by Silicon Labs [46]. The module had good RF performance, a small footprint, easy application development, and fast time to market. It included: a 2.4 GHz 802.11 radio, integrated antenna, global certifications, an MCU, an embedded Wi-Fi stack, and various Internet protocols. By integrating the 802.11 radio, antenna, MCU and Wi-Fi stack the module helped reduce development risk and accelerate the time to market. The module was useful for adding Wi-Fi to industrial M2M systems, wireless sensors, remote controls, thermostats, connected home products, automotive infotainment, point-of-sale devices, and fitness and medical equipment. The module included a scripting language to develop and host end applications. It could run in a network co-processor mode, enabling the host controller to handle applications tasks. The module enabled connectivity up to 300 m to 500 m typically and had a 2.4 GHz transceiver and an integrated high efficiency chip antenna along with peripherals including: UART, I2C, SPI, and USB. A Wi-Fi protocol stack included protocols for cloud integration. A dedicated host interface was provided to external MCUs running the end application.

A wireless Wi-Fi router is a device that performs the functions of a router and also those of a wireless access point. It can provide access to the Internet or to a private computer network. An office Wi-Fi router is shown in Figure 2.33.

In February of 2016, Silicon Labs discussed a wireless SoC with embedded Flash [47]. The wireless SoC integrated an RISC processor, a 2.4 GHz radio with up to 19.5 dBm output power, and hardware cryptography. The design included Thread and Zigbee stacks for mesh networks, intuitive radio interface software for proprietary protocols, Bluetooth for point-to-point connectivity, and tools to simplify wireless development, configuration, debugging, and low energy design. The wireless portfolio included three options for multiprotocol SoCs optimized for IoT: one had wide output power and range, one had Zigbee and Thread connectivity for mesh networks, and one offered flexible proprietary wireless protocol options for different applications.

The wireless SoCs had a multiprotocol 2.4 GHz RF transceiver in a single chip with up to 256 KB of Flash and up to 32 KB of RAM. Energy modes included: fast wake-up sleep transitions, a system that enabled autonomous operation of low power peripherals while the MCU core slept. The MCU included floating point capability and consumed 63 μA/MHz in active mode. It integrated a software programmable power amplifier and provided a scalable output power, which ranged from –30 dBm up to +19.5 dBm for applications requiring long range connectivity.

Figure 2.33 Office Wi-Fi router used in the offices at Memory Strategies International. Photo by B. Prince.

For security, the device had a built-in hardware cryptographic accelerator, which provided fast, energy efficient autonomous encryption and decryption of Internet security protocols with minimal CPU intervention. The on-chip accelerator supported algorithms such as AES with 128-bit or 256-bit keys, elliptical curve cryptography, and the secure hash algorithm (SHA)-1 and its follow-on SHA-224/256. The hardware cryptography enabled developers to meet evolving IoT security requirements more efficiently than software only techniques.

2.6.5 IoT Devices with USB Connectivity

The Universal Serial Bus (USB) interface is a fast serial bus technology that permits an electronic device to be connected to a computer. USB Flash drives are Flash memory for data storage with a built-in USB interface. Introduced in 1996, it was intended that all USB ports supply 5 V. USB2 could pull up to 100 mA while USB3 could draw from 150 mA to 900 mA.

An MCU was introduced in May of 2015 by Silicon Labs that enabled USB connectivity for a range of IoT applications including: smart metering, home and building automation, alarm and security systems, smart accessories, and wearable devices [48]. This 32-bit MCU was based on a RISC core, low energy peripherals, and a range of memory options. It had an energy management system with five energy modes ranging from active mode to deep sleep mode. In deep sleep mode the MCU had 0.9 μA standby current with a 32.8 kHz RTC, RAM/CPU state retention, brown out detector, and power-on-reset circuitry active. Active power consumption dropped to 130 μA/MHz at 24 MHz with read word code using a prime number algorithm. There was a 2 μs wake-up time from standby mode. On-chip peripherals included: analog comparator, supply voltage comparator, on-chip temperature sensor, programmable current DAC (IDAC), and a 12-bit ADC with 350 μA current consumption at a 1 MHz sample rate. On-chip AES encryption permitted secure wireless connectivity for IoT applications such as smart meters and wireless sensor networks. The chip eliminated many discrete devices by using a full-speed USB PHY, an on-chip regulator, and resistors.

IoT devices with USB connectivity for applications such as remote controls and thermostats have been introduced. In February of 2016, Silicon Labs introduced its 8-bit MCU portfolio for IoT, which included low power capacitive touch control and USB connectivity [49]. One such MCU had speeds of up to 48 MHz and 8–64 KB Flash sizes and can be used for higher end personal medical devices, remote controls, and thermostats. These have a high precision internal oscillator, a clock recovery circuit, a full speed USB transceiver, along with an on-chip battery charger detection module.

2.6.6 Single Wire Connectivity

Single wire connectivity has the advantage of minimizing the pins on the package to reduce the size of the package. In August of 2015, Atmel introduced a 2-pin, self-powered single-wire serial EEPROM for the IoT and other markets [50]. The part was self-powered by a parasitic power method over the data pin. It required just one data and one ground pin, which eliminated the requirement for external power. It had a plug and play 64-bit serial number for identification. The device's ultralow power requirements included a standby current of 700 nA, write current of 200 μA, and read current of 80 μA at 25 °C. The power scheme eliminated the need for external capacitors and rectifiers. The device had a high write endurance capability permitting more than one million cycles per memory bit. This device was targeted at IoT applications including wearables such as smart fitness and healthcare devices and consumables.

The devices followed the I2C protocol, which allows low overhead migration from existing EEPROMS and the capability of connecting up to eight devices on the same bus. The part had a security register with a 64-bit factory programmed serial number and an extra 16-bytes of user programmable and permanently lockable EEPROM storage for a unique serial number for inventory tracking and asset tagging. The data could be permanently protected. Different low voltage battery requirements were supported including low voltage applications at 1.7–3.6 V and 2.7–4.5 V for use with Li-ion polymer batteries.

2.6.7 Zigbee Interface

Zigbee is a standard interface with a mesh topology and a range of 10–100 meters. Its maximum datarate is 250 kbit/s at 2.4 GHz [51]. It is intended as a communications protocol for creating personal area networks with low power digital radios for applications such as home automation and medical device data collection. It is a low power, low data rate, close proximity, personal area wireless network that is intended to be simpler than Bluetooth or Wi-Fi.

2.6.8 ANT Interface

ANT is a proprietary interface with a range of 30 meters at 0 dBm. It is an adaptive isochronous network technology that ensures coexistence with other ANT devices. Each transmission occurs in an interference free time slot within the defined frequency band. Transmission is for less than 150 μs per message, which permits a single channel to be divided into numerous time slots [52].

2.7 Wearable Medical Devices

2.7.1 Overview of Wearable Medical Devices

Portable medical systems are a growing market due to the increasing proportion of Seniors in the population worldwide. These portable systems require large high performance memory but must have low power for battery operation. The wearable electronics part of this application includes devices for medical, fitness, and wellness, such as step tracking, heart rate and blood pressure monitors, blood glucose monitors, insulin pumps, hearing aids, and other. To service this application, suppliers of smart connected wearable devices will require low power platforms with simple basic features that include: a simple user interface, embedded processing for data collection, and sensors for environmental tracking. The portable medical system application also includes devices such as: miniature medical sensors, portable datalogging equipment, medical and pharmaceutical tracking, and implantable medical devices such as defibulators and pacemakers. Figure 2.34 is an illustration of a wireless network of wearable medical devices.

Portable medical devices are expected to use embedded EEPROMS and Flash memory as well as fast access, low power nonvolatile emerging memories such as FeRAMs, Conductive Bridge (CB) RAMS, and MRAMs. Powering these devices will require either long term batteries or energy harvesting systems. These devices will need to be networked to medical surveillance systems and medical personnel requiring networks of IoT communication devices.

These portable networked medical devices are expected to be used increasingly by the aging population, which the United Nations estimates today at 900 million including 250 million in developed countries and another 650 million in developing countries [53]. This total is expected to grow to 1.2 billion by 2025. With medical care improving worldwide, these higher value portable medical and tracking devices could provide another large market for fast, low power nonvolatile memories.

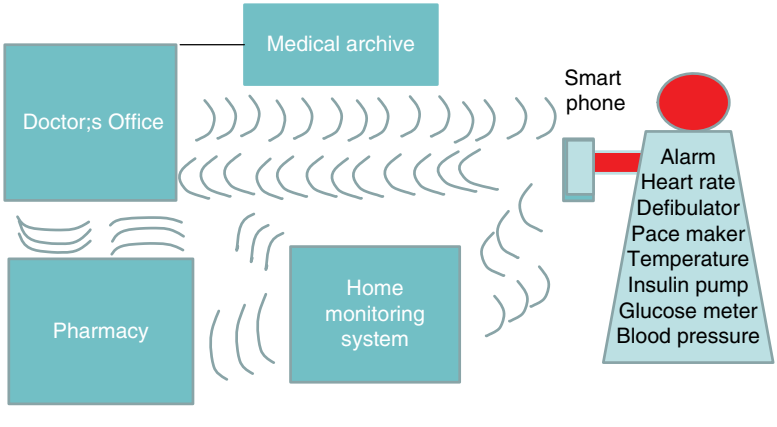


Figure 2.34 Wireless network of medical devices to medical suppliers and monitors.

2.7.2 Miniature Hearing Aids Using FeRAM Memory

The fast write speed and low power consumption during write of the 1T1C FeRAM memories compared to EEPROMs made them an obvious choice for use in battery powered medical devices such as hearing aids. In March of 2013, Fujitsu America introduced 1 Mb and 2 Mb FeRAM memories [54]. These parts were intended for miniature medical devices where the faster write speed than EEPROM eliminated undesirable noise. They consumed 92% less power during write than similar capacity EEPROMS and had a 920 times faster write speed. Endurance was $>10^{12}$ write cycles for prolonged battery life. In a system, this single chip solution reduced: component count, the size of the battery, cost, footprint, and power consumption compared to memory solutions using EEPROM or SRAM plus a battery. The fast random access of the nonvolatile FeRAM permitted more data to be retained when the power is switched off'

2.7.3 Body Sensor Node Platforms Using CB-RAM Memory

Body sensor node (BSN) networks need to operate from low levels of harvested power without a battery in the system. Low operating power nonvolatile memory is essential for the long term operation of these systems. In June of 2013, Adesto and the University of Virginia discussed a CB-RAM macro that was embedded in a BSN processing platform [55]. The device operated at voltages as low as 600 mV for write and 300 mV for read. This enabled low energy operation and compatibility with low energy digital systems. The charge pumps required for EEPROMs were eliminated. The memory macro permitted a 100 times reduction in write energy compared to Flash memory and write operation down to 0.6 V, which is compatible with other ultralow power SoCs used in wireless body sensor networks. Wireless nodes and body sensors are expected to operate at 1 V and below.

Two CB-RAM macros were embedded in a subthreshold voltage operating SoC, which provided a digital platform for body sensor nodes to process ECG, EEG, and EMG signals. The architecture included a programmable RISC processor, a custom node controller, DMA and accelerators for heart-rate extraction, and atrial fibrillation detection. The CB-RAMs acted as instruction and data memories. The chip was found to execute properly after a day of power-down. This confirmed the nonvolatile storage and operation of the low voltage arrays. An illustration of the die showing the two 64Kb macros integrated with the BSN digital platform is shown in Figure 2.35.

2.7.4 "Store Mostly" Healthcare Systems Using MRAM

An STT-MRAM can use less energy than a comparable SRAM in a system if the memory bandwidth utilization is low. In November of 2013, Kobe University and LEAP discussed a 65 nm single supply voltage 8 Mb STT-MRAM with a process-variation tolerant sense amplifier that could be used for minimal energy SRAM replacement in an infrequent access platform such as a healthcare system [56]. The sense amplifier used a boosted gate NMOS and negative resistance PMOS as loads. These maximized the readout margin in all process corners.

The STT-MRAM achieved a cycle time of 1.9 μs (0.526 MHz) at 0.38 V. The operating power was 6.15 μW at 0.38 V and the minimum energy per access was 3.89 pJ/bit at

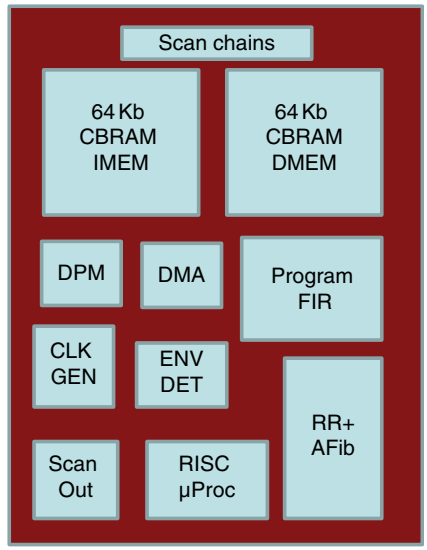


Figure 2.35 Illustration of die layout showing two 64Kb CBRAM macros integrated with the Body Sensor Node (BSN) digital platform. Based on N. Gilbert *et al.* (Adesto, University of Virginia), VLSI Circuits Symposium, June 2013 [55].

VDD = 0.44 V. The characteristics of the 65 nm CMOS test chip included: a nominal voltage of 1.2 V, operating VDD from 0.38 to 0.6 V, and operating frequency of 0.526–5000 MHz.

The STT-MRAM was found to use less energy than a comparable SRAM if utilization of the memory bandwidth is 14% or less, which means the STT-MRAM is preferable in less active applications such as a healthcare system or a sensor network. If, however, the energy consumption of the proposed STT-MRAM and an LV SRAM are compared and the ratio of read and write access is 50:50 at an operating voltage of 0.5 V, the energy consumed in the STT-MRAM is 3.03 times larger than in the SRAM. The STT-MRAM cell size was 0.203 mm^2, which compares favorably with the SRAM cell size in that technology.

2.7.5 Wearable Biomonitoring with NFC and eFeRAM Memory

A low power wearable biosignal monitoring system that communicated with smart phones using Near Field Communications (NFC) was discussed in June of 2013 by Kobe University and Omron Healthcare [57]. The NFC was used to check vital signs easily at any time using the smart phone. The system included: a battery, electrodes, a triaxial accelerometer IC, an NFC tag IC, and a biosignal processor LSI with embedded FeRAM.

The biosignal processor LSI was made in 130 nm CMOS. It included: heart-rate monitoring circuits, a 32 KB ferroelectric RAM, an accelerometer interface, and an NFC interface. This system used 38.1 µA of power for the logging application running at 320 kHz with a 3.0 V power supply.

The goal of this system was mobile wellness achieved by daily life monitoring of vital signals and physical activity of elderly people. This wearable system could log instantaneous heart rate and acceleration, along with physical activity and estimation of exercise intensity. The key factors affecting usability were miniaturization and weight reduction. Battery mass and power consumption reduction were critical. A block diagram of the

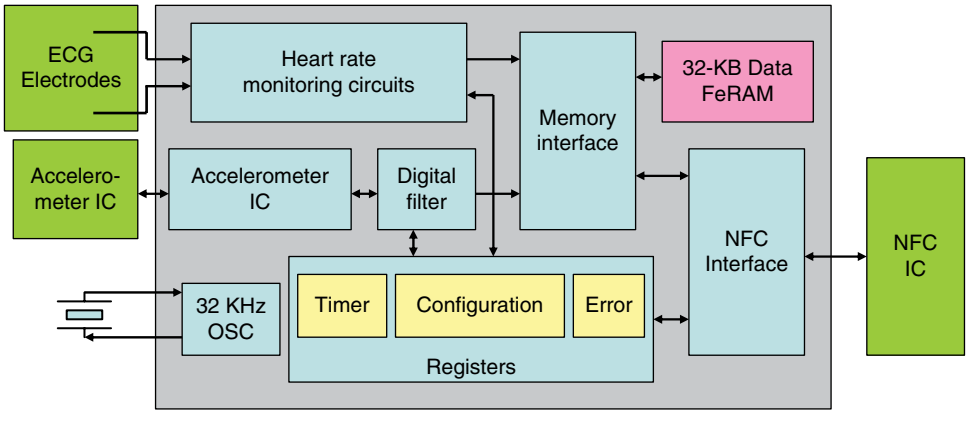


Figure 2.36 Schematic block diagram of a biosignal processor LSI. Based on K. Yamashita *et al.* (Kobe University, Omron Healthcare), NEWCAS, June 2013 [57].

biosignal processor LSI showing the embedded 32 KB data FeRAM is shown in Figure 2.36.

While a wireless transceiver consumed the most power in the system, the standby power of NFC is small compared to Bluetooth or ZigBee. NFC enabled communication to be initiated without manually configuring the communication link. NFC has high security due to its short communication range.

2.7.6 Wearable Healthcare System with ECG Processor Using FeRAM

An ECG processor intended for use with a wearable healthcare system was discussed in September of 2013 by Kobe University, Rohm, Omron, and JST CREST [58]. The wearable healthcare system included: a battery, an ECG processor, an accelerometer IC, an antenna for near field communications (NFC), and an NFC Tag IC, as shown in Figure 2.37. The system can be accessed by a smart phone or other electronic reader/writer using NFC to handshake with the smart phone.

Key factors affecting wearable system usability include: miniaturization and weight reduction. Noise can affect a wearable electrocardiograph (ECG) monitor since its electrodes are close together. An analog front end is required to counteract noise when the user is active. Battery capacity is related to battery weight, which in turn is affected by the power consumption of the system. If power consumption is low, then battery weight can be reduced. NFC has very low standby power compared with Bluetooth or ZigBee. Active communication energy is consumed by a reader/writer when using a passive NFC tag.

The ECG processor included: an analog front end, a 12-bit ADC, an instantaneous heart-rate monitor (IHR), a 32-bit RISC processor core, and a 64 KB FeRAM. A block diagram of a sketch of the chip photograph for the ECG Processor is shown in Figure 2.38. The chip used 130 nm CMOS technology. Macros included: 2 V digital 64 KB SRAM and 3.0 V 64 KB FeRAM with an area of 0.9 mm. MCU frequency was 24 MHz. The other blocks ran at 32 kHz. The ADC had 12 bit resolution. The total chip current was 12.7 μA for heart-rate logging.

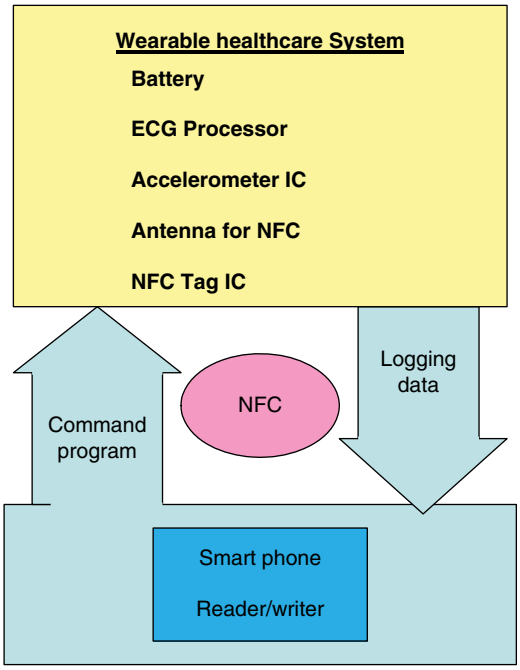


Figure 2.37 Wearable healthcare system including ECG processor, accelerometer IC, antenna for NFC, and an NFC tag IC. Based on S. Izumi *et al.* (Kobe University, Rohm Co., Omron, JST CREST), ISCDG, September 2013 [58].

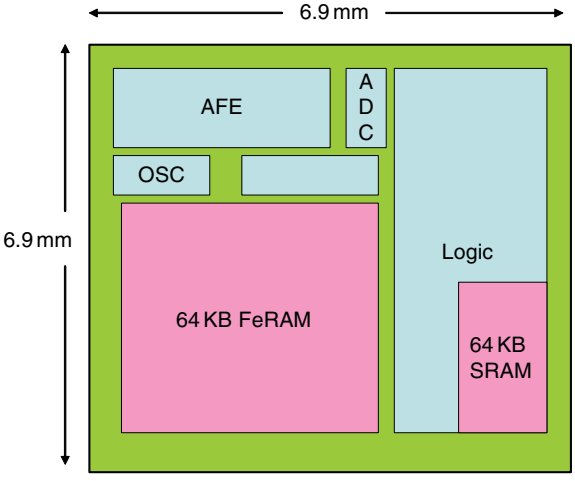


Figure 2.38 Sketch of the ECG processor chip showing 64KB FeRAM. Based on S. Izumi *et al.* (Kobe University, Rohm Co., Omron, JST CREST), ISCDG, September 2013 [58].

The heart-rate monitor used a short term autocorrelation algorithm to improve the heart-rate detection accuracy in spite of being used in noisy conditions. The instantaneous heart rate is an important biosignal used for heart disease detection, heart-rate variation analysis, and exercise intensity estimation. The ECG processor chip consumes

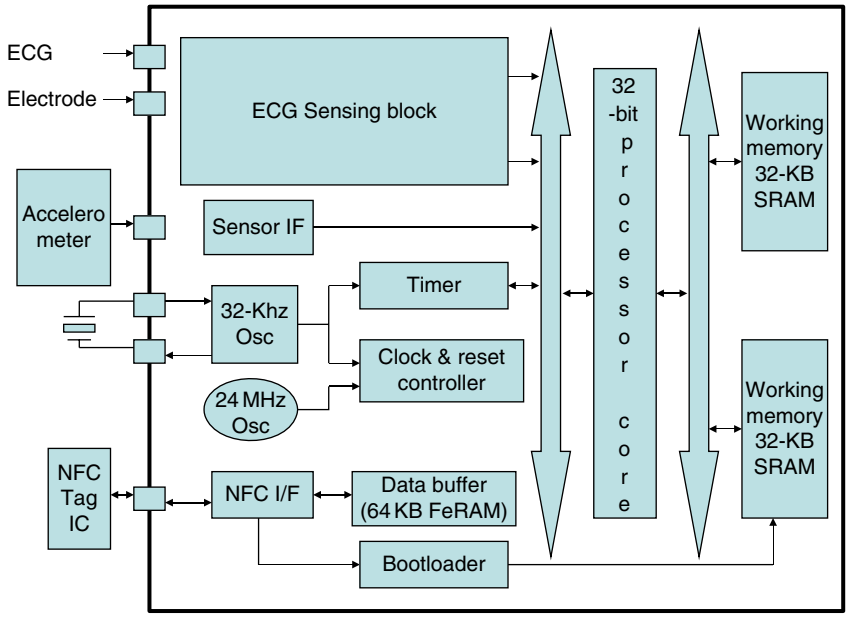


Figure 2.39 Block diagram of an ECG processor with 64KB FeRAM. Based on S. Izumi *et al.* (Kobe University, Rohm Co., Omron), *IEEE Trans. on Biomedical Circuits and Systems*, November 2014 [59].

13.7 μA for the heart-rate logging application. A block diagram of the ECG processor is shown in Figure 2.39. The data buffer with a 64 KB FeRAM is shown. In November of 2014, Kobe University, Rohm, and Omron discussed further on this wearable healthcare system with a noise-tolerant ECG processor with an embedded 64 KB FeRAM data buffer [59]. Wearable systems have size and weight constraints that limit the battery capacity and the signal-to-noise ratio of the biosignals.

In January of 2014, Kobe and Rohm discussed "normally off" technologies for healthcare appliances. The application focus was on a daily life monitoring and the ECG processor. The system used an FeRAM and near field communication (NFC) for normally-off datalogging and data retention. The limitations on wearable system usability are miniaturization and weight reduction [60]. This means battery weight and hence power consumption must be reduced. A wearable biosignal monitoring system using normally off technologies was used to reduce power consumption.

The wearable biosignal monitoring system could acquire long-term instantaneous heart rate data and an acceleration value. Physical activities were classified using a triaxial accelerometer. The instantaneous heart rate is calculated from the electrocardiogram R-waves for detection of heart disease, heart-rate variation analysis, and exercise intensity estimation. To reduce the power consumption, wearable and wireless ECG telemetry system and signal-chip ECG monitoring system LSIs were developed. The limitations on power consumption and electrode distance on wearable ECG monitors made them sensitive to noise as the subject moves around.

2.8 Low Power Battery Operated Medical Devices and Systems

2.8.1 Overview of Low Power Battery Operated Medical Devices

Portable medical systems are a growing market due to the increasing proportion of Seniors in the population. These require large memory and low power for battery operation. The characteristics of these various portable medical equipment systems are summarized in the table in Figure 2.40.

Currently there are a number of suppliers and developers with Flash MCU for portable health monitors. Details are in the next two sections. Most use RISC processors with embedded EEPROM and eFlash in amounts ranging from 64 to 128 KB of eFlash. Endurance is commonly indicated at 10^5 cycles. Operating current ranges from 35 to 110 μA/MHz. Standby current ranges from 50 nA to 40 μA. Suppliers and developers with embedded emerging memory targeted at portable health monitors offer embedded FeRAM, RRAM, and CB-RAM..

2.8.2 Low Power Battery Operated Medical Devices Using eFlash

A significant number of low power battery operated medical devices have been developed in the past few years. Many of these use the embedded Flash memory macros that are already available at various foundries to embed in SoC along with commercially available RISC cores. A number of foundries have also introduced embedded Flash memory macros for low power applications.

A 40 nm low power process with an embedded NVM intended for IoT type applications was announced in March of 2015 by NXP to be run at Global Foundries in Singapore [61]. Volume production on 300 mm wafers was expected in 2016 for this

Supplier	MCU	Speed	eMem	Bytes	Cycles	Vmin	Iop	Istd
Flash/EEPROM:								
NXP	RISC		eFlash	128 KB			110 uA/MHz	
TI	RISC	80 MHz	eFlash	64 KB	10^5			
Microchip		80 MHz	eFlash	64 KB	10^5	2.3	10 mA	10 uA
STMicro	RISC	32 MHz	eFlash	64 KB			200 uA	40 uA
Atmel	RISC	300 MHz	eFlash	2 MB			35 uA/MHz	200 nA
Si Labs	8b MCU	25 MHz	eFlash	64 KB				50 nA
Emerging Memory:								
Omron	MCU	320 kHz	FeRAM	32 KB	10^{15}	3.0	38.1 uA	
Panasonic	8b MCU		ReRAM		10^5	0.9	100 nA	60 nA
Adesto			CB-RAM	2-4 MB				

Figure 2.40 Features of various portable health monitor Flash MCU including nonvolatile memory and electrical characteristics.

process, which was intended for MCU with embedded Flash memory for applications such as ID, near field communications (NFC) and healthcare.

In September of 2014, TSMC announced an ultralow power (ULP) platform for IoT and wearable devices that included 55 nm, 40 nm, and 28 nm ultralow power processes and offered FinFET ultralow power processes down to 16 nm [62]. RF and eFlash memory were available in 180 to 40 nm ULP technology for IoT. Logic IP and libraries were available in the ULP processes with early designs in processes down to 28 ns node planned in 2015. Integration of radio and eFlash was enabled along with the energy efficient RISC CPUs from ARM Corp. A low power mixed signal design flow was offered from Cadence. Fujitsu indicated an interest in the 28 nm ULP technology platform, while Nordic Semiconductor indicated they had selected the 55 nm ULP process for their ultralow power RF SoCs. Compared with previous low power generations, the TSMC ULP processes were expected to reduce operating voltages by 20 to 30% to lower both active and standby power.

A Flash memory using a single poly 130 nm nonvolatile memory technology embedded in a 65 nm logic process was discussed by Toshiba in July of 2015 [63]. This process, which was to be run at UMC in Taiwan, was intended for the low power IoT market for applications such as wearable and healthcare related equipment. The Silicon Storage Technology third generation split gate embedded Flash in the 65 nm logic process was also being offered by UMC for MCUs, with a resulting power of about 60% of that of conventional double polysilicon stacked embedded Flash. In addition, Bluetooth low energy macros had been developed and were targeted at short range wireless products, including NFC controllers and contactless cards. For lower cost applications, an NVM embedded process using single polysilicon multitime-programmable (MTP) cells intended to run on the 130 nm logic process would be offered. This process limits the increased process steps in mask pattern lithography of conventional embedded Flash.

A number of chips with embedded Flash memory have been introduced for use in health devices. A chip with USB connectivity for use in health devices such as glucose meters and other applications was introduced in June of 2012 by NXP [64]. This 32-bit RISC MCU had up to 4 KB of EEPROM, 12 KB of SRAM, and 128 KB of eFlash with 256 bytes of the erase sector. Active power was down to 110 μA/MHz for battery operation. USB 2.0 was included. An illustration of a glucose meter is shown in Figure 2.41.

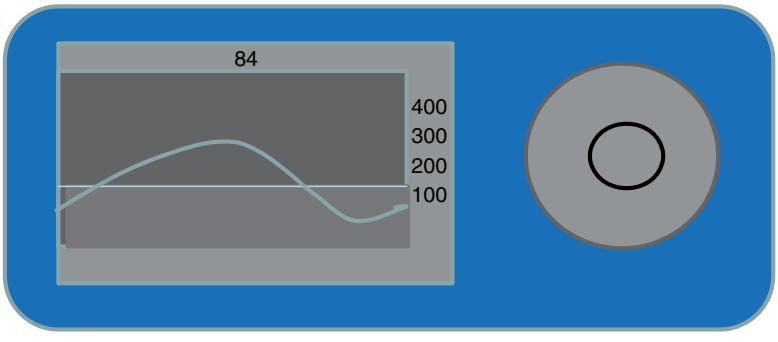


Figure 2.41 Illustration of commercial wearable blood glucose meter.

For health and fitness applications along with consumer and human interface control, TI in September of 2012 announced a kit that permitted engineers to design using an RISC MCU with a floating point operating up to 80 MHz [65]. The MCU included 64 KB of Flash with 10^5 write/erase cycles, two 12-bit ADC, and up to 27 timers.

An ionizing radiation sensor with an embedded ultralow power Flash memory was discussed by TowerJazz in November of 2012 [66]. Sensing was accomplished by measuring the accumulated V_t shift of the Flash cell after exposure to radiation. The sensor was intended for inclusion in a passive RFID for radiation measuring systems. It operated at very low power and required no power supply during the exposure sessions.

Devices intended for battery operated accessories and sensors with applications including smart meters and health monitors were introduced in February of 2013 by Microchip Technology [67]. These devices, which offered up to 8 M-bit of memory, were made using a commercial embedded Flash technology consisting of a split-gate, NOR embedded Flash with a thick oxide tunneling injector. The parts had an extended operating voltage range from 2.3 to 3.6 V. The memory was partitioned into uniform 4 KB sectors and 32 and 64 KB blocks offering seamless partitioning for program and data code in the same memory block. Sector and block erase was as fast as 18 ms and the entire chip could be erased in 35 ms. Word program time was 7 μs. Endurance was 10^5 cycles with more than 100 year data retention. The active read current was 10 mA typical at 80 MHz and the standby current was typically10 μA.

For applications that include fitting smaller batteries on portable medical equipment, such as fitness trackers and glucose meters, STMicroelectronics in February of 2014 discussed an ultralow power Flash MCU [68]. The device featured a 32 MHz RISC core, a 100 ksps ADC consuming 40 μA, and a crystal-less USBFS 2.0 [68]. The ADC consumed 40 μA at 100 ksps up to 200 μA at 1.14 Msps so a coin battery could operate a fitness tracker for two years. The CMOS technology used offered an embedded EEPROM and low variation over the 25 to 125 °C temperature range. Up to 64 KB of Flash memory was included and 2 KB of embedded EEPROM. Figure 2.42 shows a Bluetooth wireless fitness tracker. This fitness bracelet can take heart-rate readings, track sleep automatically by measuring deep, light, and REM sleep, and has activity tracking. The Bluetooth permits it to communicate with a smartphone.

The design of an NFC sensor transponder tag for wearable health and fitness applications was discussed in June of 2015 by Texas Instruments [69]. This tag combined a TI MCU core with an ISO15693 compliant NFC interface for wireless communication. It was optimized to be fully passive without a battery or for a single cell battery mode. It had integrated analog and digital sensor interfaces, which made it adaptable to wearable applications. It could be used to collect data on demand or at regular intervals.

Figure 2.42 Bluetooth wireless fitness tracker for taking heart rate readings, tracking sleep automatically, and tracking steps. Photo by B. Prince.

The TomTom Spark GPS fitness watch contained an Atmel smart RISC MCU from ARM with eFlash along with other chips and a software library to control the touch interface, as announced by Atmel in November of 2015 [70]. This ultralow power MCU helped extend the battery life of the fitness watch for up to 10 hours in GPS mode. A smart energy efficient RISC MCU from ARM ran at speeds of up to 300 MHz with up to 2 MB of embedded Flash memory. Features included: USART, SPI, I2C, 12b ADC, and DAC, and a single digital IO (SDIO). This MCU was optimized for applications requiring performance and power efficiency in a small package. Integrated on the chip were: ADC, EEPROM memory, and brown-out detectors to permit building applications without external components. They also offered Flash memory and on-chip debug. The chip could operate at 0.7 V.

An application processor intended for use in wearable devices for IoT was announced by Toshiba in September of 2015 with production scheduled for early 2016 [71]. The device could support multiple external sensors and provided a communication environment that made use of the extended hub features of Bluetooth. The device included a Bluetooth v4.1 controller and processor, which can capture data from external sensors connected by various I/Os such as I2C, UART, SPI, and ADC. It also included Flash memory to save data. Both data processing and storing of multiple sensing targets as required for IoT devices were executable simultaneously. The device could also perform as a Bluetooth hub for collecting, processing, and storing data.

The processor has a low power design feature that permits changing the supply voltage with processor frequency. This feature can be used for wearable products that require long operating hours. Software algorithms supported include: activity monitoring and ECG monitoring. This means it provides a total system solution for wearable systems. The communication function, memory, and processor for IoT are integrated in one package.

For next generation wearable IoT devices for battery operated applications that require activity and environment monitoring, Atmel introduced an ultralow power platform in January of 2016 [72]. It included an energy efficient RISC processor from ARM along with an Atmel designed Bluetooth Smart module, which had an active power down to 35 μA/MHz and 200 nA current in sleep mode. It was packaged in a 2.2 mm × 2.1 mm Wafer Level Chipscale Package, which permitted designers to build ultrasmall industrial designs for connected IoT and wearable applications. The entire platform was in a 30 mm × 40 mm form factor, which includes: the MCU, Bluetooth module, capacitive touch interface, security solution, complete software platform, real time OS, a 6-axis "Smart Hub" motion sensor and a BME280 environmental sensor. The platform could be powered by a simple coin cell battery.

A version of Silicon Lab's 8-bit MCU portfolio for IoT intended for personal medical devices and wearables was introduced in February of 2016 [73]. It included low power capacitive touch control and USB connectivity. The low energy 8-bit device offered a sleep mode power of 50 nA with full memory retention and brown-out detection, and a 2 μs wake-up time. Core speeds scaled up to 25 MHz and Flash memory sizes ranged from 2 to 64 KB. It included an on-chip battery charger detection module. The device is intended for adding USB connectivity to personal medical devices and wearables.

A wrist monitor to track physical activity, sleep quality, breathing patterns, and stress levels and offer actionable suggestions for self-improvement was discussed in August of 2016 by Nordic Semiconductor [74, 75], The device used a low energy SoC, which

automatically synched to a smart phone using Bluetooth wireless connectivity. The 32-bit RISC core chip from ARM ran at 64 MHz with a 2.4 GHz multiprotocol Bluetooth 4.2 ANT radio with −96 dB RX sensitivity. It had 512 KB of embedded Flash memory and 64 KB of SRAM. It supported up to 5 days of continuous biometric sensing and reconnection between recharges of the Li ion battery. The 2.4 GHz radio had 5.5 mA peak RX/TX current and a fully automatic power management system offering 90 CoreMark/mA. It also had an on-chip NFC tag for "touch-to-pair" communications.

2.8.3 LP Battery Operated Medical Devices Using Embedded Emerging Memories

Low power battery operated medical devices have also been developed using embedded memories of the emerging memory type, such as: RRAM and CB-RAM. A line of 8-bit MCU with embedded RRAM replacing Flash memory began production at Panasonic in August of 2013 [76]. Target applications were those that require low power operation such as sensors and portable healthcare devices. The RRAM used Ta_2O_5 to control oxide density and defects during deposition of the tantalum layer. Uniformity was obtained in the filaments by keeping the electric current constant during the forming process, which resulted in the filaments being uniform. Uniform filaments ensured stable operation. By optimizing the current, the level of filament reduction during the write operation was maximized to assure data retention. The MCU with the embedded RRAM operated at voltages as low as 0.9 V, which permitted low power consumption. The standby current was 60 nA with a 3 V power supply and the operating current was 200 nA during real time clock operation. Compared to a comparable MCU with eFlash, the power consumption was about half. In a comparable eFlash MCU, the write voltage was 2.7 V and the write speed was 30 μs/byte, while in the MCU with eRRAM, the write voltage was 1.8 V and the write speed was 3 μs/V. The rewrite cycle endurance was 100 000. The RRAM was being installed in the back-end process so it would be usable in various process nodes. The RRAM characteristics are shown in Figure 2.43.

In October of 2013, Adesto Technologies and Nordion, who make commercial gamma irradiation systems, discussed the Gamma irradiation tolerance of the Adesto CB-RAM nonvolatile memory [77]. The tests exposed standard packaged CB-RAM to doses up to 5 Mrad and showed the immunity of CB-RAM to gamma radiation. The results open medical applications, and other uses that require sterile memory components, to CB-RAM technology, since gamma irradiation is a conventional method of sterilization of single use medical devices from microbial decontamination.

MCU	8-bit
eRRAM	Ta_2O_5
Voperating (min)	0.9 V
I operating	200 nA
I standby	60 nA
Vstandby	3 V
Vwrite	2.7 V
Write speed	3 μs/V
Endurance	100,000 cycles
Process	BEOL CMOS

Figure 2.43 Features of Flash MCU with embedded RRAM for low power battery operated portable medical devices. Based on Semiconductor Industry News, August 22, 2013 [76].

A family of sterilization-tolerant low energy CB-RAM products was announced by Adesto in August of 2014 for wearable medical devices [78]. These devices have been shown to maintain data during high temperature and irradiation conditions of the sterilization processes used in medical device manufacturing. Applications include: wearable electronics, orthopedics, smart syringes, smart sample containers, and surgical devices as well as in application areas such as RFID biomaterial tracking, drug delivery, and capsule endoscopy.

A line of 1.65–4.4 V serial Flash memories using conductive bridge RAM (CBRAM) technology targeted for use in Internet of Things (IoT) applications was announced by Adesto in November of 2015 [79]. The memory was offered up to 4 Mb densities and could be used with wearable wireless sensors. It has deep power down, active interrupt, and an extended voltage range. It can be used with lithium polymer battery chemistries without added circuits. In February of 2016, Adesto described the power consumption of a CBRAM chip using an SPI interface made in a 130 nm process [80]. Core voltage is 1 V and write is at 2.5 V. Power consumption is low enough that a small capacitor can provide enough energy to write 32 bytes of data into the CBRAM. Reading 500Kb/s of data consumes 10 μW or 10 pJ/bit.

2.8.4 Security for Medical Systems

Security is essential for workable IoT applications and in particular for chips intended for medical systems. Hardware security on MCUs with an embedded Flash memory is expected to be able to meet the need for secure applications in areas such as payment applications and wearable devices.

A family of very low power MCUs was described in August of 2015 by Microchip Technology, which included an integrated hardware crypto engine with both OTP and RAM options for secure key storage and up to 256 KB of Flash memory [81]. It also included a direct drive for segmented LCD displays. A dual partition Flash with live update capability permitted the device to hold two independent software applications. It also permitted simultaneous programming of one partition while executing the application code from the other. This feature is thought to make the chip suitable for applications that require secure data transfer and storage and long battery life, such as industrial and medical applications.

A family of secure MCU was introduced in November of 2015 by STMicroelectronics. These devices included up to 2 MB of 40 nm embedded Flash program memory, fast clock speed, and a fast embedded cryptographic accelerator based on a 32-bit secure core processor from ARM [82]. The hardware architecture provided multiple strong fault protection mechanisms for CPU, memory, and buses for secure software and systems. It was expected to meet the need for secure applications in areas such as payment applications and wearable consumer devices as well as in machine-to-machine communications in factory automation and automotive in-car communications. The peripheral had a secure Data Encryption Standard algorithm and a cryptoprocessor that supported the public key algorithm. The inclusion of 2 MB of eFlash permitted the device to host additional applications.

In October of 2013, the Institute for High Performance Microelectronics (IHP) described a sensor node crypto-processor for use in wireless sensor networks with strong security requirements [83]. Wireless sensor networks consist of several randomly

deployed sensor nodes that rely on very limited power resources. The SoC used was a mixed signal processor with hardware crypto-accelerators for secure communication in the network. The system architecture combined an asynchronous processor core with a synchronous peripheral, which resulted in low power system operation. The chip integrated a 12-bit ADC and a 64KB sector embedded Flash on the chip. The improved security of the system was at the cost of an additional chip area and increased standby power. A digitally controlled oscillator was expected to be later included as an option to reduce the total power of the chip.

An MCU that combined a hardware cryptography engine, low energy modes, on-chip DC–DC converter and scalable memory for IoT applications such as wearable health and fitness devices was announced in December of 2015 by Silicon Labs [84]. The MCU addressed the need to equip IoT connected devices with security technology. The cryptography engines provided autonomous encryption and decryption for Internet security protocols such as TSL/SSL with minimal CPU intervention. The crypto-accelerator supported algorithms such as AES with 128-bit or 256-bit keys and elliptical curve cryptography. The hardware cryptography was found to meet evolving IoT security requirements more efficiently than software only techniques. These MCUs had an enhanced peripheral reflex system that allows low power peripherals to operate autonomously while the MCU core sleeps. This permitted connected devices to sleep longer and it extended battery life. The active mode current was 63 μA/MHz and sleep mode current ranged from 1.4 μA down to 30 nA. Ultrafast wake-up/sleep transitions minimized energy consumption. The MCU integrated a high efficiency on-chip DC–DC buck converter with current capacity of 200 mA, which provided a power rail for other system components while also powering the MCU. Scalable memory options included up to 256KB Flash and 32KB SRAM.

2.9 Automotive Network Applications

2.9.1 Overview of the Automotive Application

Automotive remains one of the largest Flash MCU applications. Embedded Flash memory size is growing with a more complex application code for increased functionality along with the potential to be customized for personal preferences. In automotive applications, for example, dynamic driver modes may let a driver select between Economy, Comfort, or Sport driving modes, which requires increased storage capacity in the powertrain MCU managing the engine.

Another factor increasing the complexity of applications like automotive are the requirements that different subsystems be able to share information to make better real time decisions. The evolution of Advanced Driver Assistance Systems (ADAS) is an example of this concept. The interlinking of different subsystems within a car is driving the requirement for more Flash memory in the controlling MCU. The need for datalogging is also rapidly increasing the amount of embedded Flash memory in the car. Information like odometer readings, critical sensor readings, crash logs and black box machine states are immediately written into data Flash when power is lost or crash events occur. The serviceable lifespan of a car is approaching 15–20 years and EEPROM data sets stored in data Flash are requiring 500 000 or more updates. This means that

512 KB or more of high endurance data Flash is required. A high performance automotive powertrain MCU has high bandwidth Flash memories to meet the demanding engine control applications. Embedded 28 nm Flash products are expected to enable 16 to 32 MB of Flash to be integrated into automotive MCU in support of powertrain and ADAS solutions.

Flash MCUs with embedded Flash memory are essential in automotive applications for complicated real time controls for higher fuel efficiency, better safety, and increased connectivity. High density, fast random read access is needed from the eFlash. The requirements for eFlash in automotive are secured operations and data reliability in particular reliable program/erase cycling at high temperature. Automotive has stringent operational, environmental, and reliability requirements.

Automotive ICs made after 2016 by Global Foundries are expected to include faster processors based on 22 nm FD-SOI, eMRAM and 5G wireless communications [85]. For advanced driver assistance systems (ADAS), cars will need to have up to five cameras per car and the image processors must be fast enough to react instantaneously to objects in the path of the car. GlobalFoundries expected FD-SOI technology to improve the linearity and insertion loss of RF circuits. Automotive MCUs for under-the-hood applications must operate reliably at 125 to 150 °C ambient with an even higher junction temperature. Under-the-hood applications include: powertrain, body, and safety systems. These high temperatures result in leakage being about 30% of total power for 55 nm bulk silicon technology since, in CMOS, leakage increases exponentially with temperature, whereas in 22 nm FD-SOI leakage is decreased and the technology is shrunk. High-k dielectrics will be required for automotive technology at the 28 or 22 nm rather than the replacement gate or gate last techniques used in other foundries. For in-cabin systems, such as infotainment, the temperature requirement is lower. Some examples of Flash MCU for various recently announced automotive applications are shown in Figure 2.44 along with the type of MCU and the amount of Flash expected to be embedded in the MCU. These and other examples of automotive Flash MCU are discussed in this section.

More detailed descriptions of various Flash microcontrollers that have been recently introduced for automotive applications are shown in the rest of this section as an indication of the characteristics of this application. Electronic control systems for automotive have meant that many Flash MCUs have been used in automobiles. Embedded Flash memories are used for control program code storage and for temporary data storage to provide flexibility for program update and improved control.

Requirements for automotive eFlash were discussed by Renesas in January of 2016 [86]. These included: operating and data retention junction temperature up to 170 °C for

Application	MCU	eNVM (MB)	eNVM Types
Auto entertainment	32-bit	2 MB	eFlash
Sensor	16/32-bit	2 MB	eFlash
Driver assist	32-bit	1.25 MB	eFlash.
Control	32-bit	8-MB	eMONOS

Figure 2.44 Examples of automotive applications, typical MCU, typical embedded NVM capacity for applications, and types of embedded nonvolatile memories used in these applications.

under-the-hood applications and random read access throughput, particularly for high end Flash MCU. Over 100 MHz random read access frequency is needed to achieve real time engine controls for higher fuel efficiency and for advanced driver assistance systems (ADAS). This means that fast error corrected throughput for the Flash memory is needed for advanced automotive applications. Since ECC and parity check slow down the operation, they are not applicable. This higher data reliability needs to be achieved by the memory cell itself and by very accurate read out circuits. In addition, the requirement for high speed data processing in automotive Flash MCU has required the integration of high speed CMOS logic circuits and embedded Flash and it is important to balance sufficient performance of the logic CMOS transistors and the memory cell reliability.

ADAS systems in autos include: multiple all around cameras and radar systems. An example of a car with radar systems is shown in Figure 2.45 [87]. These ultrasonic systems could be used for ultrasonic parking assist and blind spot detection. Automated radar sensors could be used for cross traffic alert, lane change assist, backup aid, side impact detection, and emergency braking, as well as adaptive cruise control.

New applications in automotive add new requirements for eFlash. For example, in-field programming updates by wireless communications and over-the-air (OTA) are expected. Low EMI noise in-field programming is necessary for wireless communications that could be affected by various sources, including by operating noise generated by charge pump circuits in eFlash.

Many ADAS functions are expected to come before fully autonomous vehicles are common. These systems will require changes in the current architecture of the automobile control systems. Today cars typically have more than 50 electronic control units (ECU) that manage many aspects of the car's performance and behavior [88]. The increasingly sophisticated controllers required for ADAS functions will require increased computing performance and connectivity as well as higher levels of safety and security. A potential future control system partitioning with ADAS, and eventually

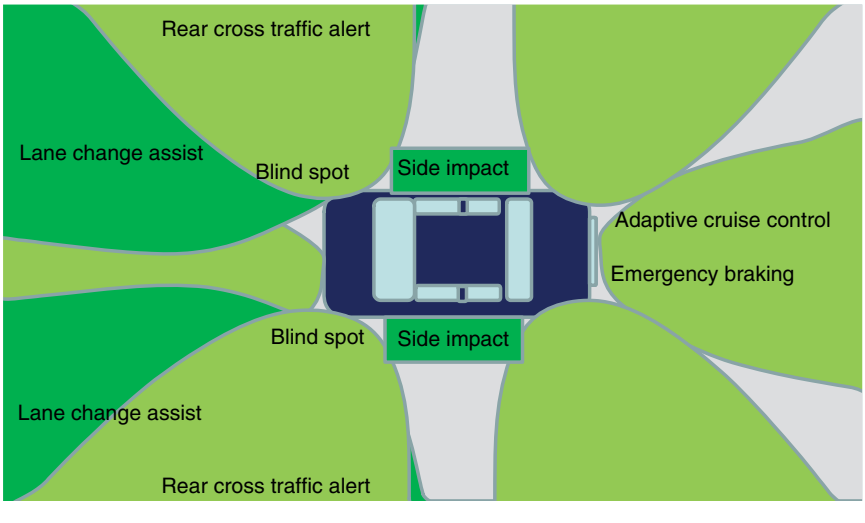


Figure 2.45 Illustration of typical ADAS system concepts in an automobile radar system. Based on Autoliv, October 3, 2016 [87].

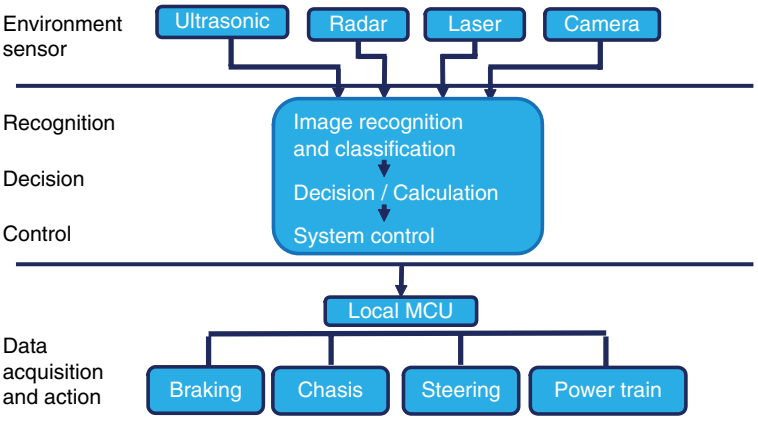


Figure 2.46 Possible future control system for ADAS and autonomous vehicles. Based on J. Scobie and M. Stachew, *EE Times*, October 29, 2015 [88].

autonomous vehicles, is shown in Figure 2.46. At the lowest level of complexity will be the need for advanced environmental sensing, which will require ultrasonic, radar, lidar, and camera/vision systems. At the next level of complexity is recognition devices and classification systems. Calculations and decisions need to be made next, followed by either driver notification and/or automated system control response. The response systems in the automobile are at the next level and consist of: braking control systems, chassis control such as air bag activation, steering control such as in response to lane recognition, and powertrain control such as needed for active collision avoidance.

Initially these systems will be driver alert systems, followed by drive assist systems, followed at some time in the future with autonomous drive systems. Autonomous cars will be the ultimate feature.

2.9.2 Early Advanced Automotive Driver Assistance Systems

Early automotive control systems included such features as headlight control, electronic braking, and anti-skid braking. Examples of such systems are shown in Figure 2.47.

The early automotive control systems began to be introduced in the new millenium. An automotive meter control for instrument clusters for two wheeled vehicles was introduced by Renesas in October of 2012. This 16-bit MCU was designed for two wheel and affordable four wheel vehicles for markets such as China, India, and Brazil [89]. They were intended to replace mechanical meters and gauges with electric meters where there is limited space for such meters. They included 256 KB of on-chip Flash memory and LCD panel display capability. Odometer data could be saved on-chip so an

Vendor	System	MCU	eFlash	Technology
Renesas	Headlight control	32-bit	8 MB	40 nm
Freescale	Electronic braking	Quad-Core	4.75 MB	55 nm
TI	Anti-skid braking	32-bit	1.25 MB	55 nm

Figure 2.47 Examples of early advanced driver assistance systems.

external EEPROM is eliminated. One to four channels of stepping motor controller/driver were included along with a sound generator.

Controllers for automotive body applications were also introduced early. In September of 2012, Renesas announced a 32-bit MCU for automotive body applications [90]. The 40 nm chips had from 256 KB to 8 MB of on-chip Flash memory. The eFlash chips were made in the MONOS charge trapping process. Power consumption for the MCU was 0.5 mA/MHz. An LIN interconnect network was on-chip along with data encryption. The MCU core could operate at 80 MHz. It is targeted at air conditioner control or LED headlight control. The embedded Flash memory was intended for storing instructions and data. It could be reprogrammed up to 125 000 times per 64-byte block, which permitted it to substitute for external EEPROM in many cases.

Electronic Braking was also an early driver assistance system. In November of 2012 Freescale announced a quad-core MCU with two pairs of redundant cores for advanced electronic braking systems [91]. The chip used Continental's fault recovery technology. The MCU contained 4.75 MB of Flash memory, 256 KB of SRAM, and Continental's fail-safe technology. The MCU was built in 55 nm technology.

Driver Assistance systems include electric power steering and anti-skid control. An MCU chipset intended for use in these systems was introduced in October of 2012 by Texas Instruments. This chipset included a safety MCU with RISC core from ARM, which operated from 80 to 180 MHz using up to 1.23 MB of Flash memory [92]. It also included a multirail safety power management IC and a safety motor driver. These chips formed a safety motor control chipset intended for failure detection and mitigation. Target applications included advanced driver assistance systems, electric power steering, anti-skid control, and other driver assistance systems.

Self-parking autos were prototyped in 2013 with the Ford fully automated self-parking car. Active Park Assist, which aided in parallel parking, was already available. Obstacle avoidance systems to actively detect objects in the road and warn drivers were being developed.

In October 21, 2014, TI announced a family of infotainment processors with added ADAS signal processing functionality [93]. The 1.4 GHz DSP had additional USB, video input, and PCLe interfaces. A second DSP core with 1.4 GHz of signal processing could be used to dynamically stitch multiple cameras into a single view. The added DSP performance included audio and speech processing and voice recognition. The processors included an RISC ARM core capable of supporting up to four 1080 displays and four 1080 video and camera inputs. High definition displays and megapixel cameras could be connected. Video, audio, and bidirectional I2C control signals could be carried concurrently over a single twisted pair cable. A power management IC was designed for seamless operation with the SoC.

2.9.3 More Recent Advanced Driver Assistance Systems (ADAS)

In January of 2016, NXP announced a single chip automotive radar sensor [94]. This 77 GHz radar transceiver provides precise data about the surrounding driving environment. Its small size permits it to be integrated invisibly anywhere in the car. Its power consumption is 40% lower than conventional radar ICs. Applications include: emergency braking, adaptive cruise control, blind spot monitoring, cross-traffic alert, pedestrian protection systems, and automated parking. For a 360 degree view a number of

radar sensors is required for robust high resolution. Future projects include a single chip automotive radar SoC, which might include the radar sensor.

Advanced driver assistance systems (ADASs) were targeted by GlobalFoundries in March of 2016 with plans for automotive ICs that used faster processors based on 22 nm fully depleted SOI (FD-SOI), embedded MRAM, and 5G wireless communications circuits [95]. These ADAS systems in a car will require either radar or up to five cameras per car and the image processors must be fast enough to react instantaneously to things in the path of the car. GlobalFoundries expects that FD-SOI technology will improve the linearity and insertion loss of RF circuits.

2.9.4 Automotive Navigation and Positioning

Satellite positioning for automotive navigation has been developed. In January of 2014, STMicroelectronics announced its Teseo III single chip positioning product family [96]. These chips are capable of receiving signals from multiple satellite navigation systems, including the Chinese BeiDou, the US GPS, European Galileo, Russia's GLONASS, and Japan's QZSS. The chip creates state-of-the-art positioning accuracy and leverages STM's sensor fusion as developed for automotive dead reckoning and assisted Global Navigation Satellite System (GNNS). By adding China's BeiDou satellite tracking capabilities to the chip, the Teseo III increases accuracy and location precision in Asia as well as throughout the world. It will permit in-car and marine navigation systems and telematics systems. It has good accuracy across urban, rural, and mountainous areas. The Teseo III is a single chip integrating RF, digital control, and Flash memory to enable simultaneous tracking of multiple satellites, even in areas with obstructions to satellite visibility. It has high indoor sensitivity. The Teseo III was expected to be available for sampling in Q12014.

Automated road access that includes RFID auto identification systems are already a reality. While the initial tags included a battery powered radio transmitter, the newer tags are adhesive stickers with a button sized microchip and a reflective antenna system that uses the windshield as part of the antenna. Removing the sticker after it is placed on the windshield separates the chip from the antenna, which invalidates the tag and helps prevent fraud. The glass windshield provides an amplification effect that is significantly reduced if the tag is not mounted on glass. While the Houston system uses an ATA communication protocol, other systems also use eGo and SeGo, which the tags also recognize along with ANSI INCITS 256-2001 and ISO 10374. The scanner range is 31.5 feet at 915 MHz. The 915 MHz RFID tag includes a 2014-bit read/write passive memory.

2.9.5 Under-the-Hood Applications

Under-the-hood automotive applications such as powertrain require nonvolatile memory. NOR Flash memory tends to be used for both code and data storage due to its high temperature reliability, 20 year data retention, and sub 30 ns random read access time. In September of 2013, Infineon and Technical University Munchen discussed a 65 nm eFlash macro for under-the-hood automotive applications [97]. A NOR Flash was used for the fast read access time. Read throughput was 5.7 GB/s and write throughput was 1.4 MB/s. The high read throughput rate used a multiple voltage domain multiplexer

that permitted a low voltage read path. It also used a local ground referenced read circuit using a time domain source side sense amplifier. This permitted a low voltage sub-50 mV swing read operation for high speed readout in the presence of more than 30 mV system noise. The triple poly-embedded Flash memory cell allowed a sub-5 μs low current write operation that enabled a high write throughput up to a junction temperature (T_j) of 170 °C.

The 4 MB eFlash macro was made in 65 nm technology. The array was split into two memory banks of 2 MB each. To achieve the high read throughput, every memory bank had its own read path and used 280 sense amplifiers, which can be read in parallel. An analog block had charge pumps and reference circuits. A hierarchical bit-line architecture was used to reduce RC delay. To reduce RC delay of the word-line, the memory array of each bank is split into two tiles of 1 MB each with word-line drivers between. The macro uses a hot source triple poly cell with dedicated select transistor and memory transistor, which permits fast write. Features for the memory macro are shown in Figure 2.48.

In April of 2014, Freescale, now NXP, announced its AEC-Q100 qualified battery sensor with CAN protocol for mission critical automotive applications [98]. The part had a 16/32-bit MCU and a CAN protocol module in a single package. The battery sensor integrated a 16/32-bit S12Z MCU with 128 KB of Flash, 8 KB of SRAM, and 4 KB of EEPROM. The part measured key battery parameters and supported 12 V lead acid batteries along with 14 V stacked cell Li-ion, high voltage junction boxes, and 24 V truck batteries. New features putting additional demand on batteries included: start–stop functionality, regenerative braking, and intelligent alternator control.

Flash MCUs with embedded Flash memory are essential in automotive applications for complicated real time controls for higher fuel efficiency, better safety, and increased connectivity. High density, fast random read access is needed from the eFlash. The requirements for eFlash in automotive are secured operations and data reliability, in particular reliable program/erase cycling at high temperature.

Automotive powertrain control applications require high performance CPU to help boost fuel efficiency, high speed network functionality to accommodate more sensors, and enhanced security to prevent unauthorized access and ensure safety. In April of 2016, Renesas introduced a 32-bit MCU for automotive powertrain control [99]. The device was intended to meet increasing fuel efficiency goals along with engine control systems to optimize ignition and fuel injection timing along with boosting processing performance to accommodate increased levels of calculations. The MCU core had improved memory access latency at an operating frequency of 320 MHz at a junction

Code flash capacity	4 MB
SA Supply	1.3 V
Operating temperature	–40°C – +170°C
Read throughput	5.7 GB/s
Write throughput	1.4 MB/s
Rewrite time	3.5 s/2 MB

Figure 2.48 Features of memory macro for automotive powertrain applications. Based on M. Jefremow *et al.* (Infineon, TU Munchen), ESSCIRC, pp. 193, September 2013 [97].

temperature of 150 °C. It used the new Single Edge Nibble Transmission (SENT) digital communication standard for sensors and CAN FD5, which accommodates the increased volume of data from the increasing number of sensors, up to 50 currently. Security functionality was improved by the inclusion of an Intelligent Cryptographic unit hardware security module, which provided functions such as data encryption and cipher message authentication and random number generation.

2.9.6 MONOS Memory for Under-the-Hood Applications

Processors for automotive under-the-hood applications were discussed in April of 2015 by Renesas using embedded charge trapping memory of the metal oxide–nitride oxide semiconductor (MONOS) type, which can withstand the high temperatures associated with this application [100]. Over 100 Flash MCUs are used in the average car for various vehicle controls such as chassis, safety, body, and powertrain. The gas engine control is a high temperature application in the powertrain that requires intensive computation with a large program memory size at a junction temperature (T_j) of 160 °C. Operating frequency for an automotive MCU by year is shown in Figure 2.49 along with frequency for the Renesas embedded MONOS Flash macro compared to an estimate of the operating frequency for floating gate Flash.

In February of 2016, Renesas discussed a 90 nm one transistor MONOS Flash memory that achieved over 100 million program/erase cycles at a high temperature of 175 °C with rewrite energy of 0.07 mJ/8 KB [101]. The 1 T-MONOS Flash technology can be used with both CMOS and DRAM. It is expected that the Flash technology will enable Flash memory to be added to automotive analog devices, improving performance and reliability. In particular, the number of electronic control units (ECUs) being added per vehicle has been increasing and automotive control is becoming more complex. As a result, making the ECUs lighter and more power efficient is increasingly important.

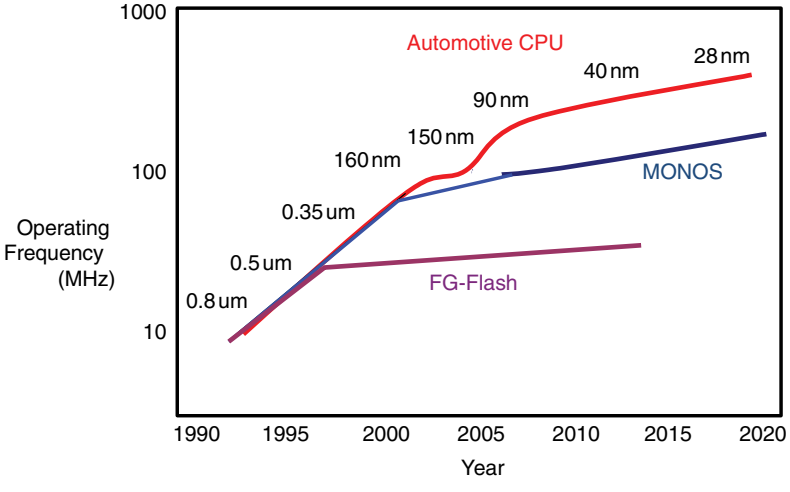


Figure 2.49 Operating frequency for an automotive MCU by year showing MONOS and FG Flash. Based on T. Yamauchi (Renesas), VLSI-TSA, April 2015 [100].

Characteristics	Code flash	Data flash
Random read frequency	200 MHz	10 MHz
Program speed	2.0 MB/s	150 μs/4 B
Erase speed	0.91 MB/s	1.5 ms/64 B
Endurance	10 k cycles	>1 M cycles
Capacity (Max)	32 MB	412 KB

Figure 2.50 Code and data Flash characteristics in 28 nm SG-MONOS embedded Flash memory.

The increase in number of motors in radiators, water pumps, and vehicle air conditioning systems has created a requirement for unification of mechanical and electrical elements so a technology that can add Flash memory to automotive analog and power controlling CPU devices is important. The high cycle endurance permits applications such as automatic calibration or status recording using high frequency sampling under actual field usage conditions. The current consumption during P/E operations is 98 μA. An idling program erase management unit function permits the Flash memory to control rewrites when the vehicle's engine is stopped by an anti-idling system. This makes it possible to stop the CPU and SRAM and reduce power consumption during idle by 99%. Trends for the code and data Flash for a 28 nm SG-MONOS embedded Flash memory is shown in Figure 2.50. The memory cell size is 0.053 μm^2 and operating temperature for the chip is specified at −40 to +170 °C junction temperature.

Automotive temperature MRAM is also required in motorcycles. In May of 2015, Everspin announced that its automotive temperature MRAM meets the environmental requirements of the BMW 1000 RR Racing Bike [102]. A family of AEC-Q100 fully qualified MRAM products for automotive applications has been introduced. The part is used in the BMW Motorrad Motorsport engine control unit, the RSM5, storing calibration data that controls the motorcycle during a race.

2.9.7 Automotive Infotainment

Devices intended for automotive infotainment require high speed interface, fast processor cores, and large amounts of embedded SRAM and NV Memory. In October of 2014, Atmel announced sampling of a family of MCUs with RISC core from ARM [103]. The devices ran up to 300 MHz and had 384 KB of SRAM as a system memory and up to 2 MB of on-chip Flash memory. Ethernet AVB and Media LB peripherals made them suitable for automotive connectivity and audio applications. The devices also target applications including: the Internet of Things (IoT) and wearable devices. The devices include USB OTG and high speed USB PHY. The automotive series has Ethernet AVB support and Media LB for combination with Cortex-M7 DSP for infotainment connectivity and audio applications.

Other examples of devices used for automotive infotainment include MCUs with video graphics. In March of 2016, Cypress Semiconductor discussed an expansion of its Automotive MCU family [104]. The MCU includes advanced graphics. The new MCU provides automotive manufacturers with a low cost platform for 2D and 3D graphics and enhanced functionality for dashboards, heads-up displays and HVAC systems in

compact vehicles. An earlier announcement of a 2D and 3D graphics MCU series for midsize cars was made in October of 2014 [105]. The graphics engine provides memory savings, increased safety features, and image capability without requiring external video RAM. This new family includes: an RISC processor core from ARM, 1 MB of embedded Flash, 1 MB of internal video RAM, enhanced secure hardware extension, and a HyperBus interface that permits connections with other HyperBus memories. These products are pin compatible with their predecessors, which provides flexibility and ease of upgrade with no changes in board layout. The chip also includes an advanced sound system and 16-bit Audio-DAC with multichannel mixer.

2.9.8 Secure Automotive

The 150–250 million connected cars expected on the roads by 2020 will be kept safe by chips with secure MCU. In February of 2016, STMicroelectronics announced a tamper-proof MCU that was qualified for automotive applications to protect data privacy and system integrity [106]. Internet connectivity was expected to bring many services to vehicles such as software updates, service packs, location-based information, social media, streaming entertainment content, and emergency support. The MCU was qualified to the automotive AEC-Q100 standard. Additional security implemented by unalterable hardware permits use as an embedded Secure Element. Robust security in autonomous and driverless vehicles will enable car sharing. The part has 1280 KB of on-chip user Flash to manage connections to multiple cellular networks.

2.9.9 Automotive Body Processors

Body controllers in automobiles include applications such as door modules, seat modules, and central body control. These body processors in automotive include airbag controllers such as a new automotive 55 nm 32-bit MCU SoC designed for entry-level vehicle safety-critical applications [106]. This chip supports airbags and anti-lock braking systems in cars and motorcycles as well as power steering and DC/DC converters for hybrid/electric vehicles. The part has clock speeds up to 80 MHz and is designed to resist soft error failures due to cosmic rays in vehicle control functions such as steering and braking.

Another example of an automotive body processor MCU was announced by NXP in April of 2016. It included high performance communication capabilities including: FlexRay, CAN, and LIN. It had up to 1.5 MB of embedded Flash memory for data storage [107]. Both the Flash memory and the SRAM included ECC. There was a 16 channel enhanced DMA and a memory management unit with a four-entry translation look-aside buffer. Operating frequency was 80 MHz.

For enhanced security, NXP also offered in July of 2016 a reliable 32-bit MCU for automotive next generation automotive central body control [108]. The 55 nm chip had a hardware security module to protect against various hack attack scenarios. The previous generation chip was made in 90 nm technology. The device contained two 160 MHz cores along with an 80 MHz core for flexible power performance. There were 6 MB of embedded NVM Flash memory and 768 KB of eSRAM. New low power modes were supported. The hardware security module exceeded the requirements of the secure hardware extension (SHE) of the Hersteller Initiative Software standard. The devices

were targeted at applications such as: door modules, seat modules, central body, vehicle body controllers, smart junction box, front module applications, and high end gateway applications.

2.10 Smart Electrical Grid and Digital Utility Smart Meters

2.10.1 Overview of the Smart Meter Market

Smart utility meters are configured for two-way communication. They record the time of day that electricity is used, permitting different prices during different times. Power quality issues and other characteristics may also be recorded and transmitted.

Security issues with the utility infrastructure have been given serious consideration. Renesas forecasted in 2016 that the number of installations per year of smart meters with enhanced security features was expected to grow from 30 million in 2016 to 110 million in 2020 worldwide. Key issues have become security of power meter data and lowering meter prices [109]. The smart meter market by 2014 was estimated by Renesas at over 100 million smart meters in operation. This was in part driven by the requirement for home energy management systems [110].

2.10.2 Smart Meter Chips with Embedded Flash Memory

Smart meter chips use a range of MCU sizes running at speeds up to 80 MHz, active power in microamps to milliamps, and standby power in the microamp range. A summary of various smart meter chips is shown in Figure 2.51. The MCU and embedded nonvolatile memory contained in these chips are shown along with active current and standby current.

2.10.3 Smart Meter Chips with Large Embedded Flash Memory

Smart electricity meters log electric power parameters, such as power consumption, active power, reactive power, load conditions, voltage, and frequency distortion at regular intervals of milliseconds to seconds. Data are stored in the meter's nonvolatile memory for periodic transmission to the system or the power grid.

Vendor	MCU	Capacity	eNVM	I(active)	I(standby)
Renesas	32-b	512 KB	Flash	1.0 μA	0.3 μA
Microchip	8-b	1 MB	Flash	1 mA/8 MHz	10.0 μA
NXP	32-b	128 KB	Flash	40 μA/MHz	1.7 μA
Silicon labs	32-b	32 KB	Flash	114 μA/MHz	0.9 μA
TI (FRAM)		128 KB	FRAM	140 μA	

Figure 2.51 Smart meter chips from various vendors showing MCU and embedded nonvolatile memory along with chip active and standby current.

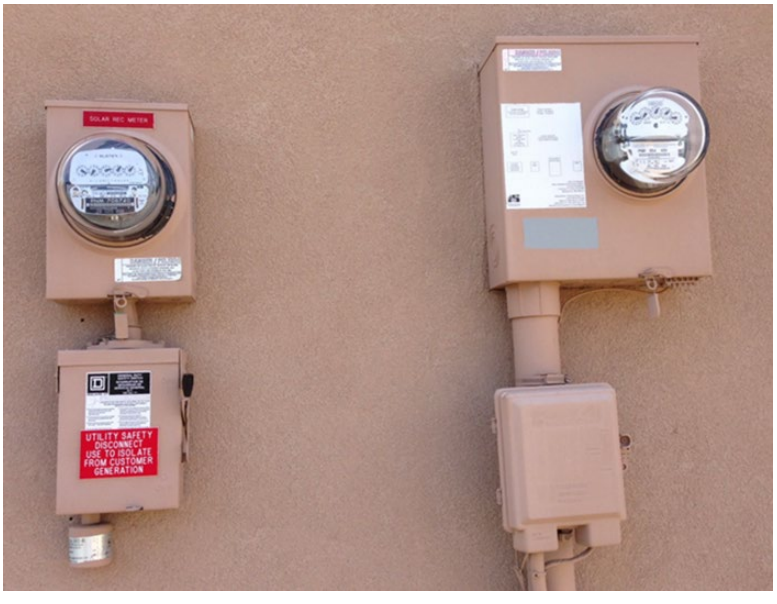


Figure 2.52 Illustration of electric meters connected between a home solar generating facility and an electric utility. The meter on the left measures electricity generated by the solar array and the one on the right shows net usage of the community and solar facilities. Photo by Dr. J.H. Ericksen, with permission.

A 32-bit MCU intended for smart meters with advanced functionality was introduced by Renesas in September of 2012 [110]. The MCU includes 512 KB of embedded Flash memory, a 32-bit CPU core, and a 24-bit delta-sigma A/D converter for high resolution measurement. The need for large Flash memory is driven by the need to accommodate larger programs. Power consumption in standby when the RTC is operating is 1.0 and 0.3 μA when the RTC is stopped.

A single chip powerline SoC for smart metering applications was announced by Atmel in October of 2012 [111]. The MCU has an RISC core from ARM, 2 MB of dual-bank Flash memory, 160 KB of SRAM, and a fast PLC transceiver. The chip was compliant with the PRIME (Power Line Intelligent Metering Evolution) PLC specification. The SoC included an integrated low power amplifier that consumed 40% less power than comparable solutions. It includes an SPI. TSI and USARTS to support various communication interfaces are required in smart metering.

A photo is shown in Figure 2.52 of the electric meters in a home associated with a home solar generating facility and a community electric utility. The meter on the left runs forward and measures electricity generated by the rooftop solar array. The meter on the right is connected to the community electric utility and indicates the electrical usage from the utility net of that generated by the solar panels. It runs both forward and backward depending on the net direction of the electricity being generated.

An energy efficient 32-bit sub-GHz wireless MCU was announced in October of 2012 by Freescale (now NXP). The MCU, which used a 32-bit low power RISC core from ARM, ran up to 48 MHz with 128 KB of Flash and 16 KB of SRAM [112]. Active power consumption was as low as 40 μA/MHz. Standby current was 1.7 μA/device with a 4.3 μs

wake-up. The device was suited for wirelessly networked smart energy applications. The chip has a 600 Kbps radio with complex modulation schemes and multiple frequencies supporting US and International protocols. It can support neighborhood area networks connecting smart meters to a common data collection point or networks of street lamps or sensors. This chip can run low level wireless protocol layers while delegating upper network protocol layers to an external host processor.

Development of a chip intended for smart electricity meters and data concentrators was announced in January of 2013 by the Atmel and Wasion Group [113]. The chip was to be based on several Atmel chips and an MCU with an RISC core from ARM; 2 Mb of Flash memory was to be included on the chip.

An MCU, targeted at the smart meter application, offering up to 8Mbits of embedded Flash memory, was introduced in February of 2013 by Microchip Technology [114]. The embedded memory was made in a split-gate, NOR embedded Flash technology with a thick oxide tunneling injector. The parts have an extended operating voltage range from 2.3 to 3.6 V. The memory is partitioned into uniform 4 KB sectors and 32 and 64 KB blocks, offering seamless partitioning for program and data code in the same memory block. Sector and block erase is as fast as 18 ms and the entire chip can be erased in 35 ms. The word program time is 7 μs. Endurance is 10^5 cycles and more than 100 year data retention. Active read current is 10 mA typical at 80 MHz and standby current is 10 μA typical.

As the rate of smart meter deployments continue to rise in Europe and Asia, the demand for smarter and better integrated meters continues. In October of 2014 [115], Atmel announced a dual core secure MCU for smart metering applications. Each of the dual cores has 2 MB of cache-enabled dual bank Flash, which are pin-to-pin compatible with existing 512 KB and 1 MB devices already in the field. The parts are built on a dual-core 32-bit RISC architecture. The devices also include: advanced security features, a low power real time clock, LCD driver, and several serial interfaces.

A low power MCU with up to 128 KB of ferroelectric nonvolatile memory replacing the Flash for low power data storage was discussed in September of 2014 by Texas Instruments [116]. The FeRAM MCU included an extended scan interface that autonomously manages and monitors four external sensors. It was targeted for flow meters for water, gas, and heat. There was a 140 μA differential input ADC. Other peripherals included: a 320 segment LCD controller, a 256-bit hardware encryption accelerator for secure data transmission, and an intellectual property encapsulation module to protect code. The part had a fast write speed and low peak currents.

In March of 2015, Silicon Labs introduced an ultralow power 32-bit MCU with a low power RISC core from ARM for Smart Metering applications [117]. The device had a short wake-up time from energy saving modes and a selection of peripherals. Features included a 24 MHz RISC MCU from ARM, up to 32kB of Flash, and a 4 KB RAM memory. A single 1.98–3.8 V power supply was included and the devices worked over the industrial temperature range from –40 to 85 °C. Flexible energy management included: 20 nA shutoff mode, 0.5 μA stop mode, with power-on reset, brown-out detection, RAM and CPU retention, 0.9 μA deep sleep mode, 48 μA/MHz sleep mode, and 114 μA/MHz run mode with code executed from Flash.

Smart electricity meters using ferroelectric RAM memory were discussed in November of 2015 by Cypress Semiconductor [20]. Smart meters have conventionally used serial EEPROM to record data such as power consumption, active power load

conditions, voltage, and frequency distortion. A 1 MHz 1 MB I2C EEPROM with 256 B page takes 28 ms to back up 1 KB of data, while an FeRAM takes 8 ms to write 1 KB of data into the FeRAM. FeRAMs consume about one-third of the active current of EEPROM while the standby current is nearly the same.

In November of 2015, Microchip announced an addition to its low power MCU portfolio which includes up to 1 MB of dual partition Flash memory with ECC and 32 KB of RAM [118]. This permits programming of one partition while executing application code from the other. The chip is targeted at industrial, computer, medical, and portable applications that require a long battery life and data transfer and storage without needing external memory such as electrical meters.

An MCU with enhanced security function and improved arithmetic operations was announced in August of 2016 by Renesas for the smart power meter market [109]. The part retained the measurement accuracy and the low power consumption of the previous chip while adding the extra security functions. On-chip memory capacity ranged from 64 KB for low end single phase meters to 128 KB for high end single phase meters and up to 256 KB for three phase meters. Security was enhanced by requiring authentication and encryption/decryption functions. An AES hardware engine was integrated for this purpose, which boosted speed by 20 times or more compared with en/decryption using software processing. The on-chip PLL boosts the maximum operating frequency from 24 to 32 MHz. The 32-bit multiply and accumulate unit reduces the load imposed by software power calculation by using a 24-bit Delta Sigma A/D converter, which results in an approximate 30% increase in arithmetic capacity. The chips can support distribution line message specification (DLMS) standard communications while processing electric power measurement. This standard was enacted for smart meters. Program code is stored in 256 KB of on-chip ROM for both electric power measurement and DLMS processing, leading to 30% reduced power consumption compared to a two chip implementation. System cost is also reduced. Standby mode current consumption is typically 0.7 μA during operation. An independent power supply and power monitoring with battery backup enabled the CPU to continue operating even in a power outage.

2.11 Consumer Home Systems and Networks

Home IoT control systems potentially include: environmental control, access control, vision systems, light control, comfort control (auto adjusting beds and chairs), information control, entertainment control, food and water control, safety alerts (tornado, fire, flood), animal access, and care and child monitoring and care. Initially remote control of these systems cannot be provided to the owner. Eventually such systems will be automated and robotic. Figure 2.53 indicates the home network of connected systems.

2.11.1 Remote Controls

In October of 2012, TI announced its low cost, value 2.4 GHz RF line of SoC for remote control applications, which supported 2.4 GHz frequency bands [119]. The SoCs integrate an MCU, 32KB of Flash and 2KB of RAM on chip. The data rate is 2 Mbps with <1 μA sleep current with the sleep timer running.

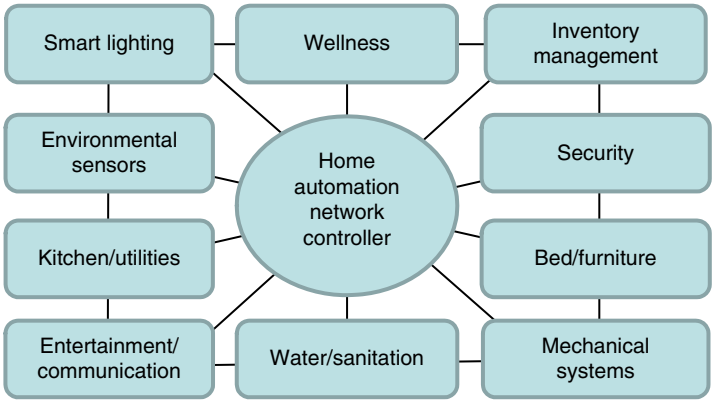


Figure 2.53 Home network of connected systems with central home automation.

In January of 2013, Energy Micro announced an MCU expected to be used in the VELUX touch screen remote control, where it keeps the handheld control asleep with a 1.2 μA low energy capacitive tough sensor interface [120]. The 32-bit MCU has 1 MB of Flash memory with the added option of embedded USB connectivity. CPU speed is up to 48 MHz. The MCU includes autonomous low energy peripherals such as AES encryption, a pulse counter, low energy UART, and others.

2.11.2 Environmental Sensors

Sensors for the internal home environment are useful in home networks and may use energy harvesting to avoid battery maintenance. In June of 2016, Cypress Semiconductor discussed a tiny light powered sensor beacon that wirelessly transmitted data it had sensed in the surrounding environment [121]. This sensor can sense temperature and humidity around its location and transmit the data using Bluetooth Low Energy. The beacon is 25 mm in diameter and needs no battery maintenance. It is powered by indoors ambient light as low as 100 lux, which is less light than a typical warehouse aisle receives. The beacon can be connected to a host PC using a USB interface for data and debugging. It comes with an app for IOS, Android, and PCs, which enables transmitting data as a graph or data. Its startup power is 1.2 μW and current consumption is as low as 280 nA.

2.11.3 Home Network Systems

Smart lighting and home automation require wireless sensor networks. In November of 2012, NXP announced a family of wireless MCU for ZigBee, JenNet, and other IEEE 802.15.4 applications [122]. The chip integrated a 2.45 GHz radio, analog peripherals with MCU on a chip. It included up to 256 KB of eFlash, 4 KB of EEPROM, and 32 KB of SRAM to support network stacks. It has an in-packet antenna diversity that chooses the best antenna on every packet received. Target application is the home networks of connected systems. This includes such things as smart lighting, home automation, and wireless sensor networks as well as other home essentials, as was shown in Figure 2.53.

2.12 Motor Control Chips with Embedded Memory

Flash MCU for motor control falls into two categories, small motors with 8-bit MCU and up to 64 KB of embedded Flash and more complex motors with 32-bit MCU and up to 1 MB of eFlash.

2.12.1 Small System Motor Control Using Embedded Memory

Small system motor control is required in household appliances and home healthcare as well as for fans, tools, battery charges, and many other consumer home appliances that can be connected.

An MCU family targeted at low power consumer, household appliance, and home healthcare was announced by Renesas in September of 2014 [123]. A motor control kit was being offered to support Internet and cloud connectivity. The kit supports three-phase and brushless DC motors. The MCU has a 140 μA/MHz RL78 core, 5 V power supply, and up to 256 KB of Flash memory.

An 8-bit MCU portfolio for IoT with one version intended for small system motor control applications such as fans and tools, power supplies, battery chargers, and sensor controllers was introduced in February of 2016 by Silicon Labs [124]. There are three versions, which include low power capacitive touch control and USB connectivity. Core speeds are up to 50 MHz and there are 2–16 KB Flash sizes. Peripherals include a 12-bit analog-to-digital converter (ADC).

In March of 2016, Toshiba announced an RISC MCU in 65 nm logic with embedded Flash memory [125]. The MCU included high performance analog and a range of functions for motor control. Flash memory ranges from 32 to 128 KB for code and 8 to 18 KB for data. Other features integrated included a high precision 12-bit ADC and an 8-bit DAC. The maximum operating frequency is 40 MHz. Peripheral circuits included on the chip include: UART, I2C, TSPI, and timers.

2.12.2 Motor Control for Multiple Motors Using Embedded MONOS Memory

In November of 2012, Renesas announced an MCU intended to provide inverter control for up to three 3-phase motors simultaneously [126]. A 32-bit RX CPU core operated at 100 MHz without a wait state using embedded MONOS Flash memory. The memory size has been increased up to 512 KB of Flash and up to 48 KB of SRAM. An additional independent 32Kb of data Flash was included for background operations to enable data to be written while a program is executing. Performance was 1.65 DMIPS/MHz at any CPU frequency without limitation from the memory.

In March of 2016, Renesas announced a 32-bit MCU group that enabled simultaneous inverter control of multiple motors with a single chip and doubled the performance of the MCU group [127]. These MCU were also targeted household appliances as well as industrial motor applications. On-chip Flash capacity ranged up from 128 KB. These MCU were ideally suited for permanent magnet AC (PMAC) motor control. They had 160 DMIPS with dedicated analog functions to increase the arithmetic operation and expand motor control peripheral functions. The MCU had on-chip FPU functionality

for faster code execution. The TXv2 CPU core had a five-stage pipeline, a floating point unit, and DSP functions. It ran at 80 MHz. The FPU functionality provided faster implementation of calculation of intensive algorithms with a wide dynamic range.

2.12.3 Motor Control with Embedded NV FeRAM

Low cost industrial and metering applications were targeted by TI in May of 2015 for their microcontroller with 64 KB of embedded NV FeRAM [128]. A new technology permitted context save and restore. The technology featured instant wake-up with intelligent system state restoration after an unexpected power loss. The chip also included integrated smart analog and digital peripherals to reduce system cost, power, and size including: a low power segment LCD controller, a 12-bit differential ADC with internal window comparator, a 256-bit AES accelerator, and LCD. The 16-bit MCU could be used for datalogging for wireless motor condition monitoring. The chip can be used with TI's energy harvesting IC and wireless MCU to operate from a coin cell battery for decades. The system can be used to monitor and analyze motor vibration frequencies to predict and schedule maintenance in automated environments.

2.13 Smart Chip Cards in Advanced Applications

For advanced inventory control applications, in March of 2014, EMM discussed a universal identity card and an encrypted tamperproof digital certificate [129]. The security measures of the device used public stream cipher technology to guard the memory data integrity. 2 K-bits of data memory were included in the device. Each device had a unique embedded EEPROM 64-bit identifier and optional 32-byte ECC-based digital certificate. An NFC enabled smartphone can read out these pieces of information and confirm the authenticity of a good by connecting to a cloud server that manages the collected data.

An advanced secure IC Flash MCU family for secure EMV (Eurocard/Mastercard/Visa) chip payment cards was announced in May of 2014 by ST Microelectronics [130]. This was intended as a family of secure payment card chips. Chip cards that embed an EMV-compliant secure chip can store and process data securely and are difficult to replicate. This family of chips was based on a secure IC in 90 nm technology. The cards used a 32-bit secure RISC core from ARM along with a contactless interface. The cards had up to 38 KB of installable EEPROM.

For payment applications, a third generation of secure MCU was announced in November of 2015 by STMicroelectronics. These chips had 2 MB of 40 nm eFlash program memory, fast clock speed, and a fast embedded cryptographic accelerator based on a 32-bit secure RISC core from ARM [131]. The hardware architecture provided multiple strong fault protection mechanisms for CPU, and memory and buses for secure software and systems. It met the need for secure applications in areas such as payment applications and wearable consumer devices, machine-to-machine communications in "Industry 4.0" factory automation, and automotive in-car communications. The peripheral had a secure Data Encryption Standard (EDES) algorithm and the NESCRYPT cryptoprocessor supported the public key algorithm. The 2 MB eFlash memory permitted the device to host a variety of applications.

For home banking and USB readers, in February of 2016, NXP announced its PN7462 family of NFC controllers that combine a low power 20 MHz ARM Cortex M0 MCU core with USB and NFC [132]. This chip had advanced power management as well as a 160 KB Flash and 4 KB EEPROM. Interfaces were: UART and ISO/IEC7816.

2.14 Analysis of Big Data Server Memory Hierarchy for Storing IoT

Items that can accessed and tracked through the Internet range from automobiles, to home security systems, to retail inventory, to factory items in process, to individual pieces of fruit in the grocery store. The current Internet protocol allows a massive number of items to be uniquely identified. Internet protocol version 6 (IPv6) uses 128 bit addresses that allow over 10^{38} separate items to be identified and located across networks. These networks require powerful processors to transfer and process huge amounts of information and large memories that can store this information.

Real time tracking can record both the flow of goods as they move through various systems over time, the changes in these goods, and also changes in measurements taken by stationary sensors over time. Flow could involve people flowing through cities, patients moving through hospital routines, goods moving through factories, energy flowing through networks, or automobiles moving along highways. Changes recorded could involve everything from processing of items being created in factories to food that is stored in refrigerators being consumed and aging.

This massive amount of "Big Data" is stored in large networks of servers, which are referred to as "the Cloud". The Internet of Things uses mainstream memories in large server farms where individual IDs for millions of sensors are stored and can be accessed from, for example, a cell phone. This involves sorting rapidly though large amounts of data using sophisticated search engines. A sophisticated memory hierarchy is required for rapid access of these large server farms. This is shown in Figure 2.54, which illustrates a cache hierarchy for servers that could be used in large server farms along with a potential search engine to be used for Big Data analysis. The circuits shown here were presented at the ISSCC Conference in February of 2016 and are intended to illustrate the state-of-the-art in these classes of circuits at a single point in time.

An example of a "Big Data" analysis search engine that operates on the data in the server farm could be a Ternary Content Addressable Memory (TCAM). SRAM-based TCAMs have been used in the past as search engines. A 256-bit 2.5 Transistor 1 Resistor (2.5T1R) Resistive RAM (RRAM) TCAM search engine was designed for this purpose in 2016 by NTHU, TSMC, and NCTU [133]. The RRAM was used to increase chip capacity by reducing cell area over that of an SRAM. An RRAM can also reduce search energy and standby power over that of SRAM-based TCAM in applications with long idle times and frequent-search-few-write operations. This RRAM cell was made in a 65 nm fully CMOS compatible process at the TSMC foundry.

Embedded SRAM has traditionally been used for cache in large servers due to its high speed, ranging from 1 ns to 10 ns in the various levels of cache. As advanced CMOS technologies have gone to lower power supply voltages, embedded SRAMs have increasingly suffered from subthreshold leakage, which affects their energy consumption and reliability. A magnetic RAM (MRAM) has sufficient capacity and switches fast

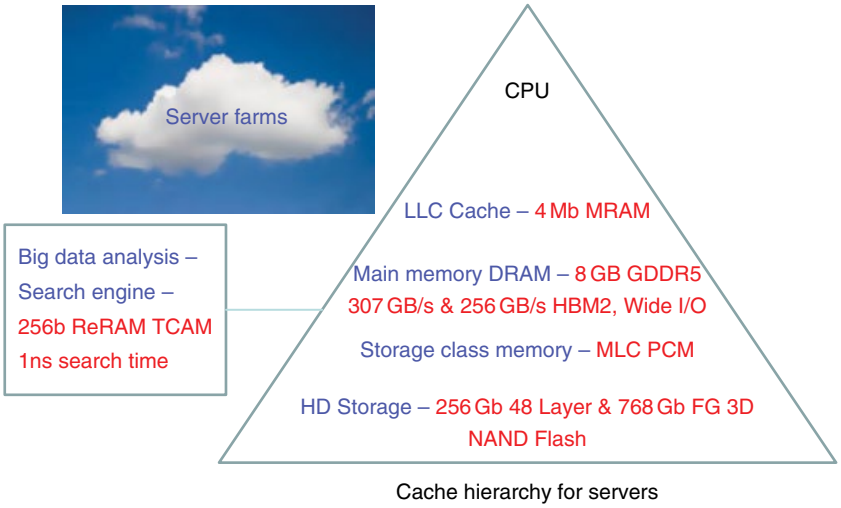


Figure 2.54 Memories for tomorrow's Cloud applications (collected from devices shown at the 2016 ISSCC).

enough to be used as a last level cache (LLC). It is nonvolatile so it consumes no power in standby, which enables it to reduce the total energy compared with a conventional SRAM last level cache (LLC). A 4 Mb magnetic RAM (MRAM) cache designed for use as an LLC was discussed by Toshiba. It was configured as a four-way SET Associative cache with a 362 K-bit SRAM TAG [134]. The MRAM cache reduced total energy by 93% compared with a conventional SRAM LLC with a read-modify-write function. Standby energy was reduced by 77%.

Below the cache in the memory hierarchy is the main memory, which consists primarily of high density banks of DRAM. The operation of the DRAM at about 100 ns is slower than the SRAM cache but much faster than the mass storage below it, which consists of high density nonvolatile Flash memory.

Capacity of a DRAM chip can be as high as 8 GB or 64 Gbits. A fast high bandwidth memory (HBM) wide I/O DRAM was shown by Samsung [135] with 8 GB capacity. It used the GDDR5 high bandwidth memory (HBM) interface. The HBM had 160 GB/s of memory bandwidth but uses up to 70% less energy per bit than other technologies.

A new category of memory being explored is the storage class memory, which is envisioned as bridging the speed gap between the main memory and the Flash memory mass storage. A concept of a classical memory hierarchy that includes storage class memory is shown in the diagram in Figure 2.55.

The access time latency goes up, that is the access time becomes slower, as we go down the memory hierarchy. The memory capacity, however, goes up. A processor can run at 1 ns, SRAM tends to run at 10 ns, and main memory DRAM at 100 ns. Mass storage speeds are around 100 μs with archival mass storage slower than a millisecond. The storage class memory is envisioned as filling the gap from 100 ns to 100 μs. A multilevel cell phase change memory (PCM) was shown by Macronix, NTHU, NCTU, and IBM and is an example of a potential storage class memory with access latency in the 1 to 10 μs range [136].

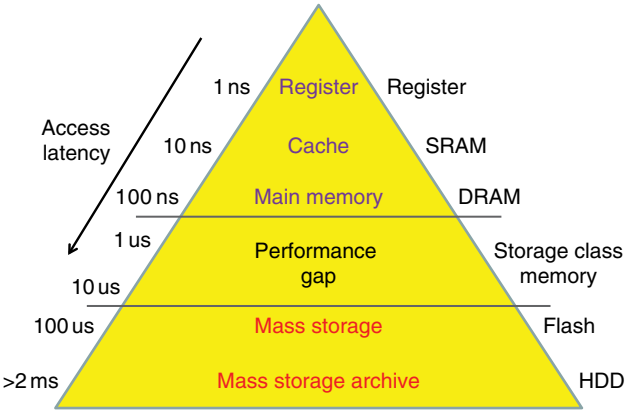


Figure 2.55 Memory hierarchy including the storage class memory.

Hard drive storage is currently provided by vertically stacked chips up to 256Gb with 48 layers of 3D Flash memory chips [137] and 768Gb 3b/cell floating gate 3D NAND Flash [138] or a 128Gb 2b/cell NAND Flash memory in 14 nm planar NAND Flash technology [139].

References

1 Kono, T. *et al.* (2014) 40-nm embedded split-gate MONOS (SG-MONOS) Flash Macros for automotive with 160-MHz random access for code and endurance over 10 M cycles for data at the junction temperature of 170 °C (Renesas). *IEEE Journal of Solid-State Circuits*, 49 (1), 154, January 2014.

2 Milani, L., Torricelli, F. and Kovacs-Vajna, Z. (2015) Single-poly-EEPROM cell in standard CMOS process for medium-density applications (Univ. Brescia). *IEEE Trans. on Elec. Dev.*, 62 (10), October 2015.

3 Kang, S.H. (2014) Embedded STT-MRAM for energy-efficient and cost-effective mobile systems (Qualcomm Technologies), VLSI Technology Symposium, June 2014.

4 Adesto Technologies and TPSCo Announce CBRAM Manufacturing Agreement, Adesto Press Release, February 25, 2016.

5 Joosting, J.P. (2015) Shanghai Huali Microelectronics looks to specialty technology to target IoT growth. *EE Times, January* 1, 2015.

6 Lee, Y.K. *et al.* (2014) A 45-nm logic compatible 4 Mb-split-gate embedded Flash with 1M-cycling endurance (Samsung), IMW, May 2014.

7 Yamauchi, T. (2015) Prospect of embedded non-volatile memory in the Smart Society (Renesas), VLSI-TSA, April 27, 2015.

8 Jew, T. (2015) Embedded microcontroller memories (Freescale), IMW, May 2015.

9 STM32F0 entry-level ARM Cortex MCUs (2016) STMicroelectronics website, August 2016.

10 Texas Instruments introduces 32-bit MSP432 Microcontrollers (MCUs): ultra-low power at its best, performance at its core, TI Press Release, March 24, 2015.

11 Jayamumar, H., Raha, A. and Rahunathan, V. (2014) QUICKRECALL: A low overhead HW/SW approach for enabling computations across power cycles in transiently powered computers (Purdue University), Conference on VLSI Design, January 5, 2014.
12 Kaneko, Y., Nichitani, Y., Ueda, M. and Omote, A. (2014) Battery-less impact-logging device consisting of a vibration energy scavenger and ferroelectric memory (Panasonic), Device Research Conference, pp. 65, June 22, 2014.
13 Ishii, T. *et al.* (2015) 0.6 V operation, 26% smaller voltage ripple, 9% energy efficient boost converter with adaptively optimized comparator bias-current for ReRAM Program in low power IoT embedded applications (Chuo University), IEEE A-SSCC, November 2015.
14 Cypress introduces the world's lowest-power energy harvesting power management ICs for battery-free wireless sensor nodes, Cypress Press Release, August 19, 2015.
15 BQ25570 ultra low power harvester power management IC with boost charger and nanopower buck converter, TI Data Sheet, 8/2016.
16 STM32 32-bit ARM Cortex MCUs, STMicroelectronics website, September 2016.
17 STMicroelectronics launches new STM32 ultra-low-power microcontrollers for consumer, health, and industrial applications, ST Micro Press Release, February 11, 2014.
18 Cypress introduces the industry's most energy-efficient ARM Cortex-M0+ MCUs for ultra-low-power Internet of Things applications, Cypress Press Release, December 2, 2015.
19 Panasonic starts world's first mass production of ReRAM mounted microcomputers, Panasonic Press Release, July 30, 2013.
20 Panasonic and UMC Partner for 40 nm ReRAM process platform, Panasonic Press Release, February 2, 2017.
21 Baumann, A. *et al.* (2013) A MCU platform with embedded FeRAM achieving 350 nA current consumption in real-time clock mode with full state retention and 6.5 μs system wakeup time (Texas Instruments), VLSI Circuits Symposium, June 12, 2013.
22 TI introduces the world's first microcontroller with a configurable low-leakage transimpedance amplifier, TI Press Release, March 22, 2016.
23 Qazi, M., Amerasekera, A. and Chandrakasan, A.P. (2013) A 3.4 pJ FeRAM-enabled D flip-flop in 0.13 μm CMOS for nonvolatile processing in digital systems (MIT, TI), ISSCC, February 2013.
24 Khanna, S. *et al.* (2013) Zero leakage microcontroller with 384 ns wakeup time using FeRAM mini-array architecture (TI, TSMC), A-SSCC, pp. 21, November 11, 2013.
25 Khanna, S. *et al.* (2014) An FeRAM-based nonvolatile logic MCU SoC exhibiting 100% digital state retention at VDD = 0 V achieving zero leakage with <400 ns wakeup time for ULP applications (Texas Instruments), *IEEE Journal of Solid-State Circuits*, 49 (1), 95, January 2014.
26 Koike, H., Sakimura, N. *et al.* (2013) A power-gated MPU with 3-microsecond entry/exit delay using MTJ-based nonvolatile flip-flop (Tohoku University, NEC), A-SSCC, November 11, 2013.
27 Sakimura, N. *et al.* (2014) A 90 m 20 MHz fully nonvolatile microcontroller for standby – power-critical applications (NEC, Tohoku University), IEEE ISSCC, February 2014.
28 Qazi, M., Amerasekera, A. and Chandrakasan, A.P. (2014) A 3.4-pJ FeRAM-enabled D flip-flop in 0.13-μm CMOS for nonvolatile processing in digital systems (Cypress, MIT, TI). *IEEE JSSC*, 49 (1), 201, January 2014.

29 Kazi, I. *et al.* (2014) Energy/reliability trade-offs in low-voltage ReRAM-based non-volatile flip-flop design (EPFL). *IEEE Trans. on Circuits and Systems. –I: Reg.Papers*, 61 (11), November 2014.
30 Vianello, E. *et al.* (2014) Resistive memories for ultra-low-power embedded computing design (CEA LETI), IEDM, December 2014.
31 Lee, A. (2015) RRAM-based 7T1R nonvolatile SRAM with 2× reduction in store energy and 94× reduction in restore energy for frequent-off instant-on applications (NTHU, NDL, National Chung Hsin University, EOL ITRI), Symposium on VLSI Technology, June 2015.
32 Liu, Y. *et al.* (2016) A 65 nm ReRAM-enabled nonvolatile processor with 6× reduction in restore time and 4× higher clock frequency using adaptive data retention and self-write-termination nonvolatile logic (Tsinghua University, National Tsing Hua University, University of California, Los Angeles), ISSCC, February 2016.
33 Levisse, A. *et al.* (2014) OxRAM-based pulsed latch for non-volatile flip-flop in 28 nm FDSOI (University of Grenoble Alpes), IEEE SOI-3D-Subthreshold Microelectronics Technology, October 6, 2014.
34 Radio-frequency identification, Wikipedia, the free encyclopedia, 1 August 2017.
35 Prophet, G. (2015) Panasonic's NFC + FeRAM technology and platform, now in distribution, www.rs-online.com, February 20, 2015.
36 Worlds first fully integrated NFC & EPC Gen2V2 dual-frequency RFID solution from EMMicroelectronic, Press Release, April 14, 2015.
37 NXP' NTAG I2C plus extends the family of connected NFC tags to specifically address the smart home market, NXP Press Release, February 23, 2016.
38 Swedberg, C. (2014) Macy's expands RFID and Beacon Deployments. *RFID Journal,* September 16, 2014.
39 EM Microelectronic now shipping FCC/CE/IC certified Bluetooth® SMART beacons, EM Microelectronic Press Release, December 9, 2014.
40 About nRF51822, Nordic Semiconductor Press Release, June 12, 2012.
41 Silicon Labs simplifies Bluetooth smart design with fully integrated blue gecko module, Silicon Labs Press Release, August 17, 2015.
42 Toshiba's new ICs for Bluetooth® smart devices feature built-in Flash ROM, Toshiba Press Release, November 17, 2015.
43 Silicon Labs advances Bluetooth smart connectivity with energy-friendly SoC and S software solution, Silicon Labs Press Release, February 23, 2016.
44 Smallest Microcontroller in STMicroelectronics' STM32 F4 Series enters production, with easy-access development support, STMicroelectronics, December 3, 2015.
45 TI's new sub-1 GHz solution spans 20 km on a coin cell, TI Press Release, November 18, 2015.
46 Silicon Labs simplifies Wi-Fi connectivity with plug-and-play module solution, Silicon Labs Press Release, February 2016.
47 Silicon Labs simplifies IoT connectivity with multiprotocol wireless Gecko SoCs, Silicon Labs Press Release, February 2016.
48 Silicon Labs introduces world's most energy friendly USB microcontrollers, Silicon Labs Press Release, May 13, 2015.
49 Silicon Labs rolls out next-generation 8-bit microcontrollers for the IoT Age, Silicon Labs Press Release, February 23, 2015.

50 Atmel introduces world's most innovative 2-pin, self-powered Serial EEPROM targeting the IoT, battery, consumable and cable identification markets, Atmel Press Release, August 10, 2015.
51 Zigbee, Wikipedia, the free encyclopedia, 26 July 2017.
52 Frenzel, L. (2012) What's the difference between Bluetooth low energy and ANT? Electronics Design, November 29, 2012.
53 Izumi, S. *et al.* (2014) A wearable healthcare system with a 13.7 μA noise tolerant ECG processor (Kobe University, Rohm, Omron). *IEEE Trans. on Biomedical Circuits and Systems*, November 2014.
54 Fujitsu releases new 1 Mbit and 2 Mbit FeRAM products, Fujitsu America Press Release, March 21, 2013.
55 Gilbert, N., Zhang, Y., Dinh, J. *et al.* (2013) A 0.6 V 8pj/write non-volatile CBRAM macro embedded in a body sensor node for ultra low energy applications (Adesto, University of Virginia), VLSI Circuits Symposium, June 2013.
56 Umeki, Y., Yanagida, K., Tsunoda, K. and Sugii, T. (2013) A 0.38-V operating STT-MRAM with process variation tolerant sense amplifier (Kobe University, LEAP), A-SSCC, pp. 249, November 11, 2013.
57 Yamashita, K. *et al.* (2013) A 38 μA wearable biosignal monitoring system with near field communication (Kobe University, Omron Healthcare), NEWCAS, June 2013.
58 Izumi, S. *et al.* (2013) A 14 μA ECG processor with robust heart rate monitor for a wearable healthcare system (Kobe University, Rohm Co., Omron, Omron Healthcare, JST CREST), ISCDG, September 13, 2013.
59 Izumi, S. *et al.* (2014) A wearable healthcare system with a 13.7 μA noise tolerant ECG processor (Kobe University, Rohm, Omron). *IEEE Trans. on Biomedical Circuits and Systems*, November 2014.
60 Izumi, S. *et al.* (2014) Normally-off technologies for healthcare appliances (Kobe, Rohm), ASP-DAC, pp. 17, January 20, 2014.
61 NXP and GlobalFoundries announce production of 40 nm embedded non-volatile memory technology, NXP and GlobalFoundries Press Release, March 24, 2015.
62 TSMC launches ultra-low power technology platform for IoT and wearable device applications, TSMC Press Release, September 29, 2014.
63 Toshiba develops two new process technologies for microcontrollers and wireless communication ICs, Toshiba Press Release, July 6, 2015.
64 NXP ships LPC11U30 USB microcontrollers with 128 KB Flash, NXP Press Release, June 25, 2012.
65 Texas Instruments unveils the Stellaris® LaunchPad – a fully-functional, flexible and low-price kit for ARM® Cortex™ – M4 developers, TI Press Release, September 25, 2012.
66 Pikhay, E. *et al.* (2012) Radiation sensor based on a floating gate device (TowerJazz), IEEE Israel, p. 1, November 14, 2012.
67 Microchip expands SPI Flash memory portfolio with three new low-power devices, Microchip Press Release, February 4, 2013.
68 STMicroelectronics launches new STM32 ultra-low-power microcontrollers for consumer, health, and industrial applications, ST Micro Press Release, February 11, 2014.
69 Develop the next generation of wearables with TI's ultimate NFC wearable design challenge, TI Press Release, June 9, 2015.

70 Atmel Smart ARM Cortex M7-based MCU and AVR powers TomTom Spark GPS fitness watches, Atmel Press Release, November 5, 2015.

71 Toshiba expands line-up of ApP Lite(TM) processor family for IoT solutions, Toshiba Press Release, September 4, 2015.

72 Atmel launches ultra-low-power connected platform for cost-optimized IoT and wearable applications at CES 2016, Atmel Press Release, January 5, 2016.

73 Silicon Labs rolls out next-generation 8-bit microcontrollers for the IoT Age, Silicon Labs Press Release, February 23, 2015.

74 Zenta from Vinaya uses Nordic semiconductors recently-introduced nRF52832 SoC to wirelessly monitor and record a range of key biological and environmental data, Nordic Semiconductor Press Release, August 10, 2016.

75 Nordic nRF52 Series redefined single-chip Bluetooth Smart by marrying barrier-breaking performance and power efficiency with on-chip NFC for touch-to-pair, Nordic Semiconductor Press Release, June 16, 2015.

76 Panasonic to manufacture embedded ReRAM microcontrollers, Semiconductor Industry News, August 22, 2013.

77 Collaboration with Adesto Technologies and Nordion confirms gamma irradiation tolerance of CBRAM® non-volatile memory (Adesto), Adesto Press Release, October 16, 2013.

78 Adesto introduces low power, sterilization-tolerant memory for medical IoT applications, Adesto Press Release, August 5, 2014.

79 Adesto Technologies introduces extended voltage range memory, Adesto Press Release, November 9, 2015.

80 Kanter, D. (2016) Adesto targets IoT using CBRAM, Microprocessor Report (Linley Group), February 22, 2016.

81 Microchip doubles flash memory and adds new security options in latest family of extreme low power PIC Microcontrollers, Microchip Press Release, August 24, 2015.

82 STMicroelectronics enables new applications with world's most advanced secure microcontroller, STMicroelectronics, November 17, 2015.

83 Panic, G., Schrape, O., Basmer, T. *et al.* (2013) TNODE: A low power sensor node processor for secure wireless networks (Innovations for High Performance Microelectronics (IHP), Frankfurt, Germany), System on Chip, October 23, 2013.

84 Silicon Labs secures IoT Nodes with new EFM32 Jade and Pearl Gecko microcontrollers, Silicon Labs Press Release, December 14, 2015.

85 Lammers, D. (2016) A technology trifecta for automotive, Foundries Files Blog, March 28, 2016.

86 Taito, Y. (2016) A 28 nm embedded split-gate MONOS (SG-MONOS) Flash Macro for automotive achieving 6.4 GB/s read throughput by 200 MHz no-wait read operation and 2.0 MB/s write throughput at T_j of 170 °C (Renesas). *IEEE Journal of Solid-State Circuits*, 51 (1), 213, January 2016.

87 Radar systems, the sound of safety, Autoliv, https://www.autoliv.com/Products AndInnovations/ActiveSafetySystems/Pages/RadarSystems.aspx, October 3, 2016.

88 Scobie, J. and Stachew, M. (2015) Electronic control system partitioning in the autonomous vehicle, *EE Times,* October 29, 2015.

89 Renesas Electronics announces the RL78/D1A group of single-chip microcontrollers for automotive instrument clusters; industry's smallest MCUs for two-wheeled vehicles, Renesas Press Release, October 2, 2012.

90 Renesas Electronics announces ultra-low power consumption RH850/F1x Series microcontrollers with on-chip 40 nm Flash memory for automotive body applications, Renesas Press Release, September 27, 2012.
91 Freescale and Continental Partner on quad-core 32-bit microcontroller for advanced stability applications, Freescale Press Release, November 13, 2012.
92 Jumpstart automotive and transportation designs with TI's new Hercules™ TMS570 ARM® safety microcontrollers, power management IC (PMIC) and motor driver, TI Press Release, October 10, 2012.
93 TI strengthens DSP and vision processing on its "Jacinto" family of infotainment processors to enhance integration, TI Press Release, October 21, 2014.
94 NXP radar technology accelerates ADAS adoption, NXP Press Release, January 4, 2016.
95 Lammers, D. (2016) A technology trifecta for automotive, Foundries Files Blog, March 28, 2016.
96 New multi-constellation satellite-location chip from STMicroelectronics adds support for China's BeiDou System, ST Micro Press Release, January 8, 2014.
97 Jefremow, M. *et al.* (2013) A 65 nm 4 MB embedded Flash Macro for automotive achieving a read throughput of 5.7 GB/s and a write throughput of 1.4 MB/s (Infineon, TU Munchen), ESSCIRC, p. 193, September 16, 2013.
98 Freescale intelligent battery sensor combines MCU and CAN with flexible three-channel analog front end, Freescale Press Release, April 10, 2014.
99 New Renesas 32-bit MCU for improved fuel efficiency in vehicle power train control applications, Renesas Press Release, April 5, 2016.
100 Yamauchi, T. (2015) Prospect of embedded non-volatile memory in the Smart Society (Renesas), VLSI-TSA, April 27, 2015.
101 Renesas Electronics develops 90 nm one-transistor MONOS Flash Memory technology to accelerate intelligence in automotive control systems, Renesas Press Release, February 3, 2016.
102 Bartoli, J. *et al.* (2014) A new non-volatile memory cell based on the flash architecture for embedded low energy applications: ATW (Asymmetrical Tunnel Window) (ST-Micro., Aix-Marseille University), CAS, October 2014.
103 Atmel samples new family of high-performance ARM Cortex-M7-based MCUs enabling next-generation IoT, industrial and automotive applications, Atmel Press Release, October 2, 2014.
104 Cypress expands Traveo automotive MCU family, Cypress Press Release, March 2016.
105 Spansion expands TraveoTM family for automotive with HMI technologies and embedded 3D graphics engine, Spansion Press Release, October 2, 2014.
106 Secure microcontrollers from STMicroelectronics bring advanced cyber safety to connected cars, STMicroelectronics Press Release, February 22, 2016.
107 MPC5510" NXP 32-bit MCU for body electronics applications, NXP Web Page, 26 April 2016.
108 MPS574xB-C-D-G" ultra-reliable MCUs for automotive and industrial control and gateway, NXP Web Page, July 28, 2016.
109 Renesas Electronics delivers RL78/I1C group of microcontrollers supporting meter International Standards (DLMS) for smart meters, Renesas Press Release, August 25, 2016.

110 Renesas Electronics introduces 32-bit RX21A group of microcontrollers with large memory capacity and built-in A/D converter enabling high-resolution measurement for highly-functional smart meters, Renesas Press Release, September 6, 2012.
111 Atmel showcases world's first ARM Cortex-M4 processor-based single-chip solution for PRIME smart metering applications, Atmel Press Release, October 9, 2012.
112 First sub-gigahertz wireless microcontroller using world's most energy-efficient 32-bit processor core, Freescale Press Release, October 9, 2012.
113 Atmel and Wasion Group to develop PRIME PLC smart meter solutions, Atmel Press Release, January 15, 2013.
114 Microchip expands SPI Flash Memory portfolio with three new low-power devices, Microchip Press Release, February 4, 2013.
115 Atmel expands metering platform for advanced smart energy applications, Atmel/Microchip Press Release, October, 2014.
116 Bush, S. (2014) TI aims at smart meters with DRAM microcontrollers, Electronics Weekly.com, September 5, 2014.
117 Silicon Labs and ARM collaborate to drive the future of low-power ARM embed IoT device platforms, Silicon Labs Press Release, March 12, 2015.
118 Microchip's new cost-effective low power PIC MCU extend battery life and eliminate external memory via 1MB of dual-partition flash, Microchip Press Release, November 24, 2015.
119 TI introduces 2.4 GHz SoCs to low-cost RF value line, bringing wireless connectivity to more consumer electronics from computer peripherals to toys, TI Press Release, October 1, 2012.
120 EFM32 Giant Cortex-M3 powered Velux smartphone interface remote, Energy Micro Press Release, January 15, 2013.
121 Cypress enables IoT developers to easily create tiny, solar-powered sensor beacons that wirelessly transmit data on their surrounding environment, Cypress Press Release, June 21, 2016.
122 NXP introduces new wireless microcontroller family for the Internet of Things, NXP Press Release, November 19, 2012.
123 Renesas Electronics announces the RL78/D1A group of single-chip microcontrollers for automotive instrument clusters; industry's smallest MCUs for two-wheeled vehicles, Renesas Press Release, October 2, 2012.
124 Silicon Labs rolls out next-generation 8-bit microcontrollers for the IoT age, Silicon Labs Press Release, February 23, 2015.
125 Toshiba launches ARM® Cortex®-M3-based microcontrollers with latest 65 nm Flash embedded logic process for motor control and consumer devices, Toshiba Press Release, March 23, 2016.
126 Renesas Electronics enhances the scalability of RX MCU family for high-performance motor control with larger pin count, Flash and RAM, enhanced safety features for IC60730 and USB 2.0, Renesas Press Release, November 12, 2012.
127 Renesas Electronics reveals RX24T group of 32-bit MCUs targeted to motor control for improved energy efficiency in industrial equipment, office equipment, and household appliances, Renesas, March 2016.
128 Lose power, not data. New ultra low power FeRAM microcontrollers from Texas Instruments revolutionize context save and restore, TI Press Release, May 20, 2015.

129 EM Microelectronic announces first secure ISO/IEC15693 compliant IC for anti-counterfeiting and authenticity safeguarding of goods, EMMicro Press Release, March 4, 2014.
130 STMicroelectronics reveals advanced secure-IC family, boosting support for US switch to highly secure EMV chip payment cards, STM Press Release, May 27, 2014.
131 STMicroelectronics announces most advanced 32 bit secure microcontroller, ST Microelectronics, November 17, 2015.
132 NXP announces the PN462 family – first all-in-one full NFC controller solutions with customizable firmware, NXP, February 2016.
133 Lin, C.C. *et al.* (2016) 256b Wordlength ReRAM TCAM, NTHU, TSMC, NCTU, ISSCC, February 2016.
134 Noguchi, H. (2016) 4 Mb STT-MRAM Cache, Toshiba, ISSCC, February 2016.
135 Sohn, K. *et al.* (2016) 18.2 A 1.2 V 20 nm 307GB/s HBM DRAM with at-speed wafer-level I/O test scheme and adaptive refresh considering temperature distribution, Samsung, ISSCC, February 2016.
136 Khwa, W.S. (2016) PCM to reduce BER in SCM (Machronix, NTHU, NCTU, IBM), IEEE ISSCC, February 2016.
137 Kang, D. *et al.* (2016) 7.1 A 256Gb 3b/cell V-NAND Flash memory with 48 stacked WL layers, Samsung, ISSCC, February 2016.
138 Tanaka, T. *et al.* (2016) 7.7 A 768Gb 3b/cell 3D-floating-gate NAND Flash Memory, Micron. Intel, ISSCC, February 2016.
139 Lee, S. *et al.* (2016) 7.5 A 128Gb 2b/cell NAND Flash memory in 14 nm technology with tPROG = 640 μs and 800MB/s I/O rate, Samsung, ISSCC, February 2016.

3

Embedded Flash and EEPROM for Smart IoT

3.1 Introduction to eFlash and eEEPROM for Smart IoT

3.1.1 Overview of eFlash and eEEPROM for Smart IoT

As the world has become more electronically interconnected, the evolution of the smart Internet of Things has increased the need for low cost microcontrollers that can control environmental sensors, store the data they gather in nonvolatile memory, control the distribution and analysis of these data, and also control feedback and use of this analysis back to the source. The embedded nonvolatile memory needs to have attributes different from the standalone memory in high volume production today.

It must first of all be low cost and producible in high volume conventional CMOS processes. Since standalone Flash memory has evolved with its own complex technology, this embedded Flash memory variation needs to be made compatible with the standard CMOS logic processes available in the various foundries. Any cost adders to the CMOS logic process, such as additional masks or process steps, can cause unacceptable cost increases to the product. Adding new or exotic materials to the CMOS logic process also can cause the cost to increase unacceptably as well as, potentially, affect the reliability of the end product. These considerations mean the first option for embedded nonvolatile memory must lie in some variation of the conventional Flash/EEPROM nonvolatile memory that is compatible with the CMOS logic process

Normal memory considerations cannot, however, be ignored. A memory must be adequate for the target application and it must be low cost. Low cost, in the case of an embedded memory, does mean, first of all, compatible with the CMOS logic process. It also, however, still means a cell size small enough that the required capacity of the memory can be achieved without unduly affecting the chip size of the microcontroller. It also means that the cell and macro size must become smaller over time as the technology advances so that the capacity can be increased to permit functionality to be added to the MCU in the future. Adequate for the application means that the speed and power requirements of the specific application must be met as well as potentially other characteristics, such as temperature, voltage, or current considerations.

Application areas important to the Smart Internet of Things include: automotive networks both within cars and between cars and highways, smart cities, smart homes, smart offices, smart factories, personal and medical device networks, banking and ID card systems, a smart electrical grid, and many other networks. The criteria for

Memories for the Intelligent Internet of Things, First Edition. Betty Prince and David Prince.
© 2018 John Wiley & Sons Ltd. Published 2018 by John Wiley & Sons Ltd.

microcontrollers for these networks can include high temperatures, high voltages, low voltages, low operating current, low or zero standby current, high speed, and high throughput.

Embedded Flash and EEPROM memories have been developed for most of these applications. Many of these embedded nonvolatile memories are available from the various foundries for logic ICs. Many of these devices will be discussed in this chapter.

3.1.2 Summary of Application Requirements for Embedded Flash in IoT

The Flash MCU requirements for the various applications differ significantly. These applications affect the type of embedded Flash that can be used in the MCU for that application. In addition, several of the emerging memories entering production have inspired creativity in using their special characteristics. Automotive remains the largest of the Flash MCU applications, but there are new applications for Flash MCU. These include: integrated GPS, the entire Internet of Things (IOT) spectrum including: beacons, energy harvesting sensors, and sensor networks. Smart cards include the rapidly growing ID bank card application. The smart electric grid needs network chips for digital utility meters. An aging population requires personal body sensor networks and portable medical devices. Secure wireless network chips need smart crypto-processors and high performance SSD requires integrated storage class memories. Analog properties of the devices have been shown to be useful in neural circuits.

Several recent options for embedded floating gate Flash are covered here. The wide array of application requirements and falling prices has resulted in a large diversity of technology options. For low cost Flash MCUs, it is necessary to reduce the process adders for the embedded nonvolatile memory so that conventional single poly-CMOS logic can be used without process adders. This has given rise to a number of potential candidates for single polysilicon Flash. These low cost, single polyCMOS logic process-compatible NVMs are covered in the early sections. The 1.5 transistor (1.5T) split gate floating gate (FG) Flash cell, widely used in foundries, is discussed in the fourth section with both early technologies covered as well as the more advanced technologies with new peripherals, and with low-*k* and copper process. Novel and unusual eFlash are covered in the following sections, including: a twin-bit cell with virtual ground architecture, an OTP for 16 nm FinFETs, an HV eMTP, an HV process eFlash for automotive uses, a FLOTOX Flash in deeply depleted channel technology, and an electromechanical OTP antifuse. Since FG eFlash is a common production technology, this early part is not divided into production and development memories.

The charge trapping Flash is discussed in the last section, including that using charge trapping materials like nitride and those using nanocrystals of charge trapping materials embedded in a dielectric. The advantages of charge trapping memory are addressed including the reduced processing cost associated with eliminating the second polysilicon.

There is currently no universal nonvolatile memory that serves the wide variety of applications well. In June of 2016, Hitachi discussed various types of nonvolatile emerging memories [1]. Various applications for microcontrollers with embedded nonvolatile memories are shown in Table 3.1. All of these applications are part of the network in the IoT.

Table 3.1 Applications for microcontrollers with embedded nonvolatile memories.

Automotive	Industry	Appliances	Security
ADAS	Robots	TV	Camera
Engine control	Sensors	Digital record	Fingerprint
Transmission control	RFID tag	Tablet	Voice recognition
Navigation	Beacons	Smartphone	Voice synthesis
Lane recognition		Digital record	

The various NVM were discussed. Charge trapping memory has been developed which is suitable for both high performance memory and low power consumption memory. Resistance change memory was discussed as superior to charge trapping memory in ultralow power operations due to its low program and erase voltage. Some magnetic RAM (MRAM) and resistance RAM (RRAM) can be programmed at a voltage of less than 2 V. MRAM can have current for program and erase reduced down to a few 10s of μA and it has endurance exceeding 10^{15} cycles, which gives it the potential to replace cache SRAM. Ferroelectric RAMs are also available in low density and low power applications. Phase change memory (PCM) has been used as a replacement for conventional NOR-type floating gate EEPROM memory. The RRAM known as the "Atom Switch" can be used as an ultralow power ROM. The ON and OFF state occurs at less than 2 V and the ON/OFF ratio exceeds three orders of magnitude at a voltage as low as 0.2 V.

3.2 Single Poly Floating Gate eFlash/EEPROM Cells for IoT

3.2.1 Overview of Single Poly Floating Gate eFlash/EEPROM for IoT

Single polysilicon CMOS logic process-compatible floating gate Flash and EEPROM are covered in this first section. For a low capacity of bits, these embedded memories have a lower process cost made in a logic process than a conventional double polysilicon Flash memory since the process adders, such as additional diffusions and mask layers, are eliminated. The cell size is generally larger for a single poly. EEPROM made in a pure CMOS logic process, but the added cost of the extra silicon is not significant if only a small number of bits are needed. The trade-off is in the reduced cost of the process compared with the added cost of the larger cell size. Many applications in IoT for embedded Flash memory, such as for trimming or ID in passive RFID tags, require no more than a few K-bits of NVM. Early automotive applications and networks required operation of low cost embedded NVM in MCU for high temperature environments.

3.2.2 Early Single Polysilicon Floating Gate EEPROMS

Several of the early single polysilicon floating gate EEPROMS are covered in this section and their characteristics are summarized in Table 3.2.

Table 3.2 Early single polysilicon floating gate EEPROM features.

Company	Date	Tech.	Capacity	V_{prog}	Supply	Cell
Philips	1985	2.5 µm	Kb	13 V	5 V	440 µm^2
Toshiba	1985	1.2 µm	Kb	12 V	5 V	86.25 µm^2
IBM	1994	0.8 µm	Kb	15 V	5 V	31 µm^2
Bell	2000	0.25 µm	Kb	12 V	2.5 V	50 µm^2

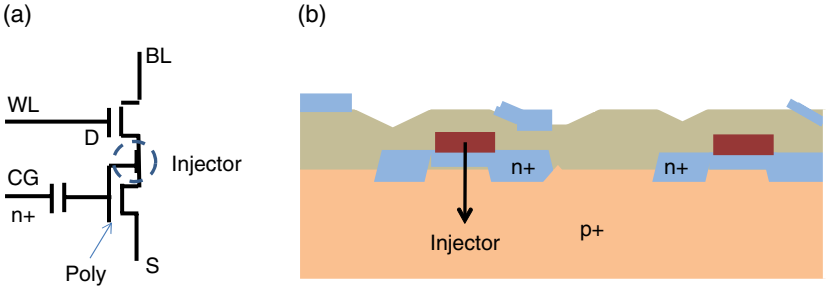


Figure 3.1 Single-poly EEPROM cell with a thin oxide coupling capacitor and with thin injector oxide: (a) schematic cross-section and (b) schematic circuit diagram. Based on R. Cuppens (Philips), *IEEE Journal of Solid-State Circuits*, April 1985 [2].

An early single polysilicon EEPROM memory made in a pure CMOS logic process was described by Philips in 1985 [2]. It was intended for use in microprocessors and custom logic circuits. The process was a 2.5 µm p-well CMOS technology. The capacity was 2048 (2 K) bits, and a 5 V power supply was used along with a 13 V programming voltage. One additional mask step was required to define the thin oxide. The 13 V program voltage was attained by decreasing the injector oxide thickness to 8 nm and increasing the capacitive coupling between the control and floating gate at the expense of increasing the cell size by about 30%. The thin 8 nm injector oxide also affected the data retention since the retention time is proportional to the thin oxide area. In active mode the 2 K-bit module dissipated 10 mW with $V_{cc} = 5$ V at 10 MHz. In standby, the power consumption was reduced to a value less than 5 µW. Figure 3.1 (a) shows a schematic circuit diagram of the EEPROM cell with the injector oxide indicated. Figure 3.1 (b) shows a schematic cross-section of the cell showing the injector oxide.

In June of 1985, Toshiba discussed a single polysilicon 256 K-bit EEPROM for integration into a microcontroller chip in a 1.2 µm process [3]. A structure similar to the Philips single polysilicon EEPROM was used and the cell area was shrunk by using a "bird's beak" isolation technology. The programming voltage was lowered to 12 V by lowering the barrier height of 2 eV between the SiO_2 and the silicon. The operating voltage was 5 V. The endurance was 10^4 write/erase cycles. Cell size was 86.25 µm^2 and chip size was 7.33×6.23 mm^2.

Another single poly EEPROM cell in a conventional CMOS process was discussed by IBM in March of 1994 [4.] The cell consisted of an NMOS and a PMOS transistor with an electrically isolated common polysilicon gate, which acted as the equivalent of a

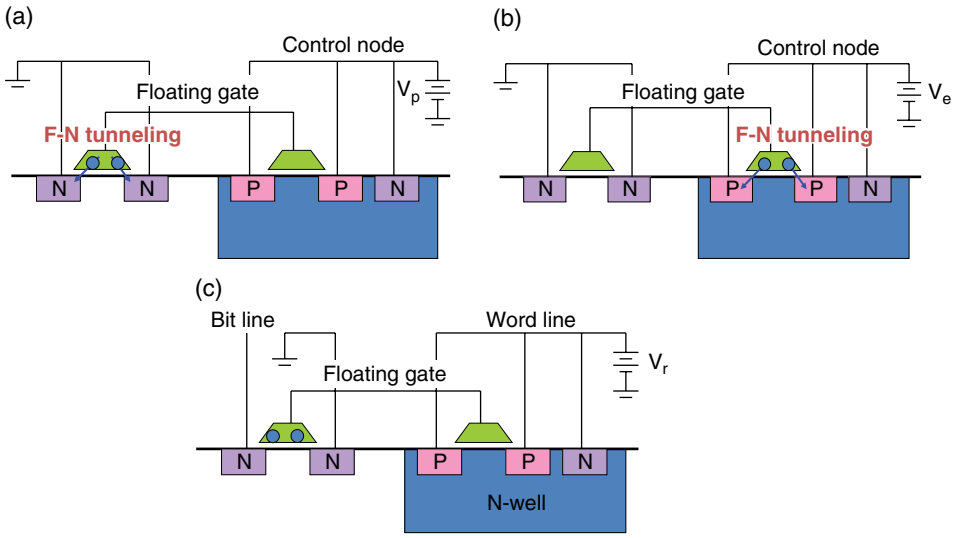


Figure 3.2 Schematic cross-sections of single poly CMOS EEPROM cell with NMOS and PMOS transistors with electrically isolated common poly gate showing Fowler-Nordheim (FN) program and erase: (a) program, (b) erase, and (c) read. Based on K. Ohsaki *et al.* (IBM), *IEEE Journal of Solid-State Circuits*, March 1994 [4].

floating gate. The inversion layer under the PMOS gate and the p + diffusions worked as a control gate. The EEPROM was intended for storage of basic input/output system (BIOS) or microcodes. The EEPROM was used for easy update of a microcode in a system configuration.

Program, erase, and read of this single poly EEPROM are illustrated in Figure 3.2. The program was by channel hot electron injection (CHEI) in the NMOS device or by Fowler-Nordheim (FN) tunneling between the gate and n + diffusion in the NMOS transistor, as shown in Figure 3.2 (a). Erase involved FN tunneling between the gate and p + diffusion in the PMOS device, as shown in Figure 3.2 (b) or erase could be by NMOS FN tunneling. Read is shown in Figure 3.2 (c). A medium voltage was used for the read voltage applied to the control gate. With no charge on the floating gate, the NMOS transistor was ON and with the gate negatively programmed, the NMOS transistor was OFF

A test chip was made in a 0.8 μm CMOS logic process. Write was done at $V_p = 15$ V for 10 ms and erase at $V_e = 18$ V for 100 ms. Cell size was 31 μm^2. An analog circuit using this cell was also developed for use in a neural circuit.

Another single polysilicon embedded Flash memory cell in CMOS logic-compatible technology was discussed by Bell Labs and Lucent Technologies in 2000 [5]. This 3 T cell was based on the IBM cell discussed in Ref. [4] but added a Fowler-Nordheim tunneling erase gate and changed the control gate design to lower the cell threshold voltage. The cell could be made in a core CMOS logic process with only one additional masking step. It had the advantage of a low VDD read operation due to a lower erased-state cell threshold and had lower gate voltage for programming. It also had higher endurance and longer data retention. Schematic circuits of the two cells are shown in Figure 3.3. Figure 3.3 (a) shows the IBM cell and Figure 3.3 (b) shows the Bell Labs cell.

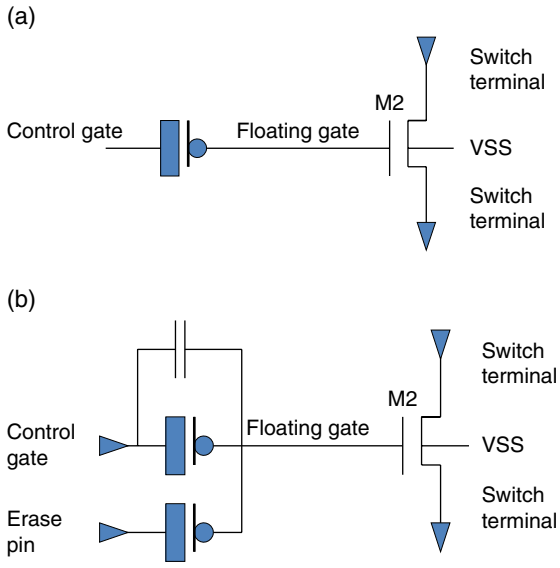


Figure 3.3 Circuit diagram of two single poly embedded Flash cells: (a) IBM and (b) Bell Labs. Based on R.J. McPartland and R. Singh (Bell Labs, Lucent), VLSI Circuits Symposium, June 2000 [5] (permission of IEEE).

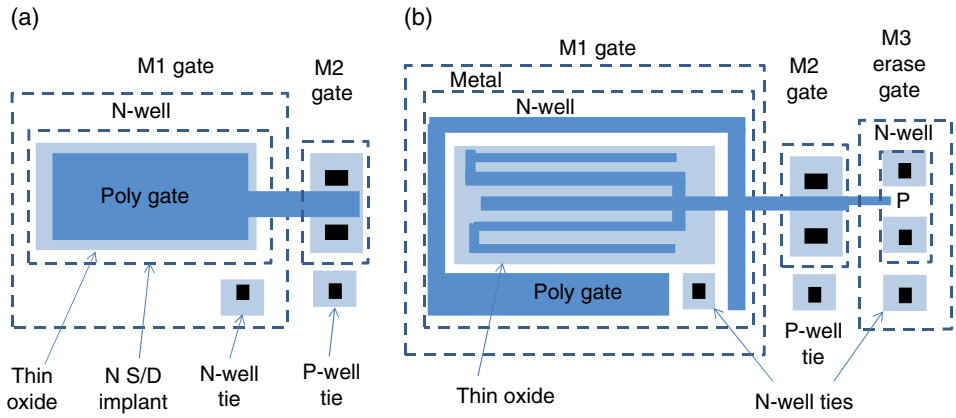


Figure 3.4 Top view of layout of (a) IBM 2T single poly EEPROM and (b) Lucent 3T single poly EEPROM with M3 erase gate added. Based on R.J. McPartland and R. Singh (Bell Labs, Lucent), VLSI Circuits Symposium, June 2000 [5] (permission of IEEE).

For the Bell Labs cell, the erase gate voltage was 10 V. For the program the control gate voltage was 6.5 V with drain at 5.5 V and source at ground. Read had V_{ds} at 1.5 V and the control gate voltage at 2.25–2.75 V. The cell was made in 0.25 μm technology using a 2.5 V logic CMOS process. Figure 3.4 shows a top-down view of the layout of (a) the IBM cell and (b) the Bell Labs cell.

In February of 2004, Virage Logic (now Synopsis) discussed a 130 nm fully CMOS logic embedded Flash memory, called a nonvolatile electrically alterable memory, intended for security code encryption [6]. This application required less than 16 K-bits of nonvolatile memory and the macro size was competitive with that of a conventional double polysilicon embedded Flash array. The requirement to embed the memory in

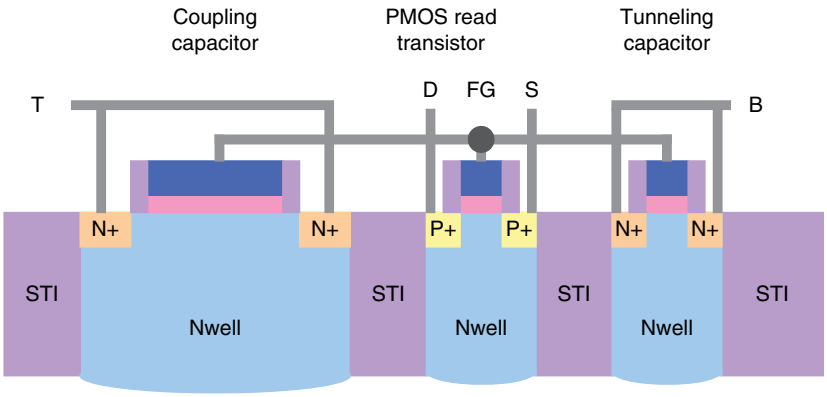
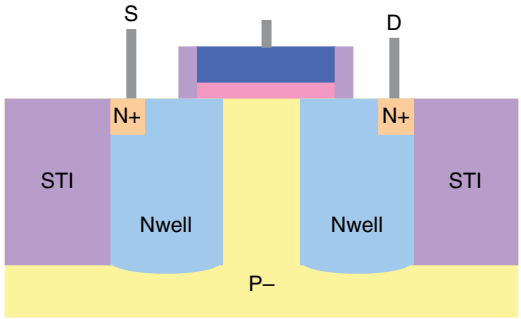


Figure 3.5 Cross-section view of 3T single poly embedded Flash cell. Based on J. Raszka *et al.* (Virage Logic), ISSCC, February 2004 [6].

Figure 3.6 High voltage NMOS device fully compatible with standard CMOS process. Based on J. Raszka *et al.* (Virage Logic), ISSCC, February 2004 [6].



the logic circuit was based on security concerns for the code in discrete memory components being intercepted when the data crossed component pin boundaries. The cell was manufactured in a standard CMOS logic process with no special masks or additional process steps. It consisted of a coupling capacitor, a tunneling capacitor, and a PMOS read transistor all connected with a common floating polysilicon gate, as shown in the cross-sectional view of the three transistor single polysilicon cell in Figure 3.5.

All elements used a 7 nm oxide grown over the active areas compared to 10 nm for a conventional double polysilicon e-Flash. There were no mask adders over a CMOS logic process or added process steps compared to seven mask adders for a standard double polysilicon Flash and 15 added process steps. Each bit was two memory cells connected in parallel and programmed with opposite data. A dedicated "compare" permitted use of a password and was used for security. The differential sensing permitted sensing signals less than 100 mV and the dual cell configuration increased yield due to the inherent redundancy. The programming voltage was supplied by an on-chip high voltage generator that uses a closed-loop regulation architecture involving three cascaded pumping states to attain 8 V and a special high voltage NMOS transistor with an NWELL source and drain, which was fully compatible with the CMOS process. The test chip contained a high voltage charge pump, which used a special high voltage NMOS device that was fully compatible with a standard CMOS process. The NMOS device is shown in Figure 3.6.

Table 3.3 Electrical characteristics of the new single poly cell and of eFlash.

Electrical characteristics	Double poly	Single poly
Program time (ms)	1.3	10
Erase time (ms)	25	10
Read access time (ns)	50	5
1 K-bit macro size (mm^2)	1.3	0.9
Retention time (years)	>10	>10

To support the greater than 10 V breakdown voltage, capacitors were used with an area efficient metal–metal finger structure. The macro size was competitive with that of a standard double polysilicon embedded Flash memory. The electrical characteristics of the new single poly cell and those of a conventional double polysilicon embedded Flash in the same 130 nm process are shown in Table 3.3.

3.2.3 Single Poly EEPROM Cells for Specialty Applications

A single polysilicon EEPROM cell for embedded memory integrated in a 130 nm RF-CMOS process without extra masks or process steps was discussed by IHP Microelectronics in 1994 together with several Italian Universities [7]. The cell used an NMOS transistor and MOS capacitor sharing a floating polysilicon layer. The process also had high speed SiGe heterojunction bipolar transistors. A ±6 V supply was used for write/erase with F-N tunneling. The cell operated at >1 ms, with endurance $>10^3$ cycles and >10 year data retention.

Early single polysilicon EEPROMs were also developed for specialty applications, such as the high temperature environments experienced by the MCUs used in automotive applications. In November of 1997, the Fraunhofer Institute in Germany discussed a single polysilicon EEPROM cell made in a silicon-on-insulator (SOI) technology known as SIMOX (Separation by IMplantation of OXygen), which used an oxygen ion beam implantation process on a silicon wafer followed by high temperature annealing to create a buried SiO_2 layer under a thin film of surface single crystal silicon [8]. A single polysilicon EEPROM cell that worked at temperatures up to 250 °C was fabricated in 2 μm SIMOX technology with a film thickness of 120 nm using 40 nm gate oxide.

The single poly EEPROM process extension required only two additional masks and a few additional process steps which had the potential of being sufficiently low cost to use as an embedded memory in the automotive application. The two transistor cell used an NMOS select transistor and a floating gate transistor with 10 nm tunnel oxide. An issue was temperature-induced leakage current in the select transistor at high temperatures. The control gate of the floating gate transistor was a separate n + diffused active area in the silicon film. Program and erase was by Fowler-Nordheim tunneling current in the tunnel oxide at the injector drain. Cell size was 650 μm^2, which was large but feasible if only a few K bits of capacity were required in the application. Only two additional masks were necessary for the EEPROM process module. The threshold window of an EEPROM cell with a capacitance ratio of 10:1 between the control gate and injector at 250 C °is illustrated in Figure 3.7.

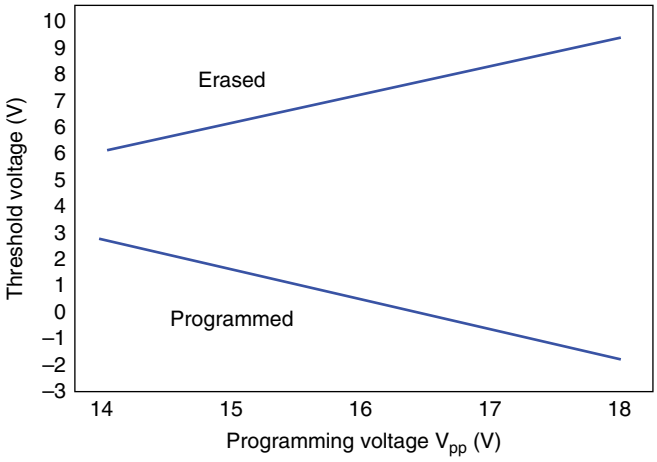


Figure 3.7 Threshold window of single poly EEPROM made in SOI Technology. Based on D. Gogl *et al.* (Fraunhofer Institute), *IEEE Electron Device Letters*, November 1997 [8].

ERASE was $V_{cg}/V_{sg} = V_{pp}$ and $V_s/V_d = 0\,V$. Program conditions were $V_d/V_{sg} = V_{pp}$ and $V_{sg} = 0\,V$ with the source floating. The threshold voltage was defined as the control gate voltage V_{cg} necessary for a current of $I_{ds} = 1\,\mu A$ at a drain bias voltage of $V_{ds} = 0.1\,V$. Endurance was about 10^5 program/erase cycles. The data retention test showed over 1000 hours at 250 °C.

In May of 2008, Tower Semiconductor discussed a low power single polysilicon logic NVM with an operating current less than 10 nA/cell and high program/erase speeds [9]. The memory was intended for use in logic with embedded memory in RFID and advanced mobile applications. No mask adders were required for a CMOS logic process. Programming and erase used Fowler-Nordheim injection. Limitations were the use of up to 10 V for a 7 nm gate oxide required for program and erase, which resulted in diode leakage due to gate-induced drain leakage (GIDL) in the periphery of the memory array.

A single poly EEPROM with a stacked MIM capacitor and n-well was discussed by Chungbuk National University in January of 2009 [10]. The oxide capacitance in the n-well region meant that the cell area and coupling ratio did not need to be sacrificed. The cell had a higher speed even though the cell size was small. Programming speed was nearly the same as a conventional MIM control gate and endurance was 10 000 program/erase cycles. The program threshold voltage shift was 1.4 V.

A 90 nm logic NVM Flash embedded in standard CMOS without a mask or process adder was described in 2011 by Genusion [11]. This cell was intended for use with RFID-type applications. The charge was stored in the Si_3Ni_4 side spacer region of the CMOS transistor so that its charge loss process was not influenced by leakage current through the gate oxide or by surface leakage current on the side spacer. It was shown that the intrinsic retention capacity is more than 10 years at 125 °C after 10K cycling. The schematic cross-sections in Figure 3.8 illustrate silicon nitride side spacer EEPROM operation for (a) the CHE program, (b) the avalanche HH erase, and (c) the read operations for the 90 nm logic NVM eFlash.

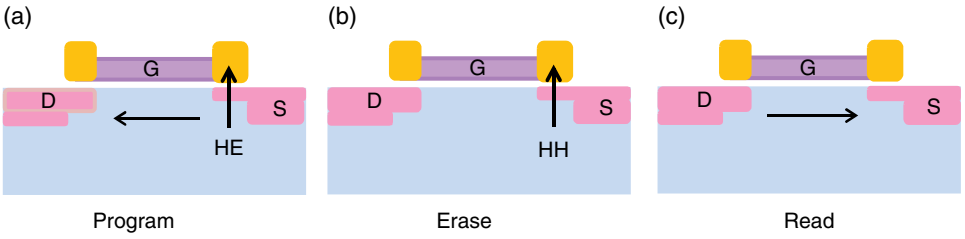


Figure 3.8 Illustration of silicon nitride side spacer EEPROM operation for (a) program by channel hot electron injection (CHEI), (b) erase by avalanche hot hole injection (AHHI), and (c) read. Based on S. Shukuri *et al.* (Genusion), IMW, May 2011 [11].

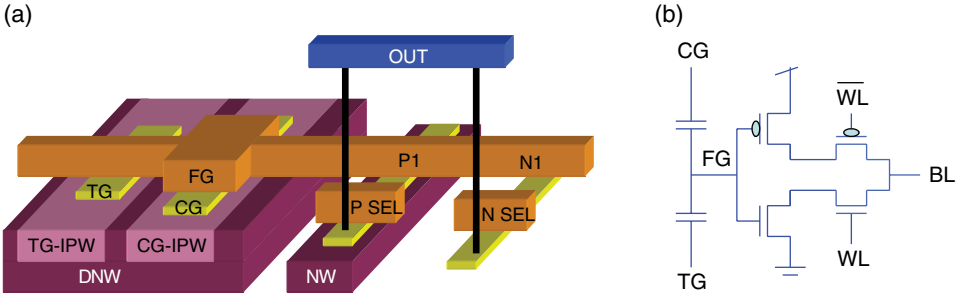


Figure 3.9 Single poly 4T cell in conventional CMOS showing (a) schematic drawing of the cell structure and (b) schematic circuit diagram of the embedded Flash memory cell. Based on H. Dagan *et al.* (Ben-Gurion University, TowerJazz), SISCAS, May, 2012 [13] (permission of IEEE).

The Flash element had an asymmetrical LDD structure constructed by using an I/O transistor in a 90 nm CMOS logic process. The source diffusion structure was optimized for channel hot electron injection (CHEI) to the side spacer for programming and avalanche hot-hole injection (AHHI) for erasure without a mask adder or any process change. Read operation was done by applying 1 V to the drain.

A fully integrated HF-band passive RFID Tag IC using 180 nm CMOS was described in 2011 by Kyun Hee University, the University of Texas, and eSilicon [12]. The target application was low cost security devices. An embedded 4 K-bit EEPROM was used to support the Advanced Encryption Standard (AES) operation. Read and write accesses were performed using a 128-bit wide buffer with self-timed bursts made in a single polysilicon six-metal low power 180 nm CMOS with a $CoSi_2$ Schottky diode and an EEPROM process.

A low power embedded nonvolatile memory for use in a low cost RFID tag was discussed in 2013 by Ben-Gurion University and TowerJazz [13]. The cell was a single polysilicon 4 T CMOS compatible Flash cell made into a 256-bit rewritable NVM array. It was made in the TowerJazz 180 nm CMOS process using only standard logic process steps and masks. Static power for this cell during operation was 3.8 μW. A schematic of the cell structure is shown in Figure 3.9 (a) and a schematic circuit diagram of the cell is shown in Figure 3.9 (b). The memory was designed using the Cadence Virtuoso IC6 tool. Each block was designed, implemented, and verified through simulation using the "Spectre" circuit level simulator.

An embedded nonvolatile memory for use in radio frequency identification (RFID) tags and near-field communication (NFC) chips was discussed in 2016 by eMemory [14]. The RFID tag has a requirement for very low power and frequently used power harvesting from the UHF signal. The embedded NVM macro uses very low power and low voltage in both read and write. A standard CMOS logic process is used without any process or mask adders to reduce manufacturing cost. The single polysilicon EEPROM macro is used for both its low operating voltage and low power [15]. Existing CMOS logic devices were used in new design implementations to lower power. The chip supported an array density up to 4Kb, μW of operating power, and 10 K cycles of endurance. It was verified in 0.11 μm logic and in a low power process platform.

3.2.4 Multitime-Programmable Single Poly Embedded Nonvolatile eMemories

Data-retention mechanisms of logic embedded NVM in multiple-time programmable (MTP) applications were discussed in 2013 by TSMC. These MTP cells have a cycling endurance that is less than the 10 000 cycles required of standalone nonvolatile memory. In April of 2013 TSMC discussed two mechanisms of data retention on logic embedded NVM in an MTP application [16]. Logic eNVM is commonly substituted for embedded Flash due to a faster time to market, logic compatibility, and the cost effectiveness of power management for MTP solutions.

Data retention of logic eNVM in bipolar, CMOS, DMOS, and HV technologies is not usually an issue due to the availability of thick tunnel oxide for the 5 V devices used in digital/analog ICs. The higher reliability requirements of the automotive application required a further understanding of the data-retention mechanism for logic eNVM. A schematic cross-section of the logic eNVM cell is shown in Figure 3.10 (a). The contact etch stop liner (CESL) covers the dielectric isolation oxide, which is on top of the floating gate. A schematic circuit diagram of the MTP unit bit-cell is shown in Figure 3.10 (b). The cell is operated by a 5 V logic device with 12 nm tunnel oxide thickness. The capacitors controlling program and erase are PG and EG. Electrons are injected into the floating gate from PG while programming and ejected to EG while erasing.

The data retention behavior of logic eNVM processed with thick tunneling oxide was studied and two new reliability mechanisms were considered. A physical data retention

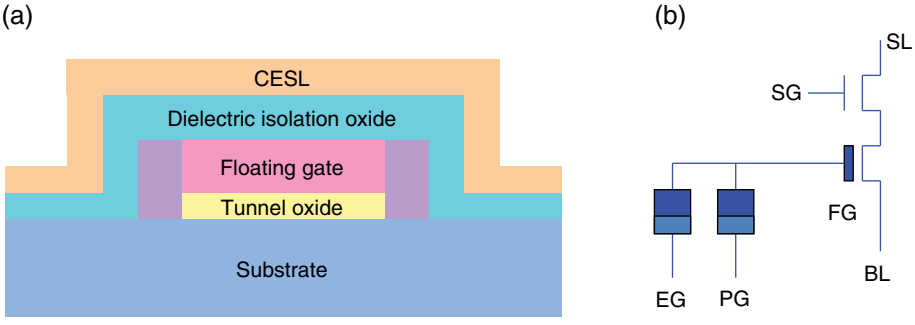


Figure 3.10 Logic eNVM MTP cell: (a) schematic cross-section and (b) schematic circuit diagram of a unit bit-cell. Based on Y.Y. Liao *et al.* (TSMC), IRPS, April 14, 2013 [16].

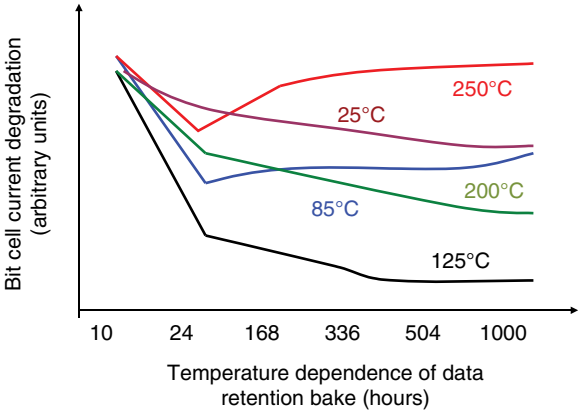


Figure 3.11 Logic eNVM cell data retention degradation trend of bit cell current in the erase state after 10,000 program/erase (P/E) cycles. Based on Y.Y. Liao *et al.* (TSMC), IRPS, April 14, 2013 [16].

model was developed for this cell. Bit cell current (BCC) data-retention degradation behaviors were studied for bake temperatures ranging from 25 to 250 °C. Larger degradation was found at higher bake temperatures or longer bake times. Retention degradation for the eNVM becomes gradually saturated with longer bake time due to the capacitive effect. The data-retention degradation trend of the erase state after 10 000 program/erase (P/E) cycles is shown in Figure 3.11. At bake temperatures greater than 125 °C, the trend degrades and then recovers in later baking. For bake temperature less than 125 °C no recovery was observed. The chart of BCC degradation versus temperature dependence of the data retention bake is shown in Figure 3.11.

The model developed to explain this phenomena was that for the high temperature bake, the transconductance, G_m, recovers faster after a 250 °C bake than after 125 °C. G_m recovery was taken to indicate that electron traps generated near the Si substrate by P/E cycles can be detrapped from the tunnel oxide interface to the silicon substrate at high bake temperatures. A lower bake temperature does not have enough thermal energy to activate electron detrapping so no current recovery is observed.

Both thermally activated electron detrapping and data code program sequence can contribute to data-retention degradation. Based on these results, low-level Si–H compositions of CESL nitride were implemented to improved DR degradation.

Another MTP cell using a single poly CMOS EEPROM technology was discussed in 2013 by STMicroelectronics and the University of Brescia [17]. It was based on a new "half-cell" structure, which was made in a conventional 130 nm CMOS process without an additional mask or process step. A test chip was made and characterized. This type of eEEPROM structure is targeted for embedding in SoC for applications such as IP security, RFID chips, analog trimming, postproduction customization, and display driver calibration. Such applications require a small amount of low cost NV memory that is programmable.

The cell is based on a new cell structure and a new design method. The new cell structure is a "half-MOS" type of device, where the drain is removed from a MOS FET device. This reduces the tunneling gate capacitance (C_{tg}). The half-MOS cell reduces area, improves program and erase performance, and improves endurance. A schematic of a "half-MOS" device is shown in Figure 3.12, with a schematic layout shown in Figure 3.12 (a) and a schematic cross-section shown in Figure 3.12 (b). "Ed" and "Es" are

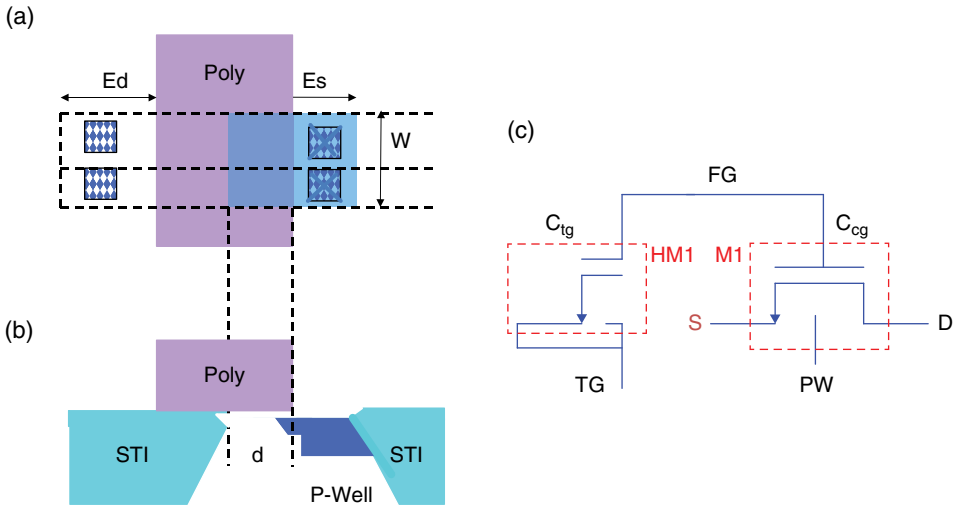


Figure 3.12 "Half-MOS" device: (a) layout, (b) cross-section and (c) cell schematic circuit diagram. Based on F. Torricelli *et al.* (University of Brescia, STMicroelectronics), *IEEE Trans. on Electron Devices*, June 2013 [17].

active extensions below the poly of the drain and source sides. "Whm" is the width of the half-MOS and "d" is the active polysilicon overlap length. In a MOSFET, Ed > 0 and Es > 0, whereas in a "half-MOS", Ed = 0 and Es > 0. A schematic circuit diagram of the "half-MOS" 1.5 transistor MTP cell is shown in Figure 3.12 (c).

During the program and erase, S, D, and PW are shorted, which forms the control gate terminal V_{cg} and M1 functions as a coupling capacitor. During read, M1 is in transistor configuration and $V_{tg} = 3.5$ V, $V_s = V_{pw} = 0$, and $V_d = 1$ V.

The "half-MOS" cell uses one of the smallest thin-oxide tunneling capacitances available in conventional CMOS technology. It has good charge retention and endurance. The required area is reduced, but the control gate coupling ratio, endurance, and data retention are the same as a full MOS cell. A 130 nm CMOS test chip showed that the memory could be programmed and erased up to 10 000 cycles. The endurance did not depend on the tunneling capacitor area. The memory cell transconductance did not degrade with cycling. Data retention at 250 °C was shown.

In December of 2013, the University of Brescia and ST Microelectronics collaborated further on their "half-MOS" single poly EEPROM cell with its program and erase per bit feature [18]. The cycling endurance was now improved to 10 000 cycles. The "half-MOS" cell could be programmed and erased by Fowler-Nordheim tunneling. The cell had a novel write inhibition method, which was enabled by a combination of the body effect of several "half-MOS" devices. A 130 nm test chip was made and characterized. The "half-MOS" bit-cell cross-section is shown in Figure 3.13.

The "half-MOS" device is an MOS transistor without the drain diffusion. An n-type cell was used for nHM and a p-type half-MOS for pHM along with two MOS transistors, nM1 and nM2. The floating gate is the contactless polysilicon layer that is shared by nHM, pHM, and nM1. The cell is compatible with conventional CMOS technology and is programmed and erased by Fowler-Nordheim tunneling. For the program (erase)

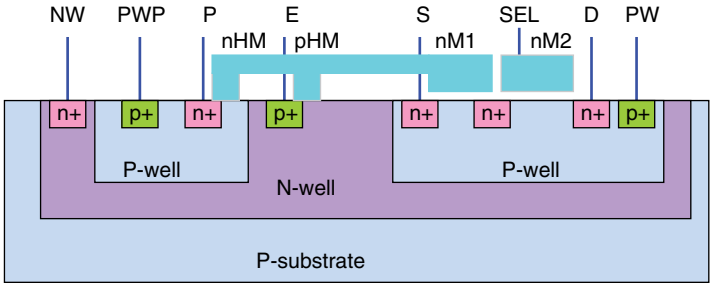


Figure 3.13 "Half-MOS" bit cell schematic cross-section. Based on F. Torricelli *et al.* (University of Brescia, STMicro), *IEEE Electron Device Letters*, December 2013 [18].

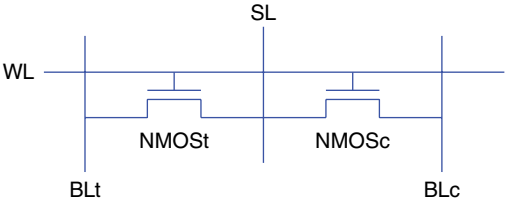


Figure 3.14 Twin transistor memory cell using HiK logic NMOS process. Based on J. Viraraghavan *et al.* (GlobalFoundries (Bangalore), IBM, UCLA), VLSI Circuits Symposium, June 2016 [19].

the nHM (pHM) acts as a tunneling capacitor while the pHM (nHM) and nM1 are the control gate capacitors. During the program, electrons are injected into the FG from the channel of nHM. During erase, electrons are removed from the FG to the channel of pHM. For read, nHM and pHM are used as coupling capacitors and the drain current of nM1 is sensed. The transistor nM2 is the select transistor for the cell.

The single poly EEPROM cells were made in 130 nm CMOS with active-poly overlap of the half-MOS devices being 250 nm and bit-cell area being 12.1 μm^2. The single-bit granularity does not use an additional area since the p-type "half-MOS" pHM is located in the n-well separating the two p-wells. Cycling endurance is 10 000 cycles and the V_t window closes after 10^5 ms of write operations.

A two-transistor embedded MTP memory cell in conventional CMOS technology was discussed in June of 2016 by Global Foundries, IBM, and UCLA [19]. An 80 K-bit logic embedded multitime programmable memory macro was made using the two-transistor cell. The cell uses a charge trapping mechanism in a 32 and 22 nm high-*k* dielectric transistor in a standard logic process with no added process complexity.

Programming is accomplished by electron injection into the HiK gate dielectric present in a standard logic NMOS transistor. This is done by applying a word-line (WL) voltage of 2 V (VPP), a source-line (SL) voltage of 1.5 V, and grounding of the bit-line (BL) at 0.0 V. This permits efficient trapping of electrons in the HfO_2 interfacial layer, resulting in a threshold voltage (V_{th}) increase. The V_{th} shift found was 200 mV with a 10 ms programming pulse. This can be increased to 300 mV with a 100 ms pulse. The cell is erased by driving the word-line (WL) to about −1 V and the SL to about 2 V. A single bit is stored in a twin cell, as illustrated in Figure 3.14.

The cell is controlled by the WL, SL, BL true (BLt0), and bit-line complement (BLc0). The V_{th} of the NMOSt or NMOSc is shifted to store the data. WL drivers include a voltage switch to select between VPP and a main voltage VDD. The column decoder

selects 1 of 4 BL pairs, which provides an 80 bit data-line (DL) readout. An SL switch coupling to VSL1, VDD, and ground (GND) is provided per DL segment of four columns so that only the SLs in the selected segment are raised to VSL1 during programming and VDD during read while grounding the CLs in unselected DL segments. A preprogrammed indicator bit supports a default bit function in the macro. The 80Kb memory array was organized into 256 rows by 320 columns. This is scalable to a 14 nm FinFET technology.

3.2.5 Recent Single Poly Fully CMOS Embedded EEPROM Devices

A scalable logic gate nonvolatile memory device fabricated in a conventional CMOS logic process was discussed by FlashSilicon in October of 2014 [20]. The device was shown in 110 nm, 55 nm, and 40 nm nodes. Cell sizes for the NOR Flash array were defined by the process design rules of the CMOS logic nodes and ranged from 0.5425 μm^2 in 110 nm technology to 0.1095 μm^2 in 40 nm technology. The EEPROM cells were made of three MOSFET single poly logic floating gate devices processed in a conventional CMOS process. The original EEPROM device used the p-MOS as the control gate, the CMOS logic gate as the charge storage floating gate, and the n-MOS as the channel device. The device source/drain electrodes shared the same bit-lines. A later version used an n-type control gate embedded in the p-type silicon substrate formed by an n-implant or separated threshold and punch-through implants from the n-well implant module of p-MOS devices in a CMOS process flow.

The minimum gate length for the logic core device gate formed the charge storage floating gate. A 6.5 to 8 nm oxide, which can hold the charges for the required retention time, formed both the tunneling oxide and coupling dielectric on top of the device channel and the n-substrate control gate. Data retention tests were done with 110 nm CMOS logic process devices. The V_t window was an average of 5.127 V. A 12 year data retention at 85 °C was achieved by baking for 24 hours at 250 °C.

The logic-compatible EEPROM devices were integrated into a 6 T SRAM cell to form a nonvolatile (nv) SRAM (nvSRAM). They were also integrated into a nonvolatile register and a nonvolatile FPGA. The foundry for the embedded NVM devices was expected to be UMC, where 1 to 8 Mb density devices were being developed in 40 to 55 nm processes.

A model for a single polysilicon EEPROM cell was developed by the National University of Defense Technology of China in October of 2014 [21]. This model provided an expression for the floating gate potential. By combining this model and the gate current data for the tunneling transistor, a method was found to do the transient simulation of the memory cell. This simulation included program and erase operations. An extension to the model was developed to enhance the model's universality. The results of simulations using the model appear to have better consistency with the TCAD simulation results than those using the traditional model. This new model is expected to be useful for the design, evaluation, and optimization of a single poly EEPROM cell.

A scalable single polysilicon eEEPROM with a tungsten (W) control gate in a full CMOS process was discussed in April of 2015 by the National Tsing Hua University (NTHU) [22]. The single polysilicon structure had a low manufacturing cost, low process defects, and short turnaround time since it was fully compatible with a CMOS process. A single poly EEPROM with a W control gate had lower control gate (CG)

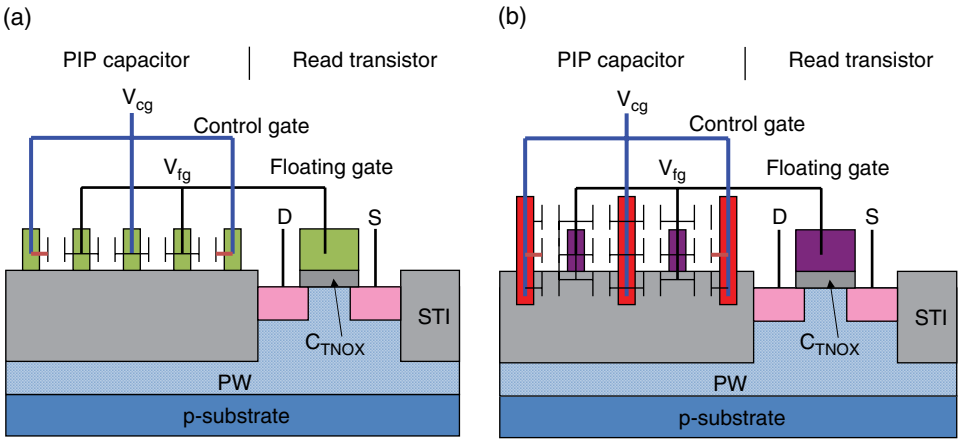


Figure 3.15 Schematic cross-section of (a) P-finger coupling cell and (b) W-finger coupling cell. Based on C.P. Chung *et al*. (NTHU), *IEEE Trans. on Electron Devices*, September 2014 [23].

resistance, no parasitic polysilicon depletion, and smaller drain-induced barrier lowering (DIBL) due to the metal CG. Its fringing capacitance enhanced the coupling ratio and increased the program/erase window compared to a cell coupled by a poly finger structure.

A metallic tungsten (W) CG cell with a select gate (SG) transistor control and various spacings from CG to floating gate (FG) was studied. It was shown that reducing the spacing from CG to FG could decrease cell size without any performance degradation. This cell could be integrated in an advanced CMOS SoC due to its scalability, simple design, and good isolation between the W CG and the FG, when made in a full CMOS process. The 130 nm CMOS cell consisted of a 3.3 V n-MOSFET for a read transistor connecting with a polysilicon finger as FG and was coupled with the surrounding W finger as W-CG. The selected bit in the W-CG EEPROM array was programmed by channel hot electron injection (CHEI) and was erased by Fowler-Nordheim (FN) tunneling.

The performance of this single polyEEPROM cell with a W-finger coupling structure was compared to that of a conventional polysilicon-finger coupling cell, also in 130 nm CMOS, in September of 2014 by NTHU [23]. A schematic cross-section of the P-finger cell and the W-finger cell is shown in Figure 3.15. The major coupling capacitances and potentials are indicated.

The W-finger coupling cell was shown to have a smaller drain-induced barrier lowering (DIBL) effect, a higher coupling ratio, and a high cell current and P/E speeds as a result of its metallic control gate and the incremental capacitance from the control gate structure. Reliability characteristics were comparable between the two cells. The W-finger coupling cell had a wider P/E window during reliability tests.

A new single poly EEPROM cell in conventional CMOS intended for medium density applications was discussed in October of 2015 by the University of Brescia [24]. A schematic cross-section of the proposed single poly EEPROM cell is shown in Figure 3.16. A pMOS tunneling device is used for programming and erase and an nMOS transistor is used for read and for selecting the memory cell. The pMOS device is a minimum sized

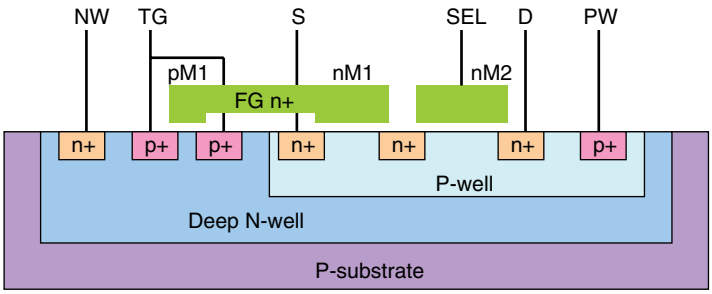
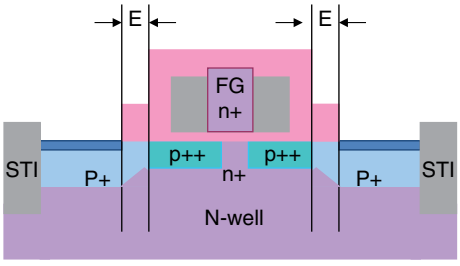


Figure 3.16 Single polysilicon EEPROM cell made in conventional CMOS with pMOS tunneling device used for programming and erase, and nMOS transistor used for read and select. Based on L. Milani *et al.* (University of Brescia), *IEEE Trans. on Electron Devices*, October 2015 [24].

Figure 3.17 Schematic cross-section of pMOS tunneling device. Based on L. Milani *et al.* (University of Brescia), *IEEE Trans. on Electron Devices*, October 2015 [24].



transistor that uses the gate oxide of the I/O transistors in the process. This improves the coupling capacitance, minimizes the area, and guarantees the retention. The cell is programmed by band-to-band hot electron injection (BBHEI) and erased by Fowler-Nordheim tunneling. The writing method used requires only a single triple well, which reduces area and ensures fast, reliable operation. The nMOS transistor (nM1) is used for sensing the drain current of the cell during the read operation and as a control gate capacitor during the write operation. The capacitive coupling ratio is CnM1/(CpM1 + CnM1).

The nMOS transistor (nM2) is used to select the memory cell. The FG is the uncontacted polysilicon gate shared by pM1 and nM1. For program, the diode is reverse biased and the BBHEI mechanism is enabled. The hot electrons generated are injected into the FG. During erase, both NW and TG are biased at high voltage, which leads to FN tunneling. During read, the drain current of nM1 is sensed; nM2 is required to select the cell in the array.

The efficiency of the write operation and the reliability are based on a new pMOS tunneling device. A schematic cross-section of this device is shown in Figure 3.17. P+ is the high doping implant. The floating gate (FG) is not silicided since it is covered by a mask, which enhances the reliability of the device.

The memory cell was programmed in 1 ms at $V_p = -5$ V and erased in 10 ms at $V_e = 18$ V. This was taken to confirm that the BBHEI tunneling is faster than the FN tunneling. BBHEI was shown to be 1000 times more efficient than CHEI. The maximum BBHEI is achieved with a programming voltage $V_p = V_{dd}$, which ensures low power operation.

The cell was made in a 180 nm CMOS process. The memory cell area is 5.91 μm^2. It could be programmed in 1 ms and erased in 10 ms. It could be cycled >10 k times with a voltage window greater than 2 V. It guaranteed the data retention and enabled reliable

programming and erase with a single triple well. The technology limited $V_p < V_{dd}$ so charge pumps are not needed for programming. The program and erase characteristics are temperature independent. This cell enables large parallel data writing, fast memory operations and test time reduction. It is suitable for moderate density and low cost applications.

3.2.6 Single Polysilicon eNVM in High Voltage CMOS

MCU for high voltage CMOS applications are also required with small amounts of low cost CMOS compatible nonvolatile memory (NVM). In March of 2014, SK Hynix discussed a select gate lateral coupling (SGLC) embedded NVM, which added no steps to a 90 nm high voltage CMOS process [25]. The select gate (SG) of the new cell is designed to function as both a control gate (CG) and an SG using only lateral capacitance coupling. The cell is relatively small, free of over-erase and has multiprogrammable features. It is programmed by channel hot electron injection (CHEI) and erased by band-to-band tunneling-assisted hot hole erase (BTBT HHE). This results in a 20 μs program and 100 ms erase. Erase time is 4.5 ms at $V_{erase} = 12$ V. Over 3 V threshold voltage window is possible after 500 cycles and 10 year data retention is estimated at 85 °C. A schematic circuit diagram of the cell is shown in Figure 3.18. The select gate (SG) controls the floating gate (FG) by lateral capacitive coupling.

The operation of the cell uses lateral capacitance coupling between the gates of neighboring cells. An SG and an FG are adjacent to each other and the coupling energy between the two is used for a floating gate. Since the select gate is also a control gate, the double poly can be eliminated along with ONO dielectric processes. C_{sg} is the lateral spacer capacitance between SG and FG and C_{ch} is the vertical capacitance between the FG and the substrate.

A test chip for the cell was made in 90 nm high voltage CMOS (HVCMOS). This process had 1.2, 6, and 32 V transistors. A 1.2 V junction was used in a 6 V gate oxide SGLC cell. Usually, the 1.2 V LDD implant is blocked on the thick gate oxide area. The cell, however, used a 1.2 V logic LDD junction as a program junction for the NVM. The cell used CHEI and BTBT HHE, which permits a lower bias than FN tunneling. The coupling ratio of C_{sg}/C_{ch} was >70%. The coupling ratio was primarily governed by the SG to FG space, with a smaller space producing a high coupling ratio. For this reason the SGLC cell was expected to be scalable.

The V_t window of the programmed and erased cell is over 3.0 V. It was shown that with the same read bias as a gate bias of 3.3 V, the programmed cell current is 2.5 nA and the erased cell current is over 27 μA. This large current difference made it possible to easily distinguish the programmed cell from the erased cell. Since the BTBT hot hole erase causes more oxide degradation than FN erase, the SGLC cell was developed for code storage applications that do not require high endurance. Using a single poly

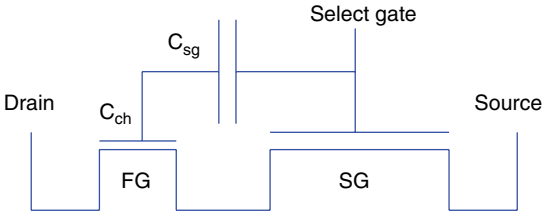


Figure 3.18 Schematic circuit diagram of select gate lateral coupling (SGLC) eNVM cell where C_{ch} is the floating gate (FG) to substrate capacitance and C_{sg} is the FG to select gate (SG) lateral capacitance. Based on S.K. Park *et al.* (Hynix), *IEEE Electron Device Letters*, March 2014 [25].

process with a select gate length of 0.2 μm the tested cell was 1.34 μm^2, which is comparable to the size of a 6 T SRAM in the same technology.

A byte alterable high voltage CMOS logic-compatible EEPROM was discussed in May of 2014 by Genusion [26]. The EEPROM was made in a 90 nm Flash process with a three transistor AND-type unit cell intended for disturb-free operation. The device used back-bias assisted band-to-band tunneling hot electron injection (B4-HE or BBHEI) for programming and Fowler-Nordheim tunneling for erase. A single pulse program and erase cycling was confirmed up to one million cycles using 10 μs program and 1 ms erase; 10 year data retention was shown at 150 °C. In an array, the unit cell size was 57 F2, which is half of conventional EEPROM at 80 F2 to 100 F2. The unit cell consists of three transistors in series using an AND array architecture.

This byte alterable BBHEI EEPROM had reliable operation and ease of implementation in a logic fab. The two transistors on both sides of the memory transistor are pass transistors, which are p-channel floating gate contacted transistors. The pass transistors control the voltage of the source and drain of the memory transistors to achieve disturb-free operation. An n-channel memory gate select transistor is connected to every 1 byte of memory cells to control the memory gate. Enhancement and depletion operation of the cell is done for the read gate voltage of 0 V.

Key features for a 1 MB device are as follows. Standby current is 1 μA. Program speed is 10 μs with current of 3 mA, erase speed is 1 ms with current of 3 mA, and read speed is 20 ns for $I_{cc} = 10$ mA. The cell is easy to make in an existing high voltage CMOS (HVCMOS) process.

3.3 eFlash Cells Using Multiple Single Polysilicon CMOS Logic Transistors

Embedded Flash cells can be made using only the CMOS logic transistors that exist in the standard CMOS process without process changes.

In June of 2012, the University of Minnesota described a five (5) transistor logic-compatible single poly embedded Flash memory that used no special devices other than a standard CMOS core and I/O transistors. A 5 T embedded Flash memory cell used a selective row refresh scheme for improved endurance [27]. It was made in a low power standard logic process with 5 nm tunnel oxide and no process adders above conventional LPCMOS. An overstress-free high voltage switch enlarges the V_{th} window by greater than 170%. A schematic circuit diagram of the 5 T cell is shown in Figure 3.19.

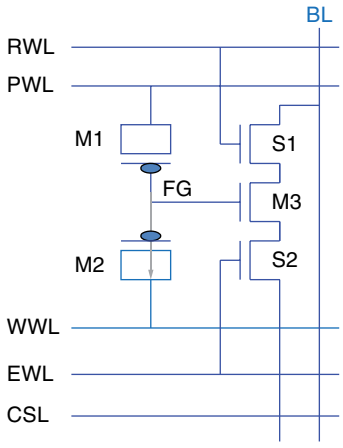


Figure 3.19 Schematic circuit diagram of 5T logic compatible single poly embedded Flash cell. Based on S.H. Song *et al.* (University of Minnesota), VLSI Circuits Symposium, June 2012 [27].

This 5 T single poly eFlash technology in conventional CMOS logic uses CMOS I/O devices and can serve as secure on-chip NVM storage for chips built in a conventional logic process. All five transistors in the cell are made using standard 2.5 V IO transistors with

an oxide thickness of 5 nm. The width of M1 is eight times wider than M2 or M3 to achieve a high coupling ratio for improved erase and program. M1 and M2 are PMOS transistors and M3 is an NMOS transistor.

The 5 T cell can be used for adaptive self-healing techniques for reducing the effects of process variation and circuit aging for system information retained for long periods of time. It can also be used for zero-standby power systems by saving critical data during power-down without adding cost to the process technology. An overstress-free high voltage switch and selective WL refresh method are used for an improved cell V_t window and higher endurance.

A new high voltage switch (HVS) was used with the 5 T cell. The previous switch had a maximum allowable program and erase voltages limited to two times the nominal I/O voltage due to gate oxide reliability concerns. An issue with the previous switch was that the internal node voltage in the PMOS cascode was sensitive to the V_t drop of the PMOS device, which made the circuit susceptible to the effects of variation and also limited the output voltage range. The new HVS had a maximum allowable program and erase voltage that was up to four times the nominal I/O voltage and provided robust output voltage levels. Technology features for the 65 nm 5 T single poly eFlash cell macro include a 5 nm oxide thickness, 1.9 V V_{th} window, erase time of 1 ms, program and read time of 10 μs and 8.62 μm^2 cell size in 65 nm technology.

Erase and program of the 5 T configuration macro are illustrated in Figure 3.20. The erase bias is shown in Figure 3.20 (a). The program bias is shown in Figure 3.20 (b), with the neighboring cell having programming inhibited using a self-boosting bias.

An earlier single poly 3 T PMOS eFlash cell was compared to the circuit diagram of the 5 T eFlash cell by the University of Minnesota in May of 2013 [28]. The circuit diagrams of the 3 T cell and the 5 T cell are shown in Figure 3.21. Both cells have been made in 65 nm conventional CMOS logic. Tunnel oxide for both was 5 nm and both used Fowler-Nordheim tunneling for program and erase. Write voltage for the 3 T cell was 8 V but 5–10 V for the 5 T cell. Write power was low. The 5 T cell offers the advantage of

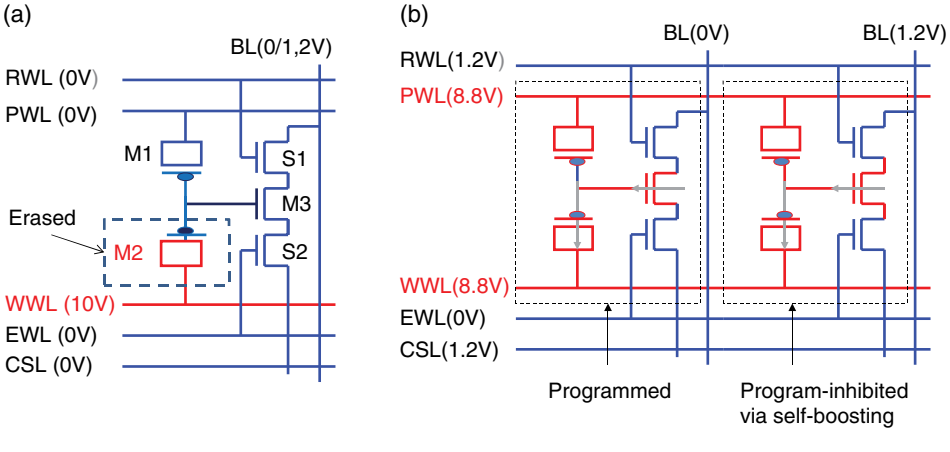


Figure 3.20 5T configuration macro: (a) erase bias and (b) program bias with neighboring cell having program inhibited via self-boosting. Based on S.H. Song *et al.* (University of Minnesota), VLSI Circuits Symposium, June 2012 [27].

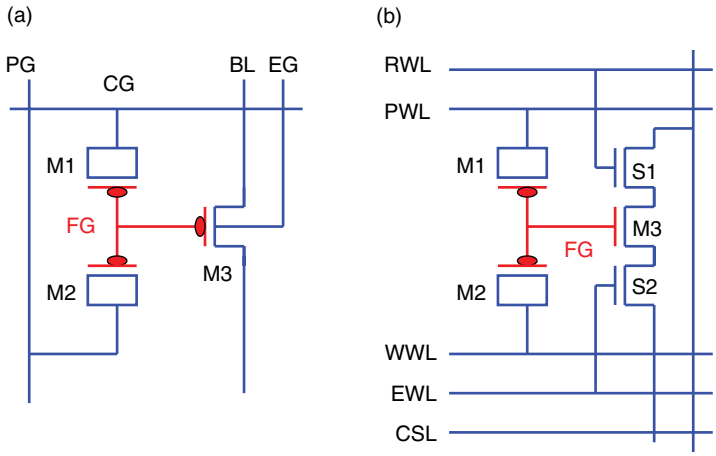


Figure 3.21 Schematic circuit diagram of 65 nm single poly (a) 3T eFlash cell compared to (b) 5T eFlash which eliminates erase disturb. Based on S.-H. Song *et al.* (University of Minnesota, Samsung), *IEEE Journal of Solid-State Circuits*, May 2013 [28].

no erase disturb of unselected WLs. Data retention for the 5 T cell was >486 hours at 27 °C and cell size was 8.62 μm^2 drop phrase "or (2111 F2)".

Compared to the single poly 3 T cell, the single poly 5 T cell has a larger V_{th} window, single word-line program, fast read time, and a much smaller cell size. The cell size of the dual polysilicon cell is smaller, but the dual poly process is more complex. The capacity of the 3 T and 5 T single poly cells are similar and smaller than that of the dual poly Flash cell.

In April of 2013, the University of Minnesota discussed the technology of the n-channel and p-channel single poly floating gate embedded Flash memory cells made in a conventional CMOS logic process [29]. These three transistor single poly devices have their floating gates coupled to net the voltage required for program and erase without a high voltage actually being applied at any point. They were made in PMOS-NMOS-PMOS in a deep n-well.

A conventional double poly or split gate eFlash requires process overhead for the floating gate and also requires high voltage (>14 V) transistors. Single poly eFlash has no process overhead since it uses conventional CMOS logic I/O devices. A schematic cross-section of a p-channel 3 T single poly eFlash that is configured for the erase operation is illustrated in Figure 3.22. For erase there is electron FN tunneling into the FG in transistor M2 under the coupling influence of –7.6 V on the WWL and 1.2 V on the PWL, while the source and drain of the FG device float. This is a net erase voltage of 8.8 V.

A schematic cross-section of the p-channel 3 T single poly eFlash configured for the program operation is illustrated in Figure 3.23. For programming there is electron FN tunneling from the FG in the PMOS M3 with the FG at 6.6 V and the PWL and WWL at –7.6 V for a net coupled programming voltage of 14.2 V.

There are several reliability effects to consider. Disturb during programming is one such reliability issue. Figure 3.24 shows schematic circuit diagrams of program and erase configurations for the p-channel 5(2 + 3)T single poly eFlash cells using electron Fowler-Nordheim tunneling.

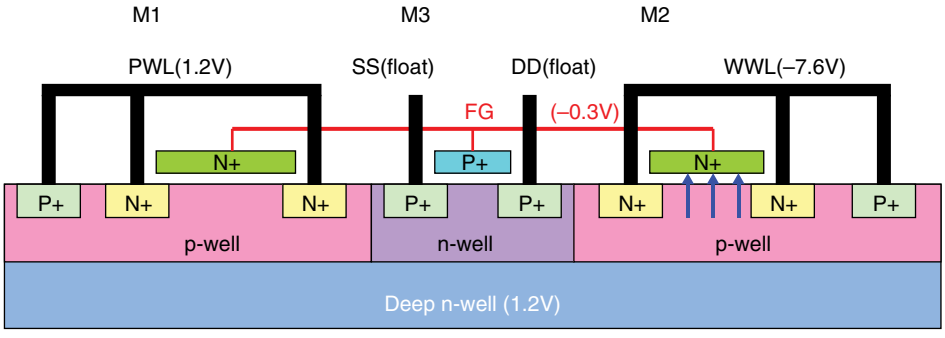


Figure 3.22 Erase of P-channel 3T single poly eFlash cell with Fowler-Nordheim electron tunneling into the floating gate of M2. Based on S.H. Song *et al.* (University of Minnesota), IRPS, April 2013 [29].

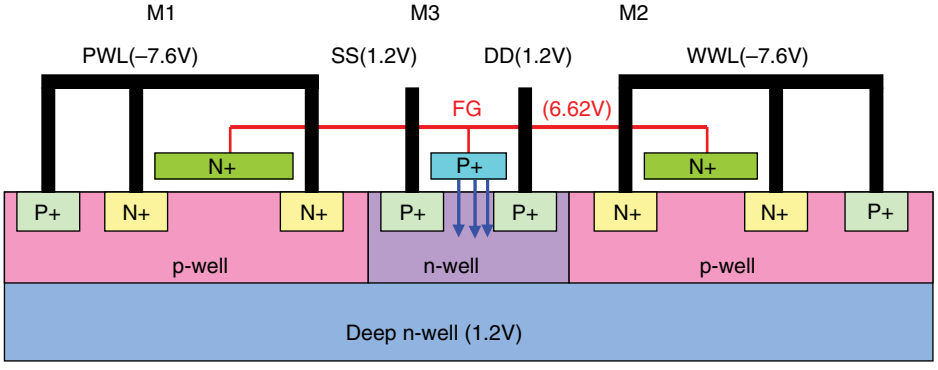


Figure 3.23 Programming of P-channel 3T single poly eFlash cell with electron Fowler-Nordheim tunneling from the floating gate of M3. Based on S.H. Song *et al.* (University of Minnesota), IRPS, April 2013 [29].

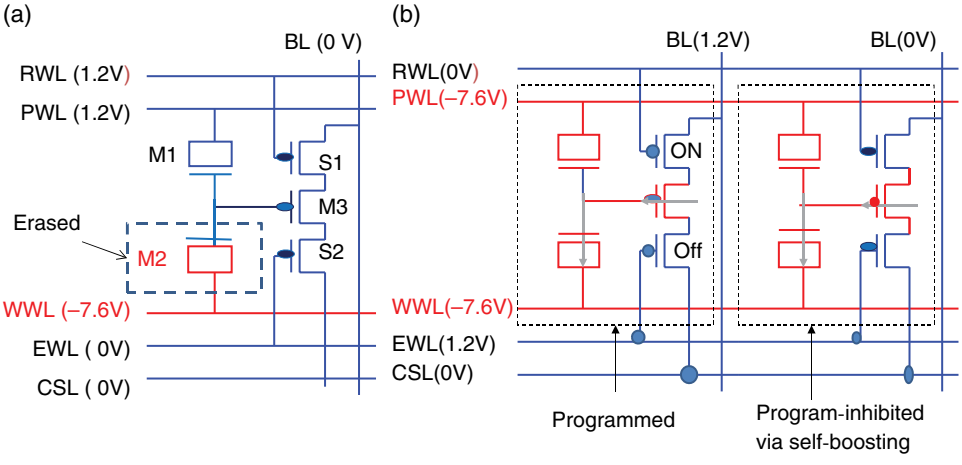


Figure 3.24 5T schematic circuit diagrams of the P-channel single poly P-channel macros with applied (a) erase bias and (b) program bias with program inhibited via self-boosting. Based on S.H. Song *et al.* (University of Minnesota), IRPS, April 2013 [29].

For the bias conditions of a cell being programmed, the boosted channel voltage should be high with suppressed subthreshold and junction leakage current to prevent program disturb in an unselected cell. The select transistor uses a longer channel length to minimize subthreshold leakage of the boosted channels. The self-boosting technique in a conventional logic technology permits row-by-row program/erase array architecture without a disturb issue in the unselected WL cells.

The feasibility of multilevel cell programming was investigated and the 5 T eFlash cells showed four distinct states with a sensing margin of 0.4 V for 150 °C baking temperature and 100 P/E pre-cycle. For the n-channel 5 T and p-channel 5 T eFlash, supply voltage was 1.2 V for the core and 2.5 V for the I/O. Tunnel oxide was 5 nm and cell size was 8.62 μm^2.

A 6 T eFlash memory cell was proposed by the University of Minnesota in August of 2014 [30] and in September of 2013 [31] for a bit-by-bit rewritable eFlash memory in 65 nm logic process. This cell improved the overall cell endurance by eliminating redundant program/erase cycles without disturbing cells on unselected word-lines. A multi-story high voltage switch used four boosted supply levels generated by a compact voltage doubler based on an on-chip negative charge pump. Schematic circuit diagrams are shown in Figure 3.25 of (a) the original 5 T eFlash and (b) the new 6 T eFlash.

In order to improve the overall cell endurance characteristics the array is capable of writing data on a bit-by-bit basis without using boosted BL voltages. To enable a bit-by-bit write, the cell boosts the FG of each cell selectively using preferential coupling. The cell does not share the source and drain node between adjacent cells in the WL direction, which permits it to have different voltage levels for each bit-cell in the same WL. The FG node voltage difference can be used for the bit-by-bit write operations.

Differences between the two cells included: cell size of 5.62 μm^2 for the 5 T and 15.3 μm^2 for the 6 T. The capacity of the 6 T at 4Kb was double that of the 5 T at 2Kb.

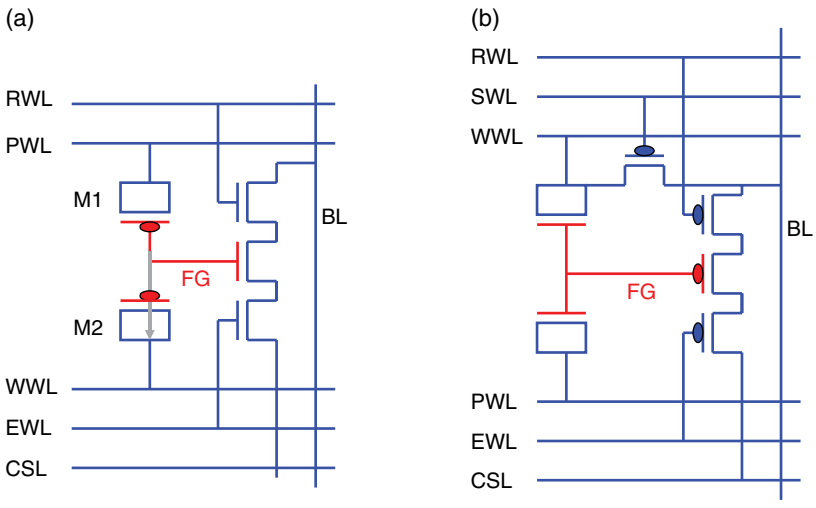


Figure 3.25 Schematic circuit diagrams of (a) original 5T eFlash and (b) 6T eFlash with improved endurance. Based on S.H. Song *et al.* (University of Minnesota), Custom Integrated Circuits Conference, September 22, 2013 [31].

Write voltage for the 5 T was 10 V and for the 6 T was –7.2 V. Both cells were in 65 nm logic and had 5 nm tunnel oxide. While the capacity of the 6 T cell macro is doubled, the write voltage is reduced but the cell size more than doubles.

A further comparative study of the 5 T single poly eFlash cell was conducted by the University of Minnesota in November of 2014 [32]. Characteristics of the different single poly eFlash memory structures compared were: memory disturbance, program/erase speed, endurance, and retention. They concluded that the 5 T eFlash memory cell structure was the most attractive option for logic-compatible eNVM.

3.4 Split Gate Technology for Floating Gate Embedded Flash

3.4.1 Early Split Gate Embedded Flash Floating Gate Technology

The 1.5 transistor (1.5 T) split gate floating gate Flash technology for use in embedded Flash memory was developed early and is widely used in foundries. The split gate has the advantage over the stacked double poly standalone 1 T Flash of a fast simplified erase at the cost of additional silicon and processing.

An early split gate embedded EEPROM application was discussed by Analog Devices in December of 1998 [33]. The SoC chip was an IEEE 1451 standard transducer interface module with an embedded Flash memory for storing transducer identification, calibration, correction data, and manufacturer-related information. The chip integrated an eight-channel 12-bit ADC, two 12-bit DAC, and an 8-bit MCU with 256bytes of SRAM and a 10.5KByte Flash EEPROM. A split gate Flash EEPROM cell was used, which had poly-to-poly erase and metal-to-poly capacitors with a calibration algorithm in the ADC [34]. The 27.5 mm^2 chip was made in 0.6 μm CMOS. It operated from 2.7 to 5.5 V and dissipated 13 mA of active current at 3 V/12 MHZ. An illustration of a schematic cross-section of the split gate Flash EEPROM cell is shown in Figure 3.26.

The conventional stacked double poly EEPROM process provides a high capacity memory device but requires: extra layers for the floating gate, high voltage devices, deep implants for HV devices, tunnel oxide, and extra drain implants for erase. Over-erase prevention could require extra logic or a state machine controller. On the other hand, single polysilicon cells simplify the memory process, making it more compatible with CMOS logic, at the expense of a much larger cell size, which makes the integration of a large number of bits unfeasible.

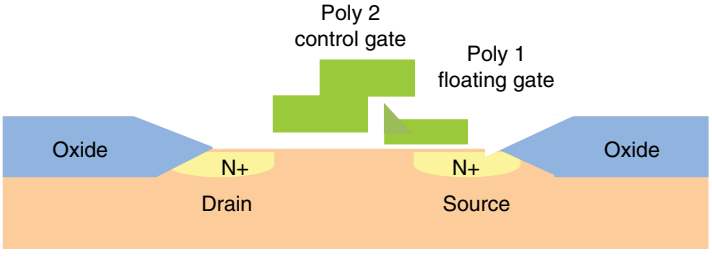


Figure 3.26 Cross-Section of split gate Flash EEPROM cell. Based on T. Cummins *et al.* (Analog Devices), *IEEE Journal of Solid-State Circuits*, December 1998 [33].

A split gate EEPROM cell with poly-to-poly erase is a compromise between the high process complexity of the conventional EEPROM and the large cell size of the single polysilicon EEPROM. The thick oxide erase mechanism means that no tunnel oxide or drain engineering is required for erase, which simplifies the process. Both the cell program and erase operations are self-limiting, which eliminates the requirement for over-erase prevention circuitry.

The split gate EEPROM cell with poly-to-poly erase was patented in 1993 by Silicon Storage Technology [34]. Erase and program operations are shown in Figure 3.27. Erase, shown in Figure 3.27 (a), is by thick oxide poly-to-poly Fowler-Nordheim tunneling of electrons from the poly1 floating gate to the poly 2 control gate and (b) program, shown in Figure 3.27(b), is by channel hot electron injection (CHEI) from the source to the floating gate.

A 65 nm embedded split gate Flash memory embedded in a conventional CMOS logic process technology was discussed by TSMC in April of 2011 [35]. A cross-sectional view of the split gate cell is shown in Figure 3.28. This cell uses source side injection (SSI) for programming and poly-poly Fowler-Nordheim tunneling for erase. The structure has multiple floating gate dielectrics. This memory does not have the limitations of the stacked gate Flash, which include scaling limitations due to the gate control oxide acting as tunnel oxide and due to drain turn-on program disturb.

This cell showed good scalability due to the erase tunnel oxide being decoupled from the floating gate oxide. It had good program disturb immunity since the select gate (WL) could turn off the channel of unselected cells. It did not need the boosted WL

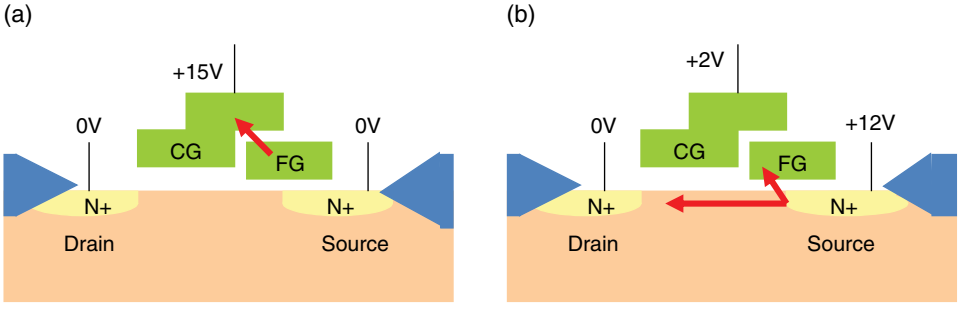


Figure 3.27 Cross-section showing (a) erase and (b) program of split gate cell. Based on B. Yeh, US Patent 5 242 848, September 7, 1993 [34].

Figure 3.28 Schematic cross-section of the 65 nm split gate Flash cell with SSI program and poly-to-poly Fowler-Nordheim erase. Based on Y.S. Chu *et al.* (TSMC), IRPS, April 10, 2011 [35].

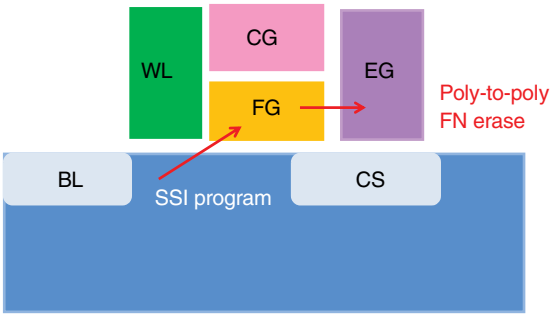


Table 3.4 Typical operating voltages for a 65 nm split gate Flash cell.

	CG	EG	WL(SG)	CS	BL
Program	11	1	1.3	5	0.2
Erase	0	13	0	0	0
Read	1.8	0	3.3	0	1.3

voltage during read access that the stacked gate Flash used. This resulted in comparative savings of power and access latency. The process was simplified by forming the FG self-aligned to the control gate (CG). Typical operating voltages for a 65 nm split gate Flash cell are given in Table 3.4.

A 256Kbit data Flash macro and a 16Mb code Flash macro were made to demonstrate that automotive grade embedded Flash could be made in 65 nm CMOS using seven added masks. The automotive grade data Flash memory was made with 100 K cycle endurance, 10 year data retention at 125 °C, and reliability at 1 ppm. The code Flash showed 10 000 cycle endurance. A comprehensive dielectric screen methodology was shown. Both erase time push-out and the dominant mechanisms of data retention were studied.

A split gate Flash memory cell embedded in a 65 nm high performance CMOS logic process with copper interconnects and low-k dielectric were discussed in May of 2012 by Infineon and TSMC [36]. The cell used a self-aligned process and had a gate spacer used during processing. The cell had good reliability and only one additional mask used for the split gate structure over the usual stacked gate process. The process flow was modular with both high voltage and Flash sections. The high voltage transistors required additional wells and thick gate oxide with special implants and LDD junctions. They were made in the same poly layer used for the floating gate. A third thick poly layer was used to form the select gate (SG) spacer. A schematic cross-section of the cell is shown in Figure 3.29.

The cell was optimized for minimum module area overhead and high endurance. It could be integrated as a module in a standard stacked gate technology. The cell was proposed for high performance automotive CMOS logic and for smart card applications. A 2MB 65nm Flash automotive product demonstrator with 2MB of SRAM and a 32KB of ROM was designed and tested. The use of a Flash cell process self-aligned to the stacked gate process allowed for scaling the cell without encountering lithographic limits or adding additional masking layers.

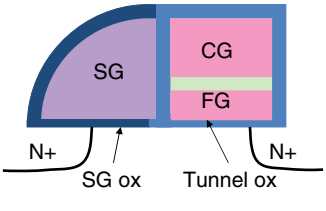


Figure 3.29 Schematic cross-section of split gate Flash cell in 65 nm high performance CMOS logic. Based on D. Shum *et al.* (Infineon, TSMC), IMW, May 2012 [36].

Low power embedded Flash in MCU is used in such applications as touch control ICs, smart meter controllers, and wireless connectivity. For such low density applications, 350, 250, and 180 nm process node technologies were in volume production at TSMC in May of 2013 [37]. The embedded Flash in these products was a first generation split gate cell. For higher density 2–16 Mb embedded Flash MCU applications, a new

split gate cell was used in 90 and 55 nm low power (LP) processes in 12 inch wafer fabs. This technology was targeted at a wider range of applications in the automotive, industrial, and commercial markets.

The new split gate cell had fast code execution and was expected to serve automotive applications better. A new single-ended sense amplifier (SA) was implemented on a 2Mb split gate eFlash in the 90 nm low power (LP) process. In 55 nm technology, the split gate Flash had a read speed up to 125 MHz using a single-ended sense amplifier (SA), but maintained a low power read at 75 μA/MHz. This cell used a five-metal eFlash IP made in a 90 nm LP process. It had up to 10 k cycle endurance in automotive applications. It was expected that the new split gate Flash cell with the new SA would primarily be used in the 55 nm technology.

High performance mobile processors and MCUs require large capacity, fast-read embedded nonvolatile memory (eNVM) for code and data storage. Since conventional current-mode sense amplifiers (CSA) cannot achieve fast random read access time due to significant summed read-path input offsets, in November of 2013, NTHU, Shou University, and TSMC proposed an asymmetric voltage-biased CSA to suppress summed read-path input offsets and enable high speed sensing without run-time offset-cancellation operations [38]. A 1Mb eFlash macro with BL-length test modes was made. A conventional split gate eFlash cell macro was used for the eNVM. This macro had 3.9 ns random read time. Features of the 90 nm low power split gate embedded Flash included: 1.2 V power supply, 512 cell bit-line length, and random read access of 3.9 ns at 1.2 V typical. Area overhead was less than 1%, read power overhead was less than 2%, and test time overhead was less than 0.5%.

A split FG Flash cell with corner enhanced poly-to-poly tunneling was announced in January of 2012 by Silicon Storage Technology, now Microchip, and SST described its first three generations of memory cell [39]. These split gate memory cells used source side hot electron injection for programming and Fowler-Nordheim poly-to-poly tunneling for erase. The first two generations had an intentionally sharpened floating gate tip formed using local FG oxidation. This made it possible to use a relatively thick tunnel oxide, which reduced such reliability issues as stress-induced leakage current and program erase cycling-induced data loss.

Schematic cross-sections of the SST split gate cell are shown in Figure 3.30. Figure 3.30 (a) illustrates the first generation 1.2 μm to 130 nm technology cell. Figure 3.30 (b) illustrates the second generation 250 to 130 nm technology cell, which used a triple self-aligned process. The FG was self-aligned to the shallow trench isolation and the source line and word line poly were self-aligned to the floating gate. The FG tip curvature was achieved by using reactive ion etching. Figure 3.30 (c) illustrates the third generation cell, which had a dedicated erase gate so the word-line oxide was not stressed during erase and could be scaled. A separate coupling gate was used for positive biasing of the FG during programming and the source line to floating gate overlap was eliminated.

The F-N poly-to-poly erase of the first generation cell is from the floating gate to the poly SG (WL) and is at 12–14 V. Programming uses SSHEI from the floating gate through the tunnel oxide to the drain and the SL is at 8–10 V.

For the third generation cell the F-N poly-to-poly erase is from the floating gate to the erase gate (EG) with the EG at 11 V and the other gates at 0 V. SSHEI programming is from the floating gate through the tunnel oxide to the drain, with the coupling gate (CG) at 10.5 V and the SL and EG at 4.5 V. The use of the CG permitted a reduction in the

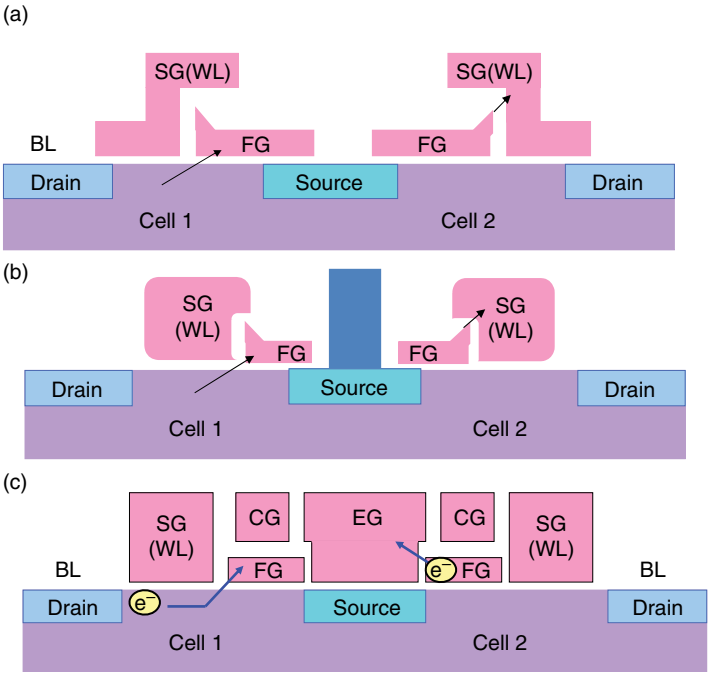


Figure 3.30 SST split gate "Super" Flash. Illustrations of (a) first generation, (b) second generation and (c) third generation technology cells. Based on Y. Tkachev (Silicon Storage Technology), *IEEE Trans. on Electron Devices*, January 2012 [39].

source voltage during programming, which reduced the programming-induced degradation of the cell channel and improved program erase cycle endurance.

Reduction of program-induced degradation for the split gate Flash was discussed in March of 2014 by SST [40]. The degradation mechanism was induced by program-erase cycling in a split gate memory cell. A test structure was used that had two cells with a common floating gate. This test structure permitted separation of the degradation mechanisms, which occurred in the floating gate oxide and in the tunnel oxide during cycling. It was shown that the program-induced floating gate oxide degradation became less significant for more advanced technologies that use lower programming voltage.

The cycling-induced degradation of erase performance was found to be caused by electron trapping in the floating gate oxide during programming and in the tunnel oxide during erase. It was shown that the third generation cell had a significantly smaller program-induced degradation than the first generation due to the lower source voltage used in programming.

3.4.2 Issues, Peripherals, and Applications-Specific FG Split Gate Memory

A 45 nm logic compatible 4Mbit triple poly splitgate cell embedded Flash was discussed in May of 2014 by Samsung [41]. This device was intended for applications such as smartphones with security features, near-field communications (NFC), banking, transportation pass, and ID. The memory capacity requirement was from 1 to 16 Mb. Some of the specifications of the various applications targeted are shown in Table 3.5.

Table 3.5 Various application specifications for split gate Flash chips. Based on Lee *et al.* [41].

Characteristics	Smart card	General MCU	Automotive MCU
Endurance (k-cycles)	100–500	10–100	100
Retention (years)	10	10	10
Temperature (°C)	−40 to +85	−40 to +85	−40 to +125/+150
Read speed (MHz)	>20	>25	>75

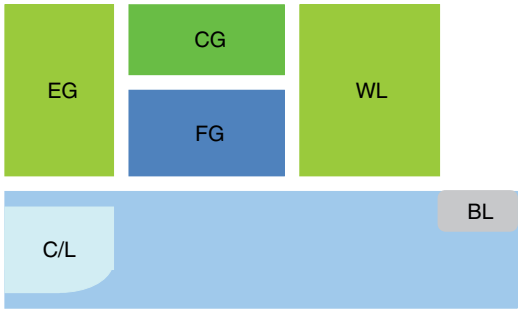


Figure 3.31 Schematic cross-section of 45 nm logic compatible triple poly split gate cell. Based on Y.K. Lee *et al.* (Samsung), IMW, May 2014 [41].

A schematic cross-section of the 45 nm logic compatible triple polysilicon split gate cell is shown in Figure 3.31. Scaling to 45 nm technology was accomplished by forming a shallow source-line junction and by using thin gate oxide in the WL transistor. A low voltage word-line (WL) transistor was used without WL boosting. The use of a separate control gate permitted source-line voltage to be reduced so that program power could be reduced. The lower program power and low voltage WL meant the charge pump design could be simplified, which resulted in size reduction. The process flow for the 45 nm 1.5 T Flash involved making the cell first. The logic transistor-related modules were processed after cell and high voltage (HV) module formation.

Process integration was designed for logic compatibility so that existing IP could be used. The chip had 40 MHz read access, program time of 25 μs, and erase time less than 2 ms. Endurance was 1 M cycles and data retention was 10 years. The good endurance resulted from strong localization of the tunneling process and suppression of hole injection in the FG corner. Good retention was shown at 150 °C. Immunity to disturbance at low program power was verified. The poly-to-poly erase operation decoupled unselected cells, resulting in immunity to disturb.

A configurable logic array test structure was discussed in March of 2013 by SST [42]. The chip was developed using a split gate Flash memory cell configuration element (SCE) in 90 nm embedded Flash technology. The SCE eliminated the requirement for special fabrication processes and SRAM circuits and reduced the configuration time for the programmable array. The configuration element was created using twin split gate Flash cells. The cell used poly-to-poly erase and source-side-channel hot-electron injection programming. The SCE technology eliminated the triple-well process requirement. A schematic circuit diagram of the split gate Flash SCE and programmable array switch is shown in Figure 3.32.

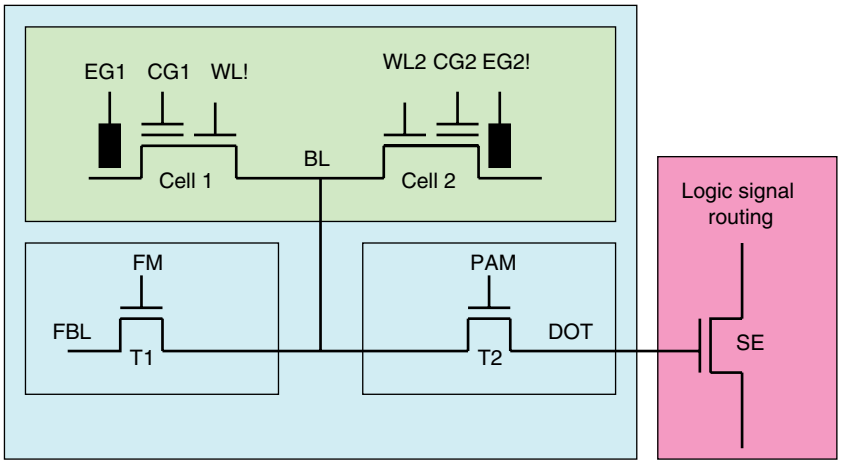


Figure 3.32 Split gate Flash SCE and programmable array switch. Based on H. Om'mani (Microchip/Silicon Storage Technology), ICMTS, March 2013 [42].

The Flash-based element used the same operation as used in the memory cell. Acceptable erased and programmed cell currents were achieved during the SCE Flash mode. The switching logic array (SLA) configuration can fully transmit the VDD signal without degradation to the output through the memory element. The cell current measured in Flash mode for the erased cell is around 27 μA and for a programmed cell is less than 10 pA.

A binary code inversion technique was used for an embedded Flash memory sense amplifier, which reduced the read current. This technique was discussed in May of 2014 by KAIST [43]. An original binary code was programmed into the Flash memory with an inverted binary code-based bit inversion technique. The de-inversion hardware was implemented with small logic gates to restore the original binary data so only logic current was consumed rather than analog sink current in the sense amplifier. The proposed technique was evaluated using a 128KB Flash memory embedded in an MCU made in 180 nm EEPROM technology. The circuit level simulation result showed that a fresh chip with the proposed sense amplifier consumed less than 22% of the operating power of conventional sense amplifiers.

A systematic study of program-disturb mechanisms in third generation 90 nm split-gate memory technology was discussed in June of 2014 by SST/Microchip [44]. A schematic cross-section of the cell is shown in Figure 3.33. The various oxides associated with programming and erase are indicated. The focus of the study was for the temperature range of –45 to +150 °C to be used in the automotive applications. For this chip at low temperatures, the dominant program disturb issue occurred with interface trap-assisted band-to-band tunneling in the split gate channel area. At high temperatures, the dominant program disturb issue occurred by surface generation in the select gate channel area.

The third generation cell improved program disturb immunity over the previous generation by having a lower source voltage (V_s) used for the program operation. This reduced hot carrier effects including the generation of secondary electrons, which are capable of producing program disturb. The high voltage erase operation in this cell was decoupled

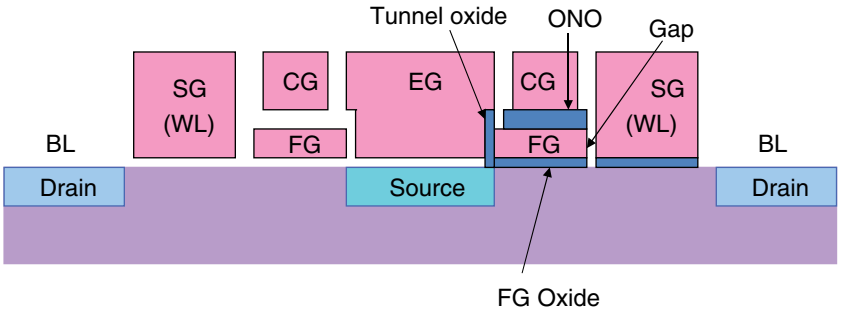


Figure 3.33 Schematic cross-section of a third generation split gate memory cell showing oxides associated with programming and erase. Based on V. Markov and A. Kotov (Silicon Storage Technology/Microchip), *IEEE Trans. on Devices and Material Reliability*, June 2014 [44].

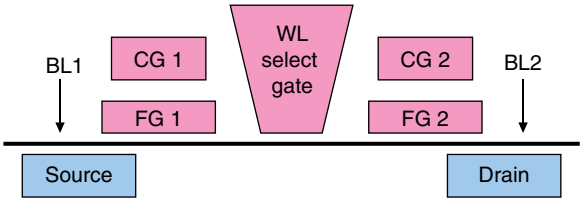


Figure 3.34 Schematic cross-section of vertical 2-bit/cell split gate embedded Flash memory with shared bit-lines. Based on L. Fang *et al.* (Chinese Academy of Science, Huahong Grace Semiconductor), *IEEE Trans. on Electron Devices*, July 2014 [45].

from the select gate (SG) transistor, which made it compatible with thin gate oxide low voltage CMOS transistors. The scaling of the select gate oxide reduced short channel effects (SCE), which suppressed program disturb induced by subthreshold current.

The effects of single interface traps on program disturb were analyzed. It was found that a split gate memory cell with a high quality Si–SiO_2 interface provided the strong program-disturb immunity that is required for high temperature and automotive embedded applications. Split gate memory cells subjected to 5×10^5 erase/program cycles did not show an acceleration of the program disturb mechanism. Electrons generated by floating gate channel interface traps could not acquire sufficient energy from a low lateral electric field in the channel to surmount the Si-SiO_2 energy barrier.

A reliable 2-bit/cell vertical split gate embedded Flash memory cell configured for immunity to program disturbs was discussed in July of 2014 by the Chinese Academy of Science and Huahong Grace Semiconductor [45]. A schematic cross-section of the cell is shown in Figure 3.34. The control gates and floating gates share BL1 and BL2. A chip using this cell architecture was fabricated in 130 nm technology with a 5 F2 bit cell area. A new metal interconnect technique was used for a virtual ground array architecture to improve program disturb as opposed to the conventional AND-type configuration commonly used. A fully self-aligned process with shallow trench isolation in the cell array was used to make this word-line shared split gate structure without having any lithography misalignment issue. A shallow trench isolation (STI) structure was used in the array to relax the lithographic alignment requirement.

A reliability issue, negative charge trapping in the select gate (SG) oxide during a conventional poly-to-poly Fowler-Nordheim tunneling erase operation was found to be an important contributor to the cycling degradation for cells with thin SG oxide. A negative control gate bias erase method was suggested for enhancing the endurance reliability. A 250 °C bake before and after cycling was performed to check the cell data retention characteristics and no extrinsic or intrinsic defects were found. Erase and program characteristics were comparable to conventional split gate cells.

A fully self-aligned split cell was shown along with a new program disturb immune virtual ground array. A reverse read scheme enlarged the operation window and high endurance reliability was found using a negative voltage-assisted erase operation. Typical operating conditions used an 8 V program and –7 V erase. For read, the SG was at 4 V. The threshold of one bit is the voltage applied to the corresponding CG terminal to obtain a fixed source-drain current of 1 μA.

Key blocks of peripheral circuits used for NOR embedded Flash memory were discussed in September of 2014 by the Chinese Academy of Science (CAS), Grace Semiconductor (GSMC), and Shanghai Huahong [46]. These included: a novel source line (SL) voltage compensation circuit and a WL voltage-generating system. A NOR Flash memory using the two techniques was developed using a GSMC 180 nm 4-poly 4-metal CMOS process. The memory used a split gate Flash cell for its fast erase speed, high program efficiency, and large process window. Average standby current for the eFlash array was less than 0.3 μA at 1.8 V and 25 °C. The common source Flash array using the split gate Flash cell are shown in Figure 3.35. The cell is programmed by source-side channel hot electron injection (CHEI).

The new SL voltage compensation circuit controlled the output voltage of the charge pump by considering the number of cells to be programmed with logic "0" data to compensate for the IR drop on the SL decoding path. This stabilized the SL voltage, providing a high program efficiency with low program disturb. To obtain low power consumption in standby as well as high speed in active mode, a WL voltage generating system was developed. A 1.8 V 64 × 32Kbit NOR eFlash was made using the two techniques. The measured standby current at 1.8 V and 25 °C was less than 0.3 μA.

A SPICE macro model of the third generation split gate cell in 55 nm CMOS technology was discussed by CEA-LETI and Microchip Technology in March of 2015 [47]. A parameter extraction procedure was proposed, which showed good agreement between the model and measurements. The split gate Flash memory was originally developed for

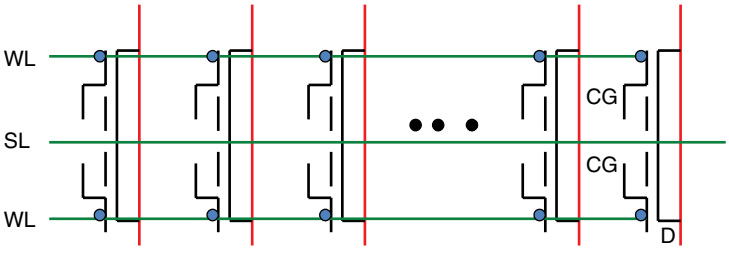


Figure 3.35 Common source Flash array with split gate Flash cell. Based on S. Zhang (Chinese Academy of Sciences (CAS), GSMC, Shanghai Huahong), *IEEE Trans. on Circuits and Systems – II. Express Briefs*, September 2014 [46].

low power embedded memory applications in order to improve injection efficiency, erase efficiency, and disturb immunity. It effectively adds a separate access transistor to the Flash memory transistor. The access, or selected gate (SG), transistor controls the current that flows in the memory during read, program, or erase operations. This lowering of the current makes the split gate memories useful in low power applications. Immunity to disturb is improved due to no current flowing when the select gate transistor is opened.

A compact SPICE model connects the process development and the circuit design. The model is intended to accurately reproduce characteristics in the design. For this purpose, model cards with the corresponding electrical characterization extraction parameters are defined. A schematic cross-section of the 55 nm embedded 2 T split gate Flash memory cell is shown in Figure 3.36.

The operating principle is: when the cell is erased, positive charges are stored in the floating gate (FG) and the threshold voltage of the storage transistor is low. When the memory is programmed by the injection of negative charges on to the floating gate, the voltage of the storage transistor becomes higher. The selected gate and storage transistors were modeled using the BSIM3v3 SPICE model. The state is determined by reading the bit-line current. Biases for erase, program, and read used in the model are shown in Table 3.6.

The macro model of the 55 nm eFlash split gate memory cell was based on a set of equations including main physical aspects of charge and discharge cycles. The model was designed for implementation in SPICE simulators. Good agreement was shown between SPICE simulations and experiments for a range of conditions during operation. This model is currently used in production to optimize 55 nm memory circuits.

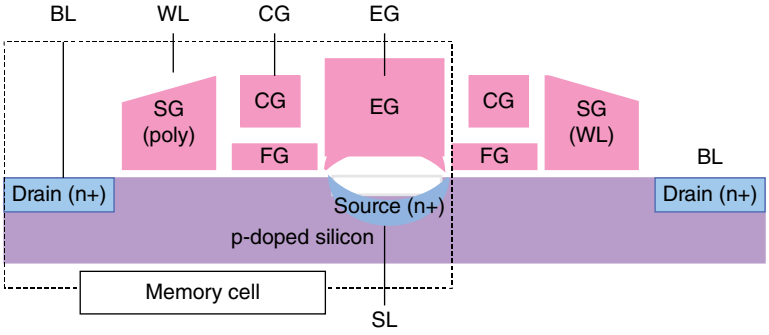


Figure 3.36 Illustration of schematic cross-section of 55 nm split gate eFlash cell used for compact SPICE model. Based on S. Martinie *et al.* (CEA-LETI, Microchip Technology), ICMTS, March 2015 [47].

Table 3.6 Biases for erase, program, and read.

Operation	WL	BL	SL	EG	CG
Erase	0	0	0	V_{eger}	0
Program	V_{wlpg}	0	V_{slpg}	V_{egpg}	V_{cgpg}
Read	V_{wlrd}	V_{blrd}	0	0	V_{cgrd}

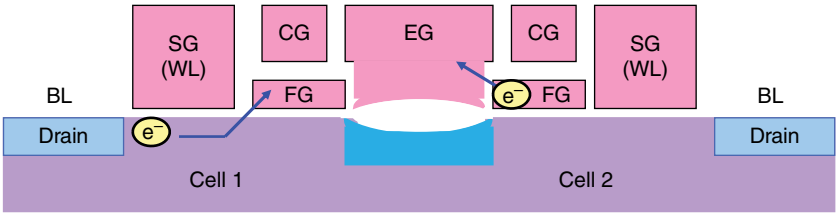


Figure 3.37 Logic compatible split gate Flash cell in 55 nm logic process. Based on N. Do *et al.* (Silicon Storage Technology/Microchip Technology), IMW, May 2015 [48].

A split gate Flash array for automotive and for smart card applications was discussed In May of 2015 by SST/Microchip [48]. This 55 nm logic-compatible split gate Flash memory, their ESG3 cell, was tested at automotive temperatures and showed fast access and good reliability. It was shown that this Flash memory macro could be embedded in multiple logic process platforms and could be scaled to 40 nm without compromising performance and reliability. A schematic cross-section of the cell is shown Figure 3.37.

The cell was scaled from 90 to 55 nm and it was shown that the programmed and erased cell currents remained comparable. A tight distribution of cell current permitted a wide window for operation. It was shown that, with a 4.5 V fixed source-line voltage, the memory array could sustain up to 12.5 V without any disturb bits after several program cycles. Similarly, with a fixed control gate voltage of 10.5 V, the memory array could sustain up to 6 V without disturbed bits after several program cycles.

A memory array underwent more than 1 million program/erase cycles at 165 °C and remained functional. The automotive specifications used required the Flash macro to have an access time of 10 ns or less and this design provided an access time of less than 7 ns for worst case operation. It was found that the same cell and process could be used to design Flash macros for smart card products that require fast program/erase speeds and high endurance. A 4 Mb Flash macro was shown to require less than 4 ns total time for sector program and erase.

It was found that a smart erase algorithm could be used to enhance endurance performance. A single pulse of erase and program was used to show the intrinsic endurance. At the 55 nm node, this cell array passed qualification at 100 k cycles and 1000 hours of high temperature operating life (HTOL) and data retention storage life (DRSL) bakes in multiple foundries. Efforts were ongoing to scale the cell to less than 0.05 μm^2 at the 28 nm node.

In October of 2014, CAS, Shanghai Huahong, and Grace Semiconductor (GSMC) discussed a 1.35 V, 16Mb twin-bit cell embedded Flash memory with virtual ground architecture [49]. A sensing current protection method was proposed to reduce the sensing margin loss caused by the side-leakage current in the virtual ground memory array. A reference voltage generating circuit for dynamic sensing window tracking was designed to maximize the sensing window under various process, voltage, and temperature (PVT) conditions. The reference voltage generating circuit and a fast sense amplifier resulted in a high speed read operation. Since four bit-lines need to be selected to read one bit, a source-drain-protection column decoding methodology was used to support the sensing current protection technique.

The eFlash was made in a GSMC 90 nm 4 poly and 4 metal CMOS process. The macro size of the eFlash was 3.2 mm^2 and the memory cell size was 0.16 μm^2. Access time was 36 nm at 1.35 V. A schematic circuit diagram of the twin-bit-cell and a schematic cross-section are shown in Figure 3.38.

The 2-bits per cell are achieved by treating both the SG and one of the FG devices as transfer gates when accessing the other bit. Due to the symmetrical structure, the CG and S/D terminals must interchange their bias conditions when switching access from Bit1 to Bit2 for both read and program. The cell acts as a conventional split gate cell and uses source side channel hot electron injection (SS CHEI) and poly-to-poly FN tunneling erase. A programmed bit has a high V_t and stores "0". An erased bit has a low V_t so it stores "1". Since each FG stores two bits, the twin bit-cell has four states. The cell size in 90 nm technology is 0.16 μm^2. Key performance parameters of the 90 nm 16Mb twin-bit eNVM include: 36 ns Tac at 1.35 V, 3.8 mA 32 bit read current, 10 µs write time for 8 bits, 1.8 mA active current, and less than 30 µA standby current. Chip size for a 16Mb array was 3.2 mm^2.

A 200 KB embedded single 1.5 V power supply EEPROM was discussed in March of 2016 by the Chinese Academy of Science [50]. The device was based on a Shanghai Huahong Grace Semiconductor Manufacturing Corporation (HHGRACE) 90 nm EEPROM process with a 4 poly and 4 metal CMOS [50]. The architecture of the eEEPROM is shown in the block diagram in Figure 3.39.

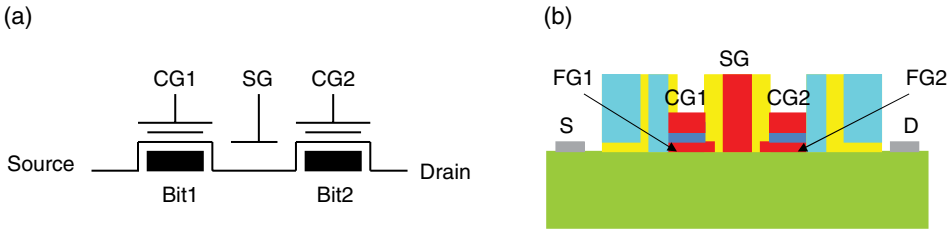


Figure 3.38 Twin bit cell: (a) schematic circuit diagram and (b) schematic cross-section. Based on S. Zhang *et al.* (CAS, Shanghai Huahong, GSMC), *IEEE Trans. on Circuits and Systems – I: Regular Papers*, October 2014 [49].

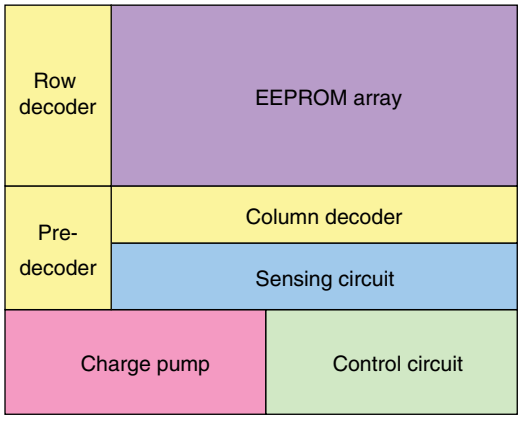


Figure 3.39 Block diagram of architecture of the embedded EEPROM. Based on Y. Xu *et al.* (Chinese Academy of Science), *IEEE Trans. on Circuits and Systems – II*, March 2016 [50].

Several key design techniques were discussed. An improved bit cell with a larger current sensing window was used in the conventional split-source EEPROM array. For high speed read a fast sense amplifier and dynamic sensing window tracking reference voltage generating circuit were proposed. The chip size of the EEPROM macro is 1.271 mm^2 and the EEPROM cell size is 0.32 μm^2. Access time was 30 ns at 1.5 V and 25 °C.

The effect of radiation and stress-induced degradation on split gate NOR Flash cells was discussed in April of 2016 by NAVSEA Crane [51]. Radiation and program/erase stress on the bit cells was shown to create both positive and negative traps in the oxide around the floating gate. The rate at which oxide traps are neutralized was affected by the annealing temperature following radiation.

3.4.3 Advanced Split Gate Floating Gate Technology below 50 nm

The properties of several embedded nonvolatile memory technologies used in microcontrollers were discussed in May of 2016 by Microchip Technology [52]. These include various types of nitride trapping and floating gate embedded NVM cells as well as several types of emerging NVM technologies.

Being able to scale an embedded technology is important to remain compatible with the CMOS logic it is embedded in. A split gate cell can be made in either charge-trapping or floating gate technology. Split gate Flash, such as the SG MONOS charge-trapping technology, has been shown to be scalable from 40 to 28 nm. A cross-section of the basic split gate SG-MONOS cell is shown in Figure 3.40 (a) and the split floating gate (SG-HS3P) cell in Figure 3.40 (b).

The SG-MONOS is a split gate cell with the nitride charge-trapping storage under the memory gate and the SG-HS3P cell uses a floating gate. In the split gate cell, the select gate, or word-line (WL) gate, can be used to turn off the channel to avoid leakage due to over-erase.

The ESF3 cells in the bit-line direction are shown in Figure 3.41. This split gate floating gate cell had the added advantage of a separate erase gate (EG). This cell in 28 nm technology functions up to 170 °C with a cell size of 0.053 μm^2. During erase the electrons tunnel between FG and EG with only a positive high voltage bias on the EG polysilicon. This cell can be used in high temperature sub 10 ns access time environments, such as automotive engines. Only the EG is subjected to high voltages. The rest of the device is used at logic voltages and can be scaled with the technology.

In operation for erase, only the erase gate has high voltage applied, with all other inputs at 0 V. For read only logic voltages are applied. For program operation, higher

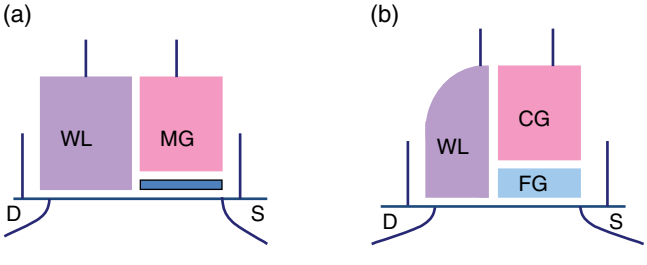


Figure 3.40 Cross-section of (a) basic split gate SG-MONOS cell and (b) SG-HS3P floating gate cell. Based on N. Do (Silicon Storage Technology/Microchip Technology), IMW, May 2016 [52].

voltages are applied to the CG and the SL. The V_t window is greater than 10 V. Program and erase are at 100–500 k cycles.

Optimizing the programming conditions and enhancing the endurance of a third generation SG-MONOS cell in split gate embedded Flash memory technology was discussed by Microchip in April of 2017 [53]. The new optimization method is based on the use of a two-step program operation with reduced programming voltage during pre-programming. This provides a reduction of the peak lateral electric field in the channel at the beginning of the program operation, as a result of a short pre-programming pulse with reduced V_{cg1}, and resulted in reduced erase performance degradation at the beginning of the program operation. This in turn reduces gate oxide degradation on the hot electron injection area. Additional improvement of endurance and operating performance was attained by reducing the programming time. This was compensated by a slight increase in programming voltage. The faster programming provides lower write energy consumption and higher program-disturb immunity. Program/erase cycling under this programming condition results in typical endurance of more than 10 million cycles.

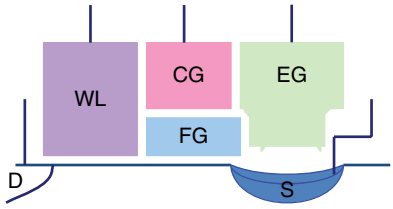


Figure 3.41 Schematic cross-section of ESF3 cell in bit-line direction. Based on N. Do *et al.* (Silicon Storage Technology/ Microchip Technology), IMW, May 2016 [52].

A second generation 45 nm split gate embedded Flash that was scaled by 40% to nearly the size of the 28 nm technology node was discussed in May of 2016 by Samsung [54]. No extra masks or processes were used nor was any advanced equipment added to the process. The device had 10 ns random access time, 25 μs write time, and a less than 2 ms erase. Reliability was 1 M cycles of endurance and 20 years data retention, and 1 to 16 Mb capacity macros were verified.

High "on-cell" current is a key cell parameter for maximizing the endurance cycling limit, which provides fast random access speed. The on-cell current of a single bit cell is output from three serially connected devices, which are a low voltage word-line select transistor, a control gate transistor, and a resistor formed in the off-set oxide spacer region between the WL and the CG. A schematic cross-section of the cell along with a device schematic showing the resistive effect of the off-set spacer is shown in Figure 3.42.

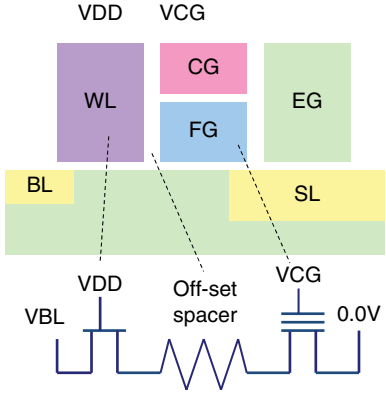


Figure 3.42 Schematic of SG cell and schematic showing resistive effect of off-set spacer. Based on Y.K. Lee *et al.* (Samsung), IMW, May 2016 [54].

For automotive applications, strict quality and reliability is required with zero failure rates at harsh temperature conditions. In May of 2016, Global Foundries and Singapore UT discussed a 16 Mb floating gate (FG) Flash macro in a 2.5 V self-aligned split gate ESG3 cell in a 40 nm CMOS logic-compatible process for embedding in automotive microcontrollers [55]. The macro had a dual power supply with operating temperatures from –40 to 150 °C. Random read access was 10 ns under worst case conditions. The technology has

Table 3.7 Operating conditions for the 40 nm ESG3 cell.

	Erase	Program	Read
Word-line	0 V	−1 V	V_{dd}
Bit-line	0 V	1 μA	0.6–1 V
Coupling gate	0 V	10.5 V	V_{dd}
Erase gate	11.5 V	4.5 V	0 V
Source-line	0 V	4.5 V	0 V

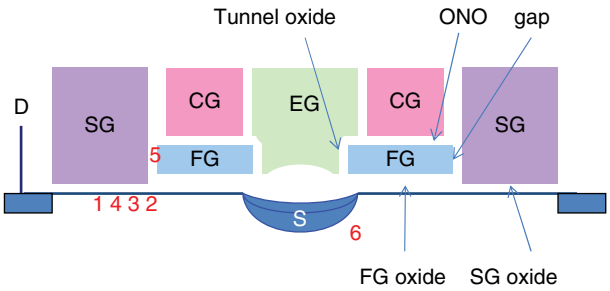


Figure 3.43 ESG3 cell with program disturb mechanisms shown. Based on V. Markov *et al.* (Silicon Storage Technology/Microchip Technology), IMW, May 2016 [56].

a large read current window, which is compatible with both automotive MCU markets and low power modes for smart card and industrial applications.

The cell and erase gate were processed with self-alignment to the gate spacer. A polysilicon chemical mechanical polishing (CMP) was used that can be integrated into the standard logic process. Code storage requires 1 ns fast random access while data storage requires 200k cycles of endurance and greater than 10 years of data retention. This cell serves both applications. The cell is an ESG3 cell, which is erased using poly-to-poly Fowler-Nordheim tunneling from a field-enhanced tunneling injector formed between the floating gate and the erase gate. It is programmed using source-side channel hot electron injection (CHEI). The operating conditions for the 40 nm ESG3 cell are given in Table 3.7. There is a dual 1.1 to 2.5 V power supply. The read IO is 32 bits wide plus 6 bits for error correction.

A study of program disturb mechanisms in the Microchip/SST ESF3 cell were discussed in May of 2016 [56]. The ESF3 cell is shown in Figure 3.43. Program disturb mechanisms are indicated with numbers.

The new program disturb mechanisms were identified as: (1) surface generation in the SG transistor with (high E_a), (2) interface trap-assisted BTB tunneling in the gap (low E_a), and (3) SG-oxide trap-assisted tunneling (low E_a). Program disturb mechanisms already identified in earlier generations include: (4) subthreshold current in SG transistor, (5) FG to SG leakage (if isolation defects are present), and (6) secondary impact ionization. A key role was found for certain interface traps located in the cell channel under the select gate (SG) and the split gate gap in acceleration of power dissipation at high and low temperatures.

3.5 Stacked Flash and Processor TSV Integration

An alternative to the embedded Flash memory in a CMOS logic process is integrating an optimized Flash memory with an optimized processor. This can be done using stacked standalone memory chip integration with TSV with the chips side-by-side on an interposer chip. An illustration is shown in Figure 3.44 [57].

3.6 OTP/MTP Embedded Flash Cells and Fuses

A conventional 350 nm CMOS foundry process without extra processing steps was used to fabricate an embedded multiple-time programmable (MTP) memory array [58]. In June of 2014 researchers from National Chung Hsing University designed this circuit using high-voltage-tolerant circuit design techniques to permit the transistors to withstand the higher biases used for program and erasing the NVM cells. A schematic circuit diagram of the MTP cell is shown in Figure 3.45.

In addition to the conventional CMOS process, the IP from the high voltage process was used for the program circuitry. Hot hole and hot electron mechanisms were used. This process has a 5 V transistor, which could be run at 8.3 volts for a few programming and erase steps as required in an MTP memory system. By using the high-voltage-tolerant design techniques with the current sense amplifiers, random access time reaches 13 ns measured after program. This was thought to be a cost-effective method for designing the high voltage parts of the eFlash chips.

Read and write capability for a one transistor one-time-programmable (OTP) anti-fuse cell macro was made in a TSMC 16 nm CMOS FinFET process in September of

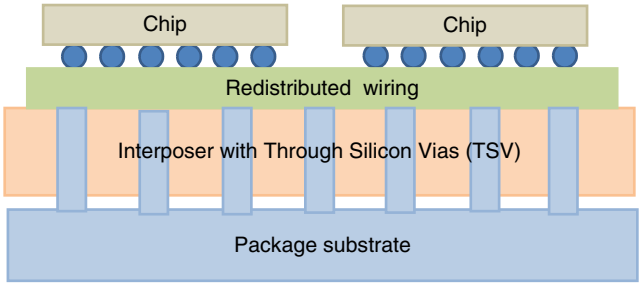


Figure 3.44 Illustration of stacked chip integration with an interposer with TSV.

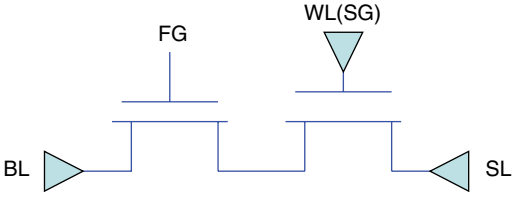


Figure 3.45 Schematic circuit diagram of the MTP memory. Based on C.Y. Huang *et al.* (NCHU), ISCAS, June 1, 2014 [58].

2014 by Sidense [59]. These OTP macros had been used for designs from 180 nm to 28 nm technology and have been proven in 20 nm technology. Early test results confirmed correct bit-cell operation on a 16 nm FinFET device with a program voltage comparable to the 1 T OTP device in 28 nm technology, but with 10 times lower leakage current. Programmed characteristics were as good as those for 20 nm cells with large margins between programmed and unprogrammed cells. Post-bake bit-cell stability was also good. The 16 nm macro was indicated as a potential option for design into devices for mobile circuits.

A logic process compatible OTP memory in Hi-K metal gate (HKMG) CMOS was made in December of 2015 by NCTU and UMC using a new dielectric fuse breakdown mechanism [60]. The breakdown observed in HKMG CMOS technology was found to be different from conventional anti-fuse dielectric breakdown. It showed a property of an open gate and could be operated at a program current of less than 50 μA and speed of 20 μs, which is an improvement over conventional fuse mechanisms. Cross-sections of 28 nm MOSFETs made to investigate the dielectric fuse mechanism are shown in Figure 3.46. Figure 3.46 (a) shows an SiO_2 polysilicon nMOSFET and Figure 3.46 (b) shows an HfON HKMG nMOSFET.

For IoT applications there is demand for reliable OTP functionality at low cost. Figure 3.47 shows the characteristics of a gate current for a conventional gate dielectric anti-fuse and the HKMG dielectric fuse mechanism discussed here.

The conventional antifuse breakdown causes a sudden increase to a high gate current; however, the new mechanism has the current suddenly shut off and the gate open. After anti-fusing, the gate current increases rapidly and after fusing it is blocked.

An embedded logic-compatible multiple time programmable NVM element was discussed in January of 2017 by IBM and UCLA [61]. The memory was intended for use in high-*k* metal gate (HKMG) CMOS technologies such as SOI and FinFET, which exhibit self-heating.

The memory device exploited an intrinsic device self-heating enhanced charge trapping in logic devices made in CMOS HKMG technology. A unique multiple time programmable embedded NVM element was developed, which was called a charge trap transistor (CTT). The intrinsic self-heating charge trapping memory is applicable to use in SOI and in bulk FinFET technology since self-heating occurs in bulk FinFETs although less than in SOI FinFETs.

Functionality was shown for fully functional memory arrays at 22 nm planar and 14 nm FinFET technology nodes. The density for 22 nm was 0.144 μm^2 per bit and for 14 nm FinFET technology was 0.082 μm^2/bit. The devices were logic voltage

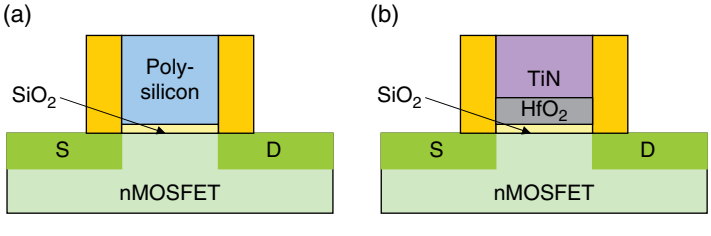


Figure 3.46 Schematic cross-sections of (a) 28 nm SiO_2 poly-silicon nMOSFET gate and (b) HfO_2 HKMG nMOSFET gate. Based on E.R. Hsieh *et al.* (NCTU, UMC), IEDM, December 2015 [60].

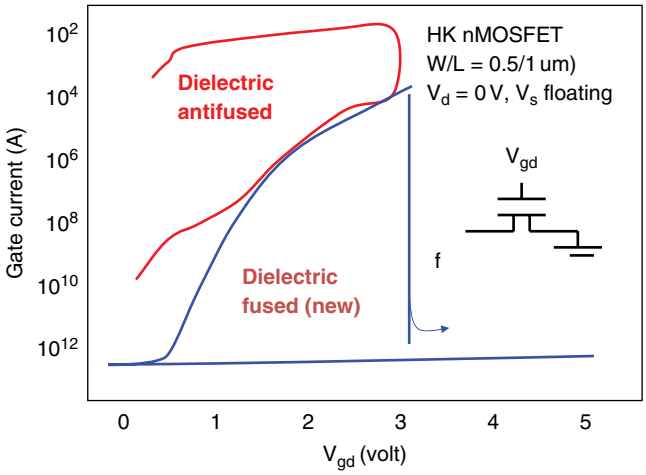


Figure 3.47 Characteristics of gate current for gate dielectric anti-fuse and HKMG dielectric fuse (Based on E.R. Hsieh *et al.* (NCTU, UMC), IEDM, December 2015 [60].

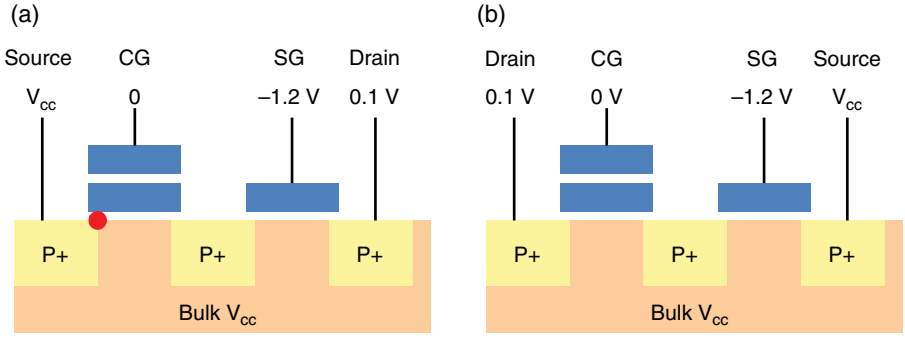


Figure 3.48 Schematic cross-section showing read operation of (a) conventional eFlash cell with location of high field stress on the tunnel oxide indicated by the dot and (b) reversed eFlash cell with reduced stress on the tunnel oxide. Based on L. Shao *et al.* (SMIC), CSTIC, May 2017 [62].

compatible. Peak power operation was about 4 mW and good data retention was shown for a fully integrated and scalable embedded NVM without adding process complexity or mask layers.

The HfO_2-based HKMG devices were developed as nonvolatile memory elements for SoC applications in HKMG CMOS technologies. Intrinsic self-heating in enhanced charge trapping HKMG devices was used to achieve large and stable device threshold voltage shifts suitable for memory applications. Multiple time programming was shown to be possible.

A new embedded Flash memory with a reversed drain-source cell was discussed in May of 2017 by SMIC [62]. The mechanism was intended to improve the read stress and the standby with power-on after more than 20 program/erase cycles, which causes conventional embedded Flash memory to fail read "0". A schematic cross-section of the cell biased for the read operation is shown in Figure 3.48 for (a) the conventional eFlash cell and (b) the new reversed eFlash cell.

The stress-induced leakage current (SILC) mechanism of an eFlash with a conventional e-Flash cell is caused by the high electric field stress on the tunneling oxide of the device at the source side during read, which causes traps formed during programming and erase to leak current. In the reversed eFlash cell, the low voltage on the drain during read reduces this leakage. The result is less SILC leakage for a multiple time programmable device.

3.7 Stacked Gate Double Poly Flash

Embedded Flash memory was made on a 55 nm deep depleted channel (DDC) platform with ultralow power of <0.9 V by Fujitsu in December of 2013 [63]. While the DDC process has a reduced thermal budget, the single bit charge loss of Flash after cycling can be optimized so it is comparable to baseline eFlash. It was verified that improved variability and low power digital performance of the DDC process was maintained in the embedded Flash flow. The DDC transistor has also been found to achieve ultralow voltage SRAM operation and to improve digital and analog performance by reduction in random dopant fluctuation.

Having an embedded NVM available in this process was expected to enable a wide range of ultralow power applications such as smart cards, energy harvesting, and sensor MCU for Internet of Things applications. A cross-section of the DDC transistor is shown in Figure 3.49. The DDC transistor has an undoped epitaxial layer and a V_t setting layer. The DDC process uses a reduced thermal budget to suppress impurity diffusion into the undoped epitaxial channel layer.

The memory cell layout of the 1 T NOR FLOTOX Flash macro used contains 512 word-lines and 1024 bit-lines, resulting in 512K cells. The memory transistor is made by adding processing to a DDC transistor. Steps in the DDC process are modified for the embedded Flash. The undoped epitaxial layer is thickened to compensate for silicon loss during the additional oxidation steps for Flash tunnel oxide and high voltage gate oxide. Temperatures for Flash-related steps are lowered significantly to suppress impurity diffusion from the screen layer to the undoped epi channel layer of DDC transistors. Channel and drain for Flash and HV transistors are modified.

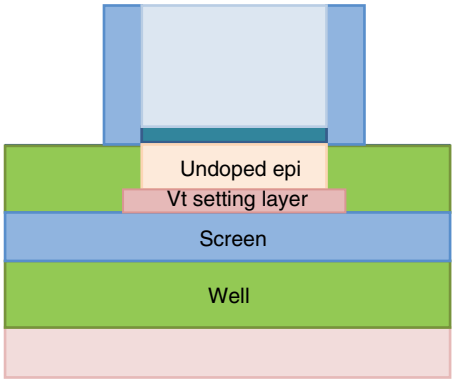


Figure 3.49 Schematic cross-section of DDC transistor showing undoped epitaxial layer and V_t setting layer Based on M. Hori *et al.* (Fujitsu), IEDM, December 2013 [63].

An embedded stacked floating gate Flash cell embedded in 65 nm logic technology was discussed in September of 2012 by Lattice Semiconductor and Fujitsu [64]. The logic transistor characteristics were optimized for low standby current for use in consumer applications. This cell was scaled from an older 90 nm stacked Flash cell technology. The gate length and active width are both 200 nm. The cell size is 0.253 μm^2, which represents a 42.6% cell area reduction compared with the 90 nm cell. It also showed an 8.6% reduction in programming voltage compared with the previous 90 nm cell. The gate-induceddrain leakage (GIDL) current was reduced by about an order of magnitude. Typical operating conditions of the Flash cell included an 8.5 V program and a 17 V erase composed of –8.5 V on the gate and 8.5 V on the well. Read was at V_{drain} of 0.45 V with V_{source} grounded and gate at 4.65 V.

Reliability requirements of the automotive industry for electronic components are challenging. In April of 2013, the University de Studi di Ferrara and Infineon discussed automotive reliability requirements for Flash memories [65]. Erratic bits are considered a major reliability issue conventionally handled by repair strategies ranging from static redundancy to dynamic correction codes. Both repair strategies need to be tailored to the product.

A signature classification methodology for erratic bits was suggested that helps choose the best repair strategy for each bit. This tends to reduce using unnecessary correction techniques and increase the chip error probability by improving the accuracy in modeling. The classification methodology is to be performed either during a wafer sort or in confidence level testing for product monitoring. It is based on the analysis of the features constituting the bit signature. This classification permits speculation of the most efficient repair strategy for each erratic signature class. It is expected that this will reduce the redundancy usage. Models were developed as a validation of the proposed approach.

The challenges of integrating a conventional nonvolatile double polysilicon NVM memory array in a base-line CMOS process were discussed by On Semiconductor in May of 2015 [66]. Process modules of ONO and floating poly were studied. The addition of new implant steps such as deep wells and high voltage LDD were discussed along with new critical etch steps such as removal of sacrificial oxide in the NVM, high voltage and low voltage ONO etch, and stack poly etch. New layers also need to be added, such as tunnel oxide, floating gate poly, and ONO. A schematic process flow for an embedded NVM integration in base-line CMOS is shown in Figure 3.50.

The basic criteria for process flow integration is that base-line process steps should not change and that any added process steps have minimal impact. While the NVM technology with floating poly had many integration issues that were addressed, the constraints of the ONO layer were found to be particularly challenging. High voltage transistors were also carefully considered.

An embedded Flash cell made on 300 mm wafers in 130 nm bipolar, CMOS, and DMOS (BCD) technology (5 to 80 V) was discussed in June of 2014 by Rohm and Genusion [67]. The eFlash cell had data retention of greater than 20 years under 150 °C testing after 100k erase/write cycles. BCD (bipolar, CMOS, DMOS) is a key technology for power ICs. BCD devices are widely used in automotive and industrial applications to improve performance and intelligence. The BCD platform permits the flexibility to choose various types and performance of devices. The reliability of BCD devices under high temperature is an important factor.

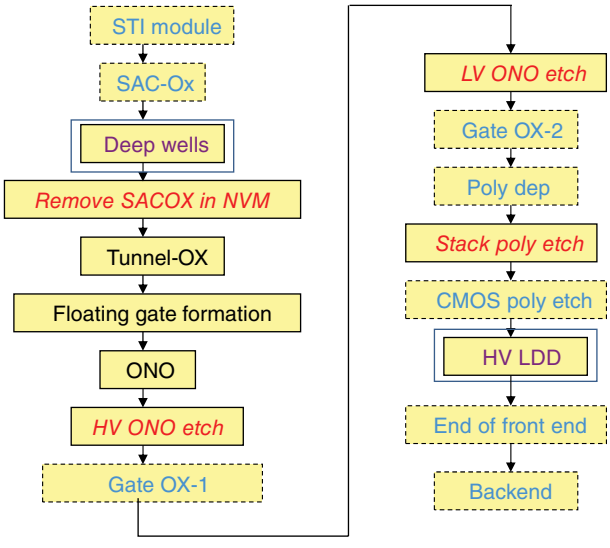


Figure 3.50 Schematic process flow for an embedded NVM in base-line CMOS. Bullets in dashed boxes are part of the conventional CMOS process. Bullets in double outlined boxes are added depositions and italicized bullets are added etches. Based on M. Agam *et al.* (On Semi.), ASMC, May 2015 [66].

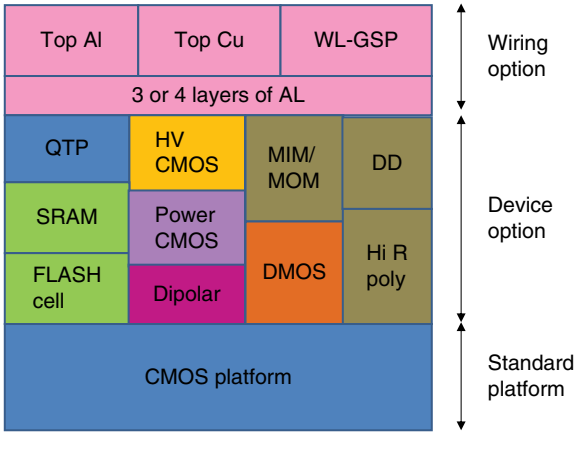


Figure 3.51 A block diagram of the 130 nm platform in bipolar, CMOS, and DMOS (BCD) is shown. Based on K. Iwamoto *et al.* (Rohm, Genusion), International Symposium on Power Semiconductor Devices and ICs, June 15, 2014 [67].

The 130 nm Rohm BCD platform is shown in Figure 3.51. The 1.5/5.5 V CMOS and IP core are standard to the platform. Device options include: DMOS, HV CMOS, power CMOS, bipolar, MIM/MOM capacitors, poly-Si resistor, Flash cell, DTI, SRAM, and OTP. There are three modules of wiring options that can be combined with the standard platform as desired.

An embedded NOR Flash memory cell was developed for the 130 nm BCD platform. The cell used a back-bias assisted band-to-band tunneling mechanism. The unit cell size was 0.267 μm^2.

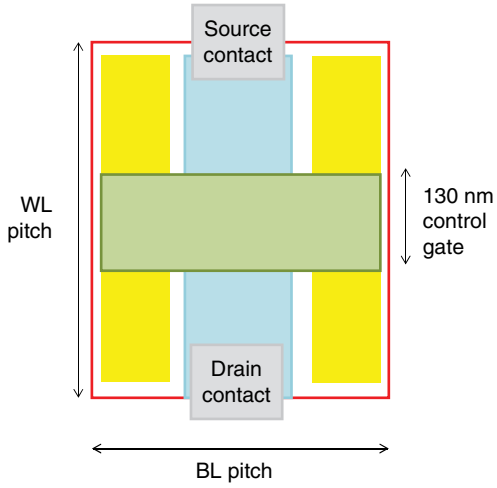


Figure 3.52 Top-down schematic view of layout of eNOR Flash cell. Based on K. Iwamoto *et al.* (Rohm, Genusion), International Symposium on Power Semiconductor Devices and ICs, June 15, 2014 [67].

The B4 Flash cell adopts the standard floating gate type structure, which fits well with both the CMOS and BCD processes. A top-down schematic view of the layout of the embedded NOR Flash memory cell is shown in Figure 3.52.

The two companies jointly developed the 90 nm B4-Flash NOR memory and have it currently in production. The drain profile is optimized to maintain its intrinsic characteristics even in the BCD process. The V_{th} distribution at 150 °C after 100 K P/E cycles was demonstrated and also a 1.5Kbit cell array, which was extrapolated to the 288Kbit production array. The V_{th} window had a good margin after 100 K P/E cycles. The chip showed good reliability under high temperature ambient. The eFlash memory cell was successfully integrated into the 130 nm BCD platform.

3.8 Charge Trapping eFlash

3.8.1 Overview of Early Embedded Charge Trapping Memory

Embedded charge trapping memories are in production at several manufacturers and the technology is also widely expected to be used in standalone vertical memories. Production examples are discussed in the first section followed by technology development in the second.

Cypress, Renesas, and Spansion, now owned by Cypress, are in production with charge trapping NV memory. Cypress is working with Foundries Shanghai Huali and UMC to produce its 55 nm SONOS technology for smart cards and IoT. UMC is already in production with the Cypress 65 nm charge trapping technology. Renesas has produced embedded split gate MONOS Flash macros for MCU for various microcontroller applications since 2004 and are currently in production in 40 nm technology. Spansion had charge trapping Flash memory devices in production from the 110 nm to 65 nm technology nodes in December of 2014.

In June of 2003, Halo LSI discussed a 180 nm twin MONOS NOR-type Flash memory that used a simple process making it suitable for embedded Flash applications [68]. A 128KB device with a 4 ns access time was discussed. The twin MONOS technology had low voltage operation and a simple process. The maximum high voltage requirement

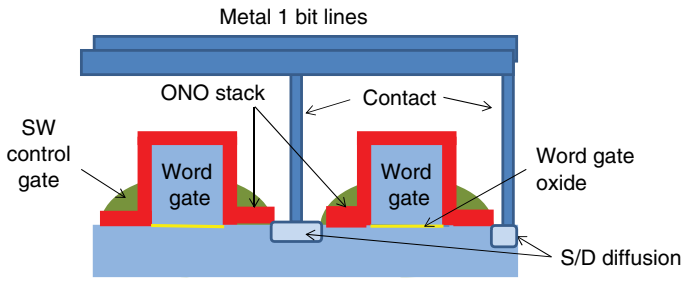


Figure 3.53 Schematic cross-section of twin MONOS cells. Based on T. Ogura *et al.* (Halo LSI), VLSI Circuits Symposium, June 2003 [68].

was limited to 4.5 V. Only three extra masking steps were required even though the cell size was around 5–7 F2, which permitted larger memory macro sizes to be integrated on a logic chip. An implementation for an embedded program and data storage required access times between 10 and 20 ns. A schematic of the twin MONOS cells cross-section is shown in Figure 3.53.

The cell consisted of a word gate, two side wall control gates, and two diffusion regions. A thin gate oxide was used beneath the word gate. An ONO film was used beneath the side wall control gates. No charge pump was required during read. For the memory array, the read voltage was 1.8 V with I_{on} >60 μA/μm and I_{off} of 3.5 μA/μm. Read access time was 20 ns. Program was at 5.5 V on the control gate and program current was <2 μA/bit for 20 μs. Erase by HHI was 4.5 V on the BL with an erase current of 2 nA/bit.

A 512KB MONOS Flash memory module made in a 180 nm CMOS logic process and intended for embedding in a microcontroller was described by Hitachi in June of 2003 [69]. The read path of the module was composed of the same low voltage transistors that were used in the CPU core. This provided a compact layout of the peripheral circuits. Read access was 34 MHz. For a 64KB block, program time was less than 4 ms and erase time was less than 11 ms. The module area was 5.4 mm^2. I_{ds} was 20 μA at V_{cg} = 1.5 V.

A local SONOS structure for an embedded NVM cell in 130 nm standard CMOS logic was discussed by Samsung in June of 2003 [70]. A localized silicon nitride storage layer gives a complete erase, low program current and high on-cell current from a low threshold voltage. The split gate local SONOS cell is shown in Figure 3.54, where (a) illustrates a top-down view of the layout showing the localized ONO structure and (b) illustrates a schematic cross-section of the cell with the halo implant junction engineering shown.

This 0.276 μm^2 local SONOS NVM cell had a 20 μs program and 2 ms erase speed using 5.5 V bias. Three added masks were required to make the SONOS cell and four more were required for the high voltage operation. This is four masks less than an embedded floating gate NVM. The local SONOS showed more complete erase characteristics than a longer SONOS layer. A halo implantation was used to suppress the drain depletion region punch-through and improve program/erase efficiency. The localized ONO and junction engineering permitted low voltage operation, where 4.5 V on the word-line resulted in sufficient threshold voltage shift. Program current was as low as 10 μA/cell. Program speed was 20 μs and erase speed was 2 ms.

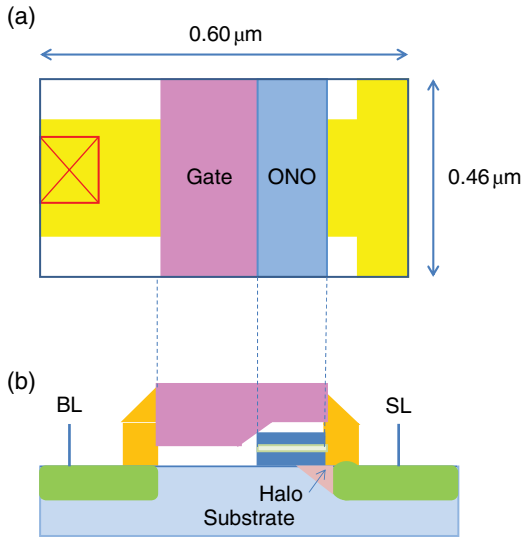


Figure 3.54 Split gate local SONOS cell in conventional CMOS logic: (a) top-down view showing localized ONO layer and (b) schematic cross-section showing Halo implant junction engineering on the source side. Based on J.H. Kim *et al.* (Samsung), VLSI Technology Symposium, June 2003 [70].

An 8Mbit macro of split gate localized SONOS charge trapping memory in a 130 nm conventional logic process with five-level copper metallization was shown by Samsung in June of 2004 [71]. Cell size was 0.276 μm^2. Read speed was up to 60 MHZ at $V_{cc} = 0.9$ V and 85 °C. Current consumption was less than 5 mA at V_{cc} −1.4 V. The cell was programmed by channel hot electron injection (CHEI) and erased by hot hole injection (HHI). The localized ONO storage technology of this cell was found to have better endurance and higher immunity to read disturb than conventional stacked SONOS due to its limited ONO layer and asymmetrical cell structure. After 100 k P/E cycles the sensing current margin is maintained due to its localized ONO layer.

This 256 K × 32 SONOS embedded Flash EEPROM macro in a 130 nm CMOS logic compatible process was discussed further by Samsung in June of 2004 [72]. Cell size was 0.276 μm^2. Program time was 20 μs and erase time was 20 ms. A new address transition detection (ATD) assisted current sense amplifier was used for the low V_{cc} application, where V_{cc} was −0.7 to 1.4 V. The read operation was done at 66 MHz with a low active current of 5 mA being typical.

In June of 2006, Genusion described a p-channel SONOS Flash memory cell technology with good scalability and high programming efficiency for NOR architectures [73]. The cell used a back bias assisted band-to-band tunneling induced hot-electron (B4-HE) injection process. If a moderate back bias is applied to the cell during programming, the bit-line voltage can be reduced below the supply voltage, which is 1.8 V. The gate length is 60 nm. Operation of the 50 nm B4-Flash cell was confirmed. The programming method provided low drain voltage programmability below 1.8 V and high program efficiency.

The back bias assisted BTBT-HE generation consists of two steps. The first is BTBT generation and the second is electron acceleration. A schematic cross-section showing the location of the BTBT generation and electron acceleration is shown in Figure 3.55.

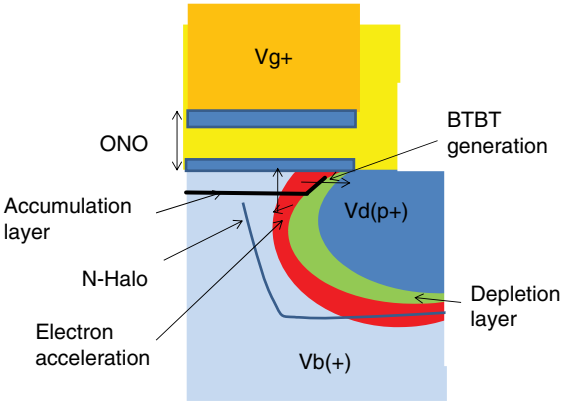


Figure 3.55 Illustration of location of BTBT generation and electron acceleration in SONOS cell. Based on S. Shukuri *et al.* (Genusion, Tohoku University), VLSI Technology Symposium, June 2006 [73].

Program has the V_g at 12 V with $V_b = 4.2$ V and erase has V_g at −12 V. Read is $V_g = -2$ V with $V_d = 1.8$ V. In embedded applications, the high program efficiency may help achieve high area efficiency due to reduction of the on-chip power supply capability. Erase time was about 100 ms by FN tunneling hole injection. Data retention of 10 k cycled cells at the maximum temperature is 10 years at around 125 °C.

3.8.2 Embedded 40 nm Charge Trapping (MONOS) Flash MCU

In May of 2012 an embedded Flash-based MCU platform was announced based on the Renesas 40 nm MONOS eFlash technology, which was expected to be licensable to other semiconductor suppliers [74]. The platform integrated the 40 nm MONOS eFlash technology and TSMC CMOS logic and analog IP. Details of the arrangement were still being discussed. The Renesas 40 nm Flash had 20 year data retention and could be read up to a 170 °C junction temperature. The code Flash supported read speed of 120 MHZ with 20 year data retention after 125 000 program/erase cycles.

Flash MCUs with embedded Flash memory are considered essential in automotive applications for complicated real-time controls for higher fuel-efficiency, better safety and increased connectivity. High density, fast random read access is needed from the eFlash. The requirements for eFlash in automotive products are secured operations and data reliability in particular reliable program/erase cycling at high temperature.

In January of 2014, Renesas discussed its 40 nm embedded split gate (SG) MONOS Flash macros for the automotive market [75]. These devices had 160 MHz random access for code. They had endurance over 10 M cycles for data at a junction temperature of 170 °C. A sense amplifier with digital offset cancellation had a fast read operation greater than 160 MHZ. Program time was decreased by an adaptable program current control method and an intelligent erase method decreased the erase time. 3D stacked capacitors provided an area-efficient charge pump. Both a 2 MB code macro and a 64KB data macro were fabricated in a 40 nm eFlash process. The code macro had 160 MHz random read at Tj = 170 °C. A 5.1 GB/s read throughput was reached by a simultaneous 256 bit read-out from two code macros. The data macro had P/E endurance over 10 million cycles at $T_j = 170$ °C without any software-assisted techniques.

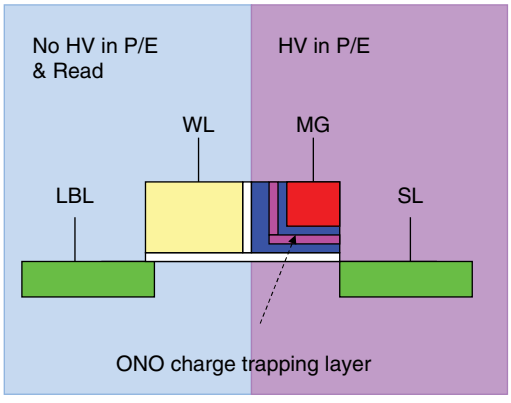
The trend of eFlash memories in high-end Flash MCUs is to have both a code macro and a data macro. The code macro is used to store the boot/application codes. These codes require large capacity and fast read at wide temperature ranges. The data macro, however, is a substitute for external EEPROM to reduce system cost. Its contents are frequently rewritten after shipment. The data macro must meet a high program/erase endurance over a wide temperature range.

Flash MCUs with embedded Flash memory are used in automotive applications for complicated real-time controls for higher fuel efficiency, better safety and increased connectivity. Both high density and fast random read access are needed from the eFlash. The requirements for eFlash in automotive products are secured operations and data reliability including reliable program/erase cycling at high temperature.

In February of 2013, Renesas discussed a 40 nm eFlash macro for automotive applications [76]. It had three key features: a 40 nm scaled split gate (SG)-MONOS cell, an over 160 MHz random read access with appropriate sense amplifier, and circuit techniques for fast reliable P/E at a junction temperature of 170 °C. The architecture of the split gate MONOS cell is shown in the schematic cross-section of the SG-MONOS device in Figure 3.56 (a). Operational voltages are shown in Figure 3.56 (b). There is no high voltage applied during program or erase on the WL and BL side of the circuit. The high voltage during P/E is applied only to the separate memory gate (MG) and SL nodes. The WL and BL side of the circuit are used during read and only CMOS logic voltages (V_{dd}) are applied to the WL and BL side.

The SG-MONOS cell uses a split gate cell with charge-trapping storage in a nitride film. The cell has two advantages over a conventional one transistor cell. The word-line voltage in the read operation is at the logic power supply voltage because there is no column current leakage with the thin split gate oxide. Fast read paths can be built with fast low voltage transistors because high voltage for P/E are not applied to them. Fast programming with small cell current around 1 μA can be realized because of the source-side injection capability. For the erase operation, hot hole injection by band-to-band tunneling is used, which has a high erase speed that enables a short erase time with a small erase unit.

(a)



(b)

OPERATIONAL VOLTAGES

	WL	MG	SL	LBL
Read	V_{dd}	0	0	V_{dd}
Program	–1.0	+HV	+HV	V_{dd}
Erase	0	–HV	–HV	V_{dd}

Figure 3.56 SG-MONOS eFlash cell: (a) cell architecture and (b) operational voltage. Based on T. Kono *et al.* (Renesas), IEEE ISSCC, February 2013 [76].

For P/E operations at 170 °C, two circuit techniques were used [76]. An adaptable program current control method was used with two data latches per main BL. One stores the program data from the Flash controller and the other stores the verification result for the SA. These data and output from the fail bit counter control the program current. An intelligent erase method was used. A simulation with a 40 nm SG-MONOS cell showed a 60% decrease in total erase pulse time. By using poly-insulator-poly (PIP) capacitors and a metal-oxide-metal (MOM) capacitor stacked over the 5 V MOS capacitors, a unit capacitor with small area and high breakdown voltage was confirmed and the area of the VSP charge pump decreased by 44%.

In the eFlash macro with the SG-MONOS cells, the eFlash was used in both the data memory and the code macro. These circuit techniques were used to develop a 2 MB code macro and a 64 KB data macro in a 40 nm eFlash process. The read frequency of the code macro was confirmed at >160 MHz at 170 °C. By reading 256 bits from two code macros simultaneously, the maximum read throughput reached 5.1GB/s. The data macro achieved the P/E endurance over 10 M cycles at 170 °C without wear-leveling.

In April of 2015, Renesas and Hitachi discussed the data retention mechanism of an embedded split gate MONOS Flash memory [77]. A data retention model was developed based on the thermionic emissions and a simple estimation method with a long lifetime and wide temperature range. A process flow for the embedded MONOS Flash was developed, which included: well/ channel implant, CG-oxidation, CG-poly etch, MG channel implant, $SiO_2/SiN/SiO_2$ deposition, MG-poly deposition, Mg- poly etch, removal of $SiO_2/SiN/SiO_2$, extension and pocket implant, side wall formation, and source/drain diffusion implant.

The retention results for the embedded SG-MONOS Flash were consistent with previous results within a temperature range of 150–450 °C. The MONOS memory with split gate had a 20 year lifetime at more than 200 °C. This SG-MONOS memory is for use in an MCU in 150, 90, and 40 nm process generations. It withstands operating temperatures higher than 170 °C and has better than 20 year data retention in automotive MCU. Operating condition voltages are shown in the schematic cross-sections of the SG-MONOS cell in Figure 3.57 for (a) program, (b) erase, and (c) read.

A data retention model and estimation method based on thermionic emissions was developed. This method can measure the trap density of nitride rapidly. The data retention lifetime could be estimated over a time and a wide temperature range. The SG-MONOS device was found capable of a 20 year lifetime at more than 200 °C.

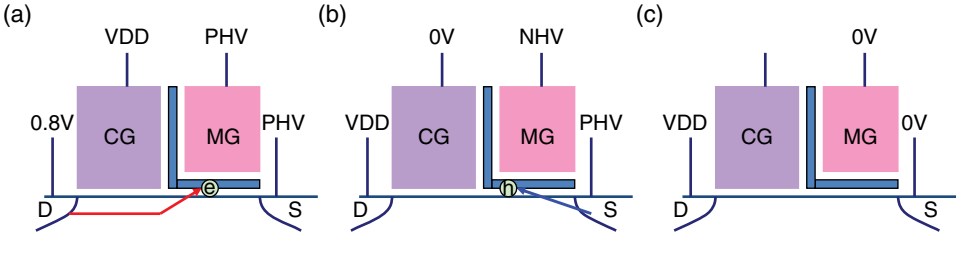


Figure 3.57 Schematic cross-section of the SG-MONOS cell with operating condition voltages for (a) program, (b) erase, and (c) read. Based on Y. Kawashima *et al.* (Renesas, Hitachi), IRPS, April 2015 [77].

3.8.3 Embedded 28 nm Charge Trapping (MONOS) Flash MCU

The development of a 28 nm eFlash MONOS technology for MCU was announced by Renesas in February of 2014 [78]. A primary target application for the technology was automotive control systems. The earlier 40 nm production technology supported up to 8 MB of on-chip Flash memory for MCU while the 28 nm technology was expected to support 16 MB of macro density. The 28 nm prototype chip could support clock frequency of 160 MHZ, data retention time of 20 years, and 250 000 cycles of endurance.

The 28 nm embedded split gate SG-MONOS Flash memory macro for automotive applications was also discussed by Renesas in April of 2014 [77]. This device had a 200 MHz read and 2.0 MB/s write throughput at a junction temperature (T_j) of 170 °C. The split gate-MONOS cell had high voltage only on the memory gate for programming and erase with only logic voltage on the select gate, as shown in the schematic cross-section of the cell in Figure 3.58.

The transistors and interconnections used in embedded Flash systems tend to effect reliability in the sub 40 nm technology generations. Thin transistor oxide films can degrade the time-dependent dielectric breakdown (TDDB) lifetime. The SG-MONOS device had a temperature adjusted overdrive word-line voltage control, which permitted both 200 MHZ random access and greater than 10 times longer TDDB lifetime for the word-line drivers. It also had a high voltage control technique to relax electrical stress on memory cells and peripheral devices by using temperature adaptive step pulse erase control. Write throughput of 2.0 MB/s was achieved by source-side injection programming with a negative back-bias, which permitted a 63% reduction of program pulse time. RF noise was reduced by 19 dB using a spread spectrum phase shifted clock generation technique. The eFlash architecture included two code Flash macros and a data Flash macro, both made up of SG-MONOS cells each using a split gate charge trapping cell. The memory cell size was 0.053 μm^2.

The 28 nm split gate embedded MONOS was discussed by Renesas in June of 2015 for use in MCUs for vehicle control systems such as power train, chassis, and brake control [79]. Powertrain applications, in particular, require a large memory capacity as well as having a 170 °C temperature rating. The application requires speed and split gate cells tend to be fast and have a low power read due to the low voltage swing of the word-line. The charge trapping structure leads to high data reliability due to robustness against

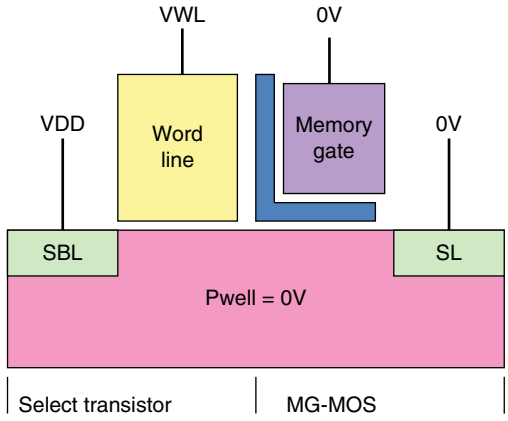


Figure 3.58 Split gate MONOS cell showing high voltage only on memory gate for P/E. Based on Y. Kawashima *et al.* (Renesas, Hitachi), IRPS, April 2015 [77].

point defects and the low cell profile assures better compatibility with the CMOS logic process than a floating gate. The 28 nm SG-MONOS macro had 200 MHz random read for code Flash and over 1 M program and erase cycles for data Flash. An 80% smaller sleep current was achieved at a junction temperature of 170 °C by using a floating bit-line and raised V_{SS} with a lowered V_{DD} for the memory cell.

In January of 2016, Renesas discussed a 28 nm embedded split gate MONOS Flash memory macro intended for automotive applications [80,81]. This device had a 6.4GB read throughput using a 200 MHz no-wait read operation and a 2.0 MB/s write throughput at a junction temperature (T_j) of 170 °C. These 28 nm eMONOS Flash macros were developed to increase memory capacity embedded in scaled microcontroller units and also to improve performance over a range of junction temperatures from –40 to +170 °C, which are the automotive under-the-hood specifications.

An issue of scaling the process is the resulting degradation of reliability characteristics. In this study, a temperature adjusted word-line overdrive method was used to improve random read access frequency by 15%, as well as have 6.4GB/s of read throughput for 200 MHz no-wait random access for code Flash macros. TDDB lifetime of WL drivers was increased by 10 times. Temperature adaptive step pulse erase control (TASPEC) was used to improve the TDDB lifetime of dielectric films between metal interconnect layers by 3 times. TASPEC was used for a data Flash macro with 1×10^6 rewrite cycles. Programming by source side injection (SSI) using negative back-bias voltage achieved 63% reduction of program pulse time and resulted in 2.0 MB/s write throughput of code Flash macros. A spread spectrum clock generation and a clock phase shift technique were developed for charge pump clock generation to suppress EMI noise due to high write throughput of the code Flash macros and reduce peak power of EMI noise by 19 dB.

An SG-MONOS cell was used that combines a split gate cell and a discrete charge trapping cell. Split gate cells have a fast random read access speed since the read path can consist only of logic CMOS transistors. Fast program speed is due to the efficiency of source-side injection (SSI). Erase operation uses hot hole injection by band-to-band tunneling (BTBT) and is faster than Fowler-Nordheim (FN) tunneling. The schematic cross-section of the MG-SONOS cell in Figure 3.59 illustrates SSI programming and BTBT erase. Figure 3.59 (a) illustrates SSI programming with the memory gate (MG) at

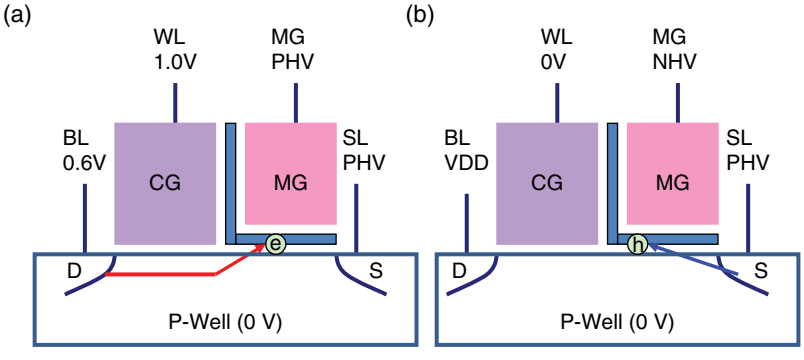


Figure 3.59 Cross-section of MG-SONOS cell showing (a) SSI program and (b) BTBT erase. Based on Y. Taito *et al.* (Renesas), *IEEE Journal of Solid-State Circuits*, January 2016 [81].

positive high voltage (PHV) and the WL and BL at appropriate logic levels. Figure 3.59 (b) illustrates band-to-band tunneling (BTBT) erase with negative high voltage (NHV) on the memory gate while the BL and WL remain at appropriate logic levels.

The memory cell height of the floating gate type of cell is twice the height of logic CMOS transistors in the 28 nm process whereas the height of the SG-MONOS cell is closer to that of the CMOS logic transistors. A high-*k* metal gate process is used, which makes it is necessary for the Flash memory cells to be polished to the same height as the logic CMOS transistors by chemical mechanical polishing (CMP). This is easier in the lower height MONOS process than in the double poly floating gate process.

In the 28 nm eFlash SG-MONOS system, the memory block contains two types of eFLash macro. Code Flash is used to store control program code and requires fast random read access and large memory capacity. The code Flash does random read at a frequency of 200 MHz. Program speed is 2.0 MB/s and erase speed in 0.91 MB/s. Its maximum capacity is 32 MB. The P/E endurance requirement is 10 k cycles. The performance of the code Flash random read access frequency is increased by 15% using a temperature-adjusted WL overdrive scheme. Write pulse time is reduced by 63% using an SSI program and negative back-bias voltage.

The data Flash is used for data storage and requires good rewrite endurance and small rewrite unit size. Rewrite endurance of the data Flash macro is 1×10^6 cycles. An SSCG and a clock phase shift technique are used to solve the EMI noise caused by charge pump operation during field programming. The rewrite unit size is 64KB while the maximum capacity is 512MB. The cell size of the 28 nm SG-MONOS is 0.053 μm^2, core power supply is 1.1 V, and IO supply range is 2.7 to 5.5 V. The operating temperature ranged from −40 to +170 °C.

3.8.4 Embedded Application-Specific 1T-MONOS Flash Macro

The computerization of automotive control has expanded the application range of MCUs. In February of 2016, Renesas discussed a 90 nm one transistor (1T) MONOS eFlash macro for use in high end automotive applications such as a high end engine control unit (ECU) [82]. A schematic cross-section of the 1T cell is shown in Figure 3.60.

This memory cell used Fowler-Nordheim tunneling for both program and erase operations. The cell had a read-disturb-free array architecture to ensure automotive grade reliability. It achieved over 100M cycles of rewrite endurance at 175 °C junction temperature (T_j) by using an adaptable slope pulse control technique. As a result total power consumption was reduced to 0.07 mJ/8KB using a 98 μA program erase current. A low power system control method using the 1 T-MONOS device resulted in a 99%

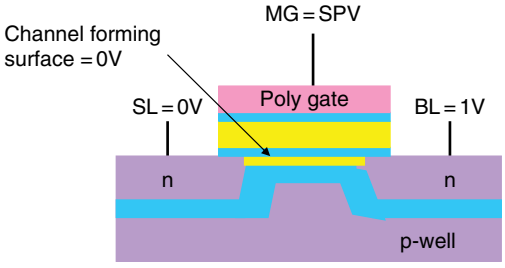


Figure 3.60 Schematic cross-section of 90 nm 1T MONOS eFlash Cell. Based on H. Mitani *et al.* (Renesas), ISSCC, February 2016 [82].

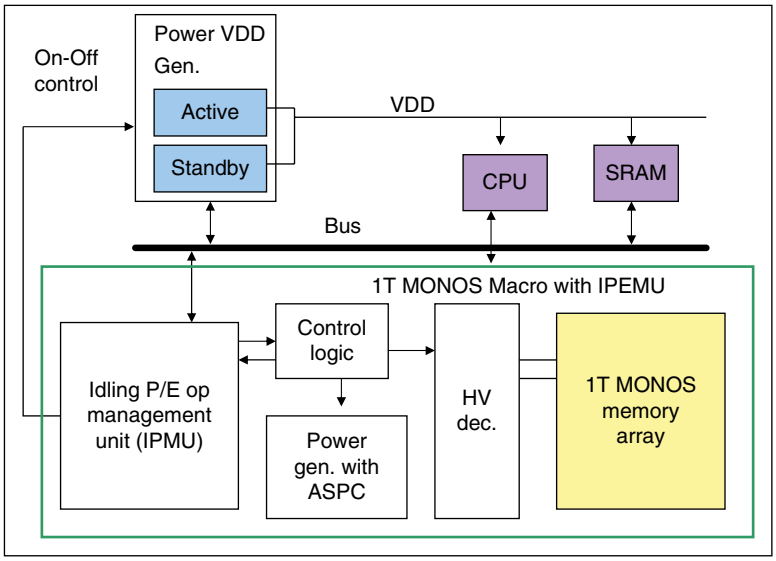


Figure 3.61 Block diagram of low power Flash MCU with 1T-MONOS aacro. Based on H. Mitani *et al.* (Renesas), ISSCC, February 2016 [82].

power reduction. A block diagram of a Flash MCU with the 1 T-MONOS integrated for a low power system is shown in Figure 3.61.

The 1T MONOS memory device was made in a three-layer aluminum process. Memory capacity of the macro was 128KB and it used a 3.3 V power supply. The operating temperature range was −40 to +175 °C. Read I/O has 32 bits + <run on text > 7 bits. The random read frequency was greater than 50 MHz. Program time was 3 ms per 128 B and erase time was 5 ms per 2KB. Program and erase current was 98 μA.

In February of 2016, Renesas announced the development of a 90 nm one-transistor (1 T) MONOS [83], which could be integrated with various processes including CMOS, bipolar CMOS, and DMOS. It provided a high program/erase endurance of over 100 million cycles with a junction temperature (T_j) of 175 °C and a low rewrite energy consumption of 0.07 mJ/8KB. The 128KB Flash memory prototypes were shown with a low rewrite current of 98 μA. It is expected that this technology can be added to automotive analog devices to improve performance and reliability.

3.8.5 FinFET SG-MONOS

For a future automotive MCU with embedded memory, in December of 2016, Renesas and Hitachi discussed a FinFET split gate MONOS embedded Flash memory in 16/14 nm technology node [84]. The subthreshold characteristics and small threshold voltage variability due to the FIN were explored. The FIN top-corner effects were found to be well suppressed by incremental step pulse programming for source side injection. Data retention at 150 °C after 250k program/erase cycles was confirmed for use in advanced automotive system applications.

In December of 2016 Renesas discussed embedded Flash technology for automotive applications such as higher fuel efficient engines and advanced driver assistance

systems (ADAS) [85]. The attributes of a charge trapping split gate type of data storage such as SG-MONOS were given. Split gate cells permit fast speed and low power read due to the low voltage swing of the word line. The charge trapping structure is robust against point defects, which could cause a floating gate to lose all stored charge. Charge trapping technologies also have a lower cell height, making them more compatible with the CMOS logic process. A 28 nm SG-MONOS is targeting over 160 MHz of random read. 1M program/erase cycles have been found for data Flash at a junction temperature of 170 °C. Temperature adaptive step pulse erase control improves the TDDB lifetime of dielectric films between metal interconnect layers by three times to improve endurance characteristics of SG-MONOS.

A 1.5 transistor SG-MONOS cell using a Fin structure was discussed along with its advantages over a one transistor cell. Bulk FinFET SG-MONOS Flash was shown. Cell characteristics using source-side injection (SSI) programming and erase using band-to-band tunneling (BTBT-HHI) were compared to that of the planar structure. Incremental step pulse SSI techniques were confirmed for suppression of Fin top-corner effects.

A schematic process flow for the integration of the SG-MONOS device into the FinFET is shown in Figure 3.62. The Fin is formed first on the bulk silicon wafer by patterning the active area and recessing the STI oxide. This is followed by the SG-MONOS process, which follows nearly the same sequence as conventional planar FG-MONOS. The FinFET SG-MONOS is very compatible with the process of the logic transistors.

The control gate (CG) and memory gate (MG) lengths of the FinFET SG-MONOS are equivalent to the planar device. Due to the fully depleted state, the subthreshold swing of the FinFET SG-MONOS is smaller than that of the planar SG-MONOS and the threshold voltage distribution is improved. The on-current of the FinFET SG-MONOS is comparable to the planar SG-MONOS even though the footprint of the Fin is 20% of the planar. The FinFET SG-MONOS shows a steeper subthreshold swing than the planar SG-MONOS whether the devices are programmed or erased. This improved the current controllability results in a larger current window for the FinFET SG-MONOS. Program and erase time are 100 μs. A 3D schematic showing the integration of the FinFET SG-MONOS and the logic FinFET CMOS is shown Figure 3.63. The two structures are the same height.

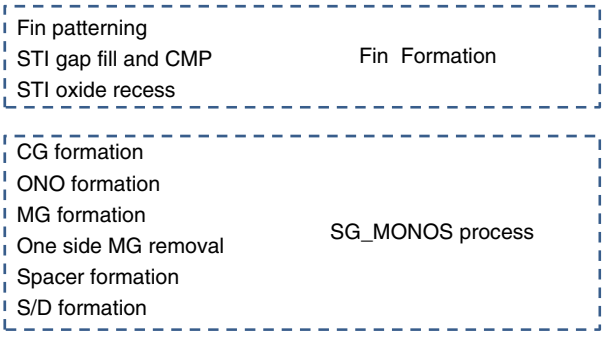


Figure 3.62 Schematic process flow of formation of the fin of the FinFET and the SG-MONOS. Based on S. Tsuda *et al.* (Renesas Electronics, Hitachi), IEEE IEDM, December 2016 [84].

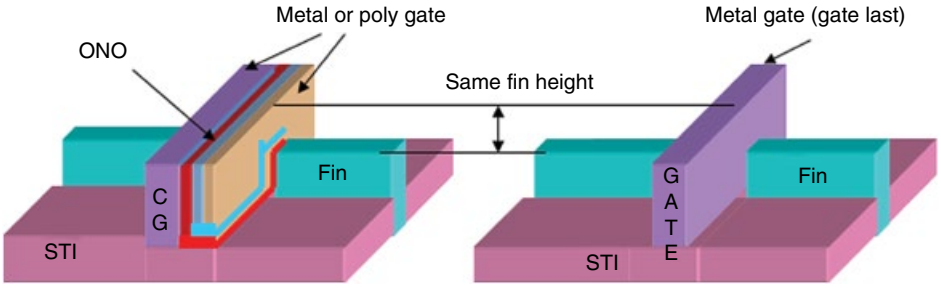


Figure 3.63 Schematic of integration of FinFET SG-MONOS and logic FinFET CMOS. Based on S. Tsuda *et al.* (Renesas Electronics, Hitachi), IEEE IEDM, December 2016 [84] (permission of IEEE).

The improved subthreshold characteristics and small threshold voltage variability due to the Fin structure are required for future low power operation. The cell current window is wider than the conventional planar SG-MONOS. Incremental step pulse programming for source side injection was effective in suppressing the Fin top corner effect. Data retention at 150 °C after 250k program/erase cycles indicates sufficient reliability for advanced automotive applications. The Fin technology permits scaling to the 16/14 nm node. The low profile of the SG-MONOS permits integration with FinFET logic CMOS with a high-*k* metal gate made by a gate last process.

3.8.6 Embedded Charge Trapping (SONOS) NOR Flash

In May of 2013, Spansion, now Cypress, discussed several types of charge trap Flash memory including embedded charge trapping (CT) NOR Flash [86]. It was expected that planar cell structures would permit continued scaling while new 3D charge-trap Flash memory technology matured. For SoC products with fast read access time, Spansion embedded CT Flash was being integrated with an advanced logic process.

To address the embedded CT Flash market, a fast, scalable embedded charge trap Flash was developed. The Spansion "MirrorBit" cell was used as the foundation for the embedded CT Flash since the "MirrorBit" cell had been proven in volume production. A low voltage select gate is paired to the memory gate of the embedded CT cell, as shown in Figure 3.64. Each cell is composed of a charge trap memory gate and a low voltage select gate.

This cell configuration and the optimized array architecture enabled fast read and enhanced program and erase. Random access time was about 5–10 ns, making the

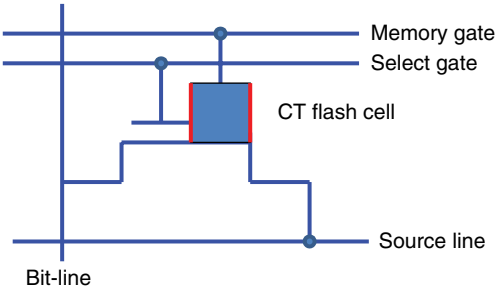


Figure 3.64 Schematic of embedded charge-trapping "Mirror-Bit" Flash cell. Based on S. Tehrani *et al.* (Spansion), IMW, May 2013 [86].

embedded CT Flash a possible solution for high end MCU and SoC. For this purpose the device was integrated with an advanced logic process at the 40 nm node.

In March of 2013, Spansion and UMC announced joint development of a 40 nm embedded charge trap NOR Flash memory technology, the Spansion embedded CT technology [87]. As part of the agreement UMC was licensed to make products based on this technology for Spansion. The technology was scalable beyond 40 nm and could be integrated into a high-*k* process. Target applications were SoC for industrial, automotive, and consumer applications.

A low voltage, low cost SONOS memory technology was announced in July of 2014 by UMC and Cypress, who stated that UMC had licensed the Cypress 55 nm eFlash memory IP for use in future Internet of Things (IoT) and wearable applications [88]. The Cypress SONOS embedded NV memory technology adds only 3–4 masks compared to 11–12 needed by other technologies. According to Cypress, the NV Flash IP does not alter standard device characteristics or models when it is added to a baseline CMOS process.

UMC qualified the Cypress 65 nm SONOS in 2013. The 55 nm technology is a scaled version of the 65 nm technology. UMC indicated that they plan to use this SONOS NV memory IP at 55 nm for various applications for embedded memory [89]. Typical applications indicated for embedded Flash include: smart cards, band cards, and wearables. This SONOS NVM technology required three to four additional masks over conventional CMOS. A schematic cross-section of the SONOS cell is shown in Figure 3.65.

A low cost 65 nm SONOS eNVM technologywas discussed in May of 2013 by Cypress and UMC [90]. It is based on integration of a SONOS-based NVM module into a foundry 65 nm CMOS process flow with only three additional masks and no additional high voltage oxide [92]. The SONOS memory cell is a 2T cell with a SONOS control gate and CMOS select gate connected in series. Both the erase and program operations use Fowler-Nordheim (FN) tunneling, which has better cycling endurance than hot electron or hot hole program and erase.

The SONOS process module was inserted between the well channel implants and the logic gate oxidation. The optimized integration ensured that the CMOS device parameters of the eNVM process were closely matched to the baseline process. Although a

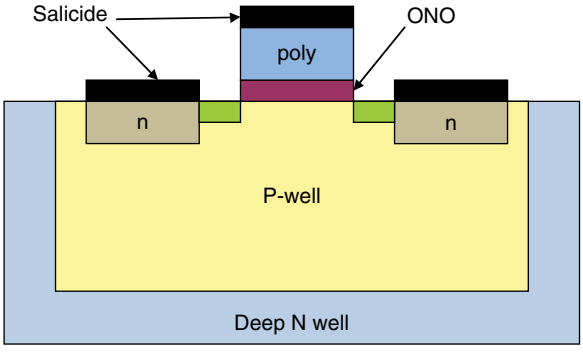


Figure 3.65 Schematic cross-section of the 55 nm SONOS eNVM cell showing the ONO layer and the Salicide. Based on R. Wilson, *Electronics Weekly*, July 15, 2014 [89].

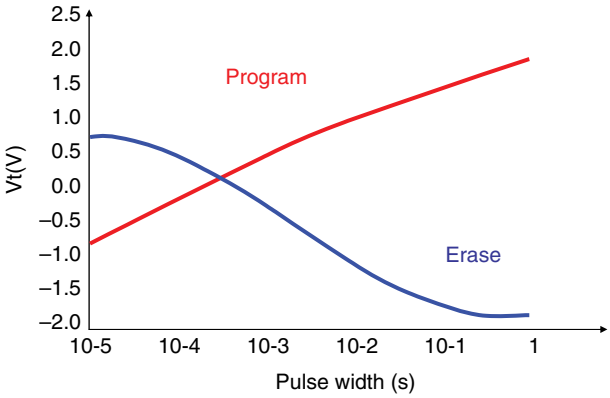


Figure 3.66 Program/erase pulse width of SONOS eNVM technology. Based on K. Ramkumar *et al.* (Cypress, UMC), IMW, May 26, 2013 [90].

7.5 V program/erase voltage was used, reliability met automotive data retention requirements and a 100k cycle endurance on a 4.5Mbit Flash memory macro.

An engineered ONO stack with effective oxide thickness (EOT) of 10 nm acted as a gate dielectric in the SONOS transistor. The engineered tunnel oxide enhanced charge trapping during P/E and minimized charge loss during data retention. The 2T cell area is about 0.154–0.200 μm^2. The program/erase was at 7.5 V differential voltage. Program times are 2–5 ms and erase times are 5–10 ms. The 7.5 V voltage differential for P/E is attained by 4.0 and 3.5 V biases, which can be designed using the gate oxide of 2.5 V I/O FETs already in the process. An illustration of the program/erase window is shown in Figure 3.66 for V_t versus pulse width in seconds. Data retention was 10 years and there was minimal degradation with endurance over 100k cycles. Read cycle time at 1.08 V was 28 ns and P/E current was 10 mA.

A 55 nm eSONOS was announced in August of 2014 by Cypress and Shanghai Huali [91]. The functioning silicon cells used the Cypress 55 nm SONOS technology embedded Flash. The silicon cells are intended for smart cards and Internet of Things (IoT) applications. Both technology and design IP were expected to be available for volume manufacturing by Shanghai Huali customers in 2H2015. Shanghai Huali licensed the 55 nm SONOS eNVM technology process in January of 2014.

The process required three mask layers to insert it into a conventional CMOS process compared to 9–12 added masks for double poly floating gate embedded Flash technologies. The reduction in masks results in lower manufacturing costs. SONOS does not affect the characteristics of models of the baseline CMOS process. Data retention was specified at 10 years and program/erase endurance was greater than 10^5 cycles. The process is resistant to soft errors. It was expected the Cypress SONOS technology could be scaled to 40 and 28 nm.

Spansion had charge trapping Flash memory devices in production from the 110 to 65 nm technology nodes in December of 2014 [92]. Operating voltages were at 1.8 and 3.0 V.

In January of 2015, Cypress announced it was licensing its 40 nm SONOS eFlash IP to UMC to manufacture next-generation MCU, IoT, and wearable circuits [93]. The agreement extended the prior collaboration on the 65 and 55 nm SONOS embedded NV memory. This is the third technology collaboration in three years between UMC and Cypress.

The Cypress 40 nm SONOS requires fewer additional mask layers than some competing technologies in a conventional CMOS process. The 5 additional masks of the 30 nm SONOS compare well with the minimum of 12 additional masks required by various floating gate embedded Flash technologies. The SONOS does not alter conventional device characteristics when added to a baseline CMOS process. It also had high yields and reliability, 10 years of data retention, 100 000 program/erase endurance cycles, and good resistance to soft errors. The 40 nm SONOS process is embedded in a low power process technology along with other 40 nm variants, which will aid in the development of power efficient products for IoT and wearable electronics markets.

3.8.7 Embedded 2T SONOS NVM in HV CMOS

An embedded 2T SONOS NVM memory intended for use in a 90 nm conventional high voltage (HV) CMOS process was discussed in May of 2014 by SK Hynix and WingCore [94]. The target application was touch-screens, which require fast read access NVM embedded in the MCU. The transmitter and receiver bias is high to provide noise immunity in the touch-screen controller. To meet this requirement, foundries would like to have the embedded process in the high voltage technology. The single poly cell can be made in several steps: ONO formation, cell junction implant, and ONO removal. These are noncritical process steps and masks. The cell programs by channel hot electron injection (CHEI) and erases using band-to-band-tunneling (hot hole injection (BTBT-HHI) erase).

A 64KB SONOS test chip was developed and operationally tested for program and erase, cycling endurance, and data retention. The test chip had an internal charge pump circuit and external bias, forcing the circuit along with the cell current read circuit. The external forcing bias was used to read the cell current. The chip had a wide window and good reliability. This is a medium density embedded memory, but process simplicity matters more than cell area for an embedded memory. The cell size is 0.2924 μm^2.

Testing of the 64KB test chip indicated that the cell was free from over-erase. It had rapid programming and fast access on-current using the CHEI program and BTBT-HHI erase. The cell met the cycling specification for code Flash, which was greater than 10 years at 85 °C after a 1K cycle. The conclusion was that the cell was adequate to serve as embedded memory in medium density applications.

Program and erase characteristics of a 2T SONOS NVM cell using simulations was discussed in October of 2014 by SK Hynix and Sogang University [95]. Program operations used channel hot electron injection (CHEI) and erase operations used band-to-band tunneling hot hole injection (BTBT-HHI). The optimized CG length was key in the 2T SONOS device. The cell used the entire channel to achieve good reliability during program and erase. It appeared that excess electrons might build up in the nitride layer near the source junction due to spatial mismatches of injected electrons and holes during program/erase cycles. Electron build-up was confirmed through device simulations and measurements of the dependence of gate length on program and erase speeds. Due to the gradual accumulation of electrons, the transconductance of the cell tended to be continually reduced. The degraded transconductance (G_m) value was found to be improved after a retention bake.

3.8.8 Self-Aligned Nitride Logic NVM

An on-chip recovery operation for self-aligned nitride (SAN) logic NVM cells was discussed in November of 2015 by the National Tsing Hua University and TSMC [96]. These cells were made in a high-*k* metal gate CMOS process [96]. This NVM cell had a merged nitride spacer sandwiched between high-*k* metal gate (HKMG) stacks in a nanometer technology full CMOS process. The storage node was defined by a merged nitride spacer decoupled from the logic transistor gate oxide. This cell was validated from 90 to 29 nm. The gate length was scaled, which enabled the SAN cell to be erased by band-to-band hot hole injection (BBHHI). This CMOS-compatible memory was used for multiple time programming operations. A schematic circuit diagram of the SAN cell is shown in Figure 3.67.

As the CMOS logic technology migrated below 28 nm, several phenomena occurred, including: soft breakdown, a metal gate was used instead of a poly gate, and stress-induced leakage current occurred. This limits the applications of these logic NVMs. The hot carrier charge injection is known to damage the Si/SiO_2 interface and degrade the I–V characteristics of memory cells. For this reason several thermal heat recovery operations were studied to extend the SAN cell life.

SAN cells are programmed by source side injection (SSI) with a small voltage applied to turn on the channel under the select gate (SG) at the same time that a high voltage V_{pp} is applied to the source line (SL) and program gate (PG) to induce a high lateral electric field. For erase the 28 nm SAN cell uses band-to-band hot hole injection. This erase used grounded SG and BL, while V_{pp} is applied on the SL to induce BBHHI in the source junction while a small negative bias on the PG attracts hot hole injection into the storage node to compensate the stored electrons. After several program and erase cycles, the oxide quality declines, caused by interface traps, which capture carriers and reduce carrier mobility. The trapped charge deteriorated the gate control and caused reliability issues. A decline of read current after P/E cycling resulted in narrowing of the ON/OFF window.

Two recovery methods were suggested to recover the ON/OFF window after cycling stress. Both AC and DC methods were used to eliminate deep trapped charges using electrical self-heating. The data indicated that the DC recovery methods provide nearly full damage anneal capability, which effectively extends the SAN cell's endurance level. In the AC recovery method an electric pulse is applied to PG and SG with SL and BL floating. The alternating current heats up the gates, which compensates the dangling bonds. DC recovery applies a forward bias voltage on the N+, -SL, and P substrate to cause a high forward PN junction current, which locally heats up the SAN cell. The on-chip heating anneals the interface damage in the bottom oxide under the merged nitride spacer. Figure 3.68 shows the decrease of charge pumping current after doing DC recovery, which indicates the reduction of interface traps. DC recovery was found to effectively anneal the damage caused by P/E stress and return the current to close to a fresh state.

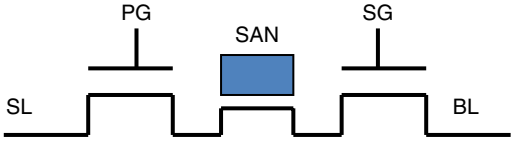


Figure 3.67 Schematic circuit diagram of the self-aligned nitride (SAN) logic non-volatile memory cell. Based on P.-Y. Lin *et al.* (NTHU, TSMC), *Journal of Electron Devices Society*, November 2015 [96].

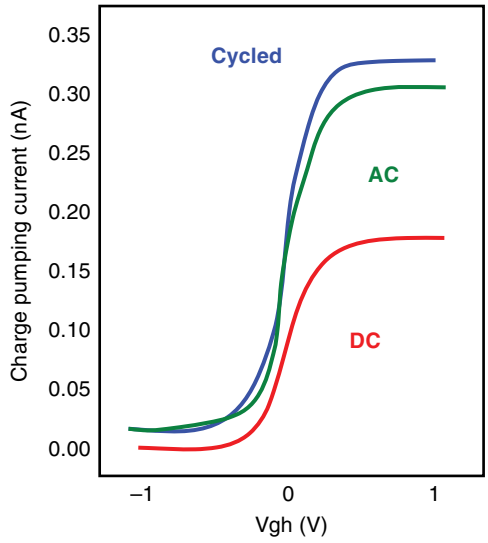


Figure 3.68 Decrease in charge pumping current after recovery due to reduction of interface traps. Based on P.-Y. Lin *et al.* (NTHU, TSMC), *Journal of Electron Devices Society*, November 2015 [96].

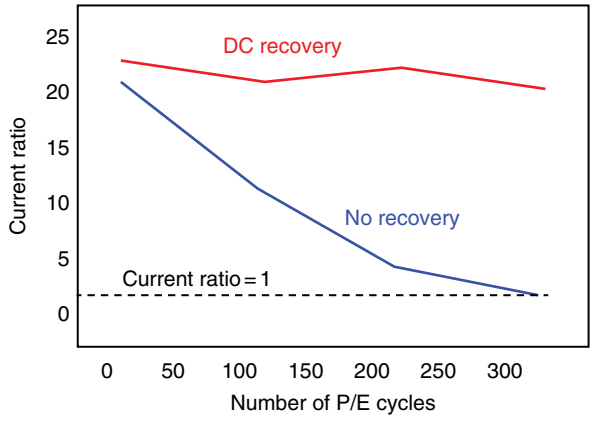


Figure 3.69 Read current ratio after 300 P/E cycles with and without periodic DC recovery. Based on P.-Y. Lin *et al.* (NTHU, TSMC), *Journal of Electron Devices Society*, November 2015 [96].

The read current ratio degraded after 300 P/E cycles for cells without DC recovery but was restored by using periodic DC recovery operations to help maintain a stable read current margin, as indicated in Figure 3.69. This on-chip recovery method was successfully shown. It indicates the feasibility for highly scaled low power SAN cells for MTP operations.

3.8.9 p-Channel SONOS Embedded Flash

In April of 2011, eMemory Technology discussed a dynamic programming method for operation of its p-channel SONOS embedded Flash memory in order to achieve high reliability and scalability [97]. The new programming scheme meant the cell could perform with better write efficiency and have less oxide degradation than with the

conventional CHE programming. In addition, the low programming current under low program bias meant a simple circuitry design with a small charge pumping area could be used. The logic-based 2T memory cell used a PMOS access transistor and a PMOS charge trapping transistor with an ONO charge trapping layer.

The dynamic programming method for channel hot hole-induced hot electron (HHIHE) injection using a programming current-clamped (PCC) scheme on a p-channel SONOS Flash memory was discussed further in May of 2012 by eMemory [98]. Improved programming efficiency and device reliability were found using this programming scheme. There was an 85% programming current reduction and an order of magnitude retention improvement. A tightened program state distribution was also found with lower programming power consumption, a smaller high voltage pumping circuit area, and a better read sensing window.

3.8.10 Charge Trap eFlash for Low Energy Applications

In October of 2014, ST-Microelectronics and Aix-Marseille University discussed a charge trap embedded Flash for low energy applications [99]. The chip was made on a 200 mm wafer in 90 nm technology. The asymmetrical tunnel window (ATW) memory cell was developed to improve programming operation during a hot carrier injection. This device has an asymmetrical tunnel oxide thickness along the channel. This is a characteristic that improves current consumption and injection efficiency for the conventional Flash floating gate memory cell. The new charge trapping memory cell is studied, which has only one gate contact rather than a split gate, which enables better scalability. This report studies the electrical characterization of the programming window, the current consumption, and the cell endurance. The simulations and measured results are compared with the 1 T Flash floating gate device measurements. A schematic cross-section of the ATW floating gate memory cell is shown in Figure 3.70.

An additional lithography mask was provided over the Flash floating gate process which enabled the etching of the high voltage oxide to obtain the tunnel oxide area. The memory gate stack was then made by sequential depositions of the following layers: polysilicon floating gate, ONO dielectric, and polysilicon control gate. The CG/FG stack was etched and then the source/drain regions were implanted. The advantage of this process flow is that it is easy to integrate in a standard CMOS process. The goal of the ATW cell lies in increasing the coupling factor, which improves the programming window, and in controlling the drain current consumption during hot carrier injection in the presence of a high voltage oxide. Experimental results showed a 20% gain in the

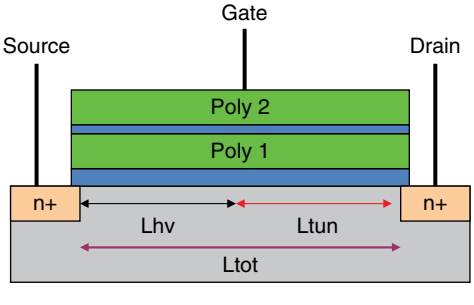


Figure 3.70 Schematic cross-section of asymmetrical tunnel window (ATW) floating gate memory cell. Based on J. Bartoli *et al.* (ST-Microelectronics, Aix-Marseille University), CAS, October 2014 [99].

programming window and a 34% gain in current and energy consumption over that of a 1T Flash floating gate.

3.8.11 Blocking and Tunnel Oxide of DT BE-SONOS Performance

A double trapping bandgap engineered SONOS was discussed in May of 2015 by Macronix and NCTU [100]. This device had a fast erase speed using a second nitride trapping layer (N3) and an additional blocking oxide (O4) added on to the basic BE-SONOS. This strategy gives a good erase performance but the additional layers increase the effective oxide thickness (EOT) and therefore increase the erase voltage, so it is needed to minimize their impact. This study investigates the effect of thinning the blocking layers. Since the incremental step pulse programming (ISPP) and high temperature retention charge loss are primarily dominated by the ONO thickness of the BE-SONOS below the blocking layers, reducing the blocking layer thickness has only a minor impact on ISPP and retention. Erase saturation is determined by the dynamic balance of channel hold injection and gate electron injection.

It was shown experimentally that reducing the thickness of the oxide between two trapping layers has a small impact on erase saturation once the gate injected electrons are efficiently suppressed by the highest level oxide in the cell. By using a high quality (HQ)-SiO_2 to replace the top tunnel ONO, the trapped electron out-tunneling was reduced, which may improve retention without increasing the effective oxide thickness. A schematic cross-section of the DT BE-SONOS stack is shown in Figure 3.71. The third nitride trapping layer, N3, and an additional blocking oxide, O4, were stacked on top of the original BE-SONOS to form the following stack: O1/N1/O2/N2/O3/N3/O4. During erase the gate injected electrons were stored in N3 and channel injected holes were stored in N2. The electric field across the top blocking oxide O4 can be further reduced when N3 stores electrons and therefore gate injection can be suppressed.

The N2 thickness effect on BE-SONOS performance has been discussed previously and it is thought that the 7 nm N2 is sufficiently thick to maintain nearly 100% capture efficiency. In this study, the N2 thickness was maintained at 6 nm to provide sufficient capture efficiency while the blocking layer, O3/N3/O4, reduced the effect on the device performance. A comparison was done of program, erase, and high temperature

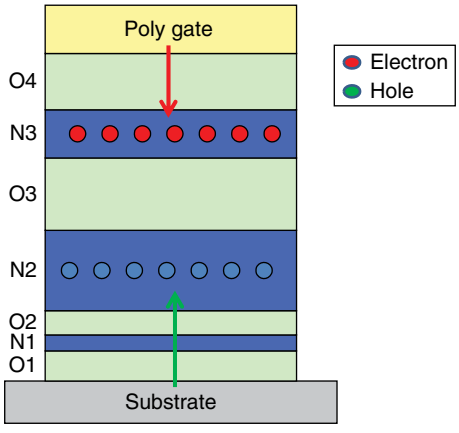


Figure 3.71 Schematic cross-section of the DT BE-SONOS stack. Based on R. Lo *et al.* (Macronix, National Chiao Tung University), IMW, May 2015 [100].

retention performance at 150 °C to find out the scaling strategy for DT BE-SONOS. An experiment with O2 was also discussed to further improve the retention performance. The effect of the blocking layer, O3/N3/O4, thickness, and O2 engineering on DT BE-SONOS performance were discussed. Blocking layer thinning showed only a minor effect on ISPP and HT retention (150 °C) since these are controlled by the O1/N1/O2/N2 thickness of BE-SONOS. Erase saturation, however, was significantly affected by a thinner blocking layer, since more gate injected electrons tunnel into the N2 to recombine with the channel injected holes. The thinnest O2/N3/O4 was determined by the required erase level and thinning the O3 showed less impact on erase saturation.

3.8.12 Novel Embedded Charge Trap Memories

A comparison of a gate-all-around (GAA) MOSFET NV memories with both SONOS stacks and with embedded nanocrystals in the gate dielectric were discussed in May of 2012 by Jadavpur University [101]. Gold (Au) nanocrystals were embedded in an SiO_2-HfO_2 stacked gate dielectric and compared with an ONO stacked GAA MOS of similar dimensions. The surface potential of the GAA MOS was evaluated using a pseudo 2D-based methodology. The fields of the different gate dielectric stacks were evaluated using a Gauss law and a Wentzel-Kramers-Brillouin approximation-based model was used to compute the Fowler-Nordheim tunneling current. Various electrical characteristics were simulated.

MONOS Flash memory using dielectrics as charge trapping layers have advantages over floating gate Flash. These advantages include: lower power consumption, higher reliability due to decreased whole charge leakage, and better scaling. Early MONOS Flash devices used Si_3N_4 with a "k" of about 7 as the dielectric. In June of 2012, the University of Hong Kong and Hong Kong University of Science and Technology discussed using rare earth metal oxides, which include: Y_2O_3 ($k = 18$), Pr_2O_3 ($k = 15$), Nd_2O_3 ($k = 16$), Er_2O_3 ($k = 13$), Gd_2O_3 ($k = 14$), and La_2O_3 ($k = 25$) for charge trapping dielectric [102]. The charge trapping characteristics of La_2O_3 were studied with and without added nitrogen (N). The MONOS stack examined was $Al/Al_2O_3/La_2O_3/SiO_3/Si$. It was found that the nitrided La_2O_3 memory window was 4.9 V at ±10 V sweep, which was larger than a nonnitrided La_2O_3. Program speed at 1 ms was also higher and the charge loss after 10 years at 27% was smaller. This was thought to be due to the NLa_2O_3 film having a less crystallized structure and higher trap density due to the N and also due to suppressed leakage from the nitrogen passivation.

In June of 2012, National Chi Nan University, TSMC, and National Tsing Hua University described a numerical examination of the read operation and scalability of multibit Schottky barrier charge-trapping cells [103]. The multi-bit Schottky barrier cell was accessed using a conventional forward and reverse read method with the "second-bit" effect improved. The two-bit programming, erase, and read schemes of the multibit Schottky barrier charge trapping cell is illustrated in the schematic cell cross-section shown in Figure 3.72.

The scaled Schottky barrier cell was found to have good short-channel immunity and tight-matched distributions of injected electron and hole carriers, and better than conventional SONOS cells, which tend to have issues with both short-channel immunity and with matched injected electron and hole carrier distributions. The uniformity of the forward and reverse read were much greater for the 2-bit Schottky barrier cell than

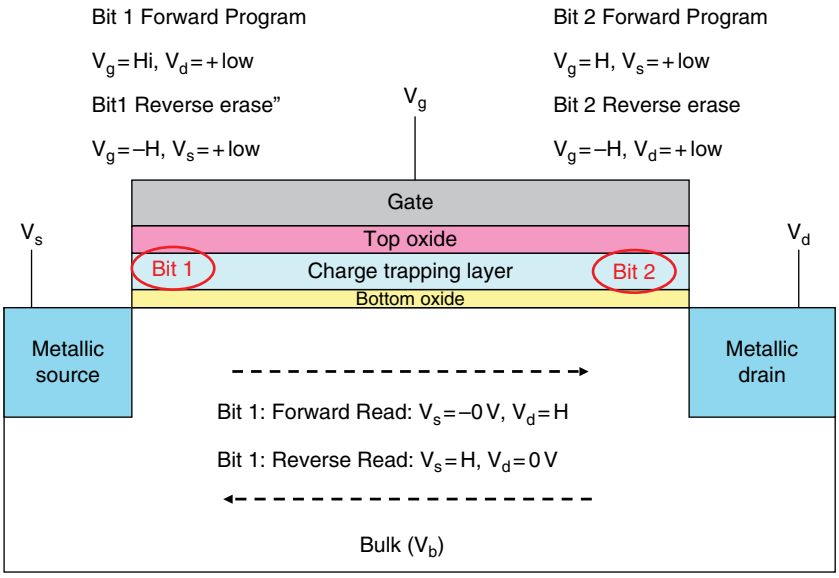


Figure 3.72 Schematic of Schottky barrier charge trapping cell with two-bit operation shown. Based on C.H. Shih (National Chi Nan University, TSMC, NTHU), *IEEE Trans. on Electron Devices*, June 2012 [103].

for the conventional 2-bit SONOS cell. The Schottky source/drain barrier permits cell reading, source-side programming and drain-side erasing to be retained when the cells are scaled. The stacked gate architecture and CMOS process compatibility were preserved using the multibit Schottky barrier charge trapping cell.

In September of 2012, NTHU and ITRI discussed $TaN/Al_2O_3/HfO_2/HfAlO_2/SiO_2/Si$ (MAHOS-type) nonvolatile memory with various Ge content in an SiGe buried channel [104]. Compared to a device with a silicon channel, both program and erase speeds were significantly improved by the SiGe buried channel. The SiGe buried channel resulted in negligible reliability degradation. Endurance was up to 10^6 program/erase cycles with a 4.1 V memory window. A schematic cross-section of the p-channel memory transistor with an SiGe buried channel with different Ge contents is shown in Figure 3.73.

The mechanism of transient V_{th} shift after erase in charge trap Flash memory and its impact on cell operation was discussed in December of 2012 by KAIST and Hynix [105]. The V_{th} after erase does not settle to its final value immediately and waiting for settlement of V_{th} increases the total erase time. The primary mechanism for the V_{th} shift was found to be hole redistribution in the charge trap layer. A new erase method to make fast redistribution of holes was found to reduce the transient V_{th} shift. This new scheme speeded up the erase process.

Prior to applying a read voltage, a small positive gate pulse was applied immediately after the main erase pulse. This small positive pulse helped fast redistribution of holes in the charge trap layer and therefore reduced the V_{th} transition period after erase. The added pulse marginally increased the erase pulse time but greatly shortened the V_{th} transition time. As a result, the total erase time was shortened. Transient V_{th} shift was also found in 3D charge trapping devices and can be reduced by properly scaling the 3D device.

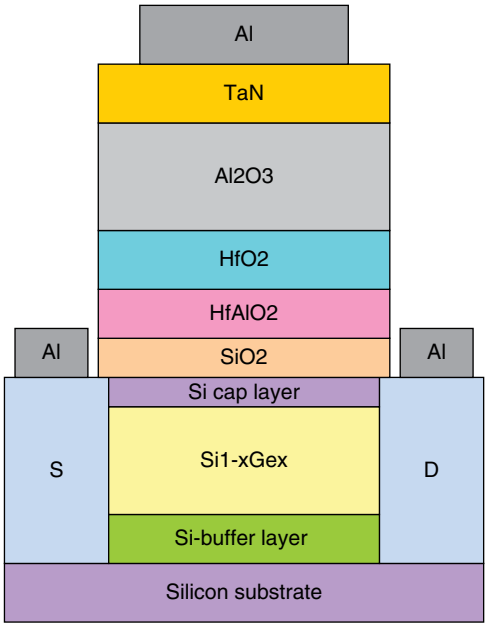


Figure 3.73 P-MOS memory transistor with SiGe buried channel with different Ge content Based on L.J. Liu *et al.* (NTHU, ITRI), *IEEE Electron Device Letters*, September 2012 [104].

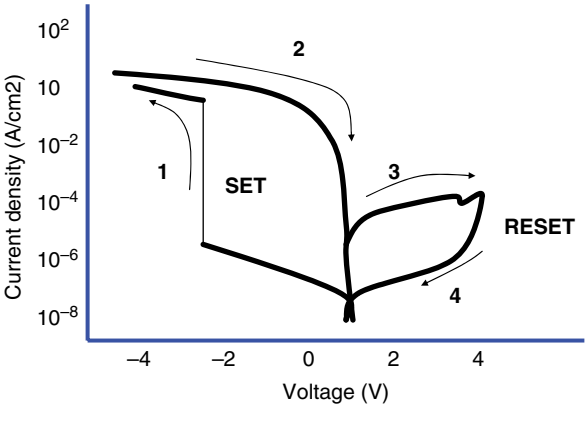


Figure 3.74 *I–V* curve for ZnO charge trap Flash using resistive switching. Based on Y. Seo (Korea University), *IEEE Electron Device Letters*, February 2013 [106].

In February of 2013, Korea University discussed a ZnO-based charge trap Flash memory using a resistive switching mechanism [106]. The MONOS stack was metal/ZnO/nitride/oxide/silicon and used the electrical transport in the ZnO resistive switching layer. This device combined the conventional SONOS structure with resistive switching materials. The device was faster than previous devices made with perovskite oxide as a conduction path. Switching speed was 10 ns, operating voltage is ±7 V for P/E states and endurance was 10^6 P/E cycles. The *I–V* curve for the device is shown in Figure 3.74.

A conducting path is made through the ZnO layer regardless of bias polarity. When a negative bias was applied to the gate and the ZnO layer was in the LR state, then electrons were provided from the top electrode to the nitride layer. When a positive bias was applied to the gate and the ZnO layer was in the LR state, then the electrons trapped in the nitride layer flowed out through the conduction path. The p-Si has few electron carriers and the Al electrode has few hole carriers, so it is similar to a p-n diode. This makes it easy to achieve self-rectification.

In March of 2014, Kyung-Hee University and Electronic and Telecom Research Institute discussed a charge trapping NVM transistor with a top gate structure made of an Al_2O_3 blocking/ZnO charge trap/IGZO active/Al_2O_3 tunnel layer [107]. The ON/OFF ratio of the memory was $>10^6$ and was obtained with 100 ms –20 V program pulses. Endurance was 10^4 cycles. The memory ON/OFF ratio $>10^3$ was found even after 10^4 s. Retention properties were affected by the read bias conditions. A schematic cross-section of the device is shown in Figure 3.75.

The top stack structure was optimized at 100 nm Al_2O_3 blocking oxide, 50 nm ZnO charge trapping layer, 20 nm IGZO active layer and 5 nm Al_2O_3 tunneling layer. The charge trapping/detrapping using Fowler-Nordheim tunneling was confirmed.

A 3D fin-channel charge trapping Flash memory using a high-*k* Al_2O_3 blocking layer for a MANOS type Flash memory was discussed in June of 2014 by AIST and NIMS [108]. To be compatible with the FinFET logic technology, a 3D fin-channel charge trapping Flash memory is needed. The cell was made by scaling gate length to 22 nm.

The electrical characteristics studied included: variability of threshold voltage, endurance, and data retention. The high-*k* effect of the Al_2O_3 blocking layer resulted in a better short-channel effect immunity and a larger memory window in the fabricated MANOS Flash memories than in MONOS ones with an SiO_2 blocking layer. A schematic cross-section of a sketch of the device fabrication process flow for the MANOS Flash memory is shown in Figure 3.76.

The electrical characteristics of the 3D fin-channel MANOS type Flash memories with different L_g values from 22 to 476 nm were investigated. It was found that better SCE immunity and a larger memory window were obtained by using the Al_2O_3 blocking layer instead of an SiO_2 blocking layer due to the high-*k* of the Al_2O_3.

In June of 2014, the University of Udine and MDM Lab IMM-CNR discussed a simulation of the trapping properties of HfO_2-based charge trap memory cells for use in

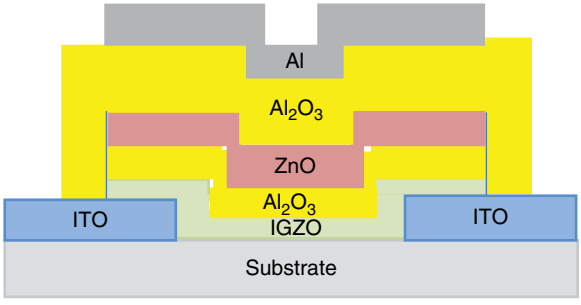


Figure 3.75 Schematic cross-section of charge trapping IGZO NVM transistor. Based on J. Bak *et al.* (Kyung-Hee University, Electron and Telecom Research Institute), *IEEE Electron Device Letters*, March 2014 [107].

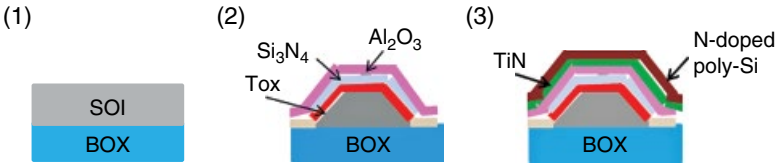


Figure 3.76 Schematic cross-section of FinFET MANOS Flash device fabrication process flow. Based on Y. Liu *et al.* (AIST, NIMS), SNW, June 2014 [108].

embedded NVM [109]. The impact of process conditions on the material structure and trapping behavior of the gate stacks was studied. Models were presented for the HfO_2 structure and for the defects responsible for the electron trapping. HfO_2 was found to have a trap density comparable with that of SiN depending on the temperature. The HfO_2 traps are shallower in energy than the SiN traps, but retention was still sufficient. HfO_2, with $k = 16$, had good trapping capability and is a material used in advanced scaled CMOS processes. The intent of this study was to assess the trapping properties of HfO_2 films used as trapping layers.

HfO_2 layers were integrated into $TaN/Al_2O_3/HfO_2/SiO_2/Si$ (TAHOS) gate stacks with a device area of 8×10–4 cm. The 4.5 nm tunnel oxide was thermally grown SiO_2, while the HfO_2 trapping layer was 6–16 nm and the Al_2O_3 was deposited 16 nm thick. The top oxide was deposited by ALD. The TaN top electrode was deposited by RF sputtering and patterned by optical lithography.

The fabricated gate stacks were characterized and the trapping properties of the fabricated HfO_2 layers were studied as a function of the fabrication conditions, particularly the HfO_2 thickness and process deposition anneal (PDA) temperature. The PDA temperature had a larger impact on cells with thin HfO_2 layers. It was shown that 1030 °C PDA ensures good HfO_2 trapping properties while preserving the insulating properties of the Al_2O_3 top layer. Although the HfO_2 traps are shallower than those of SiN, HfO_2 can ensure sufficient trapping capabilities for memory devices when considering the larger permittivity, which permits integration of thicker layers while preserving the effective oxide thickness of the gate stack.

In October of 2014, Chang Gung University discussed an MOHOS-type memory using a ZrO_2 trapping layer with nitrogen incorporated in the trapping layer [110]. This was combined with the use of rapid thermal anneal (RTA). It was found that the memory device, which included the nitrogen annealed at 900 °C, could improve memory device performance. Characteristics found included: large C–V hysteresis, faster program/erase speed, better data retention, and a 7.7% smaller charge loss.

The use of oxygen vacancies for NVM by trapping electrons in the high-k, gate dielectric layer of the n-channel device was studied by IBM and UCLA in April of 2015 [111]. Programming was done using channel carrier injection and erase was done by tunneling. 64Kb arrays were made and tested. This study discussed a fully CMOS compatible, scalable, rewritable, NVM structure without adding new processes or materials to a conventional CMOS process.

HfO_2, which is used in most advanced CMOS technologies, tends to form oxygen vacancies where carrier trapping is known to cause threshold voltage shifts. The HfO_2 dielectric tends to form trapping centers for free carriers and these trap densities are higher than in SiN and are present in the volume of the dielectric as well as at the

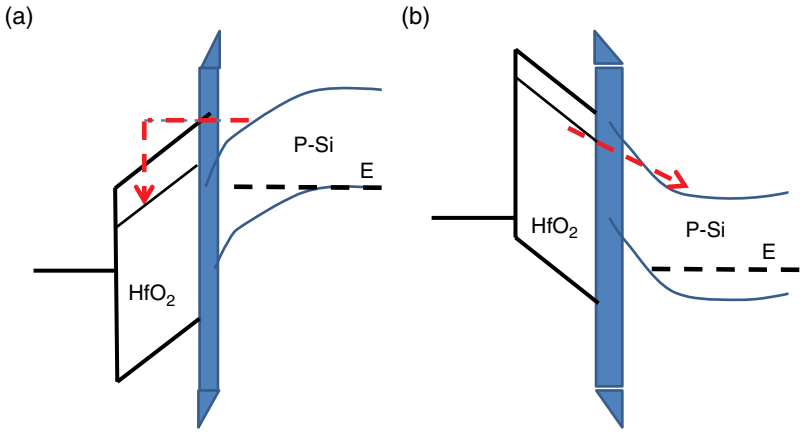


Figure 3.77 Band diagram of (a) trapping of carriers into the oxygen vacancy traps and (b) detrapping of carriers out of the oxygen vacancy traps. Based on C. Kothandaraman *et al.* (IBM, UCLA), IRPS, April 2015 [111].

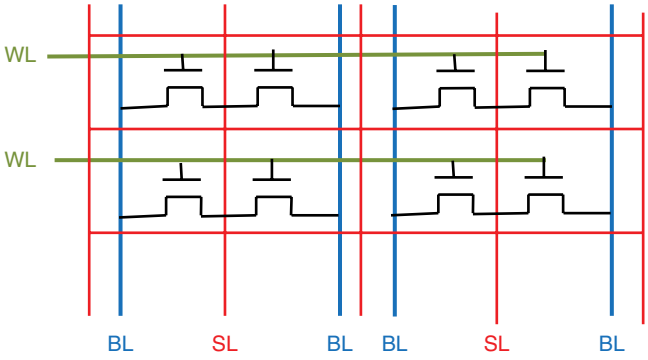


Figure 3.78 Schematic circuit diagram illustrating twin-cell method. Based on C. Kothandaraman *et al.* (IBM, UCLA), IRPS, April 2015 [111].

interfaces. A 3D high performance 22 nm SOI platform with embedded DRAM and SRAM was used and the HfO_2 in the gate was used as the memory element.

Programming was done by injecting electrons across the dielectric by applying a positive 2 V 1 ms pulse gate voltage on an nFET with the drain at 1.5 V and source at 0 V. Both the high drain field and the self-heating of the channel provided for efficient trapping of the carriers. The band diagram of the trapping of carriers is shown in Figure 3.77. Shown in Figure 3.77 (a) is the band diagram for the capture of carriers into the oxygen vacancy traps with application of a positive gate and drain voltage. Figure 3.77 (b) shows the detrapping of the carriers with reversal of the voltage.

A "twin-cell" method was used where the V_t shift of one FET is compared against a reference FET to improve the sense margin. The selected transistor was programmed by applying 2 V to the WL while the SL was connected to 1.5 V. The BL was set to 0 while the BL\ was set to 1.5 V to ensure unselected transistors were not programmed. A schematic circuit diagram illustrating the twin-cell technology is shown in Figure 3.78.

This study showed that carriers trapped in vacancy traps in high-*k*/metal gate technologies are not just a degradation phenomena but can be used to create an embedded memory element compatible with an advanced CMOS.

3.9 Split Gate CT eFlash Nanocrystal Storage

A 128KB split gate embedded Flash memory that replaced nitride storage with nanocrystal storage was discussed in June of 2008 by Freescale (now NXP) [112]. A schematic cross-section of the cell is shown in Figure 3.79. The part had thermally grown bottom oxide, deposited nanocrystals, and a high quality deposited top oxide. The dielectric layers were thinner than that used in floating gate memories. The control gate was counter-doped to eliminate read disturb resulting from the thin dielectric. The nanocrystal storage had the advantage of maintaining the charge on the gate even if some of the nanocrystals leaked charge, since there was a net mesh effect of stored charge on the gate.

An operating window of 1.5 V was maintained through 10 000 program/erase cycles. The threshold distribution of the array was well controlled. Programming was by source-side injection with a program time of 10–20 μs. Erase was by tunneling to the control gate. Good data retention was shown at high temperatures for cycled and uncycled arrays. Power was 2 to 15 μA/bit. No negative bias was required, which eliminated the need for negative charge pumps.

A commercially available MCU family built using its split gate NOR Flash memory using silicon nanocrystals for storage was discussed by Freescale/NXP in May of 2012 [113]. Their 32-bit MCU family with NC storage had a range of array sizes from 32KB to 1MB. The memory permitted fully configurable embedded EEPROM functionality, which managed wear leveling for high array endurance. The nanocrystal layer was an ultrathin film that separated naturally into individual nanocrystals. Read access time was <30 ns. Source-side injection programming time was 10–20 μs. Tunnel erase time

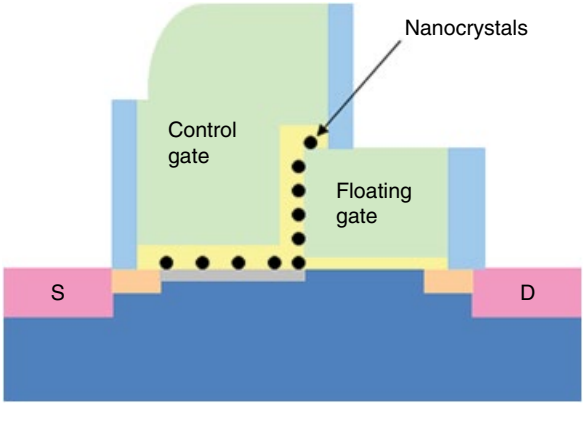


Figure 3.79 Split gate embedded Flash memory where nitride storage is replaced with nanocrystal storage. Based on G. Chindalore *et al.* (Freescale), VLSI Technology Symposium, June 2008 [112].

into the gate was 1–20 ms. High temperature data retention before and after cycling was up to 10k cycles and endurance was up to 10^6 cycles in EEPROM mode. The device operated in a temperature range of –40 to 105 °C with full operation down to 1.7 V single power supply. This made the parts suitable for some embedded automotive applications with an extended temperature range.

Scaling of a 90 nm node MCU with embedded silicon nanocrystal NV memories was discussed in May of 2012 by Freescale/NXP and Global Foundries. The devices ranged up to 4Mb in capacity [114]. The parts were made at Global Foundries. Endurance was >100k cycles in the temperature range from –40 to 125 °C. The NC memories retained good data retention and immunity to extrinsic charge loss mechanisms like SILC, even after extended cycling. The NC memory was thought to be scalable to the next generation nodes without degradation of P/E speed, endurance, or reliability. The bit-cell area was 0.11 μm^2 at the 55 nm node.

In July of 2016, Ryukoku University, NIAIST Tsukuba, and the Nara Institute of S&T discussed a type of metal nanoparticle junctionless field-effect transistor (JL-FETs) [115]. The JL-FETs were made 3.6 nm long. They were formed by anisotropic wet etching of a silicon-on-insulator (SOI) substrate to form a V-groove in the silicon. The groove defined a nanometer scale channel into which metal nanoparticles (NP) were selectively placed on to the bottom of the V-groove.

The NPs were used for charge trapping and exhibited low voltage operation and a broad threshold voltage shift during memory behavior in the channel. A schematic cross-section of the memory is shown in Figure 3.80 (a) while (b) shows a top-down view of the nonvolatile memory transistor showing the V-groove with the NPs in the 3.6 mm groove.

Characteristics of I_d–V_g with $V_d = 0.2$ V for the V-groove channel with and without the NPs are illustrated in the I–V chart in Figure 3.81. The threshold voltage change with and without the NPs is clearly shown.

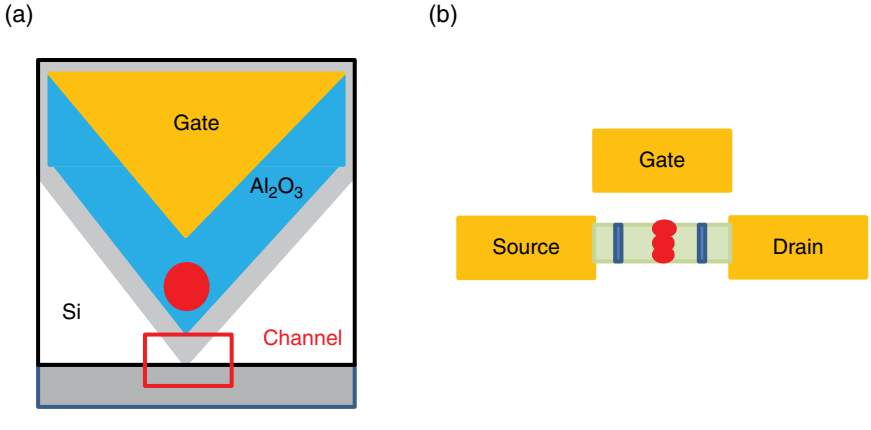


Figure 3.80 Nonvolatile memory with nanoparticles in V-groove: (a) cross-section and (b) top-down view. Based on T. Ban *et al.* (Ryukoku University, NIAIST, Tsukuba, Nara Institute of Science and Technology), AMF-PD, July 2016 [115] (permission of IEEE).

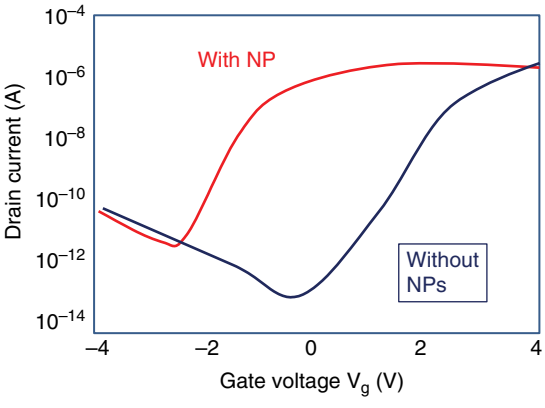


Figure 3.81 I_d–V_g characteristics with V_d = 0.2 V for V-groove channel with and without NPs. Based on T. Ban *et al.* (Ryukoku University, NIAIST, Tsukuba, Nara Institute of Science and Technology), AMF-PD, July 2016 [115].

3.10 Novel Embedded Flash Memory

An embedded NVM integrated in the back-end-of-line (BEOL) process was discussed by A*STAR and the University of Singapore in August of 2013 [116]. This one-time programmable (OTP) antifuse non-volatile memory (NVM) was made using a TaN microbeam movable arm, which fused on contact and could potentially be used in applications in a rugged environment. It required one added mask to conventional CMOS and can be integrated above the finished circuit in the BEOL. A typical fuse current is 1 mA, operating voltage is 4 V, and the measured contact resistance is <2 kohm. The memory consists of 1 transistor and 1 microbeam per bit in the memory array. A schematic cross-section of the OTP NVM microbeam is shown in Figure 3.82, where (a) is the open state ("0") and (b) is the closed/fused state ("1"). A schematic circuit diagram of the 1T1B OTP array is shown in Figure 3.82.

A typical 1 mA fusing current was required to achieve permanent fusing. The device operated at 4 V so it was usable in low power applications. It could be switched and read by an NMOS transistor. Testing under rugged conditions was planned but not yet conducted.

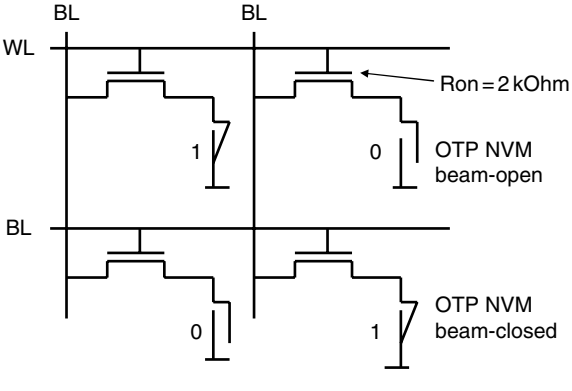


Figure 3.82 Schematic circuit diagram of the 1T1B one-time-programmable (OTP) array. Based on P. Singh *et al.* (A*STAR, University of Singapore), *IEEE Electron Device Letters*, August 2013 [116].

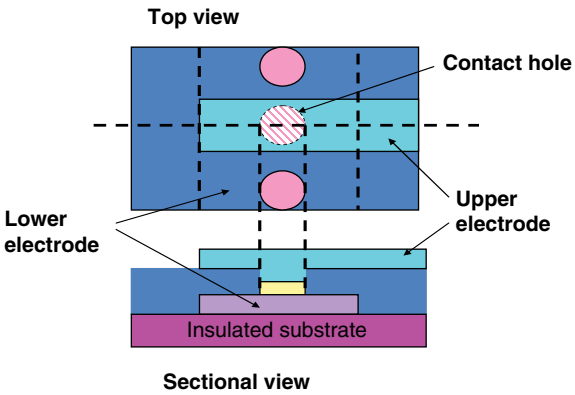


Figure 3.83 Top and cross-sectional view of nanogap memory. Based on T. Takahashi *et al.* (FEAT Research Institute, NI AIST), SNW, June 2012 [117].

In June of 2012, FEAT Research Institute and NIAIST discussed a 4kbit NV fast nanogap memory device [117]. The vertical nanogap structure was a result of an investigation showing that thin film metal electrodes with a gap of less than 10 nm on an insulating substrate have a nonvolatile memory effect in a vacuum. Such a nanogap resistance change was observed for Au, Pd, Pt, Ta, Si, and carbon nanotubes in vacuum and also in inert gases. The current between the electrodes is due to electron tunneling. The resistance between the gap can be controlled by voltage applied to the electrodes. A top and cross-sectional view of the nanogap memory is shown in Figure 3.83.

The nanogap memory phenomenon showed fast resistance switching, operation in a wide temperature range up to 200 °C, and intrinsic high bit density. In this study a nanogap memory was integrated on a CMOS LSI structure and evaluated. The programming speed from a low resistance state to a high resistance state (on to off) was 1 ns. Endurance after high speed programming and 1 ns operation to the off state were confirmed.

Various companies offer CMOS compatible commercial embedded Flash intellectual property for use in foundries. These include: eMemory, Genusion, Kilopass, Sidense, SST/Microchip, Synopsis/Virage logic, and Tower Jazz/Panasonic.

References

1 Kimura, S. (2016) Advanced non-volatile embedded memory for a wide range of applications (Hitachi), VLSI Technology Symposium, June 2016.
2 Cuppens, R. (1985) An EEPROM for microprocessors and custom logic (Philips), *IEEE Journal of Solid-State Circuits*, SC-20 (2), 603, April 1985.
3 Matsukawa, N. (1985) A high density single-poly Si structure EEPROM with LB (lowered barrier height) oxide for VLSIs (Toshiba), VLSI Tech, 1985.
4 Ohsaki, K., Asamoto, N. and Takagaki, S. (1994) A single poly EEPROM cell structure for use in standard CMOS processes (IBM). *IEEE Journal of Solid-State Circuits*, 29 (3), March 1994.
5 McPartland, R.J. and Singh, R. (2000) 1.25 Volt, low cost, embedded Flash memory for low density applications (Bell Labs, Lucent), VLSI Circuits Symposium, June 2000.

6 Raszka, J. *et al.* (2004) Embedded Flash memory for security applications in a 130 nm CMOS logic process (Virage Logic), ISSCC, February 2004.
7 Di Bartolomeo, A. (2005) A single-poly EEPROM cell for embedded memory applications (IHP Micro, University di Salerno, Chungbuk Porv. College), 2005.
8 Gogl, D. *et al.* (1997) A single-poly EEPROM cell in SIMOX technology for high-temperature applications up to 250 °C (Fraunhofer Institute of Microelectronics Circuits and Systems). *IEEE Electron Device Letters*, 18 (11), November 1997.
9 Roizin, Y. *et al.* (2008) C-Flash: An ultra-low power single poly logic NVM (Tower Semiconductor), Non-Volatile Semiconductor Memory Workshop, May 2008.
10 Cui, Z.Y. *et al.* (2009) Single poly-EEPROM with stacked MIM and n-well capacitor (Chungbuk National University). *Electronic Letters*, 45 (3), January 29, 2009.
11 Shukuri, S. *et al.* (2011) A 10k-cycling reliable 90 nm logic NVM "eCFlash" (embedded CMOS Flash) technology (Genusion), IMW, May 2011.
12 Lee, J.W. (2011) A fully integrated HF-band passive RFID tag IC using 0.18 μm CMOS technology for low cost security applications (Kyung Hee University, University of Texas (Dallas), eSilicon). *IEEE Trans. on Industrial Electronics*, 58 (6), June 2011.
13 Dagan, H., Teman, A., Fish, A. *et al.* (2012) A low-cost low-power non-volatile memory for RFID applications (Ben-Gurion University, TowerJazz), SISCAS, May 2012.
14 Chang, W.C. *et al.* (2016) An ultra low power operated logic NVM for passive UHF RFID tag applications (eMemory), ICICDT, June 27, 2016.
15 eMemory unveils EcoBit technology for RFID and NFC applications, eMemory Press Release, August 29, 2016.
16 Liao, Y.Y., Tsai, L.Y., Leu, L.Y. *et al.* (2013) Investigation of data retention window closure on logic embedded non-volatile memory (TSMC), IRPS, April 14, 2013.
17 Torricelli, F., Milani, L., Richelli, A. *et al.* (2013) Half-MOS single-poly EEPROM cell in standard CMOS process (University of Brescia, STMicroelectronics). *IEEE Trans. on Electron Devices*, 60 (6), 1892, June 2013.
18 Torricelli, F., Milani, L., Colalongo, L. *et al.* (2013) Half-MOS based single-poly EEPROM cell with program and erase bit granularity (University of Brescia, STMicroelectronics). *IEEE Electron Device Letters*, 34 (12), 1509, December 2013.
19 Viraraghavan, J. *et al.* (2016) 80Kb 10 ns read cycle logic embedded high-*k* charge trap multi-time-programmable memory scalable to 14 nm FIN with no added process complexity (GlobalFoundries (Bangalore), IBM, UCLA), VLSI Circuits Symposium, June 2016.
20 Wang, L. and Hsu, S.M. (2014) Scalable logic gate non-volatile memory (FlashSilicon), NVMTS, October 2014.
21 Li, C. *et al.* (2014) A model for single poly EEPROM cells (National University of Defense Technology), ICSICT, October 2014.
22 Chung, C.-P. and Chang-Liao, K.-S. (2015) A highly scalable single poly-silicon embedded electrically erasable programmable read only memory with tungsten control gate by full CMOS process (National Tsing Hua University). *IEEE Electron Device Letters*, 36 (4), April 2015.
23 Chung, C.P., Chang-Liao, K.S. and Chen, C.Y. (2014) Enhanced performance of single poly-silicon EEPROM cell with a tungsten finger coupling structure by full CMOS process (NTHU). *IEEE Trans. on Electron Devices*, 61 (9), 3075, September 2014.
24 Milani, L., Torricelli, F. and Kovacs-Vajna, Z. (2015) Single-poly-EEPROM cell in standard CMOS process for medium-density applications (University of Brescia). *IEEE Trans. on Electron Devices*, 62 (10), October 2015.

25 Park, S.K. *et al.* (2014) Novel select gate lateral coupling single poly eNVM for an HVCMOS Process (Hynix). *IEEE Electron Device Letters*, 35 (3), March 2014.
26 Shukuri, S. *et al.* (2014) Byte alterable embedded EEPROM with B4-HE architecture achieving 10 μsec programming and 57F2 cell size (GENUSION), IMW, May 2014.
27 Song, S.H., Chun, K.C. and Kim, C.H. (2012) A logic-compatible embedded Flash memory featuring a multi-story high voltage switch and a selective refresh scheme (University of Minnesota), VLSI Circuits Symposium, June 2012.
28 Song, S.-H., Chun, K.C. and Kim, C.H. (2013) A logic-compatible embedded Flash memory for zero-standby power system-on-chips featuring a multi-story high voltage switch and a selective refresh scheme (University of Minnesota, Samsung), *IEEE Journal of Solid-State Circuits*, 48 (5), May 2013.
29 Song, S.H., Kim, J. and Kim, C.H. (2013) Program/erase speed, endurance, retention and disturbance characteristics of single-poly embedded Flash cells (University of Minnesota), IRPS, April 2013.
30 Song, S.H., Chun, K.C. and Kim, C.H. (2014) A bit-by-bit re-writable eFlash in a generic 65 nm logic process for moderate-density nonvolatile memory applications (University of Minnesota). *IEEE Journal of Solid-State Circuits*, 40 (8), 1861, August 2014.
31 Song, S.H., Chu, K.C. and Kim, C.H. (2013) A bit-by-bit re-writable eFlash in a generic logic process for moderate-density embedded non-volatile memory applications (University of Minnesota), Custom Integrated Circuits Conference, September 22, 2013.
32 Song, S.H., Kim, J. and Kim, C.H. (2014) A comparative study of single-poly embedded Flash memory disturbance, program/erase speed, endurance, and retention characteristic (University of Minnesota). *IEEE Trans. on Electron Devices*, 61 (11), 3737, November 2014.
33 Cummins, T., Byne, E., Brannick, D. and Dempsey, D.A. (1998) An IEEE 1451 standard transducer interface chip with 12-b ADC, two 12-b DACs, 10-kB Flash EEPROM, and 8-b microcontroller (Analog Devices). *IEEE Journal of Solid-State Circuits*, 33 (12), December 1998.
34 Yeh, B. (1993) Split gate single transistor nonvolatile electrically alterable memory device, US Patent 5 242 848, September 7, 1993.
35 Chu, Y.S. *et al.* (2011) Split-gate Flash memory for automotive embedded applications (TSMC), IRPS, April 10, 2011.
36 Shum, D. *et al.* (2012) Highly reliable Flash memory with self-aligned split-gate cell embedded into high performance 65 nm CMOS for automotive and smartcard applications (Infineon, TSMC), IMW, May 2012.
37 Cho, C.Y.-S., Wang, J.C., Lin, Y.F. *et al.* (2013) A 55 nm 0.86-volt operation, 125 MHz high speed, 75 μA/MHz low power, wide voltage supply range 2M-bit split-gate embedded Flash (TSMC), IMW, May 29, 2013.
38 Liu, Y.C. *et al.* (2013) An embedded Flash macro with sub-4-ns random-read-access using asymmetric-voltage-biased current-mode sensing scheme (NTHU, Shou University, TSMC), A-SSCC, November 11, 2013.
39 Tkachev, Y., Liu, X. and Kotov, A. (2012) Floating-gate corner-enhanced poly-to-poly tunneling in split-gate Flash memory cells (Silicon Storage Technology). *IEEE Trans. on Electron Devices*, 59 (1), January 2012.
40 Do, N. and Tkachev, Y. (2014) Common-floating gate test structure for separation of cycling-induced degradation components in split-gate Flash memory cells (Silicon Storage Tech), IEEE Conference on Microelectronic Test Structures, March 24, 2014.

41 Lee, Y.K. *et al.* (2014) A 45-nm logic compatible 4Mb-split-gate embedded Flash with 1M-cycling endurance (Samsung), IMW, May 2014.

42 Om'mani, H., Tadayoni, M., Thota, N. *et al.* (2013) A novel test structure to implement a programmable logic array using split-gate Flash memory cells (Microchip/SST), ICMTS, March 2013.

43 Park, D. and Kim, T. (2014) Built-in binary code inversion technique for on-chip Flash memory sense amplifier with reduced read current consumption (KAIST). *IEEE Trans. on VLSI Systems*, 22 (5), May 2014.

44 Markov, V. and Kotov, A. (2014) Program disturb induced by interface-trap-assisted field and thermal electron emission in the channel of split-gate memory cell (SST-Microchip). *IEEE Trans. on Development and Material Reliability*, 14 (2), June 2014.

45 Fang, L. *et al.* (2014) A highly reliable 2-bits/cell split-gate Flash memory cell with a new program – disturbs immune array configuration (Chinese Academy of Science, Huahong Grace Semi.). *IEEE Trans. on Electron Devices*, 61 (7), July 2014.

46 Zhang, S., Xiao, J., Yang, G. *et al.* (2014) A novel sourceline voltage compensation circuit and a wordline voltage-generating system for embedded NOR Flash memory (Chinese Academy of Sciences (CAS), GSMC, Shanghai Huahong). *IEEE Trans. on Circuits and Systems – II. Express Briefs*, 61 (9), 691, September 2014.

47 Martinie, S. *et al.* (2015) SPICE modeling of 55 nm embedded SuperFlash® technology 2T memory cells (CEA-LETI, Microchip Technology), ICMTS, March 2015.

48 Do, N. *et al.* (2015) A 55 nm logic-process-compatible, split-gate Flash memory array fully demonstrated at automotive temperature with high access speed and reliability (Silicon Storage Technology, a subsidiary of Microchip Technology), IMW, May 2015.

49 Zhang, S., Xiao, J., Yang, G. *et al.* (2014) A 1.35-V 16Mb twin-bit-cell virtual-ground-architecture embedded Flash memory with a sensing current protection technique (CAS, Shanghai Huahong, GSMC). *IEEE Trans. on Circuits and Systems – I: Regular Papers*, 61, 2862, October 2014.

50 Xu, Y. *et al.* (2016) Design techniques for a 30-ns access time 1.5-V 200K-bytes embedded EEPROM memory (Chinese Academy of Science). *IEEE Trans. on Circuits and Systems – II: Express Briefs*, 63 (11), March 2016.

51 Duncan, A.R. *et al.* (2016) Characterizing radiation and stress-induced degradation in an embedded split-gate NOR Flash memory (NAVSEA Crane). *IEEE Trans. on Nuclear Science*, 63 (2), 1276, 20 April 2016.

52 Do, N. (2016) eNVM technologies scaling outlook and emerging NVM technologies for embedded applications (Silicon Storage Technology/Microchip Technology), IMW, May 2016.

53 Markov, V. *et al.* (2017) Optimization of programming conditions and endurance enhancement of SuperFlash memory (Silicon Storage Technology subsidiary of Microchip Technology), IRPS, April 2017.

54 Lee, Y.K. (2016) Highly scalable 2nd generation 45-nm split-gate embedded Flash with 10-ns access time and 1M-cycling endurance (Samsung), IMW, May 2016.

55 Luo, L.Q. *et al.* (2016) Functionality demonstration of a high-density 2.5 V self-aligned split-gate NVM cell embedded Into 40 nm CMOS logic process for automotive microcontrollers (Globalfoundries, Singapore University of Technology, Microchip Technology), IMW, May 2016.

56 Markov, V. *et al.* (2016) SuperFlash scaling aspects: program disturb (SST Subsidiary of Microchip Technology), IMW, May 2016.

57 Prince, B. (2014) *Vertical 3D Memory Technologies*, John Wiley, 276 pp.
58 Huang, C.Y., Lin, H. and Wu, C.Y. (2014) High-voltage tolerant circuit design for fully CMOS compatible multiple-time programmable memories (National Chung Hsing University), ISCAS, June 1, 2014.
59 Sidense demonstrates working one-time programmable (OTP) bit cells in TSMC 16 nm FinFET technology, Sidense Press Release, September 4, 2014.
60 Hsieh, E.R. *et al.* (2015) The demonstration of low-cost and logic process fully-compatible OTP memory on advanced HKMG CMOS in a newly found dielectric fuse breakdown (NCTU, UMC), IEDM, December 2015.
61 Khan, F. *et al.* (2017) Charge trap transistor (CTT): An embedded fully logic-compatible multiple-time programmable non-volatile memory element for high-*k* metal gate CMOS technologies (IBM, UCLA). *IEEE Electron Device Letters*, 38 (1), January 2017.
62 Shao, L. *et al.* (2017) A reliability study of a new embedded Flash to reduce charge-loss issue (SMIC), China Semiconductor Technology Industry Conference (CSTIC), May 2017.
63 Hori, M. *et al.* (2013) Embedded FLOTOX Flash on ultra-low power 55 nm logic DDC platform (Fujitsu), IEDM, December 2013.
64 Fong, S., Ariyoshi, J. and Ema, T. (2012) Embedded Flash on a low-power 65 nm logic technology (Lattice Semi., Fujitsu), *IEEE Electron Device Letters*, 33 (9), 1261, September 2012.
65 Zambelli, C., Olivo, O., Koebernik, G. *et al.* (2013) Erratic bits classification for efficient repair strategies in automotive embedded Flash memories (University de Studi di Ferrara, Infineon), IRPS, April 14, 2013.
66 Agam, M. *et al.* (2015) Challenges in integrating embedded non volatile memory with floating poly and ONO in base line process (On Semi.), ASMC, May 2015.
67 Iwamoto, K. *et al.* (2014) Advanced 300 mm 0.13 μm BCD technology from 5 V to 80 V with highly reliable embedded Flash (Rohm Co., Genusion), International Symposium on Power Semiconductor Devices and ICs, June 15, 2014.
68 Ogura, T. *et al.* (2003) Embedded twin MONOS Flash memories with 4 ns and 15 ns fast access times (Halo LSI), VLSI Circuits Symposium, June 2003.
69 Tanaka, T., Tanakawa, H., Yamaki, T. and Umemoto, T. (2003) A 512B MONOS type Flash memory module embedded in a microcontroller (Hitachi), VLSI Circuits Symposium, June 2003.
70 Kim, J.H. *et al.* (2003) Highly Manufacturable SONOS non-volatile memory for the embedded SoC solution (Samsung), VLSI Technology Symposium, June 2003.
71 Cho, I.W. (2004) Full integration and characterization of localized ONO memory (LONOM) for embedded Flash technology (Samsung), Symposium of VLSI Technology, June 2004.
72 Seo, M.K. *et al.* (2004) A 0.9 V 66 MHz Access, 0.13 μm 8M (256I × 32) local SONOS embedded Flash EEPROM (Samsung), VLSI Circuits Symposium, June 2004.
73 Shukuri, S. *et al.* (2006) A 60 nm NOR Flash memory cell technology utilizing back bias assisted band-to-band tunneling induced hot-electron injection (B4-Flash) (Genusion, Tohoku University), VLSI Technology Symposium, June 2006.
74 Yoshida, J. (2012) Renesas, TSMC tout licensable MCU platform using 40-nm eFlash, *EE Times, May* 28, 2012.
75 Kono, T. *et al.* (2014) 40-nm embedded split-gate MONOS (SG-MONOS) Flash macros for automotive with 160-MHz random access for code and endurance over 10

M cycles for data at the junction temperature of 170 °C (Renesas), *IEEE Journal of Solid-State Circuits*, 49 (1), 154, January 2014.

76 Kono, T. *et al.* (2013) 40 nm embedded SG-MONOS Flash macros for automotive with 160 MHz random access for code and endurance over 10M cycles for data (Renesas), IEEE ISSCC, February 2013.

77 Kawashima, Y., Hashimoto, T. and Yamakawa, I. (2015) Investigation of the data retention mechanism and modeling for the high reliability embedded split-gate MONOS Flash memory (Renesas, Hitachi Research), IRPS, April 2015.

78 Renesas Electronics develops industry's first 28 nm embedded Flash memory technology for microcontrollers, Renesas Press Release, February 18, 2014.

79 Yamauchi, T., Kondo, H. and Nii, K. (2015) Automotive low power technology for IoT Society (Renesas), Symposium on VLSI Technology, June 2015.

80 Taito, Y. *et al.* (2015) A 28 nm embedded SG-MONOS Flash macro for automotive achieving 200 MHz read operation and 2.0MB/s write throughput at Tj of 170 °C (Renesas), ISSCC, February 2015.

81 Taito, Y. (2016) A 28 nm embedded split-gate MONOS (SG-MONOS) Flash macro for automotive achieving 6.4GB/s read throughput by 200 MHz no-wait read operation and 2.0MB/s write throughput at Tj of 170 °C (Renesas), *IEEE Journal of Solid-State Circuits*, 51 (1), 213, January 2016.

82 Mitani, H. *et al.* (2016) A 90nm embedded 1 T-MONOS Flash macro for automotive applications with 0.07 mJ/8kB rewrite energy and endurance over 100 M cycles under Tj of 175 °C (Renesas), ISSCC, February 2016.

83 Renesas Electronics develops 90 nm one-transistor MONOS Flash memory technology to accelerate intelligence in automotive control systems Renesas Press Release, February 3, 2016.

84 Tsuda, S. *et al.* (2016) First demonstration of FinFET split-gate MONOS for high-speed and highly-reliable embedded Flash in 16/14 nm-node and beyond (Renesas Electronics Corp., Hitachi), *IEEE IEDM*, December 2016.

85 Yamauchi, T. *et al.* (2016) Embedded Flash technology for automotive applications (Renesas), *IEEE IEDM*, December 2016.

86 Tehrani, S. *et al.* (2013) Advancement in charge-trap Flash memory technology (Spansion), IMW, May 2013.

87 Spansion and UMC announce joint technology development and licensing agreement companies will develop integrated logic and Flash memory to enable creation of high performance and low power electronics, Spansion Press Release, March 4, 2013.

88 Brown, P. (2014) UMC licenses embedded Flash IP from Cypress, *Electronics360*, July 15, 2014.

89 Wilson, R. (2014) UMC licenses 55 nm Flash memory for wearables, *Electronics Weekly*, July 15, 2014.

90 Ramkumar, K. *et al.* (2013) A scalable, low voltage, low cost SONOS memory technology for embedded NVM applications (Cypress, UMC), IMW, May 26, 2013.

91 Manners, D. (2014) Cypress and Shanghai Huali develop SONOS embedded Flash, *Electronics Weekly*, August 26, 2014.

92 Flash Memory Guide, Spansion web page www.spansion.com, December 2014.

93 Cypress licenses 40-nanometer embedded Flash IP to UMC, enabling next-generation MCUs, IoT and wearables applications (Cypress), Press Release, January 21, 2015.

94 Kwon, Y.J. *et al.* (2014) Characterization of 64kB test chip for touch application on 90 nm SONOS technology (SK Hynix, WingCore), IMW, May 18, 2014.

95 Lee, T.-H. *et al.* (2014) Charge trap length dependence and transconductance characteristics of a 2T SONOS cell (SK Hynix, Sogang University), NVMTS, October 2014.

96 Lin, P.-Y. *et al.* (2015) On-chip recovery operation for self-aligned nitride logic non-volatile memory cells in high-*k* metal gate CMOS technology (National Tsing Hua University, TSMC), *Journal of Electron Devices Society*, 3 (6), November 2015.

97 Chen, Y.J., Liu, C.J., Lo, C.Y. *et al.* (2011) A highly reliable embedded P-channel SONOS memory using dynamic programming method (eMemory Technology), IRPS, April 10, 2011.

98 Fang, S.W. *et al.* (2012) A dynamic programming method with programming current clamped in embedded P-channel SONOS Flash memory (eMemory Technology), IMW, May 2012.

99 Bartoli, J. *et al.* (2014) A new non-volatile memory cell based on the Flash architecture for embedded low energy applications: ATW (Asymmetrical Tunnel Window) (ST-Micro., Aix-Marseille University), CAS, October 2014.

100 Lo, R. *et al.* (2015) A study of blocking and tunnel oxide engineering on double-trapping (DT) BE-SONOS performance (Macronix, National Chiao Tung University), IMW, May 2015.

101 Sengupta, A. and Sarkar, C.K. (2012) Comparative study on nanocrystal embedded gate dielectric and oxide nitride oxide stack dielectric GAA MOSFET non-volatile memory devices (Jadavpur University), ICIEV, pp. 511, May 18, 2012.

102 Huang, X.D., Sin, J.K.O. and Lai, P.T. (2012) Nitrided La_2O_3 as charge-trapping layer for nonvolatile memory applications (University of Hong Kong, Hong Kong University of Science and Technology), *IEEE Trans. on Device and Material Reliability*, 12 (2), 306, June 2012.

103 Shih, C.H., Liang, J.T. and Luo, Y.X. (2012) Reading operation and cell scalability of nonvolatile Schottky barrier multibit charge-trapping memory cells (National Chi Nan University, TSMC, NTHU), *IEEE Trans. on Electron Devices*, 59 (6), 1599, June 2012.

104 Liu, L.J. *et al.* (2012) Enhanced programming and erasing speeds in P-channel charge-trapping Flash memory device with SiGe buried channel (NTHU, ITRI), *IEEE Electron Device Letters*, 33 (9), 1264, September 2012.

105 Park, J.K. *et al.* (2012) Origin of transient V_{th} shift after erase and its impact on 2D/3D structure charge trap Flash memory cell operations (KAIST, Hynix), IEDM, December 2012.

106 Seo, Y., Song, M.Y., An, H.M. and Kim, T.G. (2013) A CMOS-process-compatible ZnO-based charge-trap Flash memory (Korea University), *IEEE Electron Device Letters*, 34 (2), February 2013.

107 Bak, J. *et al.* (2014) Nonvolatile charge-trap memory transistors with top-gate structure using In–Ga–Zn–O active channel and ZnO charge-trap layer (Kyung-Hee University, Electronic and Telecom Research Institute), *IEEE Electron Device Letters*, 35 (3), March 2014.

108 Liu, Y. *et al.* (2014) Fabrication and characterization of 3D fin-channel MANOS type Flash memory (AIST, NIMS), SNW, June 2014.

109 Driussi, F. *et al.* (2014) Simulation study of the trapping properties of HfO_2-based charge-trap memory cells (University of Udine, MDM Lab IMM-CNR), *IEEE Trans. on Electron Devices*, 61 (6), June 2014.

110 Lin, C. *et al.* (2014) The nitrogen incorporated ZrO_2 charge trapping layers in nonvolatile memory applications (Chang Gung University), ICSICT, October 2014.
111 Kothandaraman, C. *et al.* (2015) Oxygen vacancy traps in Hi-K/metal gate technologies and their potential for embedded memory applications (IBM, UCLA), IRPS, April 2015.
112 Chindalore, G. *et al.* (2008) Embedded split-gate Flash memory with silicon nanocrystals for 90 nm and beyond (Freescale), VLSI Technology Symposium, June 2008.
113 Yater, J. *et al.* (2012) First-ever high-performance, low-power 32-bit microcontrollers with embedded nanocrystal Flash and enhanced EEPROM memories (Freescale), ICICDT, p. 1, May 30, 2012.
114 Kang, S.T. *et al.* (2012) High performance nanocrystal based embedded Flash microcontrollers with exceptional endurance and nanocrystal scaling capability (Freescale, GlobalFoundries), IMW, May 20, 2012.
115 Ban, T. *et al.* (2016) Charge effects of ultrafine FET with nanodot type floating gate (Ryukoku University, NIAIST, Tsukuba, Nara Institute of Science and Technology), AMF-PD, July 2016.
116 Singh, P., Li, C.G., Pitchappa, P. and Lee, C. (2013) Tantalum-nitride antifuse electro-mechanical OTP for embedded memory applications (A*STAR, University of Singapore), *IEEE Electron Device Letters*, 34 (8), 987, August 2013.
117 Takahashi, T. *et al.* (2012) 4kb nonvolatile nanogap memory (NGpM) with 1 ns programming capability (Funai Electron Advances in Applied Technology (FEAT) Research Institute, NI AIST), SNW, June 2012.

4

Thin Film Polymer and Flexible Memories

4.1 Overview

The requirements for processors with embedded memories are very low power and very low cost. For wearable devices, the circuitry must be flexible. Circuits that can be made without the expense of conventional semiconductor processing might lower the cost. For this reason a significant amount of effort has gone into polymer circuits with embedded memories which can be made by inkjet printing or screen printing. Embedded memories that can be made by printing techniques include: ferroelectric, charge trapping, and resistance RAM memories. These efforts produce low performance, but also low power, low cost, and flexible memories. Inkjet printing of large high volume substrates with ferroelectric memory is in production and polymer RRAM and nanocrystal memories are being developed. This technology is important for low cost RFID chips and other circuits without requirements for high performance.

For other IoT applications, such as wearable recreational and medical systems, the performance of polymer circuits is not sufficient. For these applications the performance of silicon chips is required. Technologies are also being developed for integrating systems of silicon chips on flexible substrates. Silicon chips can be transferred to flexible substrates using SOI base wafers to obtain the thin layer of silicon and thin chips can be obtained by using an underlying cavity. Integration of silicon chips on flexible substrates has also been achieved in the wafer fab by placing individual chips on a tape and doing wafer level molding to embed the die in a flexible substrate. This is followed by a redistribution wiring layer and bumping of the integrated chips in a technique called Fan-Out Wafer Level Packaging.

4.2 Organic Ferroelectric Memories

4.2.1 Characteristics and Features of Organic Ferroelectric Memories

Various universities and labs have investigated properties of copolymer and terpolymer ferroelectric memories in low cost flexible transparent circuitry. An a-InGaZnO TFT with ferroelectric poly(vinylidene fluoride-trifluoroethylene) (P(VDF-TrFE)) insulator was used for a flexible memory with low power operation and mechanical flexibility. The memory window was 3.2 V and the ON/OFF ratio was 1.5×10^6. A ZnO nanowire

Memories for the Intelligent Internet of Things, First Edition. Betty Prince and David Prince.
© 2018 John Wiley & Sons Ltd. Published 2018 by John Wiley & Sons Ltd.

P(VDF-TrFE) FeRAM on glass had a 16.5 V window, ON/OFF current ratio of 10^5, and gate leakage <300 pA. Retention was >10^4 s. A technique for determining data retention without destructive read was discussed. A graphene-ferroelectric hybrid device was investigated for use in multilevel memory systems.

A fast switching protocol for P(VDF-TrFE) copolymer ferroelectric RAM (FeRAM) thin films was discussed in February of 2013 by the East China University of Science and Technology [1]. One of the issues with P(VDF-TrFE) copolymer use in FeRAMs is the slow switching speed. The influence of the internal field on the switching speed was studied.

The polarization switching behaviors were considered with different pulse processes for P(VDF-TrFE) copolymer ultrathin films. A faster switching speed with imprint time was found for films with certain switching and imprint directions. It was found upon analysis that the effective fields and charge trap states in the ferroelectric layers were responsible for the switching behaviors. This result provided a method for designing an optimum protocol for FeRAM based on P(VDF-TrFE) copolymer ultrathin film. An illustration is shown in Figure 4.1 of (a) the applied write field and (b) the resulting imprinted field followed by (c) a switching write field.

The imprint time is the time to form internal domains in response to an externally applied field. It was found that the effective fields and charge trap states in the ferroelectric layers were responsible for the switching behaviors observed. This permitted designating a protocol for copolymer FeRAM.

Using P(VDF-TrFE-CFE) terpolymer for fast NV memory was discussed in February of 2013 by Nanjing University [2]. P(VDF-TrFE-CFE) terpolymer thin film together with a small amount of chlorofluoroethylene (CFE) was considered for thin film nonvolatile memory. One of the issues with P(VDF-TrFE) copolymer is the high operating voltage required to store charges due to the high activation energy required for nucleation and domain wall motion of anti-parallel domains and the slow polarization switching. The use of the terpolymer thin film with a small amount of CFE destabilized the long range ordered and strongly correlated ferroelectric domains of the copolymer. Illustrations of the domains of the copolymer and the terpolymer are shown in Figure 4.2.

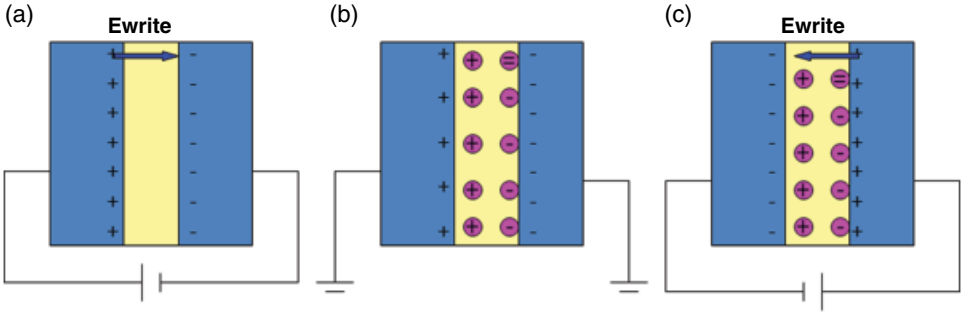


Figure 4.1 Illustration of different fields in P(VDF-TrFE) FeRAMs: (a) applied write field, (b) the resulting imprinted field, and (c) the switching write field. Based on Y. Hou (E. China University of S&T), *Applied Physics Letters*, 102 (6), 063507, February 2013 [1].

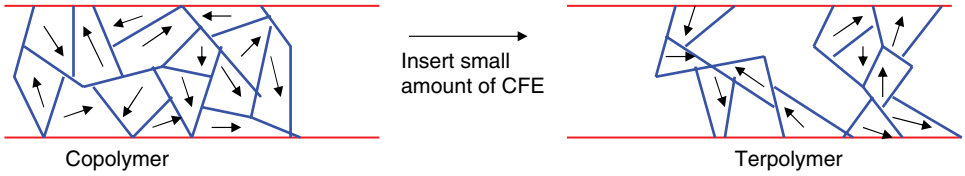


Figure 4.2 Illustrations of domains of copolymer and terpolymer. The terpolymer was obtained by inserting a small amount of chlorofluoroethylene (CFE) in the copolymer. Based on X. Chen *et al.* (Nanjing University), *Applied Physics Letters*, 102 (6), February 2013 [2].

Integration of the P(VDF-TrFE-CFE) terpolymer with a small amount of CFE can lead to a low voltage operation and a high read and write speed. A stack of Au/P(VDF-TrFE-CFE)/Au was used. The polarization switching voltage for a 50 nm thick film was down to 1 V. The writing–erasing procedure was reversible and the device had a high signal-to-noise ratio and high storage capacity. The polarization state of the terpolymer was stable, as required for a data storage device. The switching behavior was thought to be due to the reduced polar domain size and also to the energy cost of domain wall motion during electrically polarized switching decreases.

Determining data retention in thin P(VDF-TrFE) cells was discussed in July of 2013 by the University of Applied Science of Jena [3]. A technique for determining data retention in thin P(VDF-TrFE) cells and potentially for nondestructive readout of ferroelectric memory cells was discussed [4]. Data retention tests are normally based on the application of read pulses to a ferroelectric sample while the charge response is recorded. Since a read flips the cell to the opposite state continuous recording of retention after read is not possible.

A new approach was used to determine the state of a ferroelectric cell based on the non-destructive readout of the remnant polarization by measurement of small signal dielectric nonlinearities. The temporal development of the remnant polarization is directly accessible from the measured first and second harmonics in the current response to a small sinusoidal voltage signal. This technique has been used to study the retention of thin VDF-TrFE copolymer films of molar ratio 70/30 with thickness below 200 nm.

Nonvolatile memory using an InZnSiO active channel and ferroelectric gate insulator was discussed in July of 2013 by KyunHee University [4]. Thin film transistor nonvolatile memory characteristics were considered for a hybrid gate stack using a solution processed indium-zinc-silicon oxide channel and an organic Fe-(P(VDF-TrFE)) ferroelectric gate insulator. Device properties such as field-effect mobility, program speed, and retention were improved by controlling the InZnSiO channel composition of silicon. The use of silicon in the $InZnSi_xO$ channel layer modulated the carrier concentration and helped reduce defect density in the channel. It was found that the devices with InZnSiO of 2 mol% Si content showed the best performance. The subthreshold swing was 772 mV/decade, the memory window was 11.9 V, and the ON/OFF current ratio was $5.7 \times 10^{5.}$ By including the correct amount of silicon the trade-off between carrier concentration and defect densities in the channel were optimized along with the program speed, endurance, and data retention.

A graphene-ferroelectric device for multivalued memory systems was discussed in July of 2013 by the University of Texas Dallas [5]. A graphene-FE hybrid device was

considered for use as a multivalued memory system. The multilevel NVM system used a dual-gated single layer graphene FET with a polymer ferroelectric as the top-gate dielectric and a linear bottom-gate dielectric. The multiple memory states were represented by various levels of graphene channel resistance resulting from changing the doping type and the number of p–n junctions in graphene. This was accomplished by controlling the polarity of the domains in the ferroelectric thin film using a biased metal-coated atomic force microscope tip. A five level memory was demonstrated with resistance change between the lowest and highest state greater than 200%.

Switching and imprint for ferroelectric poly(vinylidene fluoride-trifluoroethylene) P(VDF-TrFE) copolymer was discussed in December of 2013 by the East China University [6]. High speed switching and improved imprint behaviors for poly(vinylidene fluoride-trifluoroetylene) copolymer ultrathin films were explored with various pulse and unipolar poling processes. The analyses showed that the properties were dependent on the direction of the switching and unipolar poling field. These results provide another method for designing an optimum protocol for FeRAM based on P(VDF-TrFE) copolymer ultrathin film.

FeRAM can be made of ZnO nanowire transistor (P(VDF-TrFE) FeRAM on glass substrate. In January of 2014, Cambridge University discussed a fast ferroelectric NVM based on a top gate ZnO nanowire (NW) transistor made on a glass substrate [7]. The ZnO NW channel was spin-coated with a ferroelectric poly(vinylidenefluoride-co-trifluoroethylene) P(VDF-TrFE) layer acting as a top-gate dielectric without a buffer layer. Electrical conductance modulation and memory hysteresis were done using a gate electric field-induced reversible electrical polarization switching of the P(VDF-TrFE) thin film. The device showed a 16.5 V memory window, a high drain current ON/OFF ratio of 10^5, a gate leakage current <300 pA, and excellent retention characteristics for over 10^4 s.

Temperature dependence of the imprint in P(VDF-TrFE) copolymer was discussed in March of 2014 by Shanghai University of Engineering Science. The temperature dependence of the imprint in P(VDF-TrFE) copolymer thin film capacitors was studied [8]. Characteristics of the imprint as a function of different temperatures and annealing processes were noted.

It appeared that the temperature-induced shift of the imprint rate was limited for all the investigated temperature conditions and that this low shift is primarily associated with the competition between the trap states increasing and the detrapping process induced by the temperature increasing. The anneal temperature-dependent imprint rates in the polymer chains have been analyzed and the annealed cell shows a low imprint rate after 10^4 switches at an annealing temperature above 100 °C. The internal electric fields for these processes were studied and it appeared that the effective fields and charge trap states in the ferroelectric layers were strongly responsible for the switching behaviors.

Flexible memory devices with amorphous-InGaZnO (a-InGaZnO) thin film memory transistor (TFT) and copolymer ferroelectric P(VDF-TrFE) gate insulator were discussed in April of 2014 by the Institute of Technology, Zurich. Future flexible electronic systems require memory devices that combine low power operation and mechanical flexibility [9]. Mechanically flexible a-InGaZnO memory TFT with a P(VDF-TrFE) gate insulator were investigated. A schematic cross-section of an IGZO memory TFT with a P(VDF-TrFE) gate insulator layer is shown in Figure 4.3.

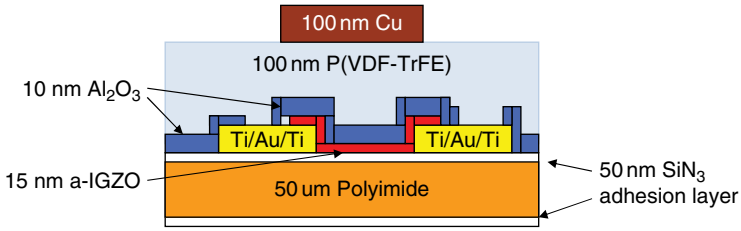


Figure 4.3 Schematic cross-section of a flexible a-InGaZnO memory TFT with a P(VDF-TrFE) gate insulator layer. Based on L. Petti *et al.* (IT Zurich), *IEEE Trans. on Electron Devices*, 61 (4), April 2014 [9].

Memory operation had a memory window of 3.2 V and an ON/OFF ratio of 1.5×10^6. The mobility of 8 $cm^2/(V\,s)$ and the ON/OFF current ratio of 10^7 were comparable with the values for reference TFTs made on the same substrate. In order to use a memory TFT in flexible applications, their behavior under mechanical strain must be understood. In this study, flexible memory and reference TFTs were characterized with bending radii down to 5.5 mm. Tensile strain caused negative threshold voltage shifts and increased the drain current while compressive strain resulted in the opposite effects.

Memory TFTs compared with reference TFTs show up to 8 times larger threshold voltage shifts and 17 times larger drain current variations. It was shown that the strain-dependent properties of a-IGZO can only explain the shifts observed in reference TFTs while the variations in memory TFTs are primarily caused by the piezoelectric properties of P(VDF-TrFE).

Memory operation was shown with a memory window of 3.2 V and a memory ON/OFF current ratio of 1.5×10^6 for a gate source voltage sweep of ± 6 V. Tensile strain caused a negatively shifted threshold voltage and an increased drain current. Compressive bending resulted in a positive threshold voltage shift and a decreased drain current. Shifts in the Fe-TFTs were caused by a superposition of the strain-induced dependence of the polarization in the piezoelectric P(VDF-TrFE) and the strain-dependent IGZO properties. These shifts should be considered when designing future flexible memories. These devices had full memory operation down to a bending radii of 5.5 mm, which makes them applicable to use as flexible memories.

In September of 2014 the University Catholique de Louvain discussed the fabrication of an organic FeFET using a ferroelectric gate dielectric made of nanostripes, which are formed by nanoimprinting P(VDF-TrFE) over a layer of semiconducting poly(triarylamine) [10]. The imprinting process results in the ferroelectric having a decreased switching voltage by a factor of about 1.5. This in turn results in a decreased operating voltage for the nanostripe material compared to a reference FeFET with a continuous ferroelectric layer. The transistor consists of a large number of nanostripe-gated transistors, which are in parallel. This offers the potential for a significant size reduction in organic FeFETS.

Flexible 2T2C FeRAM can be made with TFT and P(VDF-TrFE) copolymer. In October of 2014, Micron Technology and the University of Texas, Dallas discussed a 2T2C FeRAM in a low temperature fully integrated process for an FeRAM memory array for flexible electronics applications [11]. The cell used cadmium sulfide as the semiconductor material for n-channel TFT transistors and P(VDF-TrFE) copolymer as the ferroelectric material. At VDD = 5 V a voltage difference of about 1 volt between the

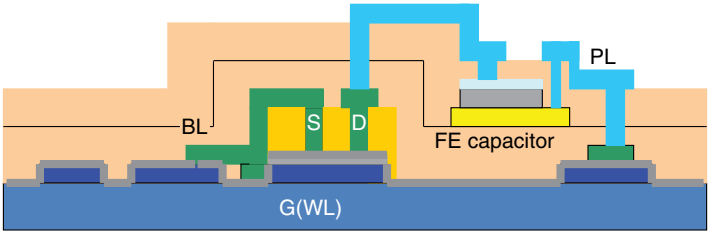


Figure 4.4 Cross-sectional view of low temperature process for 2T2C FeRAM array. Based on D. Mao *et al.* (Micron., University of Texas), *IEEE Trans. on Electron Devices*, 61 (10), 3442, October 2014 [11].

high and low states is achieved at the output of the sense amplifier. Both of the materials used in this memory are compatible with low temperature processing while maintaining good device performance. The low temperature processes have TFTs with high mobility of 9.3 $cm^2/(V\,s)$, ON/OFF ratio of 10^8 at 20 V and 0 V, and ferroelectric capacitors with high switching polarization. The low temperature processed FeRAM was based on 2T2C architecture using P(VDF-TrFE) for the capacitor and CdS for the TFT. Small FeRAM arrays of 4-bit, 16-bit, and 64-bit capacity have been made and a 1 V voltage difference output from the sense amp has been achieved on a 4-bit FeRAM array. The array is patterned using photolithographic processes compatible with flexible electronics system integration. A cross-sectional view 2T2C FeRAM array process is shown in Figure 4.4.

The FeRAM array is made on a silicon substrate with 500 nm SiO_2 on top for electrical isolation. The TFT is made using CdS as the semiconductor and FE-capacitor of P(VDF-TrFE) copolymer. The highest temperature in the process is the TFT anneal at 300 °C.

P(VDF-TrFE) copolymer organic FeFET and FeRAM characterization were discussed in October of 2014 by Slovak University of Technology [12]. The MIM structures were used to investigate the ferroelectric properties of FeFET and FeRAM. The I–V and C–V curves representing characteristics were analyzed. The MIM device fabrication and characterization was described including saturated polarization, maximum electric field intensity, critical electric field, and relative dielectric constant.

4.2.2 Printable Ferroelectric Embedded Memories

Flexible electronics has potential for use in rollable and foldable displays, smart patches and smart packaging on paper and plastic substrates. Printing sensors and transistors is an attractive idea since large area electronics can be fabricated roll-to-roll at low temperatures on plastic substrates. Due to the limitations of fully printed electronics systems in performance, today there is focus on both a fully printed electronics system and on integration of large area flexible electronics with chips. Most of the interest is in inexpensive plastic substrates although there is still some interest in metal foil and glass [13].

There has been progress in printed sensor elements, printed logic, and some printed logic and memory built with complementary n and p polymer semiconductors. Inkjet printing has been used for prototyping since short runs are possible with immediate turnaround. Many printing techniques are compatible with large area or roll-to-roll processing. Printing is valuable for those applications with a large area and for which only moderate complexity is required.

In December of 2012, Thin Film Electronics demonstrated a prototype label that replaced a disposable printed electronic tag with a low cost printed sensor system [14–16]. These tags could be used to replace bar codes in the Internet of Things". The printed system included memory and sensor, and logic was designed to detect whether temperature thresholds had been exceeded. The data was recorded digitally in a non-volatile memory for later retrieval. This technology is expected to compete with active RFID tags that have rewritable memory.

The prototype of an integrated printed electronic tag was based on rewritable ferroelectric memory. These labels could be used for tracking shipments of pharmaceuticals and perishable foods. These tags could be used to replace bar codes in the "Internet of things". One application combines a tracking tag with a temperature sensor so a shipped package with perishable goods could be confirmed to have stayed within the correct temperature range during shipment. The circuit was printed with a technology similar to the inkjet printer, which sprays electronic circuitry on to thin film surfaces. It included the electronics for memory, sensors, such as temperature sensors, and a processor. It was expected to be a few years before the system is in mass production. The tags would be at a lower cost than current RFID tags. The prototype was built by PARC, Thin Film, and PST Sensors, who provided the printed thermistor, and Acreo, who provided the electrochromic display.

The tag has multiple electronics functions including memory, logic, and temperature sensor. The devices worked together by sensing temperature thresholds and writing to memory. A polymer ferroelectric memory is used in these tags. The data triggers a display through external circuitry.

Integrating memory, logic, sensors, and displays using printed circuitry is critical for delivery of cost-effective mass produced printed electronic devices. Additional circuitry, including a timer function and wireless communication, were expected to be added to the system. The technology was expected to compete with active RFID tags that have rewritable memory.

In June of 2013, various French labs also demonstrated the required components of organic logic and complementary memory circuits [17]. The circuits were made with organic N-type and P-type transistors. Complementary NAND and NOR gates were made and characterized. An SRAM was made and an edge triggered flip-flop was made of six organic two input and three input NAND gates consisting of 26 organic transistors with an area of 170 mm^2. The maximum operating frequency of the flip-flop was 220 Hz with a ±20 V power supply. The circuits were made using an organic sheet-to-sheet process at ambient temperature. Electrical characteristics were stable. A model was developed to represent the circuit performance. Target application was low cost RFID circuits replacing barcodes. Sheet-to-sheet printing was used since it is low cost and has good alignment accuracy.

Complementary CMOS-like ICs on plastic foils have historically been difficult to make due to lack of air stable N-type organic semiconductors. The NOR and NAND gates, in this case, were made using both N-type and P-type organic transistors. They had full rail-to-rail behavior with only a small offset due to the P-type having a higher mobility than the N-type. The use of both N- and P-type transistors permitted rail-to-rail curves with no current consumption except during switching. The I_{on}/I_{off} ratio was $>10^{5.}$ No vacuum process was required during the sheet-to-sheet printing. The circuit worked with a ±20 V power supply and ran at 220 Hz. An SRAM was also made with six

organic transistors using the same manufacturing process as the rest of the sheet. The transient characteristics of the SRAM were determined. A schematic cross-section of the screen-printed complementary gate is shown in Figure 4.5.

The CMOS semiconductor devices were made on a 125 μm thick polyethylene naphthalate substrate. A 30 nm Au source/drain layer served as a first level connection and was formed by laser ablation. The P-type transistor was a polytriarylamine derivation and the N-type was acene-based, diimide. Together they were 100 nm thick. An 800 mm thick dielectric fluoropolymer was screen printed and annealed. Then the gate layer was screen printed using a 5 μm thick silver paste. This gate layer also served as a second connection layer. A final annealing was done at 100 °C. Process alignment was around ±25 μm. The process was performed in air.

The 6 T SRAM was made of the organic transistors as shown in the circuit schematic in Figure 4.6. To perform a write, the WL is set to VDD, then the BL is set to VDD, and BL\ is set to VSS. The point B is then pulled to VDD. When the WL is grounded, the SRAM is isolated with the data stored in it. This circuitry can be used to form an operational RFID tag.

An 8b general purpose MCU was made in a hybrid oxide-organic complementary thin film technology. This 8b MCU in hybrid oxide-organic thin film was discussed in February of 2014 by IMEC, KU Leuven, Panasonic, Evonik, and Holst [18]. The n-type transistors were based on a solution-processed n-type MOS and the p-type transistors used an organic semiconductor. The high mobility n-type semiconductor and the complementary logic permitted a greater than 50 times speed improvement. The robustness of the design permitted a complete standard cell library. The MPU was in two parts: a processor core chip and an instruction generator. The instruction generator has instructions stored in a write-once-read-many (WORM) memory, which is formatted using a inkjet printing step to produce a print-programmable ROM or PPROM. The entire printing process was performed at temperatures below 250 °C. A block diagram of the PPROM is shown in Figure 4.7.

The hybrid oxide-organic complementary processor had a core transistor count of 3504 and the PPROM transistor count was 852. The area was $1.2 \times 1.86\,cm^2$. Technology was 5 μm feature size p-type Pentacene with n-type metal oxide. Performance at V_{min} of 6.5 V was 2.1 kHz at 12 V VDD. Mobility was $-0.15\,cm^2/(V\,s)$ (p-type) to $20\,cm^2/(V\,s)$ (n-type).

In January of 2015, Xerox and Thinfilm announced a deal to mass-produce their thin film "printed electronics". Under the agreement, Thinfilm would be permitted to print

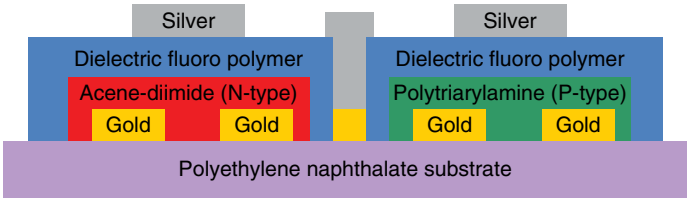


Figure 4.5 A schematic cross-section of the screen-printed complementary gate. Based on M. Guerin *et al.* (Aix-Marseille U., IM2NP, CNRS, IM2NP,LTEN-LCEI, CEA-Grenoble), *IEEE Trans. on Electron Devices*, 60 (6), June 2013 [17].

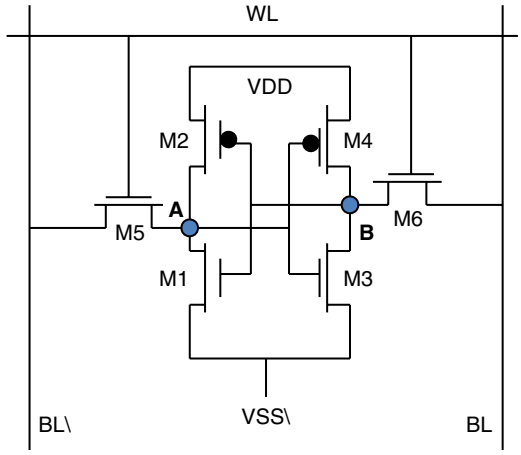


Figure 4.6 6T SRAM of organic transistors. Based on M. Guerin *et al.* (Aix-Marseille University, IM2NP, CNRS, IM2NP,LTEN-LCEI, CEA-Grenoble), *IEEE Trans. on Electron Devices*, 60 (6), June 2013 [17].

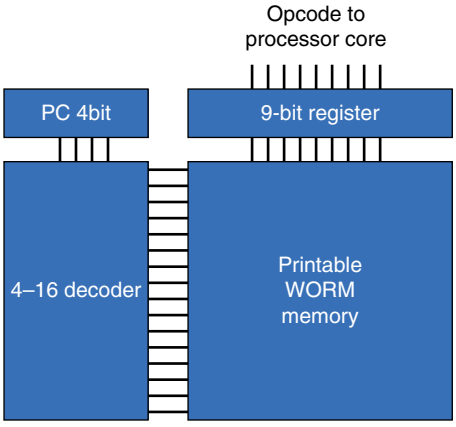


Figure 4.7 Block diagram of print programmable ROM. Based on K. Myny *et al.* (IMEC, Panasonic, KU Leuven, Evonik, Holst), ISSCC, February 2014 [18].

up to a billion chips a year for various smart connected devices such as processors, memory, and sensors for the Internet of Things [19]. The smart labels for packages were expected to sense and store data for a tenth to a hundredth of the cost of conventional electronics. Options could be a temperature sensor printed on a smart label that can indicate the maximum temperature attained. Near field communications (NCF) technology could be used to read the sensors and transfer the data to the retailer using an Internet cloud. Smart labels can also be attached to perishable food or medicines.

In September of 2015, Xerox and Thinfilm discussed bringing printed memory to market [20]. Two types of printed memory were considered with or without cryptographic security. The Xerox printed memory a the time was a small flexible label with 4 to 36 bits. It did not require a battery or a power source but did require a reading device to extract the data. Key applications considered were smart consumables, expensive fashions, and brand protection. The plan was to make the Xerox printed memory in Xerox's Webster, New York, plant. The reader was still under development.

Labels without memory could already be printed on to plastic, which could be pasted on to a consumer product as a smart label, as shown in Figure 4.8. The label can be scanned with a smartphone using an application that permits reading of the QR code on the label.

Figure 4.8 Illustration of a bottle of wine with a barcode, QR Code, and ID number, which permits the consumer to use a smartphone to ensure the authenticity of the wine and gain information about it. Photo by B. Prince.

In September of 2015, Xerox launched printed memory products intended to combat counterfeiting of branded and governmental products [21]. These labels could collect and store information about the authenticity and condition of products in 36 rewritable bits. A second product, printed memory with cryptographic security, included a unique encrypted printed code in the memory, which can only be read by an authorized person with a reader that could interface with a secure smartphone application. The cryptographic security feature was developed by Xerox PARC and is an added feature in the uniquely encoded printed memory label, which could ensure the integrity of a product from the factory to the consumer.

In January of 2016, Xerox announced it was setting up a 1.3 billion capacity smart label print line in Webster, New York [22]. Operating at full speed, this line was expected to have an annual output of over 1 billion smart labels. These labels are intended to be sold by Xerox as part of its diversification into the printed electronics market.

In June of 2016, Thin Film Electronics announced it was modifying a production line in Webster, New York, to produce memory labels [23]. These labels can add low cost intelligence to high end objects like wine bottles or perishable foods. They can store up to 36 bits of information, which enables 68 billion distinct data combinations. The label is written and read off-line and the reader needs to make physical contact. Data retention is up to 10 years. The label includes cryptographic capability and is combined with bar code technology. These are printed on the substrate. The data can be written and read using a contact reader that interfaces with a smartphone.

In March of 2017, Thin Film Electronics was making an NFC smart label that was printed and put on a bottle of Johnnie Walker Blue Label scotch [24]. This label permits communication with the consumer. The label authenticates the bottle, shows if it is counterfeit, and, using NFC, it can connect the consumer's smartphone

TFT	A-Si	MO	SAM organic	Ink-jet organic	CNT
Temperature (°C)	250	<100	<100	25	<150
Process	Litho	Roll-to-roll	Shadow mask	Ink jet print	Shadow mask
Feature size (μm)	8	5	50	50	50
Device type	N-type	N-type	N/P-type	P-Type	N/P Type
Supply voltage (V)	20	20	2	40	4
Mobility (cm^2/Vs)	1	10	0.01/0.5	0.01	25

Figure 4.9 Characteristics of the various TFT technologies including: amorphous silicon, metal oxide RRAM, self-assembly monolayer (SAM) organic and inkjet organic. Based on T.C. Huang *et al.* (Hewlett Packard, Hong Kong University.of S&T), NOCS, October 2016 [25].

to a web page with personalized promotion or a direct line to the company. The solution was in field trials in 2017. Another use of a printed smart label was a web site that sells beer to local microbreweries in the San Francisco Bay Area. The Thinfilm smart label is put on the beer bottle. The consumer taps an NFC smartphone on the label and it takes you to a video that describes the beer. The smart label would permit the beer company to interact with the customers directly rather than through expensive ads.

In October of 2016, Hewlett Packard and Hong Kong University of S&T discussed a summary overview of the design of printed circuits using thin film transistors (TFTs) [25]. A new design style, called Pseudo CMOS, was discussed, which could handle several design challenges of TFTs. Examples of Pseudo CMOS circuits designed for various applications were shown. These applications included: energy, healthcare, biomedical, and near field communication (NFC) tags.

Characteristics of the various other TFT technologies discussed and their characteristics available for printed circuits are shown in Figure 4.9. The various TFT technologies used processes that included: lithography, shadow mask, inkjet printing, and shadow mask lithography. Feature size ranged from 8 μm for amorphous silicon to 50 μm for carbon nanotubes. Supply voltage went as low as 2 V and mobility from 1 cm^2/(V s) for amorphous silicon to 25 cm^2/(V s) for carbon nanotubes.

4.2.3 IoT Applications of Thin Film Ferroelectric Memory

In June of 2015, Kyung Hee University and the Electronics and Telecom Research Institute, Daijeon, discussed a liquid crystal display pixel architecture using two oxide thin film transistors (TFTs), two ferroelectric memory TFTs, and two capacitors, as shown in Figure 4.10 [26]. Both TFTs used InGaZnO film as an active layer. The ferroelectric MTFT showed hysteresis due to the ferroelectric gate insulator permitting it to operate as a memory. A low refresh rate was achieved by using oxide TFTs with extremely low leakage currents; however, there was an issue with flicker due to the current leakage through the liquid crystal and the flexoelectric effect of the liquid crystals.

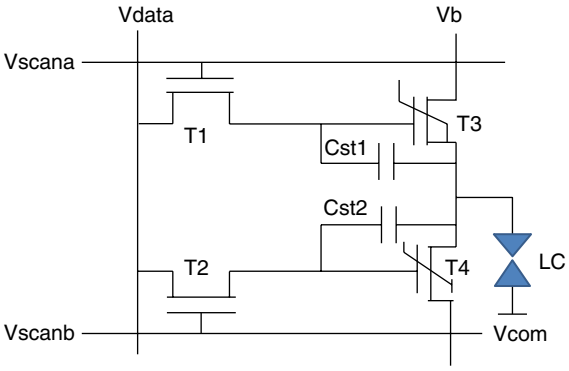


Figure 4.10 Schematic circuit diagram of liquid crystal display pixel architecture using two TFTs, two ferroelectric TFTs, and two capacitors. Based on S.H. Lee *et al.* (Kyung Hee University ETRI, Daijeon), *IEEE Electron Device Letters*, 36 (6), 585, June 2015 [26].

A 6 × 5 pixel LCD was fabricated. The memory window of the Ferroelectric MTFT was 5 V with a gate-voltage sweep from −20 to +20 V. After programming the memory, the LCD operated at a 0.5 Hz refresh rate. The ferroelectric MTFT maintained the on or off states. These states selected voltages to display black or white pixels, thereby maintaining flicker-free images regardless of LC leakage.

In October of 2015, Kobe University and ROHM discussed a ferroelectric-based nonvolatile flip-flop (nvFF) for use in wearable healthcare systems [27]. In this nvFF, coupled FE capacitors with complementary data storage were developed. By using complementary stored data in coupled ferroelectric capacitors, an 88% capacitor size reduction was achieved while still maintaining a wide read voltage margin of 240 mV at 1.5 V. This resulted in 2.4 pJ of energy with 10 year 85 °C data retention. The access speed for the ferroelectric capacitors could be changed according to the required retention time and was 1.6 μs for 10 year data retention and 170 ns for 10 hour data retention. A short data retention is okay for power gating applications.

The resulting nvFF was used in a 32-bit CPU in a wearable electrocardiogram (ECG) sensor LSI. The 32-bit CPU core had an NVFF and a 16 KB ferroelectric-based NVRAM for instruction and data caches. Since the frequency range of vital signals is low, both standby power reduction and sleep time maximization were important for reducing system level power. The standby current could be cut when the CPU core transited to the deep sleep state. In deep sleep, the data in the memory and the register values of the CPU core in the nvFF were stored sequentially in the ferroelectric capacitors. Thus 87% of the total power dissipation during measurement of the heart rate was reduced, with 64% area overhead using a 130 nm CMOS process and PZT thin films for the ferroelectric nvFF.

In January of 2016, Kyung Hee University and the Electronics and Telecom Research Institute discussed a new circuit architecture using NVM thin film IGZO transistors and oxide-based TFT for low power consumption and small size for use in dynamic driving circuit applications [28]. The top gate used ferroelectric poly(vinylidene flouride-triflouroethylene) gate insulator along with Al_2O_3/InGaZnO active layers. The memory device characteristics were confirmed for the fabricated nonvolatile memory TFT. The feasibility of a low power display driver circuit using embedded nonvolatile memory TFTs was confirmed.

4.3 Polymer Ferroelectric Tunnel Junctions

Ultrathin $BaTiO_3$ ferroelectric tunnel junctions were studied in 2015 by the National University of Singapore and Trinity College Dublin using interface engineering [29]. States were changed using voltage in ferroelectric tunnel junctions (FTJs) since this permits lowering the switching energy of the memory. Enhanced tunneling electroresistance in FTJs can be achieved by using asymmetric electrodes or by introducing metal–insulator transition layers. The effect of each interface on FTJ performance was studied using varying asymmetries of the electrode and ferroelectric interfaces. High tunneling electroresistance (TER) of about 400% could be achieved with $BaTiO_3$ layer thicknesses down to about 0.8 nm. The results showed that band offsets at each interface in the FTJs control the TER ratio. The off-state resistance (Roff) was found to increase much more rapidly with the number of interfaces than the on-state resistance (Ron). These results were found promising for future low energy memories.

Polymer ferroelectric tunnel electroresistance was discussed by CAS, University of Paris Saclay, University of Paris South, and East China Normal University in May of 2016 [30]. The application being considered was organic electronics for large area applications such as rollable displays or electronic paper that requires a simple, low power organic memory that can be written, erased, and read electrically. The NVM used ferroelectric polarization of an organic tunnel barrier. The device encoded the stored information and set the read out tunnel current. Piezoresponse force microscopy showed that films as thin as one or two layers of ferroelectric poly(vinylidene fluoride) (PVDF) remained switchable with low voltages. Submicrometer technology junctions based on these films displayed tunnel electroresistance (TER), reaching 1000% at room temperature. This was explained by ferroelectric switching and by electrostatic effects of direct tunneling.

An opportunity was provided to develop low-cost, large-scale arrays of organic ferroelectric tunnel junctions on silicon or flexible substrates. The electronic transport mechanism was more fully explored by observing room temperature TER in organic FTJs using ultrathin films. It was demonstrated that the ferroelectric switching could be sustained down to one layer and that charge transport across monolayers and bilayers of PVDF-based FTJs proceeds by direct quantum mechanical tunneling transport.

4.4 Types and Characteristics of Polymer Resistive RAMs with Flexible Substrate

4.4.1 Overview of Polymer Resistive RAMs with Flexible Substrate

RRAMs can be made with low cost materials and simplified manufacturing procedures. Use of the RRAM for IoT applications is discussed in terms of process variation produced performance and yield limitations. Flexible polymer RRAMs are being developed based on Parylene-C, which have good chemical stability and CMOS process compatibility. Copper atom switches can be made with a polymer solid electrolyte and with chalcogenides and exhibit a high ON/OFF current ratio. Inorganic thin film RRAMS have the potential for integration on flexible substrates with a low voltage read and multilevel operation.

4.4.2 Parylene-C-Based Resistive RAM

Parylene-C based RRAMs in an 8 × 8 cross-point array were discussed in December of 2012 by Peking University and Princeton University [31]. They discussed the resistive switching behavior of a single component polymer resistive memory using polychloro-*para*-xylylene (parylene-C) [31]. A sandwiched structure of parylene-C was made as an 8 × 8 crosspoint array of RRAMs integrated in a CMOS back-end process. Parylene-C was expected to be useful in the future for transparent, low cost, flexible NV memory applications. The ON/OFF current ratio of the RRAMs was 10^7. SET/RESET speed was in the nanosecond range and switching voltages were low. Retention was good. The switching mechanism was studied using Al, Cu, and Ag as active electrodes and Pt as the inert electrode. The read voltage was ±0.2 V. Ion was 10^{-4} and I_{off} was 10^{-30}. Cell size was 25 μm^2. The ON/OFF current ratio was about 10^7. The currents at different read voltages showed a distinct difference between the ON and OFF states. For V(read) of 0.2 V, I(on) was 10^{-4} A and I(off) was 10^{-30} A. For V(read) of –0.2 V, I(on) was 10^{-3} A and I(off) was 10^{-30} A.

The RRAM showed good cycling behavior and retention. The resistive switching of the device was attributed to the formation and annihilation of a metallic filament. Since this material can tolerate conventional photolithography and is compatible with a CMOS back-end process, the device was considered as a potential for hybrid integration with CMOS.

An RRAM with sub-20 nA reset current was discussed in February 2013 by Peking University [32]. The organic RRAM was based on double layer parylene-C technology. These devices were fabricated and showed stable bipolar resistive switching behavior and good data retention. Reset current was <20 nA and set current was 0.15 μA. The switching ratio was nearly 10^5 times that of single layer parylene-C cells. A schematic view of the Al/parylene-C/parylene-C/W cell is shown in Figure 4.11.

The bottom electrode (BE) was sputtered W film on a silicon substrate. Parylene-C film was deposited by room temperature polymer CVD. After an 18 hour exposure to air a thicker parylene-C film was deposited. An Al top film electrode was deposited and patterned. Bias is put on the top electrode and the bottom electrode is grounded.

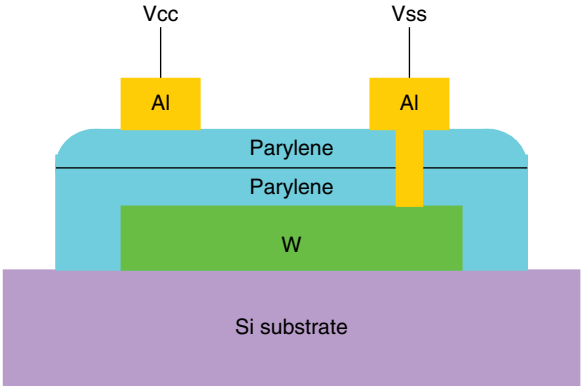


Figure 4.11 Schematic cross-section of Al/parylene-C/parylene-C/W cell. Based on W. Bai *et al.* (Peking University), *IEEE Electron Device Letters*, 34 (2), February 2013 [32].

The resistive switching mechanism for the single layer parylene-C device was interpreted as the formation and rupture of metallic filaments. The double layer parylene-C with thin parylene-C layer showed a high resistance state (HRS) current due to Poole-Frenkel emission. The low resistance state (LRS) current was attributed to traps in dielectrics of very thin parylene-film. The thin layer showed a negative temperature coefficient as opposed to the positive temperature coefficient in single layer metallic filaments. Hopping conduction was also shown in the double layer device, leading to a large internal resistance and a resulting small self-compliance SET current. The resistance ratio (R_{hrs}/R_{lrs}) > 10^4. The device showed good retention for 10^4 s at 85 °C although degradation of the resistance ratio and retention ability was found at 210 °C.

Low cost, low power flexible parylene-C RRAM was discussed in June of 2014 by IM Peking University, who studied organic resistance memories for future flexible applications [33]. The focus was on a polymer resistance memory device based on parylene-C. This device had good chemical stability and CMOS process compatibility. It reduced operational current as needed for information storage in many flexible systems.

The switching layer of flexible RRAM can be organic, inorganic, or organic–inorganic. Organic RAMs can be based on various organic materials such as polymers or polymer composites containing nanoparticles. In the parylene-C RRAM, the device was made on an SiO_2/Si substrate or plastic substrate in order to be flexible. The parylene-C layer is deposited by CVD. This process could be integrated into the BEOL in CMOS technology without contamination. The RRAM device had good NVM characteristics with low switching voltages and a high ON/OFF current ratio larger than 10^7 at 0.01 V read. The *I–V* curve of a typical Al/parylene-C/W device is shown in Figure 4.12.

This device can be made by a conventional photolithography process. It can switch in nanoseconds and has a high ON/OFF current ratio greater than 10^7. A double layer parylene RRAM was proposed for future investigation. A parylene-C RRAM device was shown to be flexible, have low cost and low power, while remaining CMOS compatible.

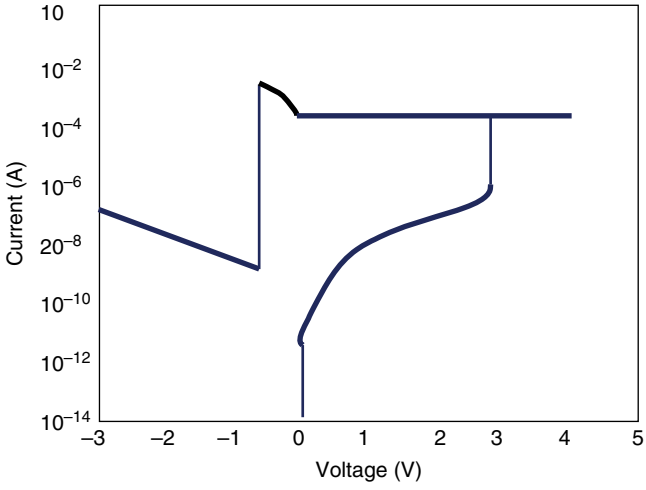


Figure 4.12 Illustration of *I–V* curve of typical Al/parylene-C/W RRAM device. Based on R. Huang *et al.* (Peking University), ISCAS, June 2014 [33].

4.4.3 Cu Atom Switches

The conduction mechanism in a Cu atom switch using a polymer solid electrolyte was discussed by LEAP in December of 2012 [34]. LEAP discussed low temperature characterization of electrical transport in an atom switch, which is also called a conductive bridge RAM. Atom switches and RRAMs both consist of metal–insulator–metal (MIM) structures. For the atom switch, the insulator works as a solid electrolyte that is sandwiched by an oxidizable electrode, of either copper or silver, and an inert electrode, either Ru or Pt. Chalcogenides such as GeSE, Cu_2S, and Ag_2S have also been used as the electrolyte for an atom switch.

The low resistive ON state with resistance <400 ohms shows metallic conduction along with high residual resistance. In the high resistive OFF state with resistance $>10^8$ ohms, the resistance exponentially increases with decreasing temperature due to Poole-Frenkel conduction. In the intermediate resistance range of about 10^5 to 10^7 ohms, the resistance has a small temperature dependence since the electron tunneling is due to Cu residues in a solid electrolyte and is dominant. Metal oxides such as Ta_2O_5 and TaSiO achieve a reliable OFF state and a high ON/OFF conductance ratio, as needed for programmable switches in reconfigurable LSIs.

The polymer solid electrolyte (PSE) enables the complete collection of the Cu residues without degrading the electrolyte. This results in forming free operation and a high ON/OFF conductance ratio. In the atom switch, the resistance state reversibly switches between dielectric and metallic. Complete collection of the Cu residue from the solid electrolyte was shown to be key for forming free operation and a high ON/OFF conductance ratio.

In September of 2015, LEAP discussed the relationship between speed and low voltage programming of a Cu atom switch [35]. The effect was studied of the composition in the $Al_{1-x}Ti_xO_y$ buffer layer, which is sandwiched between the Ru and Cu electrodes below the polymer solid electrolyte (PSE). A schematic cross-section of the Cu atom switch stack is shown in Figure 4.13.

To improve the voltage dependence of the switching slope (SS), which determines the time-to-on-state, the Cu ionization rate was increased. This was done by changing the composition ratio of the $Al_{1-x}Ti_xO_y$ buffer and PSE to attain a steep SS of 56 mV/decade by eliminating metallic Al residue on the Cu electrode. The operation of an atom switch, a type of conducting bridge RRAM, is illustrated in Figure 4.14.

Figure 4.14 shows two electrodes with the bottom electrode made of Cu and the top electrode made of Ru alloy sandwiching an AlTiO layer. When a positive voltage is placed on the Cu electrode and the Ru is grounded, a conducting Cu filament forms in

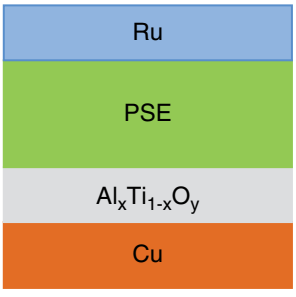


Figure 4.13 Copper atom switch stack with PSE and $Al_{1-x}Ti_xO_y$ layers sandwiched between Ru and Cu electrodes. Based on N. Banno *et al.* (LEAP), *IEEE Trans. on Electron Devices*, 62 (9), September 2015 [35].

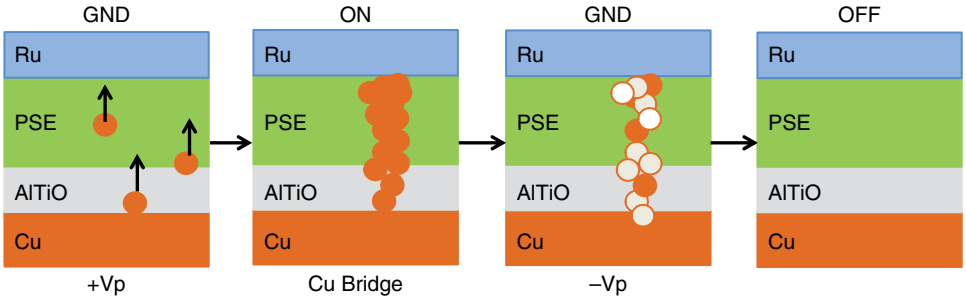


Figure 4.14 Atom switch physical operating mechanism. Based on N. Banno *et al.* (LEAP), *IEEE Trans. on Electron Devices*, 62 (9), September 2015 [35].

the AlTiO between the two electrodes. This is the ON state. If the voltage is reversed, the filament dissolves back into the PSE.

This buffer has 10 ns 2 V programming in a 1 Mb array. Cycle endurance was greater than 10^3 cycles with an ON/OFF resistance ratio $>10^4$. The steep SS is required for conducting bridges used in low power nonvolatile field-programmable gate arrays (FPGA). The switch was demonstrated in an FPGA, which is an array that can change its circuit configuration by programming switches and their connecting logic blocks.

The two terminal Cu atom switch (CAS) is used with a control terminal. The CAS is made up of two atom switches connected in the opposite direction to each other through a control gate. T1 and T2 are for signal transfers and the control gate (C) is for programming. Figure 4.15 illustrates the signal current for the switch in the (a) ON state and in the (b) OFF state.

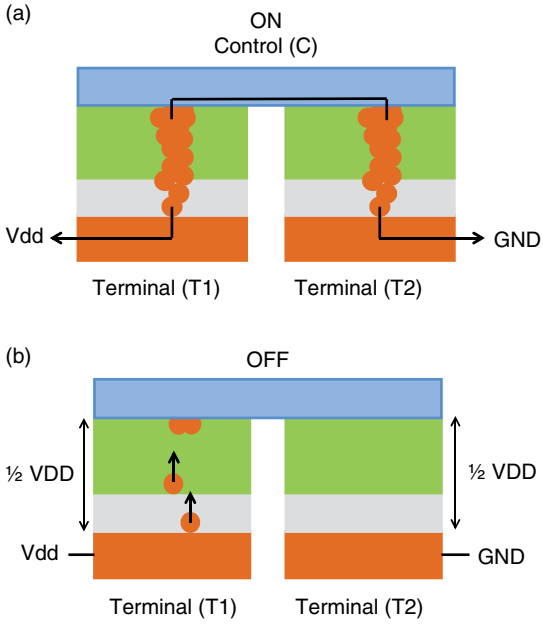


Figure 4.15 Signal current for two terminal Cu atom switches (CASs) in (a) ON state and (b) OFF state. Based on N. Banno *et al.* (LEAP), *IEEE Trans.* on *Electron Devices*, 62 (9), September 2015 [35].

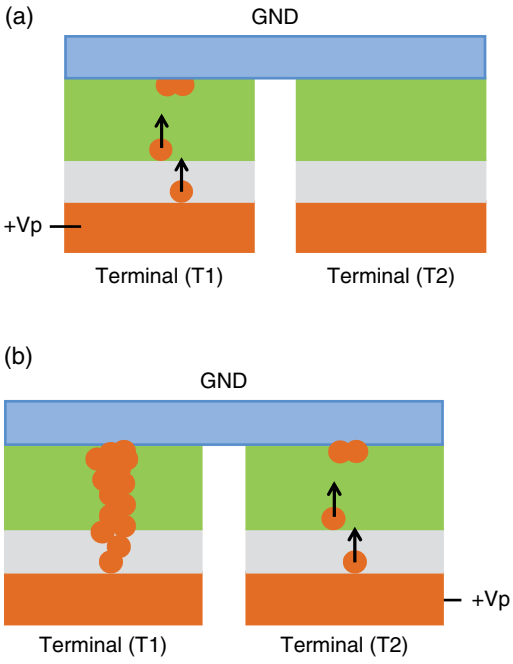


Figure 4.16 Programming the two terminal Cu atom switches (CASs): (a) left side and (b) right side. Based on N. Banno *et al.* (LEAP), *IEEE Trans. on Electron Devices*, 62 (9), September 2015 [35].

For programming, the terminals are individually connected to + Vp and GND, as shown in Figure 4.16. To program the T1 terminal, +Vp is connected to T1 and GND to C. To program the T2 terminal +VP is connected to T2 and GND to C. The CAS was integrated in a 65 nm CMOS process with a seven-layer Cu BEOL using two additional masks. The CAS was formed on the edge of two M4 lines.

A 1 Mb array was made. The two 512Kb arrays were surrounded with peripheral circuits and decoders. The arrays consisted of CAS elements placed at the crosspoints of the bit lines and word lines. The entire Cu bridge dissolves into the PSE and the Cu ions return to the Cu electrode, indicating a forming free operation of the atom switch. More than 1000 cycles of endurance with a 10^4 ON/OFF resistance ratio were confirmed.

In September of 2015, NEC discussed the ON-state reliability of a Cu atom switch under current–temperature stress [36]. The DC current–stress tolerance of the ON-state Cu atom switch was evaluated at high temperatures. It was found that stress due to the current in the RESET direction caused time-dependent failures as a result of E-field driven diffusion of the Cu in the conducting bridge. The maximum current per atom switch was found to be 115 μA. The reconfigurable switch block was operated at 500 MHz at 125 °C.

In June of 2016, NEC discussed a Cu atom switch with over 400 °C thermally tolerant polymer solid electrolyte (TT-PSE) for nonvolatile programmable logic [37]. The Cu atom switch was made in a standard Cu back-end-of-the-line (BEOL), as shown in Figure 4.17. To achieve switching, hydrocarbons with weak chemical bindings in the PSE were selectively eliminated, resulting in higher thermal stability. Data retention characteristics were confirmed after thermal cycle stress at

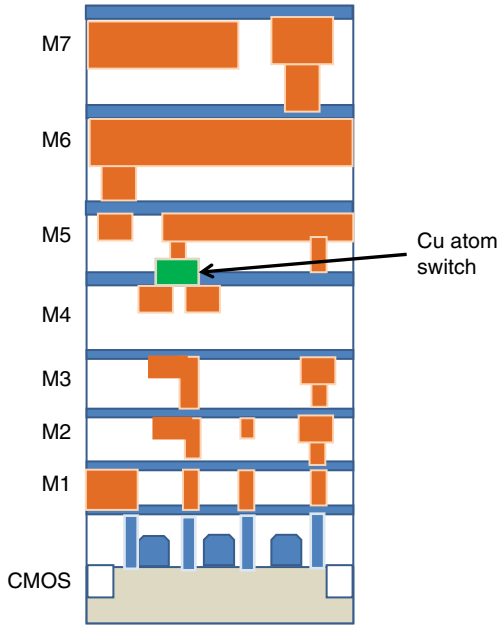


Figure 4.17 Illustration of polymer solid electrolyte BEOL Cu atom switch (CAS), which is between M4 and M5. Based on K. Okamoto *et al.* (NEC), VLSI-Technology Symposium, June 2016 [37].

temperatures ranging from –65 to 150 °C for 1000 cycles. The atom switch was intended as a technology component of reliable reprogrammable logic for future robotic and vehicle applications in high temperature systems.

A SET/RESET switching model of a Cu atom switch based on electrolysis was discussed by NEC in April of 2017 [38]. A nanometer thick layer of solid electrolyte gave an electric field of 3.3 MV/cm in the solid electrolyte at 2 V. This enabled ionization of Cu and the formation of a Cu bridge. It was found that the switching voltage depended on the resistance of the electrolyte. A compact model was used for estimating the resistance of the Cu atom switch programmed by various conditions.

4.4.4 Inorganic Thin Film Resistive RAMs on Flexible Substrates

An Ni/GeO_x/TiO_y/TaN RRAM on flexible substrate was discussed in April of 2013 by NCTU, NTNU, and IM-CAS [39]. The dielectric substrate was GeO_x/TiO_x. The top and bottom electrodes were Ni/N-rich TaN. Performance characteristics were 30 μW switching power (9 μA at 3 V, –1 μA at –3 V), 10^5 cycles of endurance, and good retention at 85 °C. These properties were thought due to: bulk transport properties by tunneling between traps, forming-free resistive switching, and low power switching. It is difficult to integrate a floating gate or charge trapping Flash memory on a flexible substrate due to the degraded gate oxide quality at low temperature. The RRAM, however, can be made using a low temperature process and simple structure that can be used in the flexible electronics application.

A typical *I*–*V* sweep cycle curve of the Ni/GeO_x/TiO_y/TaN RRAM device on polyimide with an area of 11 300 μm^2 is illustrated in Figure 4.18. A V_{read} at 0.5 V is shown with the low resistance state (LRS) and high resistance state (HRS) indicated.

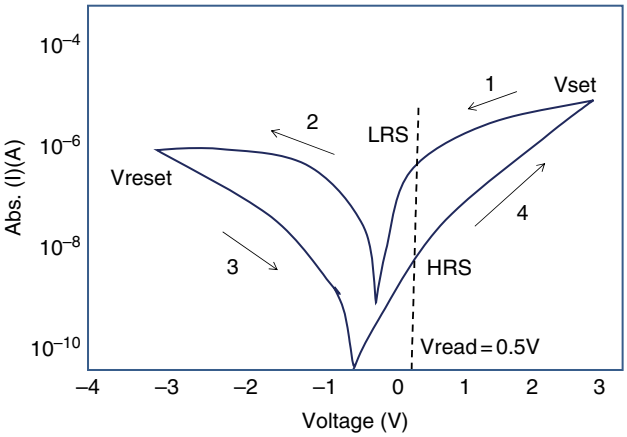


Figure 4.18 Illustration of swept *I–V* curve of Ni/GeOs/TiO_y/TaN flexible RRAM device on flexible polyimide. A read is shown at 0.5 V with low resistance state (LRS) and high resistance state (HRS) shown. Based on K.I. Chou *et al.* (NCTU, NTNU, IM-CAS), *IEEE Electron Device Letters*, 34 (4), April 2013 [39].

The NVM device had adequate performance on a low cost flexible substrate. The device data showed I_{set} at V_{set} of 9 μA at 3 V and I_{reset} at V_{reset} of −1 μA at −3 V. Switching power was 30 μW. The ON/OFF current ratio was 30. Electrical endurance was 10^5 cycles and retention time was 10^4 seconds at 85 °C. Resistive switching was forming-free and set/reset currents were low.

Inorganic–organic RRAM with uniform, multilevel cell (MLC) operation was discussed in April of 2014 by IM Peking University [40]. The concern was for improving uniformity and multilevel operation for an HfO_x/parylene-C hybrid RRAM device. Measurements showed that the switching parameter uniformity of this hybrid device was significantly improved compared to pure parylene-C RRAM memories. A current/voltage (I–V) curve of a Pt/HfO_x/parylene-C/Al structure is shown in Figure 4.19.

The resistive switching mechanism was attributed to formation and rupture of the metallic filament in the parylene layer. By inserting an HfO_x layer, the filament is localized and good uniformity and stability are achieved. The organic hybrid devices are capable of low voltage operation with $V_{set} = 2$ V, $V_{reset} = -0.6$ V, and a high ON/OFF current ratio $>10^3$. Since the device is highly uniform and has a large ON/OFF current ratio, the MLC storage capability is good and good retention was demonstrated. It was concluded that the device has potential in flexible, high density NV memory applications.

Flexible electronics devices are interesting due to characteristics such as low cost, low fabrication temperature, light weight, and flexibility. In July of 2014, National Dong Hwa University discussed resistive switching memory with good performance to be used with flexible electronics [41]. In this study an inorganic Al/Al_2O_3/ZrO_2/Al flexible resistive switching memory was made at room temperature. Inorganic materials tend to have relatively high mechanical strength, high humidity resistance, high durability, and good stability compared with organic materials.

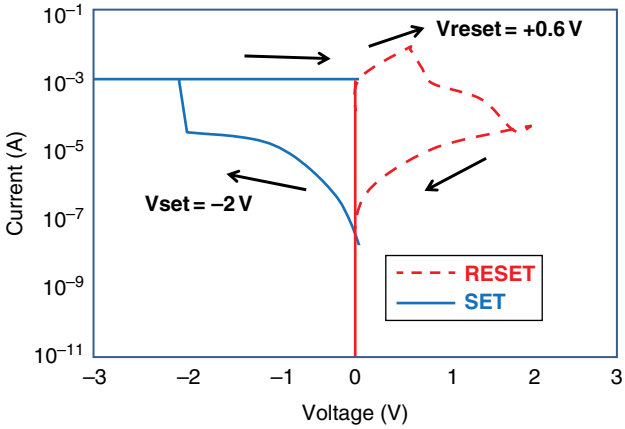


Figure 4.19 Illustration of *I–V* curve of Pt/HfO_x/parylene-C/Al structure. Based on Y. Liu *et al.* (IM Peking University), VLSI-TSA, April 28, 2014 [40].

To make the switching memory, an Al bottom electrode was deposited on a polyethersulfone plastic substrate; then ZrO_2 and Al_2O_3 resistive switching layers were deposited on the bottom electrode. Al top electrodes were made on the Al_2O_3 layer through a shadow mask. All deposition steps were done at room temperature by RF sputtering. A resistive switching filamentary model of the flexible device was developed, as illustrated in Figure 4.20 showing (a) SET and (b) RESET. Resistive switching endurance was over 100 times with stable resistive switching characteristics. Flexibility and nonvolatility characteristics of the device were shown.

4.4.5 IZO and IGZO Resistive RAM Memories

High density and high performance logic with embedded NVM using amorphous InGaZnO (a-IGZO) TFTs was discussed in July of 2013 by NCTU [42]. High performance logic and NVM applications were served using the same device structure. This opened the opportunity of low cost embedded memory for system-on-plastic integration. The use of a-IGZO TFT embedded NVM could permit several compact memory array architectures with high bit density.

Logic circuits have been shown using a-IGZO TFT including logic gates, ring oscillators, shift registers and SRAM. The a-IGZO TFT technology is therefore a potential platform for monolithic system-on-plastic (SoP) integration, which offers low cost, light weight, high density, large area, and mechanical flexibility. Embedded NVM devices have been suggested for use with flexible a-IGZO TFT circuits. Charge storage NVM using both floating gate and charge trapping have been shown. RRAM can be implemented in back-end metallization processes but it requires extra process steps and masks, which lower yield and add to cost.

An RRAM in the standard a-IGZO process was shown to combine the advantages of low cost integration, high bit density, and flexible memory characteristics, making it suitable for SoP applications. A four mask a-IGZO TFT with a top gate top contact

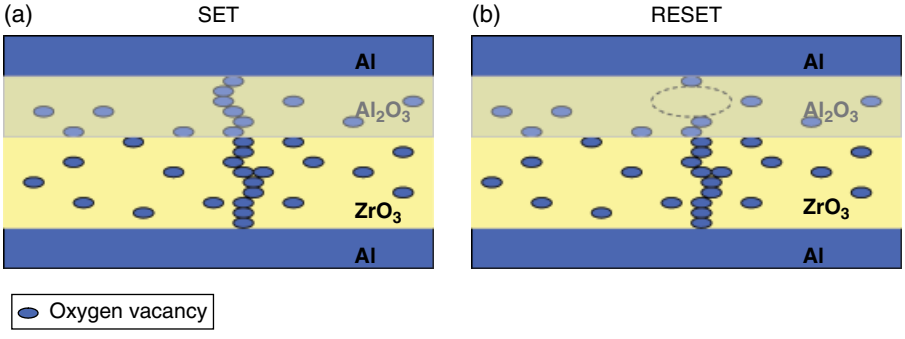


Figure 4.20 A resistive switching filamentary model of an Al_2O_3/ArO_3 flexible device: (a) set and (b) reset. Based on C.C. Lin (NDHU), *IEEE Trans. on Magnetism*, 50 (7), July 2014 [41].

structure was made on a flexible polyimide substrate in both coplanar and staggered configurations. The coplanar configuration is shown in Figure 4.21 (a) and the staggered in Figure 4.21 (b).

The staggered a-IGZO TFT used a reversed patterning sequence for the active regions and the S/D contacts. Also in the staggered TFT, the top gates and S/D contacts were replaced by Ni and Ti respectively. The resistance memory at the drain bit was activated by applying a 4 V forming voltage to the drain with the gate grounded. A stable bipolar resistance was attributed to the connection and rupture of Ni filaments in HfO_2 localized at the gate drain overlap region. The logic transistors and resistance memory device had good characteristics in both flat and flexed states, making it potentially for low cost flexible embedded memory applications.

The logic transistors before forming had SS for the coplanar device of 120 mV/decade while the staggered one had SS of 140 mV/decade. Mobility of the coplanar device was 21 $cm^2/(V\,s)$ while the staggered one was 2 $cm^2/(V\,s)$. The forming voltage for the coplanar device was +4 V and for the staggered one was −6.5 V. Reliability characteristics differed only in retention, which was >10 years at 85 °C for the coplanar device and $>10^4$ s at 85 °C for the staggered one. The pulse speed was 50 ns for the coplanar and 10 μs for the staggered device.

A three-state nonvolatile resistive memory for a-IGZO TFT circuit was discussed in October of 2013 by National Chiao Tung University [43]. A logic compatible flexible amorphous indium–gallium–zinc oxide (a-IGZO) TFT was made at low temperature. Before electrical forming, the a-IGZO TFT showed good transistor characteristics such as an ON/OFF current ratio of 8.8×10^6, steep subthreshold slope of 0.14 V/decade, V_{th} of 0.55 V, and maximum field-effect mobility of 2 cm/(V s). After electrical forming, a three-bit-per-cell resistive switching memory was obtained using localized multilevel resistance states. Good transistor and NVM characteristics were found using a staggered a-IGZO TFT structure.

The device was made at low temperature on a flexible substrate and offered low cost integration and good characteristics during flexing. This offers the potential for system-on-plastic applications. The staggered a-IGZO TFT was made using the four mask process flow. The three distinct resistance states in the a-IGZO TFT RRAM memory

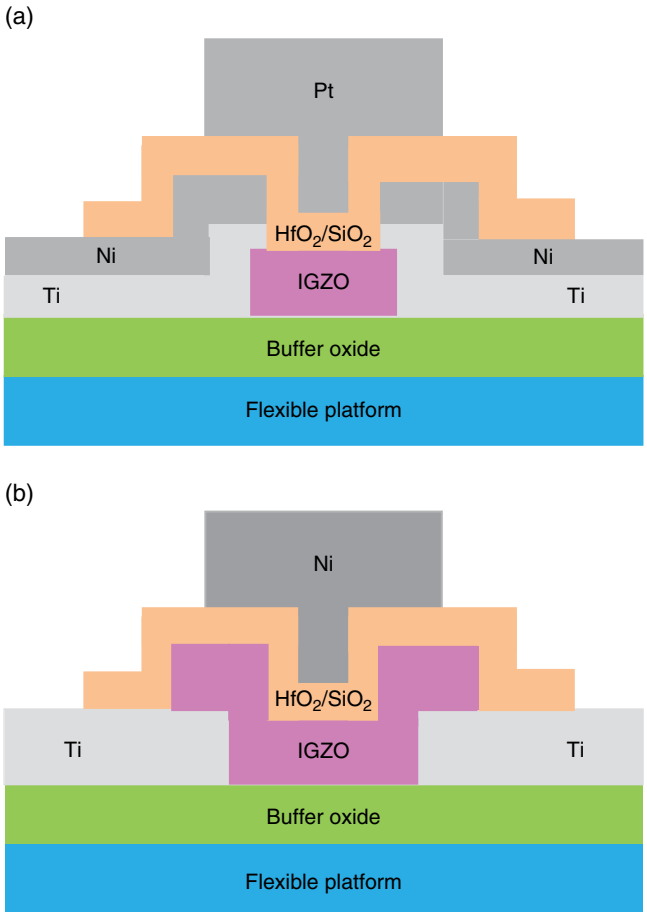


Figure 4.21 Four mask a-IGZO TFTs with top gate top contact on flexible polyimide substrate in (a) coplanar configuration and (b) staggered configuration. Based on T.H. Hou *et al.* (NCTU), AM-FPD, July 2, 2013 [42].

are illustrated in the *I–V* curve in Figure 4.22. Voltage was applied to the drain while the gate was grounded. SET pulses were −5 V for 10 μs and RESET pulses were 5 V for 500 μs. Good retention characteristics at 85 °C for the three states were shown to 10^4 cycles. Bending endurance was good over 10^4 s with a 10 nm bending radius.

In July of 2015, Kyun Hee University discussed the nonvolatile memory performances of transparent flexible thin film memory transistors using IGZO channels and ZnO charge trap layers [44]. An ON/OFF ratio greater than 7 orders of magnitude was found for the fully transparent charge trapping TFTs made on glass substrates when the width of the program pulses were set as 500 nm and the amplitude set at ±20 V. The memory window was 25.8 V wide.

To enable transmittance to visible wavelength, the gate electrodes were made of transparent In–Sn–O layers, which permitted the overall device structure to be made with all-oxide gate stacks. Memory operations of the flexible charge trap memory TFTs on plastic polyethylene naphthalate substrates were also discussed.

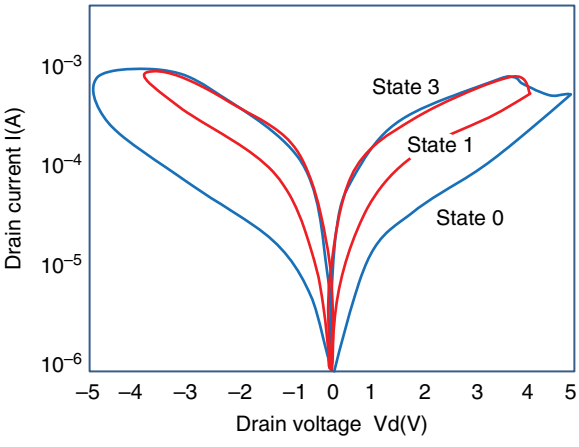


Figure 4.22 *I–V* curve of a-IGZO RRAM memory showing three resistance states. Based on S.C. Wu *et al.* (NCTU), *IEEE Electron Device Letters*, 34 (10), October 2013 [43].

4.4.6 Other Polymer Resistive RAMS with Flexible Substrates

Dynamic response of Al_2O_3/polymer diode RRAM was discussed in April of 2013 by the Institute de Telecommunications Lisboa, Portugal, who demonstrated dynamic response of a bistable resistive NVM made of Al_2O_3/polymer diodes [45], These diodes were probed in both the OFF and ON states with triangular and step voltage profiles. The OFF-state response followed the predictions based on a classical, two-layer capacitor description. As voltage scan rates increase, the model predicts that the fraction of the applied voltage, V_{ox}, appearing across the oxide would decrease.

Device responses to step voltages in both the OFF and ON states show that switching events are characterized by a delay time. Coupling these delays to the lower values of V_{ox} found during fast scan rates, the observation in the on-state that device currents decrease with increasing voltage scan rate can be explained. If it is assumed that a critical current is required to turn off a conducting channel in the oxide, then a model can be formulated to explain the shift in the onset of negative differential resistance to lower voltages as the voltage scan rate increases. The fundamental limitations on the speed of operation of a bilayer resistive memory appear to be the time and voltage dependance of the switch-on mechanism rather than the switch-off process. This result sheds light on the apparently anomalous behavior of the on-state, which is the disappearance of the negative differential resistance region at high voltage scan rates.

Resistive switching in multiferroic thin film was discussed in September of 2013 by various Chinese and German labs, which found resistive switching properties in multiferroic $BiFeO_2$ and $YMnO_3$ thin films grown by pulsed laser deposition [46]. Both materials when sandwiched between Au top and Pt/Ti bottom electrodes showed nonvolatile resistive switching when an electric field was applied. $BiFeO_3$ switched in bipolar mode when a positive and negative bias was applied. The resistance ratio between high and low resistance states was larger than 100. The switching mechanism was described by a model of the flexible Schottky-like barrier at the BFO/Pt interface, which may be related to the drift of oxygen vacancies (ions). An illustration of the bipolar *I–V* curve of the Au/$BiFeO_3$/Pt/Ti is shown in Figure 4.23.

$YMnO_3$ showed a unipolar resistive switching with the resistance ratio between the high resistance state and low resistance state greater than 10^5. In this case the origin of

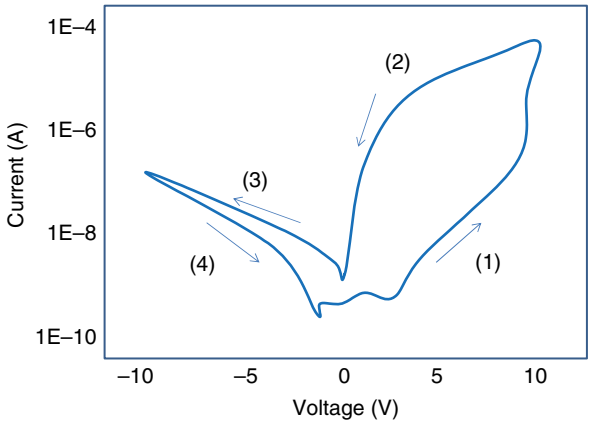


Figure 4.23 Illustration of the bipolar *I–V* curve of a flexible Au/$BiFeO_3$/Pt/Ti RRAM. Based on A. Bogusz *et al*. (Chemnitz Tech, University of ES&T Chengdi, IIN Dresden), ISCDG, September 26, 2013 [46].

the behavior was thought to be the formation and rupture of conductive filaments. Both of the films were considered potential for future NV RRAMs.

Ag/polystyrene/Ag conductive bridge RAM (CB-RAM) nanogap devices were discussed in November of 2013 by the University of Pennsylvania [47]. A CB-RAM was made with silver/polystyrene/silver nanogap devices. This reversible resistive switching used metal filament formation as the switching mechanism. The devices were encapsulated in polystyrene. The switching showed a high ON/OFF current ratio of $>10^3$ during cycle switching tests over many cycles. The gap was observed to evolve after extensive testing. Reversible electrical bistability was accomplished with an electrically inactive polymer, which expands the range of polymers suitable for organic digital memory applications.

A switchable diode effect in a polycrystalline thin film BiNdTiO RRAM was discussed in December of 2013 by Xiangtan University [48]. A switchable diode effect with a residual polarization of 55 $\mu C/cm^2$ was shown in the BiNdTiO RRAM when fabricated on Pt/Ti/SiO_2/Si substrates by chemical solution deposition. Consistencies of the *P–V* and *I–V* curves showed that the switchable diode effect was primarily caused by polarization modulated Schottky-like barriers. The ON/OFF current ratio based on these switchable diodes was greater than three orders of magnitude during the retention capacity measurement. The polycrystalline BNT thin films were thought promising for the resistive memory applications.

Resistive switching in Au/pentacene/Si-nanowire (NW) arrays was discussed in February of 2014 by National Changhua University of Education [49]. Resistive switching memory devices were fabricated including: Au/pentacene/n + Si, Au/pentacene/Si NW and Au/pentacene/H_2O_2 treated SiNWn + Si. The Au/pentacene/SiNW/n + Si devices showed hysteresis type behavior. H_2O_2 treatment could potentially lead to degradation of the hysteresis curve. No hysteresis current–voltage characteristics were observed for Au/pentacene/n + Si devices, which may indicate that the resistive switching characteristic is sensitive to silicon nanowire characteristics and that, potentially, the charge trapping effect originates from the SiNWs. The exploration of SiNWs as organic memory devices was thought promising.

Resistive memory based on PEDOT:PSS polymer was discussed in September of 2014 by the Changchun Institute of Applied Chemistry. Resistive memory devices based on PEDOT:PSS doping with polyvinyl alcohol were studied [50]. The resistive switching characteristics of the memory device were found to be dependent on the treatment of the polymer blend film by UV-ozone. The UV-ozone treated device showed improved performance. The ON/OFF current ratio was greater than 100. ON and OFF state retention was over 96 hours without deterioration. The resistive switching behavior in the UV-ozone treated device was attributed to the formation and rupture of PEDOT-PSS filaments as well as to narrow conducting paths through the native aluminum oxide.

In January of 2016, the University of Rome, Technical University Munchen, CNR-ISMN, NanoTecCenter Weiz, and Graz University of Technology discussed modeling of filamentary conduction in organic thin film memories [51]. A semiclassical drift-diffusion model was developed of electrical conductivity in the filament. It was shown that the global behavior of a memory device and the total current can be explained by fully formed and well connected filaments.

Resistive switching has been observed in many organic materials and two main mechanisms have been proposed to explain the bistable behavior. One mechanism is nanoparticles in the material that can act as trap sites. The bistability is an effect of the trapped charge, which can be trapped and detrapped under the control of an applied voltage, thereby switching the device. The other mechanism is the formation of highly conductive localized filaments, which provide pathways for current. These ionic pathways are formed under the influence of an applied voltage. The I–V characteristics of light exposed devices show a different open circuit voltage for each resistive state. This behavior is typical of an organic photovoltaic cell coupled with a shunt resistance, which is the filament.

A heating model of the filament was used to check if assumptions on the number of filaments and their radii were physically plausible. This permits an organic ITO/Alq3/Ag memory device to be pictured as an equivalent circuit under light where the conductive filaments can be represented as a shunt resistance R_{sh}. The total current is $I = I_{org} - V/R_{sh}$. Different resistive states correspond to different open circuit voltages. The equivalent circuits of an organic ITO/Alq3.Ag RRAM memory device is shown in Figure 4.24. Alq3 is chemically $C_{27}H_{18}AlN_3O_3$ and ITO is indium–tin–oxide.

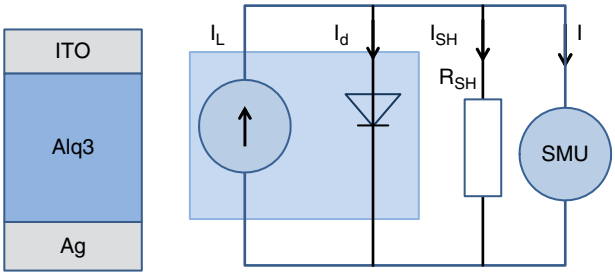


Figure 4.24 Equivalent circuit of an organic ITO/Alq3/Ag Memory Device where Alq3 is chemically $C_{27}H_{18}AlN_3O_3$ and ITO is indium–tin oxide. Based on F. Santoni *et al.* (University of Rome, TU Munchen, CNR-ISMN, NTC Weiz, Graz Tech.), *IEEE Trans. on Nanotechnology*, 15 (1), January 2016 [51] (permission of IEEE).

The ITO is used as an optically transparent bottom electrode. For the photovoltaic measurements, the device was illuminated with an inorganic blue emitting LED driven with a constant current. A symbolic view of the switching process showing the forming process is shown in Figure 4.25 (a). The ON/OFF switch intermediate state is shown in Figure 4.25 (b) and the OFF/ON switch is shown in Figure 4.25 (c).

A semiclassical drift-diffusion model of the electrical conductivity of the filament was developed. The model correctly accounts for the behavior of the device observed by the photovoltaic measurements. Both trap-assisted Poole-Frenkel emission and Miller-Abrahams hopping can be used to explain the observed behaviors. Poole-Frenkel emission described the motion of electronics in an insulator due to random thermal fluctuations, which provide the energy to move electrons out of a localized state and into the conduction band before falling back into another localized state. The Miller-Abrahams hopping is based on a phonon-assisted tunneling mechanism. A simple thermal analysis was used to estimate the number of formed filaments and their radii.

A new type of resistance RAM memory device using graphene oxide (GO) Al/GO-PEDOT: PSS/Pt was discussed in August of 2016 by Tsinghua, Beijing, and Fuzhou Universities [52]. The RRAM was nonvolatile and bipolar switching. The GO-PEDOT:PSS composite showed potential for use in high density flexible nonvolatile

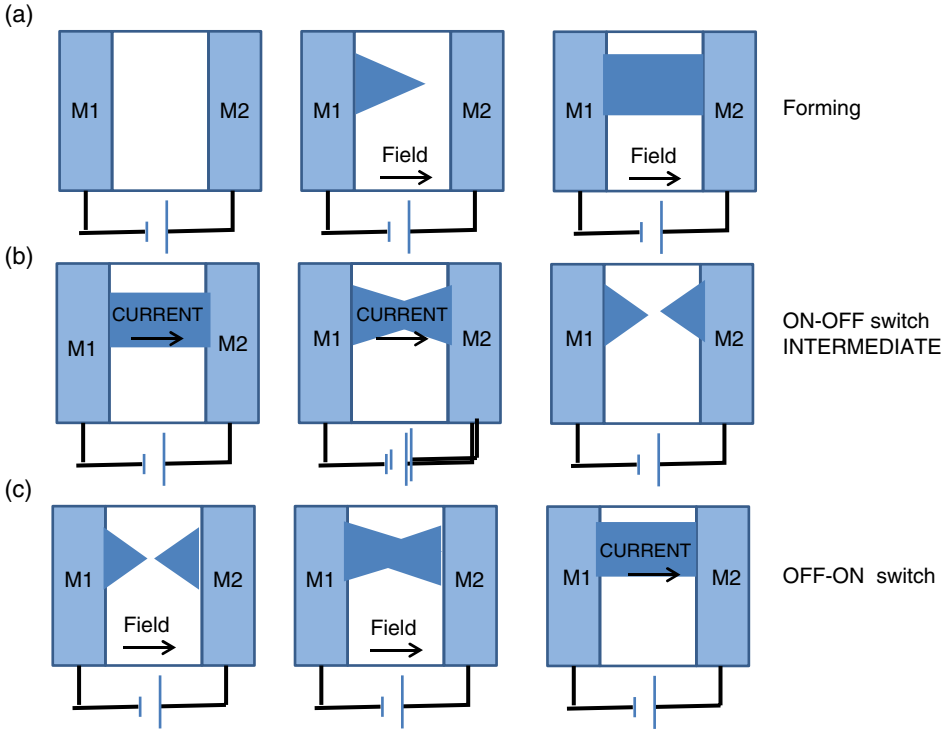


Figure 4.25 Schematic showing the switching process showing: (a) forming, (b) ON/OFF, (c) OFF/ON. Based on F. Santoni *et al.* (University of Rome, TU Munchen, CNR-ISMN, NTC Weiz, Graz Tech.), *IEEE Trans. on Nanotechnology*, 15 (1), January 2016 [51] (permission of IEEE).

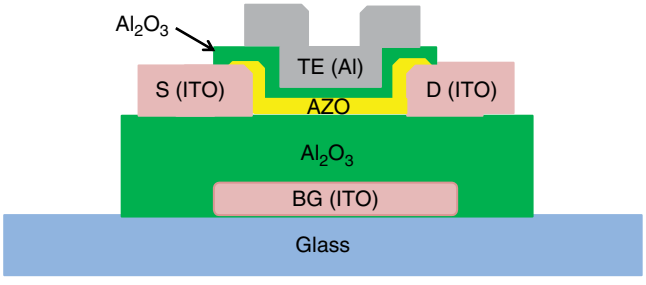


Figure 4.26 Three terminal resistive change devices in TFT configuration with AZO channel. Based on W.H. Lee *et al.* (Kyong Hee University), *IEEE Electron Device Letters*, 37 (8), August 2016.[53].

RRAM memory. It showed repeated high speed switching, 1–3 V tight distribution of HRS and LRS and retention greater than 10^4 s. The ON/OFF current ratio was about 200.

In August of 2016, Kyong Hee University discussed a multilevel RRAM with a thin-film transistor (TFT) structure [53]. A three-terminal device structure was used, resulting in integration of a resistive change memory into the thin film transistor configuration. The three terminals were source (S), drain (D), and top electrode (TE). The device was formed on either a glass or a plastic substrate. Figure 4.26 shows a schematic cross-section of the three-terminal resistive change device.

Both the functions of a transistor and a nonvolatile memory could be performed. Al-doped ZnO (AZO) was used for the active channel of the TFT and for the resistive change material in the memory device.

The resistive change memory operations in the AZO could be found between the source and drain and the top electrode by the formation and collapse of the conductive filaments. The initial state is from the first forming process at the TE and the S/D regions. The first SET state was programmed by switching events at the TE/source region and the second SET state was programmed by a SET switching events at the TE/drain region. The final RESET state was programmed by the RESET switching event at both regions. The three states consisted of the RESET state, a single source or drain SET state, and a SET state for the source and drain.

The multilevel memory operations were realized by controlling the number of conducting filaments formed in the channel between the source, drain, and the top electrode, which effectively modulated the Ids levels in three different states. The programmed Ids values could be switched in a stable manner among the RESET, a first SET, and a second SET state. There were no variations in the programmed Ids at each memory state during a 10 hour retention time. The programmed current ratios for three-level memory states were 1:3.8:25. Retention was 10^4 s and program endurance was 100 cycles.

An illustration of the bipolar resistive switching *I–V* characteristics of the Al/Al_2O_3/AZO/ITO structure between the source and drain and the top electrode is shown in Figure 4.27.

An indium–-tin-oxide (ITO) RRAM with an ultralow switching voltage due to an inserted SiO_2 layer was discussed in October of 2016 by National Sun Yat-Sen University, University of Texas, Austin, National Cheng Kung, University, National Kaohsiung Normal University, and Xiamen University [54]. The amplitude of the switching voltage

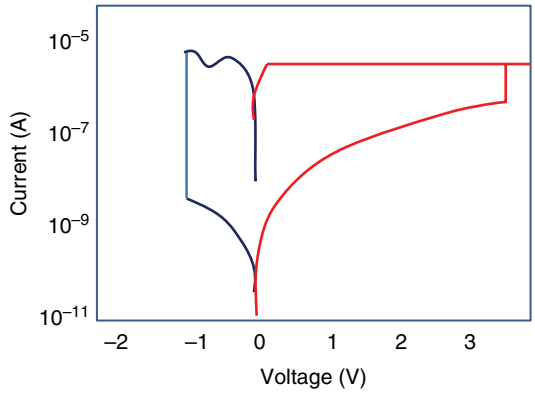


Figure 4.27 Bipolar resistive switching *I–V* characteristics of $Al/Al_2O_3/AZO/ITO$ structure between top electrode (TE) and the source (S) and drain (D). Based on W.H. Lee *et al.* (Kyong Hee University), *IEEE Electron Device Letters*, 37 (8), August 2016 [53].

of the device was below 0.2 V whether measured by direct or by alternating current sweep operation. The observed RESET voltage increased with temperature while the Set voltage did not notably change under temperature variation. To determine the switching mechanism, conduction current fitting and switching voltage statistics were applied to explore the regular voltage variation dependence on temperature.

The switching mechanism of the device was discussed along with the current conduction mechanism. An ohmic conduction mechanism was found to be caused by original carriers at a low electric field. As the operating voltage became larger, the current conduction in the LRS was dominated by hopping conduction. A reaction model was proposed to explain the oxygen concentration gradient induced between the inserted SiO_2 and the ITO electrode on the ITO-based RRAM device.

Inkjet printed high performance RRAMs based on a high-k HfO_2 dielectric were implemented in April of 2017 by the Universities de Barcelona and Autonoma de Barcelona [55]. The technology had a high ON/OFF current ratio and low switching voltage, which could permit low power applications. The inkjet printed dielectric layer resulted in a flexible substrate that could enable low integration density memory in portable devices. The simple vertical metal–insulator–metal structure of the RRAM appeared suitable to this application. The use of a crossbar architecture permitted scalability.

Inkjet printing permitted a versatile, maskless, noncontact digital fabrication method that could be done at ambient room conditions. One memory cell that was found flexible and inexpensive was $Cu/Cu_xO/Ag$ for working at low voltages, but lack of resolution and of available inks were drawbacks. The study focused on inkjet printed HfO_2 as a flexible thin film dielectric layer for an HfO_2 based RRAM device. Inkjet printed RRAMS were shown to have comparable features to conventional CMOS compatible RRAMs.

An illustration of an RRAM with Ag top contact, HfO_2 dielectric, and Au bottom contact is shown in Figure 4.28. The device is about 1 mm square. The Au bottom electrode is 50 nm thick, HfO_2 was 200 nm thick, and the Ag top electrode was 700 nm thick.

An HfO_2 nanoparticle ink was printed over the bottom electrode while the substrate temperature was kept at a constant 30 °C. A 3 hour postdeposition annealing process at 240 °C in a vacuum was needed to eliminate the solvents. A 700 nm thick silver top electrode was printed on the dielectric followed by annealing in vacuum at 200 °C to

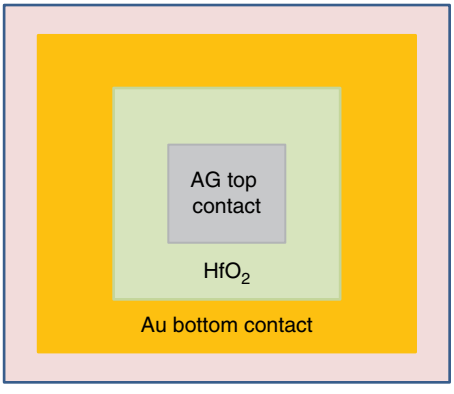


Figure 4.28 Illustration of 1 mm^2 HfO_2-based RRAM with 50 nm thick Au BED, 200 nm thick HfO_2 dielectric, and 700 nm thick Ag top electrode. Based on G. Vescio *et al*. (University de Barcelona, University A. de Barcelona), *IEEE Electron Device Letters*, 38 (4), April 2017 [55].

remove organic solvents and attain resistivity of about 16 μohm cm. A minimum thickness for each printed layer was specified to avoid pinholes. A forming process with current limitation was used to create a switchable conductive filament. A bipolar resistive switching process was used with positive bias for SET and negative bias for RESET. The I_{on}/I_{off} ratio was around 1000 in most cycles.

In May of 2017, National Cheng-Kung University, National University of Kaohsiung, and Air Force Institute of Technology, Kaohsiung, discussed the performance of flexible RRAM devices based on simple spin-coated sol gel derived strontium titanate nickelate (STN) thin films on a polymer substrate [56]. The stack of the device is shown in Figure 4.29. These devices had an ON/OFF current ratio of 10^5 and showed a uniform current distribution.

Issues with metal oxide-based thin film RRAMs include: a low ON/OFF ratio ($<10^3$), cracking issues during repeated bending, and poor current distribution. The stability of the STN thin film is in part due to the strong bonding between bidentate ligands of nickel acetylacetone and titanium metal ions, which enable the chelation effect. The Ni ions can act as clusters and play a critical role in the formation and rupture of cone-shaped conductive paths during resistive switching. The formation and rupture process promotes local alignment of oxygen vacancies among metallic Ni clusters, which enhance the distribution of switching parameters. Figure 4.30 illustrates the resistive switching mechanism of the STN-based RRAM in the (a) pristine, (b) LRS, and (c) HRS conditions. The devices were made on a flexible plastic substrate and were durable under a repeated bending test.

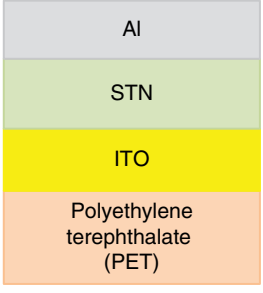


Figure 4.29 Stack of flexible RRAM devices based on simple spin-coated Sol Gel derived strontium titanate nickelate thin film on polymer substrate. Based on K.J. Lee *et al.* (NCKU, NU and Air Force IT, Kaohsiumg), *IEEE Trans. on Electron Devices*, 64 (5), May 2017 [56].

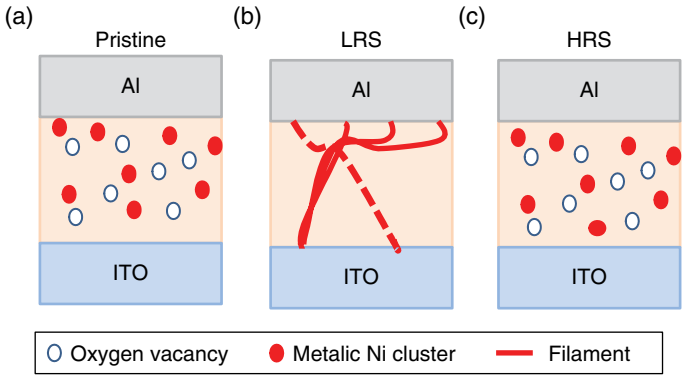


Figure 4.30 Illustration of the resistive switching mechanism model of the STN-based RRAM in the (a) pristine, (b) LRS, and (c) HRS conditions. Based on K.J. Lee *et al.* (NCKU, NU and Air Force IT, Kaohsiumg), *IEEE Trans. on Electron Devices*, 64 (5), May 2017 [56].

It was found that the STN thin film prepared by a solution process could be used effectively in making flexible memory devices with forming-free resistive switching. The ON/OFF ratio was 10^5 after 90 days under an atmospheric environment. Current distribution showed good uniformity. The ON/OFF ratio showed no obvious degradation after 100 bending cycles. An increased barrier height, after adding the Ni, meant a lower HRS current and higher ON/OFF ratio. The metallic Ni clusters were used to interpret the switching mechanism in the STN devices.

In May of 2017, the Imperial College London and King Abdullah University of S&T discussed resistive switching memories based on empty aluminum nanogap electrodes [57]. Adhesion lithography was used to make the dense arrays of RRAMs along with a low temperature, large-area-compatible nanogap fabrication technique. These arrays were made on both rigid and flexible plastic substrates. The devices showed a nonvolatile memory operation. Endurance was stable. Resistance ratios were $>10^4$ and retention time was several months. The controlled resistive switching was found to be due to migration of metal from the electrodes into the nanogap when an external electric field was applied. This technology had potential in large area electronics and also in electronics for harsh environments such as: space, high and low temperatures, radiation, vibration, and pressure.

4.5 Charge Trapping Nanoparticle (NP) Memory on Flexible Substrates

4.5.1 Overview of Charge Trapping NP Memory on Flexible Substrates

A polymer carbon nanotube charge trapping memory was made using a thin parylene layer separating an SiO_2/nanotube interface from the polymer coating. Graphene nanocrystal (NC) memory in a polymer matrix was studied where the graphene quantum dots acted as charge traps.

Inkjet printed IGZO thin film transistors with nanoparticle storage was described. Al/Au nanoparticles embedded in polystyrene were made with a good memory window for multilevel operation. RRAMs with NC of different diameters were found to suppress switching fluctuations. Variation of the resistive switching parameters for NP memories was studied.

4.5.2 Carbon Nanotube Charge Trapping Memory with Flexible Substrates

Polymer carbon nanotube (CNT) charge trapping memory was discussed in January of 2013 by CEA, IRAMIS, CNRS, and LEM. These devices were used as optoelectronic nonvolatile memories [58]. The detailed mechanism of charge trapping of mixed organic/inorganic optoelectronic devices was investigated by intercollating layers of parylene at different interfaces in the device structure. A thin parylene layer separated the SiO_2/nanotube interface from the photosensitive polymer coating. The structure was suggested as an optimized solution for charging, stability, and robustness. A schematic cross-section of this structure is shown in Figure 4.31.

To be used in reconfigurable circuits, the memory devices must support multiple reprogramming and read cycles. The robustness of the geometry was verified by performing more than 200 cycles. The ON/OFF ratio remained approximately constant while the ON and OFF states showed some variability. The metallic CNTs were suboptimal but it was expected that single tube devices would not have an issue.

Memory properties of graphene quantum dots in semiconducting poly93-hexlthiophene polymeric matrix were discussed in August of 2014 by the National Institute for R&D in Microtechnology, Romania [59]. These graphene quantum dots act as charge trapping particles. The I–V measurements of thin films made from this nanocomposite deposited on gold interdigitated electrodes indicated that the whole transport I–V characteristics showed a strong nonlinear behavior and a hysteresis curve representative of a memristive response.

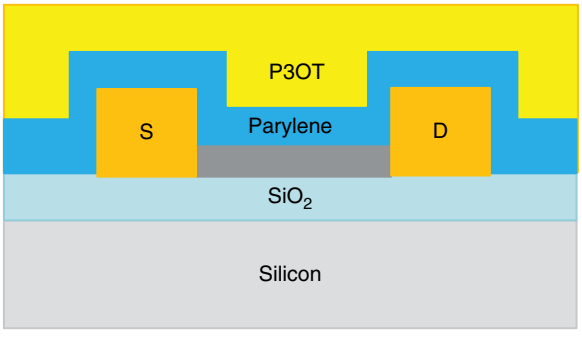


Figure 4.31 Schematic of a polymer carbon nanotube charge trapping memory using a thin parylene layer separating the SiO_2/nanotube interface from a photosentive polymer coating. Based on B. Brunel *et al.* (CEA, IRAMIS, CNRS, LEM), *Applied Physics Letters*, 102 (1), 013103, January 2013 [58].

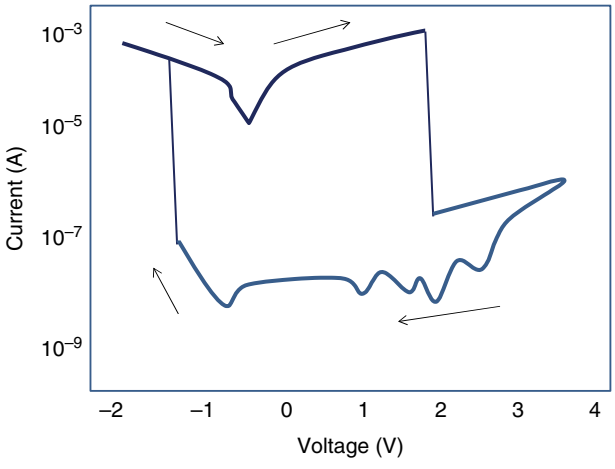


Figure 4.32 *I–V* switching characteristics on a semilog scale showing hysteresis curve of MIMb consisting of Ag NWPMSSQ/GQD/PMSSQITOPET. Based on J. Lin *et al.* (Fuzhou University, Hanyang University), *IEEE Electron Device Letters*, 36 (11), November 2015 [60].

In November of 2015, Fuzhou University and Hanyang University discussed a solution involving processed, flexible, transparent nonvolatile memory with embedded graphene quantum dots in polymethylsilsesquioxane layers made on a transparent flexible substrate [60]. The hysteresis window was found from the current–voltage plots. The nonvolatile memory devices were reprogrammable and stable up to 1×10^4 s with an ON/OFF current ratio of 10^4. The memory showed a stable hysteresis window with no obvious degradation upon bending under different radii of curvature.

Graphene quantum dots were used as the charge trap medium due to properties such as chemical inertness, low toxicity, and higher work function than other reported nanoparticles. The quantum confinement effect and boundary effect also made them a good charge trapping medium. Two MIMs were used. The structures were: MIMA Ag NW/PMSSQ/indium–tin oxide and MIMb Ag NWPMSSQ/GQD/PMSSQITOPET. The *I–V* characteristics of MIM-B on a semilog scale are indicated in Figure 4.32. A representative hysteresis curve is shown.

An illustration of the band diagrams for the (a) write process and the (b) erase process are shown in Figure 4.33. The device was programmed when electrons were trapped in the graphene quantum dots and erased when the electrons in the GQDs were detrapped.

4.5.3 Inkjet Printed Nanoparticle Memory

An inkjet printed thin film transistor (TFT) with silicon NP storage was discussed in January of 2013 by the Technical University of Singapore. An NP memory was produced with an inkjet printed In–Ga–Zn oxide (IGZO) thin film transistor [61]. An SiO_2 gate dielectric was used that had embedded silicon nanoparticles (SiNP). The SiNP acted as charge trapping sites. A clockwise V_{gs}–I_d hysteresis curve was found. The curve was attributed to the charging and discharging of the silicon nanoparticles. The printed IGZO-based memory had an ON/OFF current ratio of 10^3.

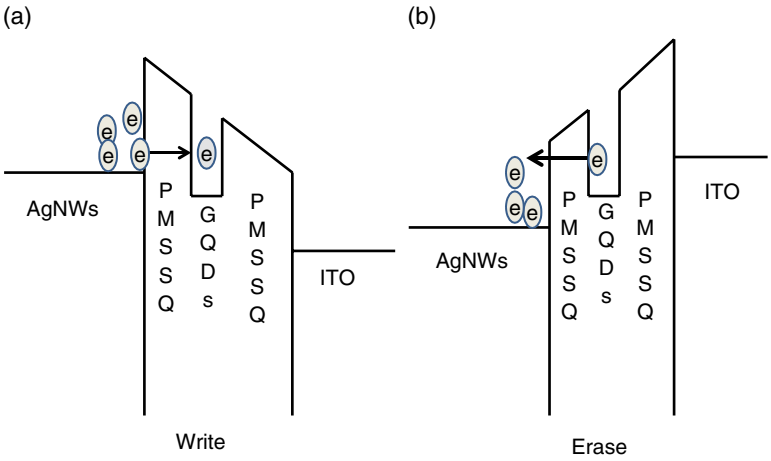


Figure 4.33 Band diagrams for (a) the write and (b) erase processes. Based on J. Lin *et al.* (Fuzhou University, Hanyang University), *IEEE Electron Device Letters*, 36 (11), November 2015 [60].

4.5.4 Other Nanoparticle Charge Trapping Memories on Flexible Substrates

Operation of Al/Au nanoparticles in a polystyrene layer was discussed in January of 2014 by Hanyang University, who discussed capacitance–voltage curves for Al/Au nanoparticles embedded in polystyrene layer/p-Si devices at 300 K [62]. These curves showed a metal–insulator–semiconductor (MIS) behavior with flat-band voltage shifts of the *C*–*V* curves due to the presence of charge trapping. Memory windows between 2.6 and 9.9 V were observed at different voltages, which indicated the potential for multilevel behavior.

Ct measurement showed that the charge trapping capability of Au nanoparticles embedded in a polystyrene layer was maintained for retention times larger than 10^4 s without major degradation. The multilevel charging and discharging mechanisms were models using the experimental results.

In September of 2015, the National Technical University of Athens discussed resistive switching characteristics of TiN/Ti/TiO_{2-x}/Au RRAM devices containing Pt nanoparticles (NP) with different diameters [63]. A primary concern was suppressing the fluctuation of switching parameters in the RRAM devices. Since the resistive switching effect originates in the creation and annihilation of conducting filaments within the metal oxide layer, the variations of the switching characteristics would be expected to be due to the stochastic nature of the conductive filament formation. While it is difficult to control the complex dielectric breakdown process, it is simple to indicate the regions in which it occurs. A significant improvement in switching parameters resulted from embedding metal nanoparticles in the dielectric matrix. This study looked at the influence of Pt NP size and density on the resistive switching properties of room temperature deposited TiO_2 thin films. These operate without an electroforming process and showed large switching ratios. It was found that the devices with Pt NC embedded had improved uniformity and reliability as well as multilevel capability from increasing the NC size. Pt NCs assisted the formation of oxygen vacancy-based filaments near the location of the NCs. This is important for increased

uniformity since it limits the creation of clustered oxygen vacancy regions that show uncontrolled switching patterns.

Physical vapor deposition (PVD) was used with the process integrated in a single vacuum system. This permitted formation of a metal oxide thin film with embedded nanocrystals made under high purity vacuum conditions.

The RRAM device structure was: TiN/Ti/TiO_2/Pt NCs/TiO_2/Au/SiO2/Si. The MIM capacitors were patterned using lift-off lithography. Each square electrode was about $100 \times 100\,\mu m^2$. Electrical characterization was done by applying all signals to the top electrode while keeping the bottom electrode grounded.

A typical DC current–voltage hysteresis loop for an RRAM is shown in Figure 4.34. The solid line represents the RRAM with 3 nm Pt NP and the black dotted line indicates the RRAM without the NP. The continuous transition between the HRS forward sweep and the LRS backward sweep is indicated. The arrows indicate the switching direction. The sweep is from –4 to +4 V.

A significant improvement of the resistive switching parameters was attained by embedding Pt NPs of different sizes and densities in the dielectric matrix. Variability was reduced by the enforcement of the possible percolative networks where switching took place. The bigger NPs induce very large ON currents with 3 nm Pt NP considered the optimum solution for spatial variability and power consumption. The small size of the NP permits integration into thin dielectric films, permitting device scaling. Multilevel capabilities were demonstrated.

Transparent memories using nanoparticles (NP) of polymer/ZnO nanocomposites were discussed in February of 2016 by National Taiwan University, National Taipei University of Technology, and National Central University [64]. These floating gate memories with metal NP covered with insulating polymer are studied for next generation nonvolatile organic memory. At issue is the fact that the transparency of the device with metal NPs is restricted to 60–70% due to the light absorption in the visible region

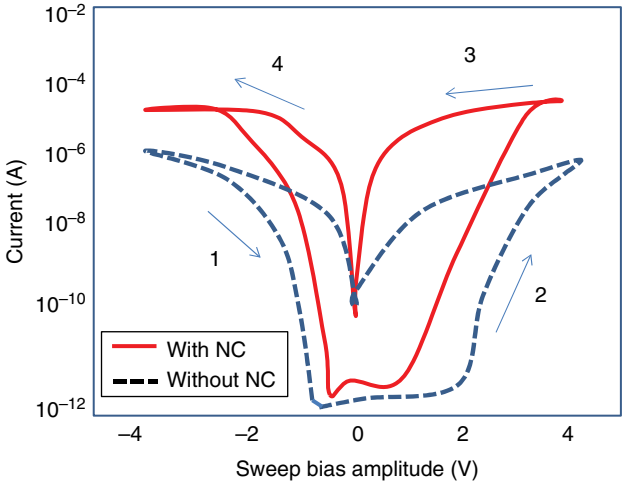


Figure 4.34 Typical DC current–voltage hysteresis loop for a TiN/Ti/TiO_{2-x}/Au RRAM device (a) with 3 nm Pt NP (solid line) and (b) without NP (dotted line). Based on P. Bousoulas *et al.* (NTU Athens), ESSDIRC, September 2015 [63].

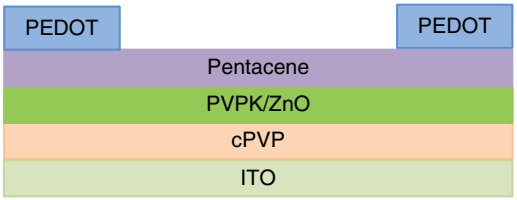


Figure 4.35 Transparent memory stack with polymer/ZnO nanoparticles for storage. Based on C.C. Shih *et al.* (National Sun Yat Sen University), Scientific Reports, February 1, 2016 [64].

caused by surface plasmon resonance effects of the NPs. A new memory used blends of hole trapping poly(9-(4-vinylphenyl) carbazole) with electron trapping ZnO NPs for charge storage. The memory stack is shown in Figure 4.35.

These devices showed a programmable memory window of 60 V during program/erase due to trapping and detrapping of charge carriers in the ZnO NP/PVPK composite. The retention time was $>10^5$ s and endurance was greater than 200 cycles. The devices had a transparency of 90% at 500 nm wavelength for the spray-coated PEDOT:PSS electrode.

Flexible thin film transistors using the IGZO channel and ZnO charge-trap layers made on polyethylene naphthalate (PEN) substrate were discussed in April of 2016 by Kyung Hee University and the Electronics and Telecomm Research Institute [65]. The fabrication process temperature was below 180 °C. An organic/inorganic hybrid barrier layer was introduced to improve the surface roughness and water vapor transmission rate of the PEN substrate. The gate stack was all oxide layers with an InGaZnO (IGZO) active channel, ZnO charge-trap layer, Al_2O_3 blocking/tunneling layers, and In–Sn–O (ITO) transparent electrode in which the double layer tunneling and top protection layers were designed. A wide memory margin of 25.6 V, fast programming speed of 500 ns, and retention time of greater than 3 hours were achieved at 25 °C and at 80 °C. These memory characteristics were not degraded after delamination of the PEN substrate or under bending with a radius of curvature of 3.3 mm.

The switching behavior of TaN/Al_2O_3:Ag:ZnO/ITO resistive RAMs made on flexible substrates using embedded Ag nanoparticles was discussed by Fudan University in July of 2016 [66]. Variation in the resistive switching parameters is a critical issue in RRAMs. The embedded Ag nanoparticles improved device yield and reduced resistance variability from more than 160% to 30%. The Ag NP also reduced variability in SET voltages from 35 to 18% and in RESET voltages from 40 to 11%. Synaptic behavior of this flexible device was shown opening potential applications in neuromorphic computing applications.

The RRAM devices were made on flexible substrates. A 10 nm thick ZnO film was deposited on an ITO-coated PET substrate using atomic layer deposition (ALD) at 130 °C with various precursors. The Ag nanoparticles were spin-coated on the ZnO layer and baked at 100 °C to evaporate the solvent. Then a 10 nm Al_2O_3 layer was deposited using ALD at 130 °C with precursors. A 50 μm^2 top electrode was defined using photolithography followed by a PVD deposited layer of TaN followed by a lift-off process. The effect of the nanoparticles can be seen in the bipolar switching *I–V* curves of the TaN/Al_2O_3:Ag:ZnO/ITC RRAM devices (a) with and (b) without the nanoparticles in Figure 4.36. The average surface density of the Ag nanoparticles was estimated at $1.3 \times 10^8/cm^2$.

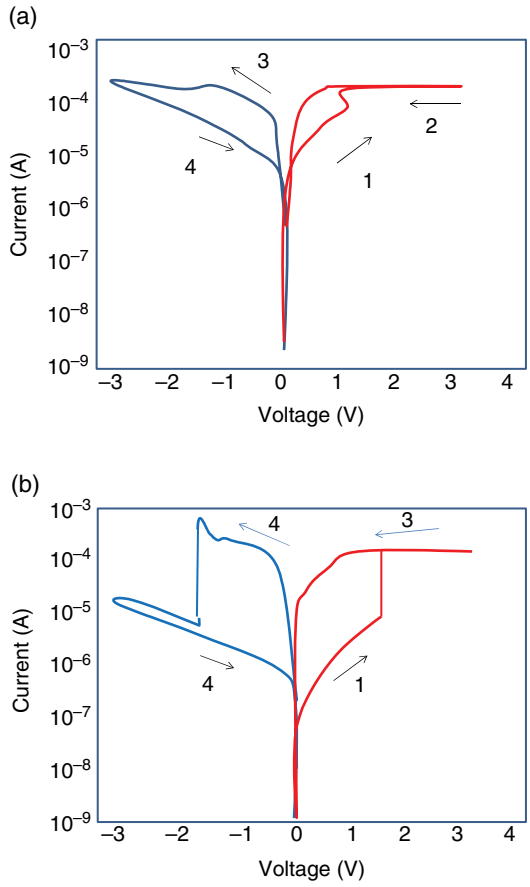


Figure 4.36 Illustrations of *I–V* curves of (a) TaN/Al_2O_3:Ag:ZnO/ITO RRAM devices with Ag nanoparticles (NP) and of *I–V* curves of (b) TaN/Al_2O_3/ZnO/ITO RRAM devices without Ag NP. Based on D.T. Wang *et al.* (Fudan University), *IEEE Electron Device Letters*, 37 (7), July 2016 [66].

The reset current of the sample with the Ag NPs is smaller than that without the NPs. The set and reset processes were abrupt without the Ag NP, while with the Ag NP gradual set and reset processes were found. The retention after folding the material hundreds of times showed only a slight shift after 10^4 s. The 10 year memory window was still good. For the set process the switching speed of the Ag NP embedded device reached 100 ns while the control device was 10 ns. For reset, the switching speed with the Ag NP reached 1 μs and without the Ag NP was 10 ns. This was attributed to the slower speed of Ag ions compared to oxygen vacancies. The device with Ag NP showed gradual set and reset processes. The Ag NPs enhanced the device yield and uniformity, lowered the operating voltage, and reduced power consumption. This improvement was attributed to Ag ions providing an easy and controllable path to the formation of a conductive filament and suppressing random filament formation and rupture.

In September of 2016, KAIST discussed a floating gate memory using an MoS2 channel with a metal nanoparticle charge trapping layer and polymer tunneling dielectric [67]. MoS2 is a 2D planar material, like graphene, which has a layer dependent bandgap from 1.4 to 1.8 eV with high mobility of about 150 $cm^2/(V\,s)$ and a high ON/OFF current ratio of 10^9. The device has a Ti/Au source and drain, 90 nm SiO_2 dielectric, and heavily p-doped bottom gate. A schematic cross-section of the device is shown in Figure 4.37.

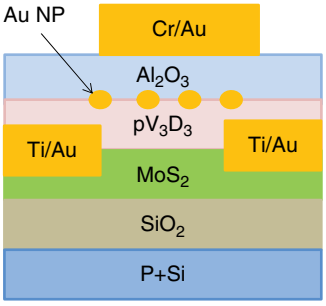


Figure 4.37 Schematic cross-section of a floating gate memory with an MoS2 channel with an Au nanoparticle charge trapping layer. Based on M.H. Woo *et al.* (KAIST), ESSDERC, September 2016 [67].

For fabrication, a conformal, stable polymer insulator layer was deposited via initiated chemical vapor deposition (iCVD), which ensured the memory could endure a significant electrical stress. Different thicknesses of gold nanoparticles, achieved by means of thermal evaporation, were used to control the density and distribution of nanoparticles in the charge trapping layer, while the Al_2O_3 blocking dielectric was deposited using an atomic layer deposition (ALD) process to increase the gate coupling ratio for low power operation. The floating gate memory device was fabricated and showed a tunable memory window with a high ON/OFF ratio after applied programming and erase pulse. This permitted multibit data storage with a long retention current ratio (I_{on}/I_{off}).

The trap and detrap of charges from the MoS2 channel to the gold NP layer determined the program and erase states during memory operation. The ON/OFF ratio was greater than 10^8 and the mobility was $18.6\,cm^2/(V\,s)$. The device showed a stable memory operation with a 6.8 V memory window and more than 10^3 s of data retention with greater than 10^3 current ratio.

4.6 Transfer of Conventional Memory Chips on to Flexible Substrates

Flexible electronics is important for IoT applications due to its low weight, high portability, low fragility, and options for variable structure. Flexible processors and memories are also important for data storage and processing and network communications. For these applications the flexible devices developed using organic and polymeric materials frequently have insufficient performance and low integration density. For these high performance applications, flexible electronics using single substrate integration of silicon electronics devices is needed. Several techniques have been developed for integration of silicon chips on flexible substrates. These are outlined in this section.

4.6.1 Transfer of Silicon Chips Using SOI Base Wafers

Using SOI wafers to create thin layers of memory chips has the advantage of having a thin film of single crystalline silicon, which is compatible with full CMOS. The underlying buried oxide (BOX) layer can be etched away leaving the ultrathin silicon layer or maintained to protect the silicon.

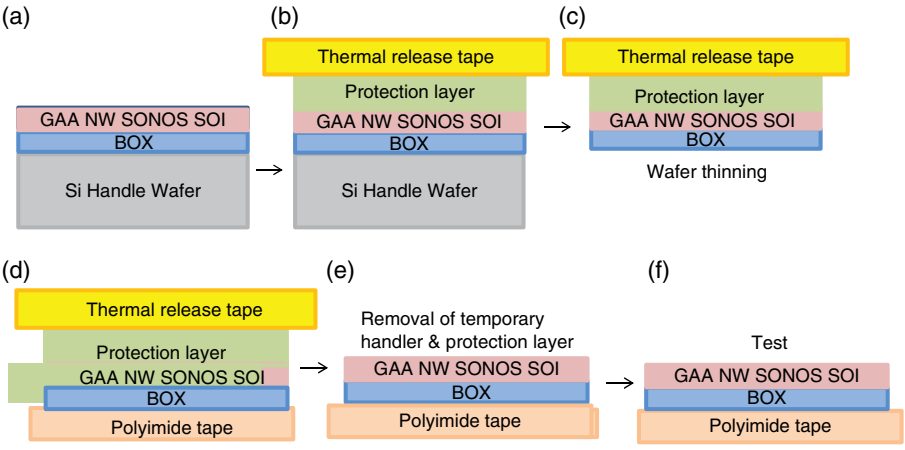


Figure 4.38 Schematic representation of the wafer thinning and transfer process. Based on J.M. Choi *et al.* (KAIST), IEDM, December 2014 [68].

High performance gate-all-around (GAA) Si nanowire (NW) SONOS on plastic substrate was discussed in December of 2014 by KAIST [68]. The ultrathin film of single crystal Si-NW GAA SONOS memory devices were transferred on to a plastic substrate from a host silicon wafer. A schematic representation of the wafer thinning and transfer process is shown in Figure 4.38.

The process uses (a) a GAA NW SONOS silicon-on-insulator (SOI) over a buried oxide (BOX) and silicon handle wafer; then (b) a protection layer coating and thermal release tape is bonded on, (c) the wafer is then thinned down to the buried oxide, and (d) transferred on to a plastic/polyimide tape; (e) the thermal release tape and protection layer is then removed ending with (f) a GAA Si-NW SONOS memory device on a flexible plastic substrate. This device can be tested and packaged. The final film layer is about 1 μm thick and the polyimide flexible backing is about 35 μm thick.

To form the GAA SONOS FETs, an SOI wafer with top silicon thickness of 110 nm was used. The top silicon was thinned to 50 nm by oxidation and etch and patterned to form a 50 nm diameter SiNW. The SiNW was reduced to 30 nm by sacrificial oxidation. The NW was suspended by wet etching the buried oxide beneath it while supported at the ends to the source and drain probe pads. The ONO layer was formed with thicknesses of 3/7/12 nm and implantation of source and drain done. SiO_2 passivation was deposited and via holes cut, followed by Al deposition and patterning. No device degradation was observed during the film transfer.

There was negligible degradation when the substrate was flexed. Typical I–V curves were found. Program time was 10 ms and erase time 100 ms. The V_{th} window was about 4 V at $V_{prm} = 14$ V and $V_{ers} = -14$ V. The result was a high performance nonvolatile memory chip on a flexible polymer substrate.

In December of 2015, KAIST discussed an ultrathin Si-based flexible NAND Flash memory made without using a conventional transfer process [69]. The device and the flexible substrate are interconnected by flip-chip thermocompression bonding with

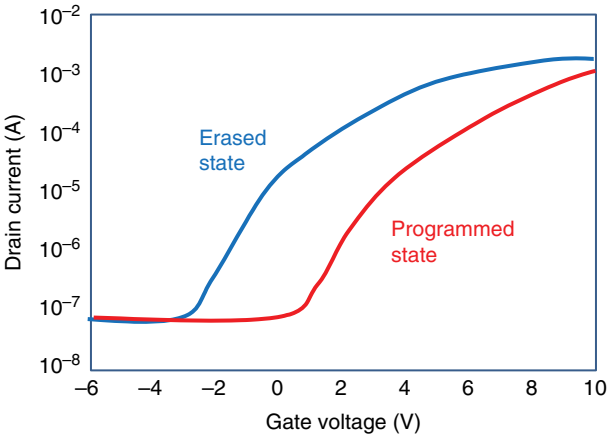


Figure 4.39 *I–V* curves of programmed and erased states of a cell in the flexible NAND Flash. Based on D.H. Kim, *et al.* (KAIST), IEEE IEDM, December 2015 [69].

anisotropic conductive film (ACF). The bottom sacrificial silicon of the SOI wafer was gently etched away, which left flip-chip bonded devices, which were sufficiently thinned to roughly 1 μm to make a very flexible fully packaged NAND Flash memory without cracks or wrinkles.

Using SOI wafers to create flexible memory chips has the advantage of having a piece of single crystalline silicon that is compatible with full CMOS. Other advantages are: that an ultrathin chip is known to be highly flexible, which makes the chip resilient under various stress conditions, it provides a direct interconnection between the device and the outer electrode without requiring an additional wiring process, and it is unnecessary to transfer the device to a flexible substrate since the flip-chip bonded device is already attached to the flexible printed circuit board substrate.

Retention characteristics of the flexible Flash memory showed that the threshold window exceeded 10^4 seconds. Endurance of more than 2000 program/erase cycles was shown. The *I–V* curves of the programmed and erased states of a cell in the NAND Flash memory measured at $V_d = 0.5$ V is shown in Figure 4.39. Programming was at $V = 12$ V and erase at $V = -16$ V.

The final device had reproducible memory operation with good mechanical stability on a flexible substrate. Addressing tests were performed successfully under various conditions. A roll-to-roll process was in development to permit mass production of packaged silicon circuits on flexible substrates.

4.6.2 Creating Thin Chips Using an Underlying Cavity

An additive technique for growing ultrathin silicon chips was discussed in February of 2013 by the Institut für Mikroelecktronik, Stuttgart [70]. This method involved growing layers of crystalline silicon one at a time on a foundation laced with sealed cavities. This resulted in the crystalline silicon layer being attached to an array of small vertical pillars, ensuring that the foundation would be strong enough to support the chip throughout

the processing but weak enough to snap the finished chip off the top of the wafer. The wafer could then be polished and reused as a substrate.

The process involved etching a 1 μm layer of porous silicon into a solid wafer and then a second, 200 nanometer, layer of more coarsely porous silicon beneath that. Both layers were then sintered at high temperature causing the nanopores in the coarse layer to merge into larger pores. The result was one continuous cavity interrupted by vertical pillars. The surface layer then serves as a seed for the crystalline silicon, which is grown over the entire surface of the wafer to the desired thickness. The chip then goes through typical integrated circuit processing. After fabrication is complete, the surface layer is still firmly attached to the thick silicon wafer by the array of pillars within the buried cavity. A deep trench is then etched at the edges of the chip and down into the cavity, which leaves the chip supported by the pillars alone. A pick-and-place tool using a vacuum gripper is used to grab the chip and tug it, snapping the vertical pillars with mechanical force. The tool then places each chip on to a stack or on to a flexible substrate with other thin film components.

In February of 2016, Arizona State University discussed the systematic design of a flexible system [71]. Hybrid flexible systems and system-on-polymer (SoP) are described that combine the advantages of flexible electronics and traditional silicon technology. Flexibility of the chip is considered as a new design metric along with power, performance, and area metrics. A new approach to optimization of placement of rigid components on to a flexible substrate is discussed, which can result in minimizing the loss in flexibility.

Flexible electronics include: mechanically bendable, rollable, conformal, or elastic circuits. Flexible electronics are lighter, thinner, and cost less to make. They can be used in wearable systems as part of clothing or labels. Flexible electronics have already been used in displays, sensors, photovoltaic cells, batteries, and small MCU and RF transmitters. Drawbacks of flexible electronics include: lower degrees of integration and performance and larger variations than in current silicon technology. The solution proposed is a form of hybrid flexible electronics where silicon integrated circuits are combined with flexible substrates. These hybrid devices are designated systems on polymer/plastic/paper (SoP) and flexibility is proposed as a new design metric along with usual performance and area metrics.

In June of 2017, The University of Oulu, Meyer Burger BV, and Tampere University discussed using inkjet printed silver traces and interconnections made with the print-on-slope technique for connection to an RFID chip operating in the ultrahigh frequency range [72]. An underfill material was used to attach silicon RFID chips on to the flexible polymer substrates. The cured underfill was also used as a sloped surface for printing interconnection traces from the chip to the plastic substrate's radiators. A single phase continuous roll-to-roll compatible process was used. The compatibility of various low temperature nanoparticle inks was tested with different substrate materials. The structures were exposed to thermal/humidity tests at 85 °C/85% relative humidity for up to a 2000 hour period. The samples were characterized by a read range measurement device and by optical imaging and field emission scanning electron microscopy. The bulk conductive traces were characterized electrically by measuring their resistances. The conclusion was that inkjet printing is feasible for producing conductive traces for RFID structures and the print-on-slope technique is a cost-effective method for making interconnections between chip and substrate.

Modeling devices and circuits on flexible ultrathin chips were discussed in February of 2017 by the University of Glasgow [73]. Compact models are needed for simulation and prediction of changes in response to device and circuit bending induced stress to make designing circuits for bendable electronics easier. Compressive and tensile stress studies were done on transistors and simple circuits like inverters with different channel orientation of transistors on ultrathin chips. Two fabricated chips were thinned to 20 μm using conventional dicing-before-grinding steps of CMOS processing to obtain sufficient bendability, which was defined as a 20 mm bending radius. Electrical characterization was done of the thinned chip on a flexible substrate. The results showed a change of carrier mobilities in various transistors and threshold voltage of the inverters during different bending conditions. A compact model was developed that predicts the mobility variation, and threshold voltage in compressive and tensile bending stress conditions and orientations. The predictions agreed with the experimental measurements.

4.6.3 Fan-Out Wafer Level Packaging for Assembling Silicon Chips on Flexible Substrate

Fan-out wafer level packaging (FO-WLP) technology extends conventional WLP to provide a greater number of external package contacts while permitting a smaller chip area. The goal is to avoid using expensive silicon for accommodating more I/Os by embedding an individual die in an epoxy-like material that extends the dimensions of the package beyond the silicon die.

The interconnects are redistributed over the base of the extended package, as shown in Figure 4.40. Figure 4.40 (a) illustrates a single layer of redistribution while Figure 4.40 (b) illustrates a two-layer redistribution.

In March of 2017, Sumitomo Bakelite Co. discussed the different types of packaging structures found in more advanced wafer level packaging (WLP) [74]. Advanced FO-WLP is made in a package-on-package (PoP) or single inline package (SiP) with multiple side-by-side chips. An illustration is shown in Figure 4.41 (a) of fan-out PoP and in Figure 4.41 (b) of fan-out SiP with multiple side-by-side chips.

Some changes of process and package structure are required for use of FO-WLP with epoxy molding compound. For fan-out PoP a grinding process is required to expose the face of the Cu pillar on the chip. If the die top thickness is below 70 μm, thicker epoxy molding compound is required to control the optical transparency. An application mentioned is a smartphone with an application processor that includes an FO-WLP as a bottom package in PoP.

In June of 2017, Orbotech discussed an overview of the fan-out wafer level packaging process and its challenges [75]. An illustration of the process is shown in Figure 4.42.

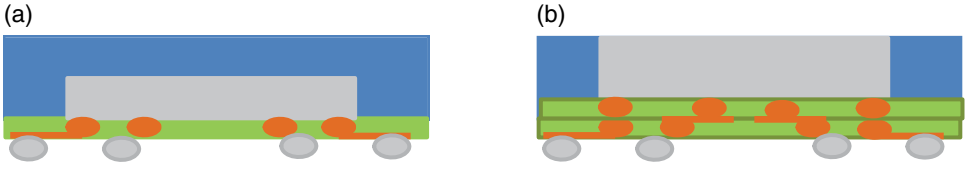


Figure 4.40 Redistribution of interconnects over base of extended fan-out WL package: (a) single layer of redistribution and (b) two layers of redistribution.

The chips are initially placed on a tape-plus-carrier using a placement tool and then the chips are embedded in a mold compound. The cured mold with the embedded chips is released from the carrier. Redistribution layers for the I/Os are formed and bump processing is done. The individual chips are then singulated.

The FO-WLP devices can have thousands of I/O points that are seamlessly connected using finely spaced lines as thin as 2 to 5 μm. Heterogenous devices including different types of chips can be included on the substrate and the system connected in the redistribution layers.

Fan-out wafers feature singulated chips embedded in the epoxy molding compound. Spin-on dielectrics surround the redistribution layers. Advantages include significant package height reduction by elimination of the substrate. Issues include: moisture absorption, outgassing, and limited tolerance to elevated temperatures. While silicon circuits can withstand heat up to 400 °C and can be degassed quickly, the EMC and dielectrics used in FO-WLP withstand heat up to about 120 °C and degassing takes

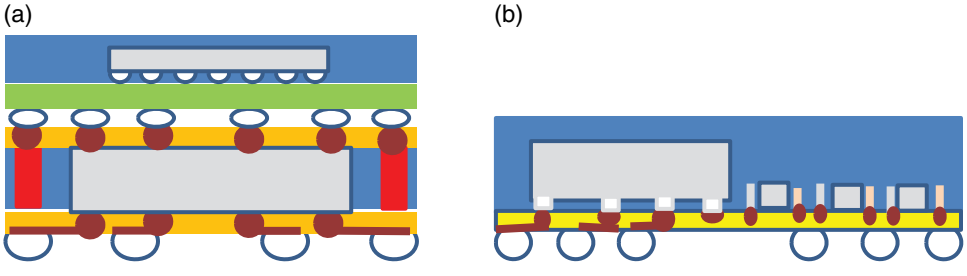


Figure 4.41 Advanced fan-out–wafer level packaging (FO-WLP) in (a) fan-out package-on-package (FO-PoP) and (b) fan-out single-in-line package (FO-SiP). Based on I. Watanabe *et al.* (CSTIC), March 12, 2017 [74].

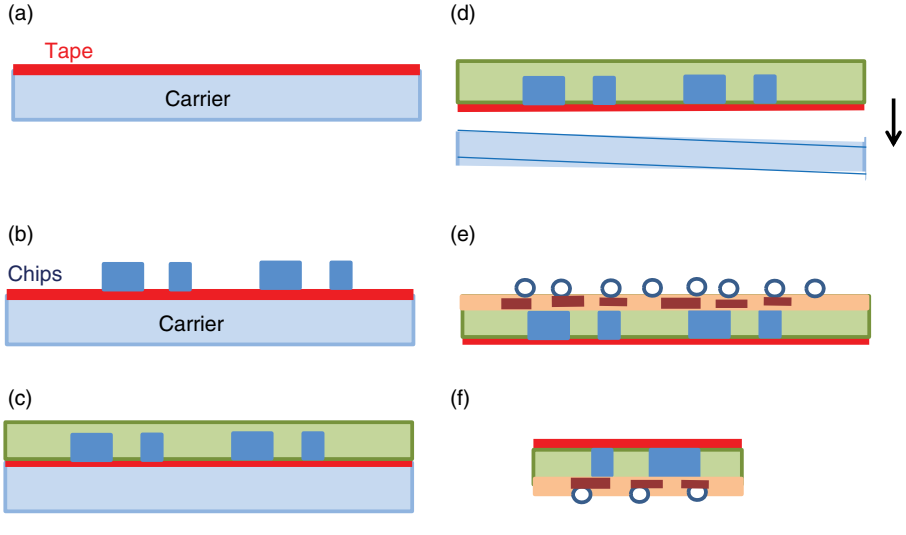


Figure 4.42 Fan-out wafer level packaging process: (a) application of tape, (b) chip placement, (c) wafer level molding, (d) release from carrier, (e) redistribution and bumping, and (f) singulation. Based on SPTS Technology Press Release, August 2016 [75].

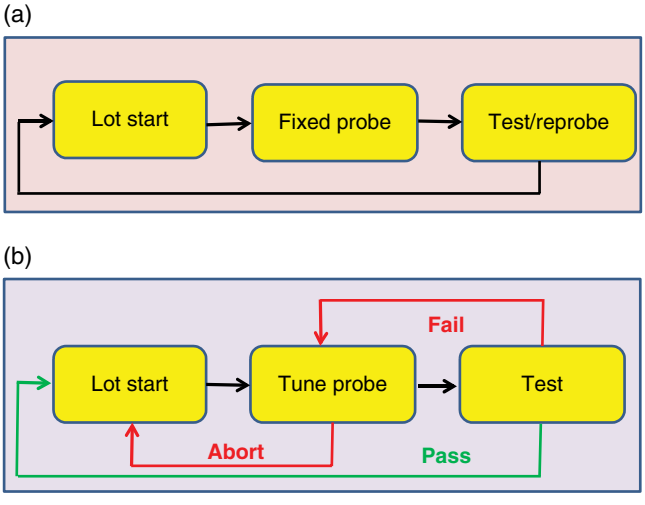


Figure 4.43 Concept of adaptive probing: (a) traditional fixed probing and (b) adaptive probing with probing depth set according to initial test results. Based on M. Lee *et al.* (TSMC), eMDC&ISSM, September 2015 [76].

longer, which reduces the throughput of a conventional sputter system. Multiwafer degas has helped.

In 2015, TSMC discussed probing for the integrated fan-out wafer level chip scale package (FO-WLP) [76]. This package is more cost effective than other current 3D IC packages for consumer mobile wearable and IoT markets. A study was undertaken to improve the quality, yield, and cost of the integrated FO-WLP by probing copper pillars without solder caps. Adaptive probing was developed to improve traditional fixed probe methods so variations of copper pillar, wafer warping, and noncoplanarity of high pin count probes could be tolerated.

The proposed adaptive probing technique can handle varieties of critical probing cases. Adaptive and online tuning for each die are required. Additional probing depth was used based on the depth of the first passed die using a manual trial. Multiple touch-downs on each die are used in probing to be sure the CuO has been penetrated to reduce the electrical resistance. An adaptive programing algorithm was generated. An example of the concept of adaptive probing is shown in Figure 4.43.

In 2016 Kulicke and Soffa Industries discussed thermocompression bonding for FO-WLP using new equipment with much higher throughput than before, which significantly lowered cost per unit [77]. This equipment was developed to support next generation fine pitch, 2.5D and 3D assemblies using Cu interconnects. It was adapted to support die placement for fan-out wafer level packaging (FO-WLP) and also for high accuracy flip-chip placement. Although there was initial concern about the differential cost of the process compared to an earlier process, as the units per hour approached 1400, the difference became negligible.

The first assembly step in an FO-WLP process is the placement of known good die on to a carrier. These dies can be either face up or face down. In single die face down FO-WLP

packages placement accuracy is important but not critical and the required resolution for redistribution layers is typically 10 μm. In more advanced multidie SiP FO-WLP products, the requirement is 5 μm or less and die placement accuracy becomes critical for assembly yield. Underfills applied prior to bonding can be a nonconductive paste (NCP) or a nonconductive film (NCF) that is applied to the wafer prior to dicing. High volume NCF processes are expected to dominate production as the units-per-hour rate increases.

Several fatigue-related mechanisms connected to standalone package reliability or solder joint capability have been investigated successfully for FO-WLP. In 2017, MediaTek discussed a failure mechanism induced by chip-to-board interaction for FO-WLP [78]. This is a failure mechanism specifically related to this packaging technology. Circular cracks were detected on passivation and on the redistribution layer (RDL) due to mismatched thermal expansion among the modules. This is primarily due to the mismatch in thermal expansion between the molding compound and the die, as shown in the fan-out wafer level CSP in Figure 4.44 (a). Figure 4.44 (b) for a fan-in wafer level CSP did not show the problem. The figure illustrates thermal gradient induced stress concentration on the junction among the die edge, molding, and RDL on the FO-WLP. No indication of this type of stress is shown on the fan-in WLCSP, which lacks the effect of the material change from molding compound to silicon die.

The chip-to-board interaction was studied and improved by various passivation techniques, tuning the process window, optimizing the redistribution layer patterns, and by using suitable solder ball material. A simulation model was developed with design guidelines for reducing the problem.

In 2017, Nanyang TU and ChipPAC discussed the effect of a high temperature storage (HTS) test on FO-WLP package strength [79]. There were three different structures of FO-WLP studied, as shown in Figure 4.45.

High temperature storage (HTS) for 1000 hours was done according to the JEDEC standard for high temperature storage life (JESD22-A103). A three-point bending test was used to evaluate the flexing strength. It was conducted at room temperature at three readout points in the high temperature storage test – at 0, 500, and 1000 hours. Results indicated that FO-WLP flex strength increased with an increase in high temperature storage test time. The thick FO-WLP showed the highest flex strength. The back-side

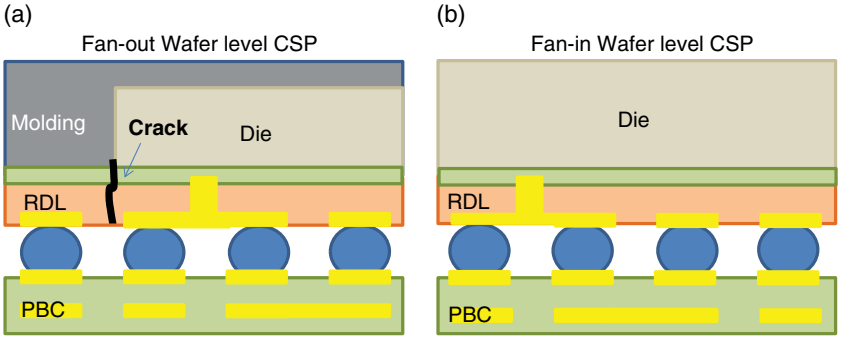


Figure 4.44 Failure mechanism for fan-out but not for fan-in wafer level packaging: (a) example of cracking on fan-out wafer and (b) illustration of no cracking on fan-in wafer. Based on T. Strothmann and H. Clauberg (Kulicke&Soffa), CSTIC, March 2016 [77].

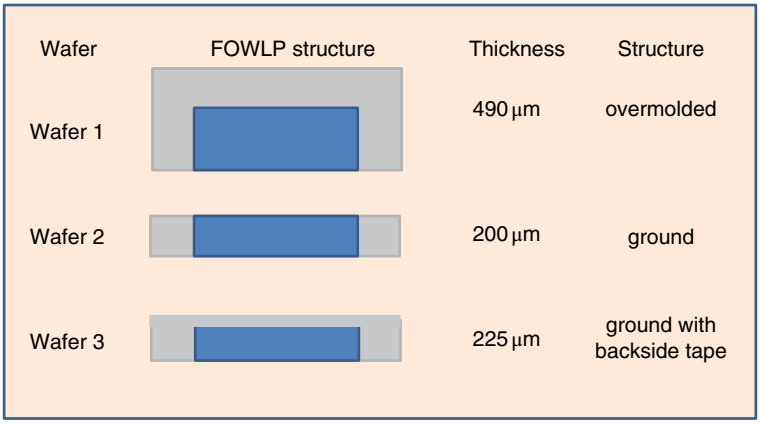


Figure 4.45 An overmolded FO-WLP with 490 µm thickness and two thin FO-WLP both ground to 200 µm thickness. On wafer 3 a thin backside protection tape was laminated. Based on C. Xu *et al.* (Nanyang Technology University), CSTIC, March 2017 [79].

protection tape did not increase the flexure strength but was shown to have a tighter flexure strength distribution.

In 2017, NCTU and Kingyoup Optronics discussed release layers in a temporary bonding scheme involving mechanically supported thin wafer handling for applications such as FO-WLP, 2.5D interposers with TSV, and 3D-IC high density integration with TSV interconnects [80]. The release layer studied was an inorganic amorphous silicon together with a type of polyimide adhesive based on high 355 nm wavelength laser absorption.

An illustration of the process flow for temporary bonding is shown in Figure 4.46 where (a) shows the device wafer and the glass handler, (b) shows the use of the spin-coated adhesive layer applied to the device wafer and the release layer deposited by HDP-CVD to the glass wafer, (c) shows the layers temporarily bonded, and (d) shows the separation using a room temperature laser release process.

Bonding was done under a temperature of 210 °C using a 1 MPa bonding force. Also checked for the bonded structure were: chemical resistance for BEOL processes, mechanical bonding strength between the carrier wafer and an ultrathin device with thickness below 100 µm, and thermal stability up to 300 °C. Release process temperature and high throughput are also significant factors.

In May of 2016, the Fraunhofer Institute and TU Berlin discussed a foldable FO-WLP [81]. In FO-WLP, known good bare dies are embedded into a mold compound to form a reconfigured wafer. A redistribution layer is applied on the reconfigured die, which routes the die pads to the space both around and on the die. Bumps are then formed and the system is singulated. This technology can be used for multichip packages, referred to as system-in-package (SiP) or package-on-package (PoP), which is usually accomplished by stacking and using TSV. This study looks at a foldable FO-WLP that uses a single sided planar process to achieve a stacked 3D package by folding alone. The folding can be done by using a flexible redistribution layer and a singulation process that cuts through only the molding compound leaving the redistribution layer untouched. The feasibility was checked using a multichip package with the dies connected by

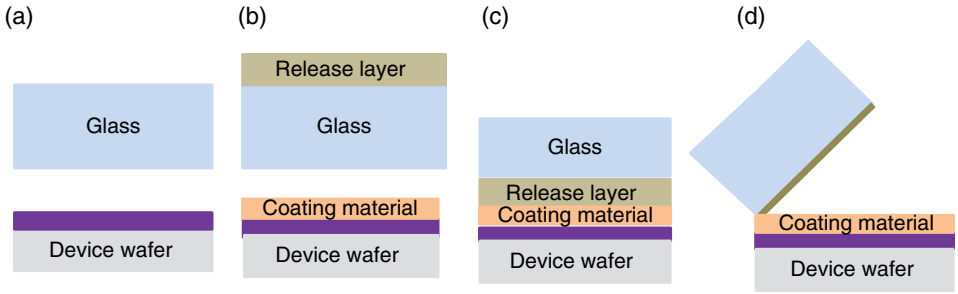


Figure 4.46 Illustration of process flow for temporary bonding before and after the laser ablation process. Based on C.-A. Cheng *et al.* (NCTU, Kingyoup Optronics), *Journal of Electron Devices Society*, 5 (2), March 2017 [80].

wiring done by laminating a polyimide film over the embedded chips. Microvias were then drilled to the die pads using a UV laser and metallized by Cu electroplating. The Cu was etched to form wire and pads. The wafer was diced in two steps: bending cuts were done by dicing through only the molding compound and then the package was folded. This technique can also be used to integrate multiple die packages on to free form surfaces.

To be able to make high performance circuits using ultrathin silicon chips in flexible electronic circuits, simulation and modeling of changes on device and circuit characteristics in response to bending induced stress is needed. In 2017, the University of Glasgow discussed issues in developing a compact model of CMOS devices and circuits on flexible thin chips [82]. Compressive and tensile stress studies on transistors and simple circuits were done. Different designs of devices and circuits in standard 180 nm CMOS were made in two separated chips. The chips were thinned to 20 μm using a conventional dicing-before-grinding technique followed by post-CMOS thinning required to obtain a bendability of 20 mm bending radius. The thinned chips were packaged on a flexible substrate. Results showed a change of carrier mobilities in various transistors and change of switching threshold voltage of inverters during different bending conditions. A compact model was developed that predicts the variation in mobility and threshold voltage using compressive and tensile behind stress conditions and orientations. The model showed agreement with the experimental measurements.

References

1 Hou, Y., Lu, Z., Pu, T. *et al.* (2013) Fast switching protocol for ferroelectric random access memory based on poly(vinylidene fluoride-trifluoroethylene) copolymer ultrathin films (East China University of S&T). *Applied Physics Letters*, 102 (6), 063507, February 2013.

2 Chen, X. *et al.* (2013) P(VDF-TrFE-CFE) terpolymer thin-film for high performance non-volatile memory (Nanjing University). *Applied Physics Letters*, 102, (6), 063101, February 2013.

3 von Nordheim, D. *et al.* (2013) Retention of thin ferroelectric VDF-TrFE copolymer films evaluated from dielectric non-linearities (University of Applied Science, Jena), ISAF/PFM, pp. 208, July 21, 2013.

4 Bak, J.Y. and Yoon, S.M. (2013) Nonvolatile memory characteristics of thin-film transistors using hybrid gate stack composed of solution-processed indium–zinc–silicon oxide active channel and organic ferroelectric gate insulator (Kyung Hee University). *Journal of Vacuum S&T*, 31 (4), 040601, July 2013.

5 Jandhyala, S. *et al.* (2013) Graphene-ferroelectric hybrid devices for multi-valued memory system (University of Texas Dallas). *Applied Physics Letters*, 103 (2), 022903, July 2013.

6 Hou, Y. *et al.* (2013) Unipolar poling-induced high switching speed and improved imprint behaviors for poly(vinylidene fluoride-trifluoroethylene) copolymer ultrathin films (East China University). *Applied Physics Letters*, 103 (26), 263503, December 2013.

7 Nedic, S. *et al.* (2014) High performance non-volatile ferroelectric copolymer memory based on a ZnO nanowire transistor fabricated on a transparent substrate (University of Cambridge). *Applied Physics Letters*, 104 (3), 033101, January 2014.

8 Zhang, X. *et al.* (2014) Temperature dependence of imprint mechanism in poly(vinylidene fluoride-trifluoroethylene) copolymer ultrathin films (Shanghai University of Engineering Science). *Applied Physics Letters*, 104 (10), 103505, March 2014.

9 Petti, L. *et al.* (2014) Influence of mechanical bending on flexible InGaZnO-based ferroelectric memory TFTs (Institute of Technology, Zurich). *IEEE Trans. on Electron Devices*, 61 (4), 1085, April 2014.

10 Cai, R. *et al*. (2014) An organic ferroelectric field effect transistor with poly(vinylidene fluoride-co-trifluoroethylene) nanostripes as gate dielectric (Université Catholique de Louvain). *Applied Physics Letters*, 105 (11), 113113, September 2014.

11 Mao, D. *et al.* (2014) Low temperature processed two-transistor-two-capacitor-based ferroelectric random access memory (Micron., University of Texas Dallas). *IEEE Trans. on Electron Devices*, 61 (10), 3442, October 2014.

12 Micjan, M. *et al.* (2014) Ferroelectric polymer films for flexible memory devices (Slovak University of Technology), ASDAM, October 20, 2014.

13 Bringans, R.D. and Verese, J. (2016) Challenges and opportunities in flexible electronics, IEEE IEDM (PARC), December 2016.

14 Printed sensor system rolls off the press, Thinfilm Electronics Press Release, December 20, 2012.

15 Thinfilm demonstrates first integrated printed electronic system with rewritable memory, Thin Film Press Release, December 20, 2012.

16 Takahashi, D. (2012) Thin Film Electronics demonstrates a way to make disposable interactive smart labels that replace bar codes, Thin Film Electronics, December 19, 2012.

17 Guerin, M. *et al.* (2013) Organic complementary logic circuits and volatile memories integrated on plastic foils (Aix-Marseille Université, IM2NP, CNRS, IM2NP, LTEN-LCEI, CEA-Grenoble). *IEEE Trans. on Electron Devices*, 60 (6), June 2013.

18 Myny, K. *et al.* (2014) Thin-film microprocessor using a hybrid oxide-organic complementary technology with inkjet-printed P2ROM memory (IMEC, Panasonic, KU Leuven, Evonik, Holst), ISSCC, February 2014.

19 Takahashi, D. (2015) Thinfilm teams with Xerox so it can print a billion chips a year for the Internet of Things, https//VentureBeat.com, January 21, 2015.

20 Savastano, Xerox, Thinfilm look to bring Xerox printed memory to market, http//www.printedelectronicsnow.com, September 23, 2015.

21 Xerox launches printed memory products to combat counterfeiting, Xerox Press Release, September 15, 2015.

22 Xerox to set up 1.3 billion capacity Thinfilm smart label print line, Plastic Electronics, https//www.plusplasticelectronics.com, January 21, 2016.

23 Hilson, G. (2016) Printed memory to secure IoT, *EE Times,* June 13, 2016

24 Takahashi, D. (2017) Thinfilm's tiny printed NFC labels enable brands to connect with consumers via smart packaging, March 23, 2017.

25 Huang, T.C. *et al.* (2016) Printed circuits on flexible substrates: opportunities and challenges (Hewlett Packard, Hong Kong University of S&T), Symposium on Networks-on-Chips (NOCS), October 2016.

26 Lee, S.H. *et al.* (2015) Pixel architecture for low-power liquid crystal display comprising oxide and ferroelectric memory thin film transistors (Kyung Hee University, Electronics and Telecommunications Research Institute, Daijeon). *IEEE Electron Device Letters*, 36 (6), 585, June 2015.

27 Izumi, S., Kimura, H. *et al.* (2015) A ferroelectric-based non-volatile flip-flop for wearable healthcare systems (Kobe University, ROHM), Non-volatile Memory Technology Symposium, October 2015.

28 Kyeong-Ah Kim *et al.* (2016) Read-out modulation scheme for the display driving circuits composed of nonvolatile ferroelectric memory and oxide-semiconductor thin-film transistors for low-power consumption (Kyung Hee University, Electronics and Telecommunication Research Institute), *IEEE Trans. on Electron Devices*, 61 (1), January 2016.

29 Li, C. *et al.* (2015) Ultrathin $BaTiO_2$-based ferroelectric tunnel junctions through interface engineering (NUS, Trinity College Dublin). *ACS Nanoletters, Nano Lett.*, 15 (4), 2568–2573, March 2015.

30 Tian, T.T. *et al.* (2016) Tunnel electroresistance through organic ferroelectrics (CAS, Université Paris-Saclay, East China Normal University). *Nature Communications*, 7, 11502, May 4, 2016.

31 Huang, R., Tang, Y., Kuang, Y. *et al.* (2012) Resistive switching in organic memory device based on parylene-C with highly compatible process for high-density and low-cost memory applications (Peking University, Princeton University). *IEEE Trans. on Electron Devices*, 59 (12), 3578, December 2012.

32 Bai, W., Huang, R., Cai, Y. *et al.* (2013) Record low-power organic RRAM with sub-20 nA reset current (Peking University), *IEEE Electron Device Letters*, 34 (2), 223, February 2013.

33 Huang, R. *et al.* (2014) Resistive switching in organic memory devices for flexible applications (Peking University), ISCAS, June 2014.

34 Sakamoto, T., Tada, M., Okamoto, K. and Hada, H. (2012) Electronic conduction mechanism in atom switch using polymer solid electrolyte (LEAP). *IEEE Trans. on Electron Devices*, 59 (12), 3574, December 2012.

35 Banno, N. *et al.* (2015) Cu atom switch with steep time-to-ON-state versus switching voltage using Cu ionization control (LEAP). *IEEE Trans. on Electron Devices*, 62 (9), September 2015.

36 Tada, M. *et al.* (2015) ON-state reliability of Cu atom switch under current–temperature stress (NEC). *IEEE Trans. on Electron Devices*, 62 (9), September 2015.

37 Okamoto, K. *et al.* (2016) Robust Cu atom switch with over-400 °C thermally tolerant polymer-solid electrolyte (TT-PSE) for nonvolatile programmable logic (NEC), VLSI-T, June 2016.

38 Tada, M. and Sakamoto, T. (2017) Set/reset switching model of Cu atom switch based on electrolysis (NEC). *IEEE Trans. on Electron Devices*, 64 (4), April 2017.

39 Chou, K.I., Cheng, C.H., Zheng, Z.W. *et al.* (2013) Ni/GeO_x/TiO_y/TaN RRAM on flexible substrate with excellent resistance distribution (NCTU, National Taiwan Normal University, IM-CAS). *IEEE Electron Device Letters*, 34 (4), 505, April 2013.

40 Liu, Y. *et al.* (2014) Inorganic–organic hybrid switching memory with high uniformity and multilevel operation (IM Peking University), VLSI-TSA, April 28, 2014.

41 Lin, C.C. (2014) Resistive switching behavior of Al/Al_2O_3/ZrO_2/Al structural device for flexible nonvolatile memory application (National Dong Hwa University). *IEEE Trans. on Magnetics*, 50 (7), July 2014.

42 Hou, T.H. *et al.* (2013) Low-cost embedded RRAM technology for system-on-plastic integration using a-IGZO TFTs (NCTU), AM-FPD, July 2, 2013.

43 Wu, S.C. *et al.* (2013) Flexible three-bit-per-cell resistive switching memory using a-IGZO TFTs (NCTU). *IEEE Electron Device Letters*, 34 (10), 1263, October 2013.

44 Kim, S.J., Won-Ho Lee and Sung-Min Yoon (2015) Nonvolatile memory performances of transparent and/or flexible memory thin-film transistors using IGZO channel and ZnO charge-trap layers (Kyung Hee University), AMFPD, July 2015.

45 Rocha, P.R.F. *et al.* (2013) The role of internal structure in the anomalous switching dynamics of metal-oxide/polymer resistive random access memories (Institute de Telecom, Lisboa, Portugal). *Journal of Applied Physics*, 113 (13), 134504, April 2013.

46 Bogusz, A. *et al.* (2013) Resistive switching in thin multiferroic films (Chemnitz University of Technology, University of Electronic S&T Chengdi, Institute of Ion-Beam Physics and Materials Research Dresden, Institute of Integrated Nanosciences Dresden), ISCDG, September 26, 2013.

47 Mutiso, R.M. *et al.* (2013) Resistive switching in silver/polystyrene/silver nano-gap devices (University of Pennsylvania). *Applied Physics Letters*, 103 (22), 223302, November 2013.

48 Song, H.J. (2013) Switchable diode effect in polycrystalline Bi3.15Nd0.85Ti3O12 thin films for resistive memories (Xiangtan University). *Applied Physics Letters*, 103 (26), 262901, December 2013.

49 Tsai, H.Y. *et al.* (2014) Resistive switching behaviors of Au/pentacene/Si-nanowire arrays/heavily doped n-type Si devices for memory applications (National Changhua University of Education). *Applied Physics Letters*, 104 (5), 053501, February 2014.

50 Huang, J. and Ma, D. (2014) Electrical switching and memory behaviors in organic diodes based on polymer blend films treated by ultraviolet ozone (Changchun Institute of Applied Chemistry). *Applied Physics Letters*, 105 (9), 093303, September 2014.

51 Santoni, F. *et al.* (2016) Modeling of filamentary conduction in organic thin film memories and comparison with experimental data (University of Rome, TU Munchen, CNR-ISMN, NanoTecCenter Weiz, Graz University of Technology). *IEEE Trans. on Nanotechnology*, 15 (1), January 2016.

52 Chao, H., Chongjie Wang, Wei, R. *et al.* (2016) Bipolar resistive switching in Al/GO-PEDOT: PSS/Pt memory devices (Tsinghua University, Beijing University of Chemical Technology, Fuzhou University), EDSSC, August 2016.

53 Lee, W.H., Kim, E.J. and Yoon, S.M. (2016) Multilevel resistive-change memory operation of Al-doped ZnO thin-film transistor (Kyong Hee University). *IEEE Electron Device Letters*, 37 (8), August 2016.

54 Shih, C.C. *et al.* (2016) Ultra-low switching voltage induced by inserting SiO_2 layer in indium-tin-oxide-based resistance random access memory (Nat.ional Sun Yat-Sen University, University of Texas, Austin, National Cheng Kung University, National Kaohsiung Normal University, Xiamen University). *IEEE Electron Device Letters*, 37 (10), October 2016.

55 Vescio, G. *et al.* (2017) Inkjet printed HfO_2-based ReRAMs: First demonstration and performance characterization (University de Barcelona, University Autonoma de Barcelona). *IEEE Electron Device Letters*, 38 (4), April 2017.

56 Lee, K.J. *et al.* (2017) Effects of Ni in strontium titanate nickelate thin films for flexible nonvolatile memory applications (National Cheng-Kung University, National University of Kaohsiung, Air Force Institute of Technology, Kaohsiumg). *IEEE Trans. on Electron Devices*, 64 (5), 2001, May 2017.

57 Semple, J. *et al.* (2017) Semiconductor-free nonvolatile resistive switching memory devices based on metal nanogaps fabricated on flexible substrates via adhesion lithography (Imperial College London, King Abdullah University of S&T). *IEEE Trans. on Electron Devices*, 64 (5), 1973, May 2017.

58 Brunel, B., Levesque, P.L., Ardiaca, F. *et al.* (2013) Control over the interface properties of carbon nanotube-based optoelectronic memory devices (CEA, IRAMIS, CNRS, LEM). *Applied Physics Letters*, 102 (1), 013103, January 2013.

59 Cosmin Obreja, A. *et al.* (2014) Charge transport and memristive properties of graphene quantum dots embedded in poly(3-hexylthiophene) matrix (National Institute for R&D in Microtech, Romania). *Applied Physics Letters*, 105 (8), 083303, August 28, 2014.

60 Lin, J. *et al.* (2015) Solution-processed, flexible, and transparent non-volatile memory with embedded graphene quantum dots in polymethylsilsesquioxane layers (Fuzhou University, Hanyang University). *IEEE Electron Device Letters*, 36 (11), November 2015.

61 Wang, Y. *et al.* (2013) Ink-jet printed In-Ga-Zn oxide nonvolatile TFT memory utilizing silicon nanocrystals embedded in SiO_2 gate dielectric (Technology University of Singapore), Nanoelectronics Conference (INEC), pp. 222, January 2, 2013.

62 Yun, D.Y. (2014) Multilevel charging and discharging mechanisms of nonvolatile memory devices based on nanocomposites consisting of monolayered Au nanoparticles embedded in a polystyrene layer (Hanyang University). *Applied Physics Letters*, 104 (2), 023304, January 20, 2014.

63 P. Bousoulas, P. *et al.* (2015) Improving the resistive switching uniformity of forming-free TiO_{2-x} based devices by embedded Pt nanocrystals (National Technical University of Athens), ESSDIRC, September 2015.

64 Shih, C.C. *et al.* (2016) High performance transparent transistor memory devices using nano-floating gate of polymer/ZnO nanocomposites (National Sun Yat Sen University), Scientific Reports, February 1, 2016.

65 Kim, S.J. *et al.* (2016) High performance and stable flexible memory thin-film transistors using In-Ga-Zn-O channel and ZnO charge-trap layers on poly (ethylene naphthalate) substrate (Kyung Hee University, Electronics and Telecomm Research Institute). *IEEE Trans. on Electron Devices*, 63 (4), April 2016.

66 Wang, D.T. *et al.* (2016) Resistive switching and synaptic behaviors of $TaN/Al_2O_3/ZnO/ITO$ flexible devices with embedded Ag nanoparticles (Fudan University), *IEEE Electron Device Letters*, 37 (7), 878, July 2016.

67 Woo, M.H. *et al.* (2016) Floating gate memory based on MoS2 channel and iCVD polymer tunneling dielectric (KAIST), ESSDERC, September 2016.
68 Choi, J.M., Han, J.W. and Choi, Y.K. (2014) Flexible high-performance nonvolatile memory by transferring GAA silicon nanowire SONOS onto a plastic substrate (KAIST), IEDM, December 2014.
69 Kim, D.H. *et al.* (2015) ACF-packaged ultrathin Si-based flexible NAND Flash memory (KAIST), IEEE IEDM, December 2015.
70 Burghartz, J.N. (2013) Make way for flexible silicon chips (Institut fur Mikroelektronik Stuttgart), IEEE Spectrum, February 25, 2013.
71 Gupta, U. *et al.* (2016) Can systems extend to polymer? SoP architecture design and challenges (Arizona State University), IEEE SOCC, February 2016.
72 Putalla, J. *et al.* (2017) Capability assessment of inkjet printing for reliable RFID applications (University of Oulu, Meyer Burger BV, Tampere University). *IEEE Trans. on Device and Materials Reliability*, 17 (2), 281, June 2017.
73 Vilouras, A. *et al.* (2017) Modeling of CMOS devices and circuits on flexible ultrathin chips (University of Glasgow). *IEEE Trans. on Electron Devices*, 64 (5), May 2017, 2038, February 23, 2017.
74 Watanabe, I. *et al.* (2017) Latest material technologies for fan-out wafer level package (Sumitomo Bakelite Co.), Semiconductor Technology International Conference (CSTIC), March 12, 2017.
75 Fan-out wafer level packaging: Breakthrough advantages and surmountable challenges (SPTS Technology), Press Release, August 2016.
76 Lee, M. *et al.* (2015) Novel adaptive probing for wafer level chip scale package (TSMC), eMDC&ISSM, September 2015.
77 Strothmann, T. and Clauberg, H. (2016) High throughput thermocompression bonding enabled with a flexible manufacturing platform (Kulicke&Soffa), CSTIC, March 2016.
78 Yu, C.K. *et al.* (2017) A unique failure mechanism induced by chip to board interaction on fan-out wafer level package (MediaTek), IRPS, April 2017.
79 Xu, C., Zhong, Z.W. and Choi, W.K. (2017) Effect of high temperature storage on fan-out wafer level package strength (Nanyang Technology University), CSTIC, March 2017.
80 Cheng, C.-A. *et al.* (2017) Feasibility investigation of amorphous silicon as release layer in temporary bonding for 3-D integration and FOWLP scheme (National Chiao Tung University, Kingyoup Optronics). *Journal of Electron Devices Society*, 5 (2), March 2017.
81 Braun, T. *et al.* (2016) Foldable fan-out wafer level packaging (Fraunhofer Institüt Berlin, TU Berlin), ECTC, May 2016.
82 Vilouras, A. *et al.* (2017) Modeling of CMOS devices and circuits on flexible ultrathin chips (University of Glasgow). *IEEE Trans. Electron Devices*, 64, (5), May 2017.

5

Neuromorphic Computing Using Emerging NV Memory Devices

5.1 Overview of Resistive RAMs and Ferroelectric RAMs in Neuromorphic Systems

Neural network systems, imitating the human brain, are also a type of interconnected network and use many of the same low cost, low power types of devices as other smart networks. Neural networks can require multilevel, analog like, memory devices as synapses connecting neurons. These single neuromorphic chips can be used for data processing in applications such as voice and vision recognition at the edge of the Internet of Things. Local intelligent nodes can analyze local data and send the results on providing an extra level of data security.

Various types of resistive RAMs (RRAMs) and ferroelectric RAMs (FeRAMs) have been used in artificial neural networks. These technologies can emulate synaptic plasticity and learning rules such as spike-timing dependent plasticity (STDP) in which the synapses, which are connections between neurons, are altered in response to stimulus. The persistence of this change enables learning, that is, synaptic changes are persistent so the network state can be reactivated and stimulated further later. The multivalue property of RRAMs can be increasing or decreasing. When used as a synapse in a neural net, if the conductance is increasing, the property is called potentiation, and if the conductance is decreasing, the property is called depression.

Direct mapping can be provided using learning algorithms. The low power consumption of RRAMs and their plasticity or multivalue properties make them promising for implementing synaptic behavior in complex systems.

5.2 Various Resistive RAMs for use as Synapses in Neuromorphic Systems

5.2.1 Metal Oxide Resistive RAM (MO-RRAMs) as Synapses

Metal oxide RRAM cells consist of a metal–insulator–metal (MIM) structure where the insulator can undergo abrupt switching into a conductive state and gradual switching into a nonconductive state. This resistive switching is related to the reversible formation and disruption of a conductive filament (CF), which provides a shunt for the top and bottom electrodes through the insulating layer. The CF formation and dissolution can be related to the displacement of oxygen atoms.

Memories for the Intelligent Internet of Things, First Edition. Betty Prince and David Prince.
© 2018 John Wiley & Sons Ltd. Published 2018 by John Wiley & Sons Ltd.

In June of 2014, Arizona State University discussed using a metal oxide RRAM (MO-RRAM) as a synaptic device for a hardware implementation of neuromorphic computing [1]. The gradual resistance modulation capability in RRAM can emulate analog synapses and the stochastic switching behavior of the RRAM could emulate binary synapses. Neuromorphic computing is of interest since it uses massive parallelism from distributed computing and localized storage in neural networks. Its error tolerance is interesting for use in vision and speech recognition. Neuromorphic computing can be run in a conventional digital computer enabled by software. Two terminal RRAMs can be used as artificial synapses. Metal oxide RRAM is interesting for neuromorphic systems due to low energy consumption, compatibility with CMOS technology, and potential for Gbit density through 3D integration.

A hybrid neuromorphic system with CMOS neurons and RRAM synapses was demonstrated. During RESET a conductive filament (CF) is ruptured and a variable tunnel gap forms between an electrode and the residual filament. The variation in tunnel gap results in multilevel resistance states. The SET is abrupt. Resistance versus number of voltage pulses is shown Figure 5.1 for (a) an abrupt SET and (b) a gradual RESET.

The RRAM can function as the analog synapse using the gradual RESET and the binary synapse using the stochastic SET for competitive learning. A simulation of a winner-take-all network for an orientation classification function was discussed. Comparable system performance was shown between the analog synapse and the binary synapse.

Any resistive memory with multilevel characteristics can be used as a synaptic device in a neuromorphic computing system. In 2016 POSTECH [2] investigated abruptly formed filaments and found this abrupt formation degrades analog switching behavior during potentiation. By using a balanced SET/RESET pulse operation and modulating the switching gap to control the filament growth, the conductance behavior could be gradually changed.

An HfO_2-based RRAM array with multilevel states under a constant pulse was shown to have a controlled filament. The RRAM cell had a TiN/HfO_2/Ti/TiN stack made on top of a transistor in an 8 inch wafer process. An I–V curve under 100 μA with 1.2 V was found. A conductive filament (CF) of oxygen vacancies (ions) was formed under positive bias and gradually dissolved with negative bias. This resulted in a switching gap between the CF and the bottom electrode.

Synaptic behavior was evaluated by applying successive identical pulses. The multilevel states of conductance were observed in depression and a binary state was shown during potentiation. Synaptic behavior was demonstrated. A schematic illustration of a plausible mechanism is shown in Figure 5.2. Figure 5.2 (a) shows the initial filament (SET) while Figure 5.2 (b) shows a RESET with a partially ruptured conductive filament. Figure 5.2 (c) represents the second SET showing a thicker conductive filament.

In August of 2016, POSTECH discussed improving synaptic behavior of a neuromorphic system using identical pulses on a filamentary AlO_x/HfO_2 bilayer RRAM array to emulate the synaptic function [3]. It was found that the multilevel states of conductance were attained by varying measurement conditions for the formation and rupture of a conductive filament.

A TiN/HfO_2/Ti/TiN stack was used for the RRAM. Multiple current states were found by either varying the RESET voltage from −0.7 to −1.9 V with a step voltage of −0.05 V negatively or varying the gate voltage from 0.6 to 1.2 V with a step voltage of 0.05 V positively.

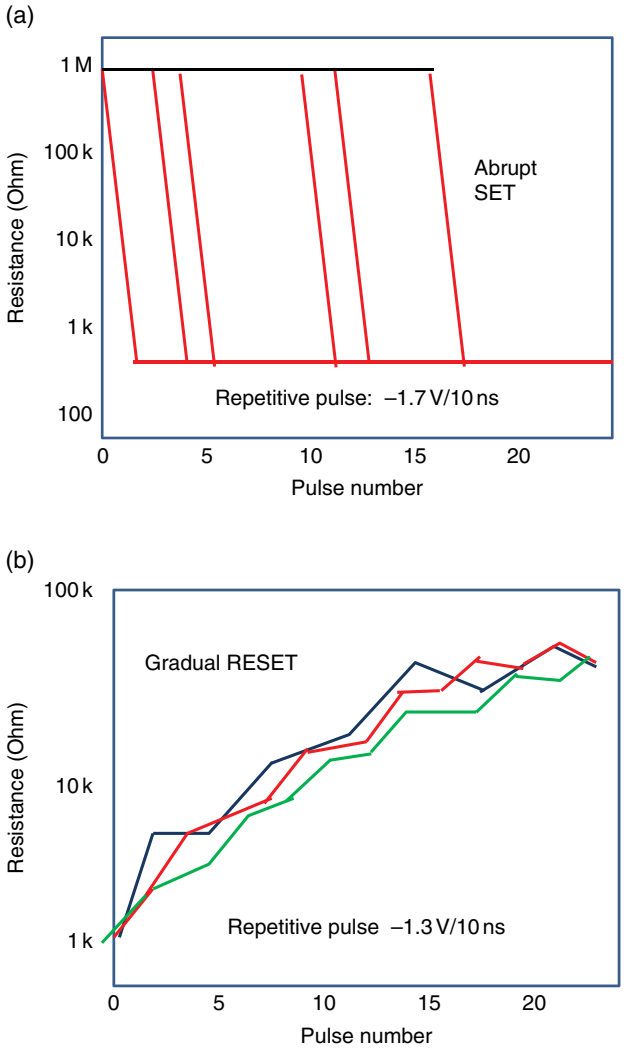


Figure 5.1 Resistance versus pulse number for an MO-RRAM: (a) an abrupt SET and (b) a gradual RESET operation. Based on S. Yu (ASU), ISCAS, June 2014 [1].

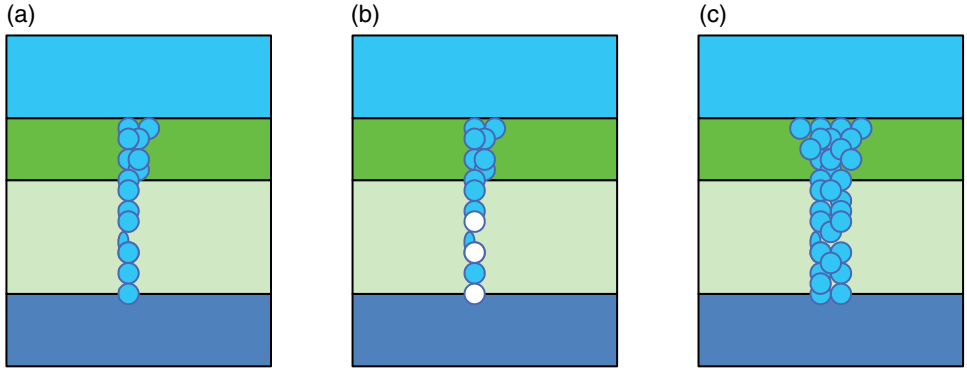


Figure 5.2 Schematic illustration of plausible mechanism of multilevel states of conductance: (a) initial filament (SET), (b) RESET with partially ruptured conductive filament, and (c) second SET with thicker conductive filament. Based on J. Woo *et al.* (POSTECH), SNW, June 2016 [2].

Abrupt switching was found to lead to an unchanged conductance state causing degradation in the accuracy of pattern recognition. A linear potentiation or depression behavior of conductance was shown by applying identical pulses using the effect of a barrier layer on switching. This was done by making an RRAM on top of an Al electrode. It was found that when the range of conductance was symmetrically controlled at both polarities improved accuracy was achieved for pattern recognition using a neural network.

An optimized programming method for linear potentiation in a filamentary HfO_2 RRAM synapse intended for use in a neuromorphic system was discussed by POSTECH in December of 2016 [4]. Various programming methods were used to linearly change the resistance state with either SET voltage/pulse ramping or Gate voltage ramping. The results showed that the multilevel cell (MLC) was related to the size of the conductive filament where the oxygen vacancies moved under the influence of applied pulses. By optimizing the pulse for a given condition, such as identical pulses, it was possible to linearly increase the MLC behavior, which enabled pattern recognition accuracy in neuromorphic systems.

For proper operation of an RRAM synaptic device, a monotonic increase or decrease of the MLC characteristics should be expected. Recent studies have, however, focused on the use of gradually transited MLC in depression with abrupt formation of conductive filament in potentiation. This asymmetric MLC behavior tends to lead to a degradation of pattern recognition accuracy. A possible solution is two identical synaptic devices serving as a single synapse, but this consumes a large amount of power per area.

An alternative solution is linearly increased MLC behavior in potentiation and in depression. This was achieved by using different pulse methods for potentiation and for depression. A CMOS compatible $TiN/HfO_2/Ti/TiN$ RRAM structure was used with a 0.35 μm transistor for a 1T1R configuration in a 200 mm wafer process. During potentiation, the gate voltage is ramped at a fixed set voltage. The increased currents are proportional to the large gate voltage due to the thickness of the filament formation. Ramping the gate voltage at a fixed SET voltage results in an MLC operation during potentiation due to thickening of the filament.

For depression, an optimized RESET pulse can lead to dissolving the conductive filament with a particular switching gap, which enables intermediate states as a result of varying partially dissolved conductive filament gaps. As a result, MLC operation with varied programming schemes during potentiation resulted in MLC states during potentiation and depression, which improved the accuracy for pattern recognition applications in neuromorphic systems.

5.2.2 Conductive Bridge RRAM (CB-RRAM) as Synapses

In the case of conductive bridge RRAMs (CB-RAM) cells, the CF formation and dissolution is related to metal ions. Both CB-RAMs and MO-RRAM cells have low operating voltages, fast programming, nanoseconds of read time, low power consumption, scalability, and endurance that is higher than conventional Flash memory.

Various trade-offs exist between the memory window and endurance and the memory window and switching time. The low resistance state (LRS) when the CF is fully formed is modulated by the compliance current, which is applied so the higher the programming current the thicker the CF and the lower the resistance state. In 2015, CEA LETI described an example of an MO RRAM and a CB-RAM for use as synapses [5]. The MO RRAM was formed by two electrodes, in this case, one Ti and one TiN with

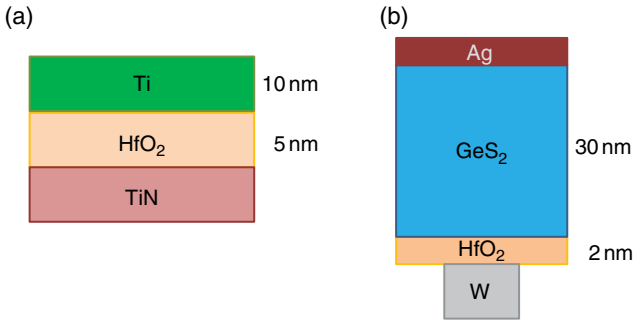


Figure 5.3 Schematic cross-sections of (a) a metal oxide MO-RAM in 28 nm CMOS and (b) a CB-RAM in 130 nm CMOS. Based on B. DeSalvo *et al.* (CEA LETI), IMW, May 2015 [5].

an HfO_2 insulating layer between them. The CB-RAM was formed with a W/HfO_2 electrode and an Ag or Cu electrode with a GeS_2 electrolyte between them. In the case of the CB-RAM, the CF will be formed of Ag or Cu atoms. Schematic cross-sections of the MO RRAM and a CBRAM are shown in Figure 5.3.

The multilevel capability of a solid electrolyte RRAM, a form of conductive bridge RAM (CB-RAM), was explored for use as a synapse device by several Chinese and Taiwanese Universities in August of 2015 [6]. The device used a Pt/GeSO/TiN structure. The cut-off voltage was varied during the I–V cycles, resulting in continuous multilevel conductance states. It was also possible to vary the AC pulse programming and voltage amplitudes to obtain multilevel states. It was noted that arithmetic operations based on analog memories may also be an interesting computational paradigm to complement digital von Neumann systems. An electronic synapse device was made with a Pt/GeSO/TiN thin film stack. The 16 nm thick switching layer was RF magnetron sputtered on to a $TiN/Ti/SiO_2/Si$ substrate. The multilevel bipolar resistance switching I–V curves of the Pt/GeSO/TiN device are shown in the diagram in Figure 5.4.

A series of gradually increasing set pulse voltages and pulse widths were applied to the Pt/GeSO/TiN device and the conductance gradually increased with increasing pulse voltage and pulse width. The multiple levels are shown as nearly continuous. This multilevel effect can be useful in systems emulating biological synapses.

Certain characteristics are important when investigating the capability of the device to emulate adaptive learning in biological synapses. In a biological synapse, when the pre-spike signal precedes the post-spike signal, the strength of the synapse undergoes long term potentiation (LTP) or increase; otherwise it undergoes long term depression (LTD) or decrease. Closer timing spikes lead to a larger change of feedback. This is called the spike timing-dependent plasticity (STDP) learning rule, which has been shown in various brain circuits for many species of creature. STDP learning was confirmed in this CBRAM device.

5.2.3 Phase Change Memory (PCM) as Synapses

Technologies for brain-type computing intended to overcome implementation issues for neural networks can also use resistive memory such as phase change memory (PCM). These programmable resistive RAMs were used in March of 2014 by CEA-LETI, CEA-LIST, and InRIA to obtain low area synapses [7]. PCM are made of

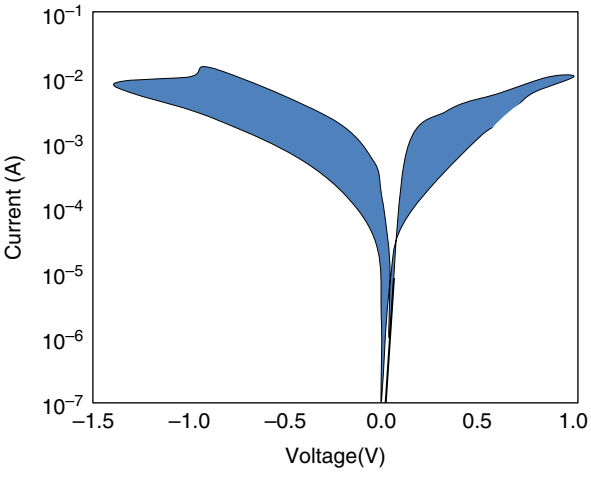


Figure 5.4 *I–V* curve for electronic synapse CBRAM device using Pt/GeSO/TiN Stack with gradually increasing set pulse voltages and pulse widths resulting in a continuous multilevel effect, Based on W. Zhang (Xidian University, NCKU, National Sun Yet-sen University, Xiamen University, NCTU), *IEEE Electron Device Letters*, August 2015 [6].

chalcogenide glass material, which can be either in an amorphous or a crystalline state depending on the rate of heating and cooling of the material. The crystalline phase has a lower resistivity than the amorphous phase, which results in a resistive RAM with multiple states that can be used as variable state synapses.

Through-silicon-vias (TSV) were used with the PCM in this demonstration to improve analog neuron connections, reduce their area, and improve connectivity. The conductance of the PCM is modulated by the heat from the current flowing through the PCM devices. These devices were used to produce arrays of artificial synapses.

A 2-PCM synapse was discussed that emulated synaptic functions in large scale neural networks. The two PCM devices had opposite contributions to the neuron integration. Partial crystallization of the phase change material of one PCM was used to increase the equivalent weight of the synapse and partial crystallization of the other to decrease it.

5.2.4 PCMO RRAM as Synapses

A Pr0.7Ca0.3MnO3 (PCMO) RRAM-based synaptic device for neuromorphic computing was developed in 2014 by Pohang University of Science and Technology (POSTECH) and the Gwangju Institute of Science and Technology [8]. Thin film elements of PCMO can be made with thicknesses down to a few tens of nanometers. A wafer scale device was made and showed good switching uniformity and analog memory behavior. Figure 5.5 shows bipolar I–V curves at different sweep voltages, illustrating the possibility of analog memory behavior.

The PCMO-RRAM was used to implement analog artificial synapse characteristics in neuromorphic systems. A face recognition function was used to show varying potentiation and depression characteristics of the RRAM synapse. Figure 5.6 shows potentiation and depression curves. Potentiation voltage is –3 V and depression voltage is +3 V.

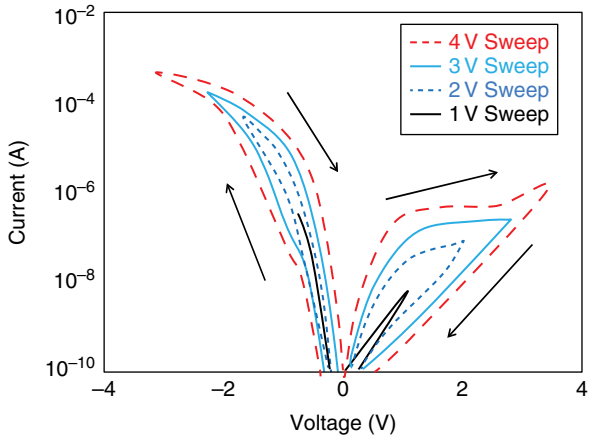


Figure 5.5 Illustration of RRAM *I–V* curves at different sweep voltages showing potential for analog behavior of a PCMO RRAM. (Based on J. Jang *et al.* (Pohang University, Gwangju Institute of Science and Technology), ISCAS, June 2014 [8].

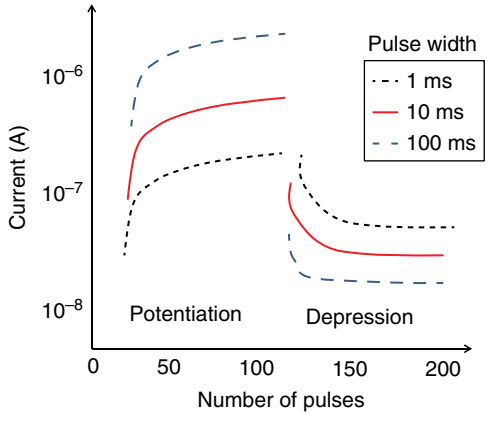


Figure 5.6 Illustration of potentiation and depression curves for PCMO-RRAM. Potentiation voltage is −3 V and depression voltage is +3 V. Based on J. Jang *et al.* (Pohang University, Gwangju Institute of Science and Technology), ISCAS, June 2014 [8].

Simulations were performed for time series prediction and face recognition applications based on a proposed RRAM synaptic device model. A multilayer neuromorphic architecture was designed. The architecture consisted of 408 CMOS neurons and 83 200 RRAM synapses. For negative synaptic weight implementation, two RRAM synaptic devices were used as one synapse.

In May of 2015, POSTECH, Gwangju Institute of S&T, and IBM discussed the optimization of conductance change behavior in analog resistive memory synaptic devices based on $PrCaMnO_3$ (PCMO) [9]. In this study, resistive memory based on PCMO is applied in a neural network application, which is classification of the MNIST handwritten digits using a multilayer perceptron trained with back propagation under a variety of simulated conductance change behaviors. Linear and symmetric conductance changes during increasing and decreasing device conductance are shown to offer the highest classification accuracies. Additional improvements could be obtained using nonidentical training pulses at the cost of requiring measurement of individual conductance during training. Using PCMO-based synaptic devices, a generalization accuracy on a previously unseen test set was 90.55%.

PCMO RRAM memory arrays of 1-Kbit were made to evaluate synaptic devices. The device stack had a 50 nm thick Pt bottom electrode, with 30 nm deposited polycrystalline PCMO film patterned with conventional lithography and reactive ion etching (RIE), followed by an 80 nm thick SiN_x layer. The switching behavior of an RRAM based on this PCMO stack was determined and neural network simulations were done using its conductance change behavior. Handwriting digit classification accuracies were found to be high for RRAM-based synaptic devices with symmetric switching behavior. They are maximized when that response is linear.

In August of 2016, POSTECH, Gwangju Institute of S&T, and IMEC discussed an analog synapse PCMO-based interface switching device with varying electrode materials [10]. It was compared with a filamentary switching device having 1-bit storage and variability issues. The interface switching devices showed good electrical properties, which included 5-bit (32 level) multilevel cell characteristics along with wafer scale switching uniformity and scalability of switching energy with device area. An Mo electrode was proposed to improve data retention of the interface switching device. The Mo electrode increased the oxidation barrier height to 0.4 eV, which improved the retention time and pattern classification accuracy of the neural networks.

A 10 nm thick polycrystalline PCMO film was deposited on a $Pt/Ti/SiO_2/Si$ wafer using RF sputtering. A 100 nm SiN_x vertical side wall was deposited by plasma-enhanced CVD and then dry etched using RIE with a 150 nm via-hole structure. Various top electrodes were used including: Al, Ti, Ta, and Mo. The data retention characteristics and switching uniformity of the basic interface RRAM were studied with varying electrode materials for synapse applications. Using neural network simulations it was confirmed that data retention characteristics were critical to guarantee the pattern classification accuracy.

An Mo/PCMO synapse device was reported with good wafer scale switching uniformity, multilevel data storage capability, and good retention characteristics. The improved LRS retention characteristics were explained by the high activation energy of the oxidation process and high electronegativity of the Mo electrode. It was confirmed that an improvement in data retention directly affected the pattern classification accuracy of the neural network.

5.2.5 RRAM with Simultaneous Potentiation and Depression

A homogeneous barrier modulation (HBM) RRAM that was capable of a simultaneous potentiation (P) and depression (D) was described in October of 2015 by National Chiao Tung University [11]. A simulation of pattern learning algorithm was performed based on the winner-take-all rule and on experimental synaptic characteristics. A circuit diagram of the simulated neuromorphic system and applied signal waveforms is shown Figure 5.7.

The winner-take-all learning rule was applied to update the synaptic weight of each synapse through training. This clock-synchronized system included two phases: (1) training during which one set of training voltage patterns was sent into the synapse array by pre-neurons (column) while post-neurons (row) sum up the current through the synaptic cells in every column and (2) fireback during which voltage patterns of pre-neurons were inverted and only the post-neuron that sums the largest amount of current during (1) fired a large but shorter pulse back into the same column. The

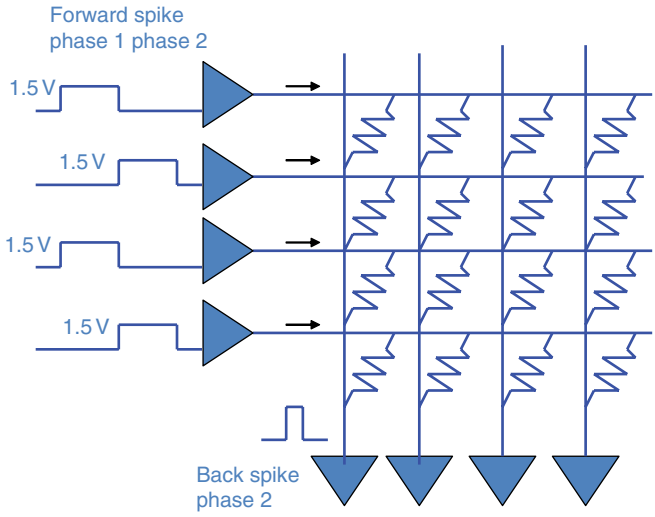


Figure 5.7 Schematic circuit of simulated neuromorphic system with applied signal waveforms. Based on Chou *et al.* (NCTU), VLSI-TSA, April 2015 [11].

combinational effects are designed to potentiate or depress synaptic weights of the cells on the fired column toward lower or higher resistance according to the 1/0 of the training patterns.

The potentiation plus depression (P + D) method improved the contrast development of pattern learning and immunity to input noise compared with the P-only method. The recent fast development of RRAMs has inspired several attempts to use them as the high density electronic synapses in artificial neuromorphic systems. This study reported on a homogeneous barrier modulation (HBM) RRAM that was capable of a simultaneous P and D operational scheme. The RRAM stack was SiO_2/Ti/TiO_2(30 nm)/TaO_2(10 nm)/Ta/BE.

5.2.6 Other Nonvolatile Memories with Analog Properties

An example of potentiation and depression in a memory using a measurement of capacitance rather than resistance was illustrated in February of 2015 by Seoul National University and Samsung [12]. A silicon-based charge trap memory with an Al/HfO_2/Al_2O_3/Si_2N_4/Si stack was used, which emulated that memory function of a biological synapse. The quantity of the trapped charge in the proposed Al_2O_3/HfO_2/Si_3N_4 stack was estimated by measuring the capacitance over time with different repeated input pulses, which are similar to the behavior of a biological synapse and were regarded as representing synaptic weight changes. By applying repeated voltage pulses at periodic intervals at different times, reliable short term plasticity and long term potentiation properties were found along with their transition behavior. This architecture was found to be compatible with a CMOS process and had promise of being possible to implement in an electrical neuromorphic system.

The fabricated device consisted of an Al gate electrode and an Al_2O_3/HfO_2/Si_3N_4 gate stack formed on a p-type silicon (100) wafer. Initially an n^+ region was formed with

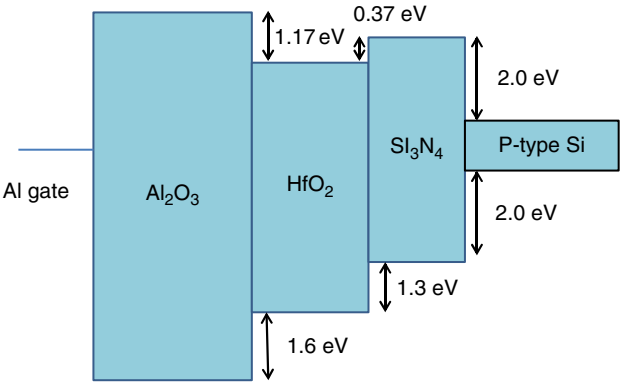


Figure 5.8 Energy band diagram of $Si/Si_3N_4/HfO_2/Al_2O_3/Al$ cell. Based on M.-S. Lee *et al.*, *IEEE Trans. on Electron Devices*, February 2015 [12].

As^+ ions and annealed to supply electrons to the gate stack during programming. Then HfO_2 layers and Al_2O_3 layers were formed by ALD at 350 °C. An Al layer was then deposited and patterned to form the top electrode. The area of the Al electrode is about $100 \times 100\,\mu m^2$. Programming was done at 7 V for 1 ms and erased at −7 V for 1 ms. An energy band diagram of an $Si/Si_3N_4/HfO_2/Al_2O_3/Al$ cell is shown in the diagram in Figure 5.8. CV measurements were used in an effort to imitate synaptic plasticity. The competing effect of memory loss and memory strengthening was observed in repeated simulations.

An amorphous silicon-based metal–semiconductor–metal (MSM) device can be used as a resistive switching device or a neuron in a neuromorphic computing system. In 2015 Carnegie Mellon University showed that CMOS compatible Pt/a-Si/Cu stacks show negative differential resistance (NDR) that enables their operation as oscillatory elements and nonlinear threshold switches [13]. To demonstrate oscillations in amorphous silicon (a-Si), this study showed frequency tunability from 5 kHz to 80 MHz by modulating the series ballast and changing the source voltage. The devices showed <1.2 V and <100 μW operation. As threshold switches, these devices showed an ON/OFF resistance >800 with a peak current density of > 0.3 MA cm^2 at 1 V. The change in modes of operation of the device was studied as a threshold switch and an oscillator.

5.3 3D Neuromorphic Memories

5.3.1 Neuromorphic Architectures as Dense TSV 3D Structures

Neuromorphic architectures are 3D structures with massive parallelism. They exhibit a high level of connectivity. Deep computation and cognitive functions reach thousands of neurons and millions of synapses. It is assumed, therefore, that the neural network must be multilayer and densely connected using point-to-point connectivity. 3D architectures are therefore expected to reduce routing congestion, reduce critical path length, and save area and power.

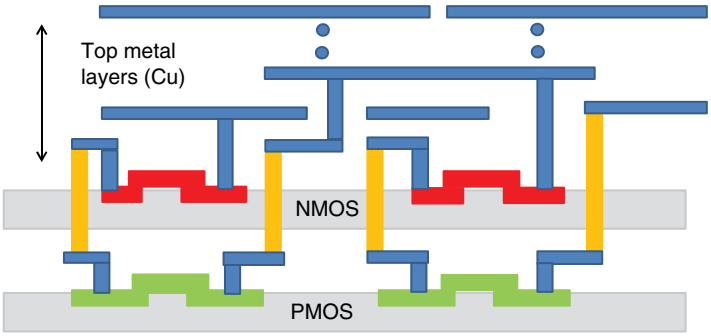


Figure 5.9 Example of 3D transistor level monolithic integration. Based on F. Clermidy *et al.* (CEA-LETI, CEA-LIST, INRIA), ASP-DAC, January 2014 [7].

The network studied by CEA and associated labs in 2014 was a neural processor with two layers of densely connected networks using through silicon via (TSV) intended to recognize objects appearing in a 1000 frame/s video stream [7]. The target operation was executing complex tasks. Two separate circuits were designed for face-to-face connections. Short TSVs could be used to bond the last metal layers of both circuits and longer TSVs could be used that cross all metal layers. In 3D monolithic integration, transistor layers are made one on top of the other on the same die. An example of transistor level monolithic 3D integration is shown in Figure 5.9.

Another dense 3D synaptic architecture using a self-rectifying bipolar Ta/TaO_x/TiO_2/Ti RRAM stack was described in December of 2014 by NCTU [14]. This RRAM device showed good analog synaptic features. Ultralow energy consumption was shown, which was comparable to that of a biological synapse at <10 fJ/spike. Several synaptic features were demonstrated using a Ta/TaO_x/TiO_2/Ti RRAM synaptic element. The analog synaptic characteristics were modeled using physical and compact models. SET controlled and RESET controlled multilevel cell (MLC) operations for the 2D Ta/TaO_x/TiO_2/Ti RRAM device were shown. By varying V_{set} or V_{reset}, multiple resistance states could be read out at −2 V. A very low current was found at a positive voltage bias.

5.3.2 3D Vertical RRAMs as Synapses Connecting Neurons

A number of vertical 3D RRAM arrays have been described for use in neuromorphic applications. The RRAM element that exhibits analog conductance behavior acts as a synapse connecting parallel input and output neurons. In July of 2014, the University of Michigan discussed a 3D vertical two-layer RRAM with well characterized analog switching characteristics for use in neuromorphic computing applications [15]. The dual-layer RRAMs had a WO_x switching layer, which was formed at the sidewall of horizontal electrodes. The devices showed well characterized analog switching characteristics and a very small mismatch in electrical characteristics between devices formed in the two layers, as shown in Figure 5.10, which is a schematic of current output versus time for the upper and lower device.

The 3D vertical device structure allowed higher storage density and bigger connectivity for use in neuromorphic computing. The vertical devices were shown to have potentiation and depression characteristics similar to planar RRAMs. The two layers could be programmed independently with no crosstalk seen between the layers.

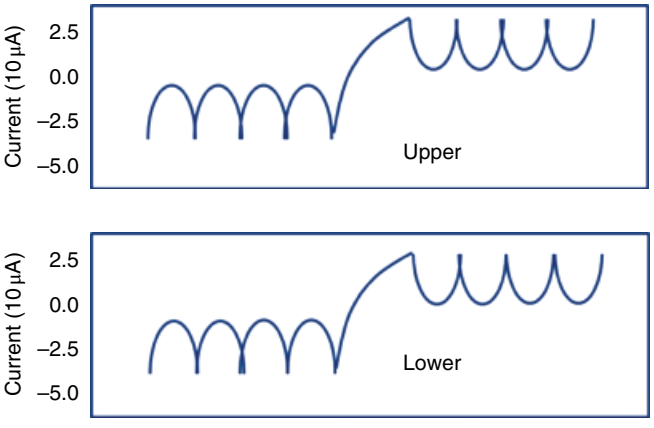


Figure 5.10 Current output versus time for upper and lower devices illustrating good matching between the two. Based on S. Gaba *et al.* (University of Michigan), *IEEE Trans. on Electron Devices*, July 2014 [15].

The devices could be programmed with <10% mismatch. The architecture was considered scalable and suitable for development of analog memory and neuromorphic systems. During write and erase a V/2 method could be used to avoid disturbing unselected bit-lines and word-lines. During READ, grounding methods could be used to avoid sneak path currents. The *I–V* characteristic was nonlinear, which was expected to improve the read margin as the array size scaled. The conductance change ratio and switching speed may be improved further by optimizing the various process steps.

In December of 2015 various French labs investigated the vertical pillar RRAM (VRRAM) concept for use in neuromorphic applications [16]. In this case, one synapse was emulated by one VRRAM pillar. The basic functionality of the HfO_2-based MO-VRRAM was given as 20 ns switching time, up to 10^7 cycles with stable 200 °C data retention. Specific analyses were made of the resistance and switching voltage variability to conveniently address the various potentiation applications. A correlation effect was found to exist between adjacent cycles. This was taken to mean that the conductive filament keeps a memory of its shape in the previous state, which leads to reduced cycle-to-cycle variability in a sufficiently short cycle range. A compact model and circuit simulations were used to propose the VRRAM for convolutional neural network applications. Reliability was good and of lesser circuit complexity than planar approaches.

The VRRAM device used a TiN/SiO_2 double layer deposited on a vertical W plug. The TiN thickness was 10 nm. The cylinder was formed using a plasma etch to pattern it. An nMOS transistor was used as the selector for the memory. An analog-like conductance behavior was due to the variability of an oxide-based RRAM. A model was used to simulate a synapse composed of stacked VRRAMs with from 2 to 24 levels. The VRRAMs were addressed simultaneously and the output signal corresponded to the sum of all the VRRAMs in parallel. A simulated neuromorphic array is shown in Figure 5.11. A synapse is composed of "N" VRRAM cells in a vertical pillar, which are addressed in parallel. The output neuron collects the contributions of all the synapses.

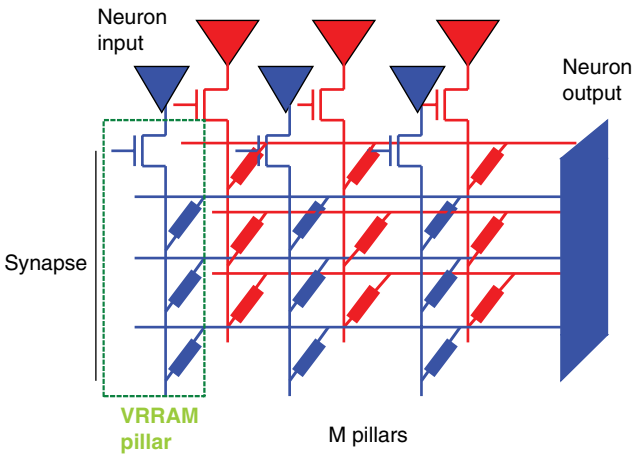


Figure 5.11 Simulated neuromorphic network showing synapse of "*N*" VRRAM cells. Based on G. Piccolboni *et al.* (CEA, LETI, Minatec, IM2NP, IMEP-LAHC), IEDM, December 2015 [16].

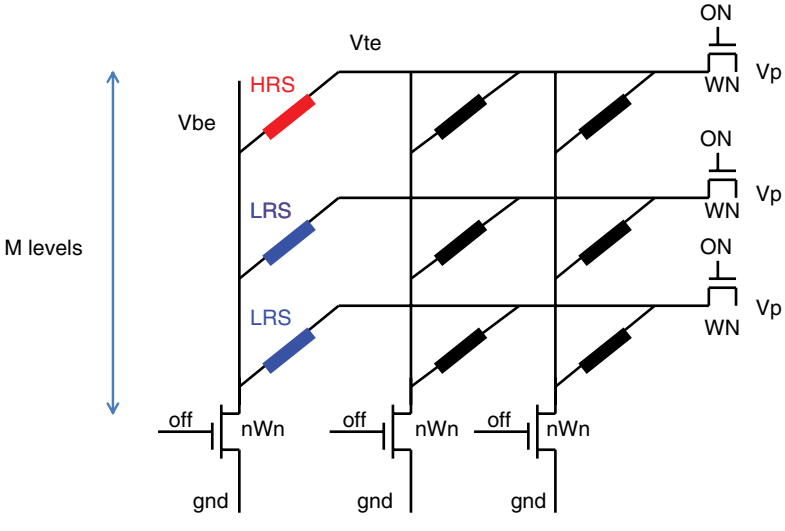


Figure 5.12 Neural circuit with V-CBRAM pillars having *m* levels sharing a common transistor. Based on G. Piccolboni *et al.* (CEA, LETI, Minatec), IMW, May 2016 [17].

The vertical pillar RRAM used for neuromorphic applications was shown again in 2016 by CEA and various associated labs [17]. An Al_2O_3/Cu CB-RAM was used in a vertical configuration for neuromorphic applications. Each cell constituted a vertical synapse. A selector and array configuration was tuned to reduce the variability of voltage seen by each cell. A circuit was designed to implement a neuromorphic system. The neural circuit scheme is shown in Figure 5.12, where vertical-CBRAM pillars have *m* levels. All levels share a common transistor. Each pillar has a transistor with an $n \times$ wider channel compared to plane select transistors.

The synapses are emulated by 1TnR VRRAM pillars, which share a transistor per level. One transistor per pillar is used to select the pillar itself and no back-end selector is required. In a worst case, a given CBRAM on a selected pillar has to be SET while all the other CBRAM on the selected pillar are LRS. All other pillars are left floating to cancel any sneak path current on other CBRAMs so their influences can be neglected. The main issue is the voltage difference $V_{te}-V_{be}$ seen by the V-CBRAM in the LRS state serially connected to the line access transistor on an m-levels selected pillar. The SET condition can be studied as a function of the number m of VRRAM on a pillar and the oversize factor n of the access transistor.

The actual implementation in silicon of such a vertical pillar structure was shown in 2016 by Stanford University, NNDL, and Narlabs, who discussed a four-layer 3D vertical RRAM integrated with a FinFET [18]. A schematic cross-section of the 3D RRAM is shown Figure 5.13.

The 3D architecture was intended for dense neuron-in-memory computing and implemented 10^{11} address and read cycles on 3D RRAM. The RRAM stack was used as a 3D neuromorphic vision system based on a $32 \times 32 \times 4$ 3D array. The input and output neurons were connected by the 3D RRAMs acting as 3D variable resistance synapses. A winner-take-all network with a stochastic learning rule was simulated for orientation detection. Four-layer synapses were self-organized into distinct resistance maps. The balanced array configuration and unique common pillar structure made this vertical 3D RRAM an energy efficient versatile computing unit.

A manufacturable two-layer 3D implementation of a metal oxide RRAM crossbar array circuit was described in 2016 by the University of California, Santa Barbara, which could be used for both high density nonvolatile memory applications and also for neuromorphic implementations [19]. Analog timing of the memory state combined with high integration density of resistive RAM was required for compact implementation of a synapse. These are the most frequently encountered devices in artificial neural networks and would be essential for low energy implementations of large scale

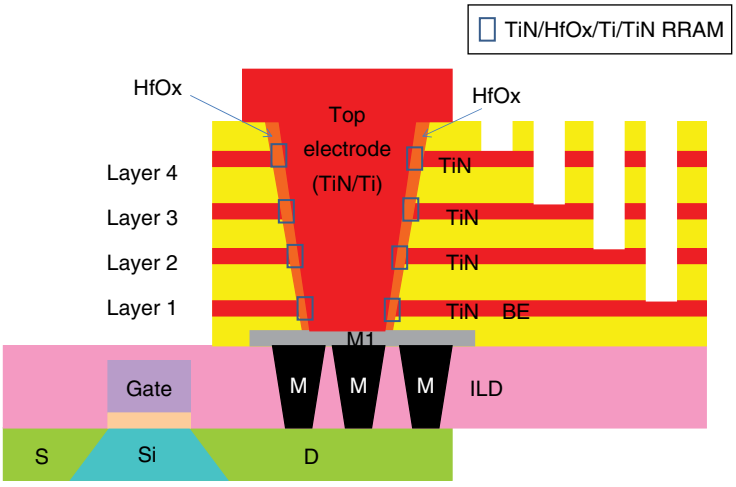


Figure 5.13 Schematic cross-section of 3D vertical RRAM integrated with FinFET. Based on H. Li *et al.* (Stanford University, NNDL Narlabs), VLSI, June 2016 [18].

neuromorphic circuits. One method of increasing the effective RRAM density is by vertical monolithic integration of RRAM crossbar circuits. A bilayer stacked metal oxide RRAM crossbar circuit was fabricated and characterized.

These stacked crossbar arrays can be made using 1T1R configurations, but these three terminal devices have low scalability. The alternative is to use RRAMs without a selector device, which have highly nonlinear I–V characteristics to avoid sneak paths and half select issues. Experimental results showed good uniformity for the RRAMs in both crossbar layers. The process flow used was low temperature and CMOS compatible and could be extended to multilayer stacking.

Another issue is thermal crosstalk, which determines the high maximum temperature of switching. This was estimated to be >700 K for MO-RRAMs. This means the switching in the bottom device could cause disturbances in the devices stacked above. More needs to be understood about device behavior through the entire crossbar system to know if successful analog computing applications are feasible.

The 3D configuration used in this device monolithically integrated two TiO_x-based passive crossbar arrays. The fabrication method required no high temperature processing, showed no thermal crosstalk, and could be repeated with additional layers. In this process, two fully passive TiO 10 × 10 RRAM crossbars were monolithically integrated using an in-situ low temperature reactive sputtering deposition followed by a planarization step. The middle electrodes were shared between the bottom and top layers, as shown in the cross-section of the process in Figure 5.14. The metal layers were deposited as a blanket using reactive sputtering along with highly selective hard masks such as Al_2O_3. The hard mask was patterned and the metal etched using an Ar ion beam. The metal feature shape was controlled using tilted etching. Chemical mechanical polishing and etch-back was used. The maximum temperature during fabrication was 175 °C during sacrificial SiO_2 deposition, which permits this crossbar array to be made in the back-end-of-the line (BEOL) over a conventional CMOS process.

Electrical characterization showed good uniformity and analog programmability for the 3D crossbar circuits. The devices were formed using a floating biasing configuration and quasi-DC current ramps without external or internal compliance. The devices were formed one by one and then immediately switched to the low conductive state. The bottom devices were done first followed by the top layer devices. Bottom devices have an average forming voltage of 2.5 V and top devices 2 V. No thermal crosstalk was seen during forming. Good similarity in device behavior between bottom and top devices was observed.

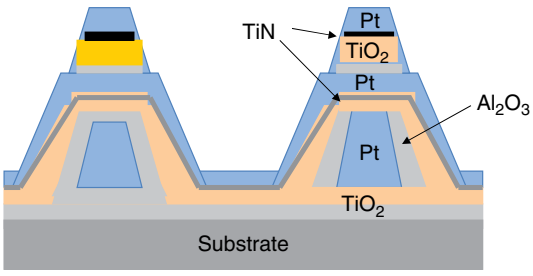


Figure 5.14 Monolithic integration of RRAM crossbars with shared middle electrodes. Based on G.C. Adam *et al.* (UC Santa Barbara), ESSDERC, September 2016 [19].

An ON/OFF current ratio of 10 was observed with a RESET voltage of –1.8 V and SET current of 300 μA. At a RESET voltage of –2.4 V the ON/OFF ratio was about 100. A 74% planarization yield was achieved across the wafer. This work was expected to be a key step toward multilayer RRAM crossbars integrated on CMOS for neuromorphic applications.

5.4 Modeling and Characterization of RRAMs as Synaptic Devices

Since different switchable nonvolatile memory devices can be used as variable synaptic devices in neuromorphic computing, it would be useful to have a test to identify such devices. In 2015, the University of Southampton and the Imperial College, London, described an algorithm intended to identify switchable RRAM devices. These are defined as devices that can be programmed in distinct resistive states and change their state predictably and repeatedly in response to input stimuli [20]. Since RRAM devices can be highly nonlinear, they are useful for applications such as neuromorphic computing. Devices were identified that switch between highly distinct resistive states such as a low resistive state and a high resistive state when subjected to a given biasing protocol. No assumption was made about definition of levels of resistance as "1"s and "0"s. This method extrapolates the statistical significance of difference between two distinct resistive states. These states are measured from devices subjected to standardized bias protocols. A typical testing procedure is illustrated in Figure 5.15.

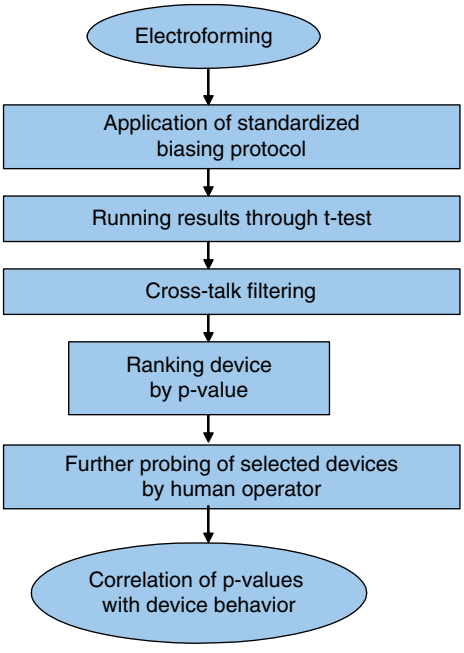


Figure 5.15 A typical testing procedure for RRAM devices that switch between highly distinct resistive states. Based on I. Gupta *et al.* (University of Southampton, Imperial College, London), *IEEE Trans. on Circulation and Systems*, July 2015 [20].

Modeling, characterization, and optimization of an oxide-based analog RRAM synapse was discussed in 2016 by Tsinghua and Peking Universities and Southern University of Science and Technology [21]. Such RRAM-based neural networks have the advantages of low energy consumption, fast computing speed, and low fabrication cost in addition to nonvolatile storage characteristics. Multilevel switching behavior, however, is not easy to control in filamentary RRAM, particularly in the SET process where devices tend to show abrupt switching, which blocks the implementation of on-line training. Nonfilamentary RRAM devices are possible, but analog synapses tend to have reliability issues. A simulated electric potential is shown in Figure 5.16 (a) for a filamentary device and (b) for a nonfilamentary type device. A voltage drop can be observed in the filament gap region.

This experiment develops a physical model based on oxygen vacancy migration to describe the analog switching behavior of nonfilamentary RRAM devices. Simulation analysis results in an optimization method to achieve on-line training in an oxide-based

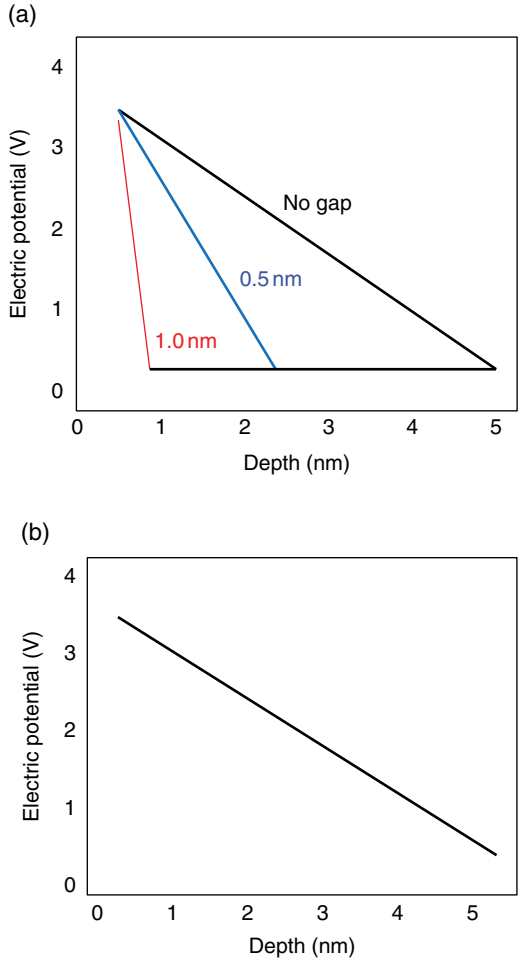


Figure 5.16 Simulated potential for different types of RRAM devices: (a) filamentary showing electric potential drop with increasing gap and (b) nonfilamentary RRAM device. Based on B. Gao *et al.* (Tsinghua University, Peking University, Southern University of Science and Technology), IEDM, December 2016 [21].

analog synapse where a thin oxide switching layer is formed between the electrode or oxygen vacancy (Vo) rich layer and another oxide or semiconductor layer.

During SET, oxygen ions migrate from the switching layer to the Vo reservoir layer, while in RESET, the oxygen ions migrate toward the switching layer. The interface barrier Ei between the reservoir layer and switching layer is modeled from the reaction of oxygen with the Vo layer. A model was based on solid-state defect theory to describe the migration of Vo. The current was calculated using a Poole-Frenkel emission model. The resulting physical model describes the analog switching behaviors in the oxide-based synapse. Analog switching and retention loss were correlated with key material parameters. Optimization guidelines were formed for implementing a high performance neural network system with on-line training ability.

A new physical model was developed to quantify the analog weight modulation behaviors in the oxide-based analog synapse. The analog SET, RESET, and retention loss processes were simulated and verified experimentally using an HfO_x-based synapse.

In the oxide-based analog synapse, a thin oxide switching layer is formed between the electrode or vacancy (Vo) reservoir layer, and another oxide or semiconductor layer. The device resistance is determined by the concentration of Vo in the switching layer. During SET, oxygen ions migrate from the switching layer to the Vo reservoir layer, while, in the RESET process, the oxygen ions migrated from the Vo reservoir to the switching layer.

It is difficult for the ion drift diffusion model to describe the Vo migration during analog switching due to the nonlinear effect and the large concentration of Vo. A new model was developed based on the basic solid-state defect theory to quantify the migration of Vo. The current was calculated using a modified Poole-Frenkel emission model.

An HfO_x/SiO_y-based analog synapse was made. An HfSiO switching layer was shown to form at the HfO_2–Si interface with a 4 nm thickness. The resulting I–V curve showed good analog switching in both SET and RESET processes. The resistance was between 10^7 and 10^8 ohms, which is higher than filamentary RRAMs. A schematic flowchart for the simulation of analog switching is shown in Figure 5.17.

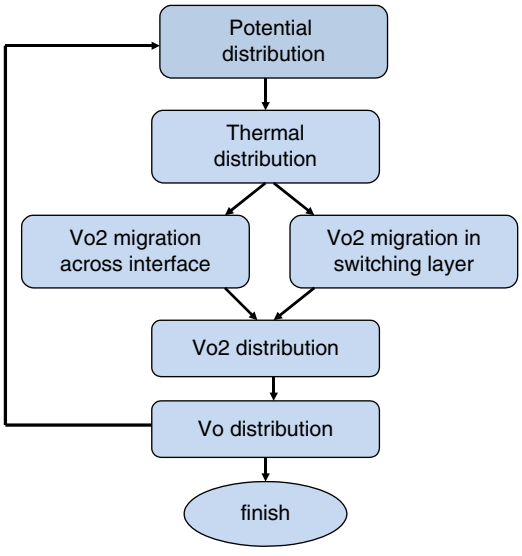


Figure 5.17 Schematic flowchart for simulation of analog switching in an oxide-based nonfilamentary RRAM. Based on B. Gao *et al.* (Tsinghua University, Peking University, Southern University of Science and Technology), IEDM, December 2016 [21].

Analog switching and retention loss were correlated with key material parameters and captured and verified with experiments. Optimization guidelines were provided for boosting the performance and on-line training ability of the neural network.

5.5 Spiking Neural Nets, STDP, Potentiation, and Depression

5.5.1 Introduction to Spiking Neural Networks

Neuromorphic computing can be achieved using variable resistance memories as synapses in spiking neural networks. Each synapse can have its own conductance profile and these profiles can be altered during evolution of the system. A synaptic array can be configured using self-rectifying multivalued RRAMs. It can also be configured as a hybrid CMOS/RRAM array using diodes or transistors to eliminate sneak current with multivalued RRAM.

5.5.2 Hybrid RRAM/CMOS STDP Neuromorphic Systems

Resistive memory devices can be used as synapses in neuromorphic cognitive systems. Hybrid MO-RRAM/CMOS synapses can be based on resistive RAM and transistor circuit design and operation. Real time spike-based learning and recognition of visual patterns was demonstrated in 2016 by the Politecnico di Milano and IU.NET [22]. The learning accuracy and power consumption of the synapse blocks was discussed. Resistive memories have high speed and small cell size due to the two-terminal structure and they have high endurance. Hybrid RRAM/CMOS synapses for neuromorphic circuits were shown capable of on-line pattern learning and recognition. Synaptic schemes can be based on 2T1R and 1T1R structures. The spike-timing dependent plasticity (STDP) characteristics of the 1T1R synapse with bipolar RRAM was shown along with unsupervised learning with a fully connected neuromorphic network using simulations. Learning efficiency and power consumption of the synaptic blocks was shown.

Hybrid RRAM/CMOS synapses were used in the circuit design, operation, and demonstration of real time spike-based learning and recognition of visual patterns. The learning accuracy and power consumption of the synapse blocks was discussed. A main requirement of synaptic circuits is being able to tune their resistance in response to incoming neuron spikes according to the STDP protocol, which has been observed in biological systems. Circuit diagrams of a 2T1R synapse and a 1T1R synapse are shown in Figure 5.18.

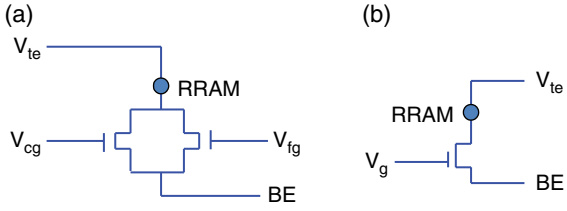


Figure 5.18 Circuit schematics of (a) a 2T1R synapse and (b) a 1T1R synapse. Based on D. Ielmini *et al.* (Politecnico di Milano, IU NET), ISCAS, May 2016 [22].

In both of these circuits, the connections with the PRE and POST permit overlaps between spikes, which enable STDP. The resistive element (circle) can be any two-terminal memory including a unipolar or bipolar RRAM or phase change memory (PCM).

In the 1T1R design, the BE terminal is connected to the POST to sustain both communication and STDP functionalities. A PRE synapse–POST circuit for a 1T1R synapse is shown in Figure 5.19. The PRE spike applied to V_g together with a constant bias sustained by the POST at V_{te} excites a transient current in the 1T1R synapse. Figure 5.19 shows this as the negative constant V_{te}, which induces a negative spiking current corresponding to the PRE spike applied to V_g. The spiking current enters the POST using the BE and is integrated, causing a staircase-type increase in the internal potential V_{int} in the POST. As V_{int} exceeds a threshold, the POST delivers a forward spike to the next neuron layer and a backward spike to the V_{te}. The backward spike acts as the STDP function and may have various shapes depending on the various specific RRAM used.

Waveforms of V_g and V_{te} for delta $t > 0$ and delta $t < 0$ are shown in Figure 5.20 (a) and (b) respectively. The PRE circuit drives V_g, which induces a current spike in the 1T1R, which is integrated by the POST circuit. The membrane potential V_{int} increases by integration, eventually causing a fire event as it exceeds a threshold. At the fire event, a backward POST spike is applied to V_{te} to induce potentiation (depression) in case of overlap with the PRE spike with $\Delta t > 0$ ($\Delta t < 0$).

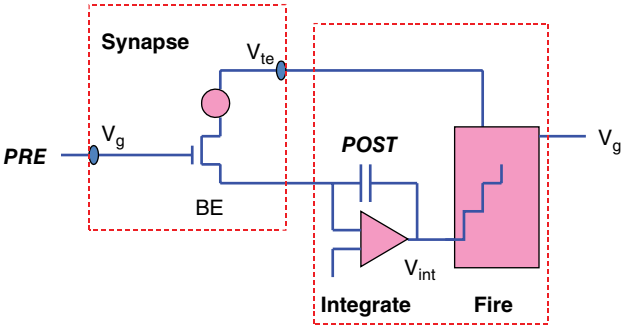


Figure 5.19 Circuit of 1T1R synapse connecting PRE and POST in a spike-timing dependent plasticity (STDP) integrate and fire network. Based on D. Ielmini *et al.* (Politecnico di Milano, IU NET), ISCAS, May 2016 [22] (permission of IEEE).

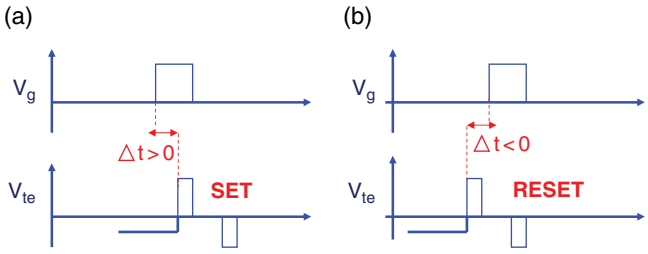


Figure 5.20 Waveforms of V_g and V_{te} for (a) potentiation ($\Delta t > 0$) and (b) depression ($\Delta t < 0$). Based on D. Ielmini *et al.* (Politecnico di Milano, IU NET), ISCAS, May 2016 [22] (permission of IEEE).

The results of the analysis supported on-line pattern learning and recognition by STDP in RRAM-based synapses. A critical parameter for on-line learning and recognition is power consumption. It was found that, by reducing the compliance current with filament size, downscaling was possible and that STDP energy could be decreased by reducing the pulse width for potentiation and depression. The compliance current is set to control the resistance value of the low resistance state (LRS) during the SET operation. It was estimated that the energy per synapse during a fire event could be as small as 1 pJ. The communication energy also needed to be reduced, which involved the spike duration. It was found energy could be reduced below the fJ range.

A 1T1R (RRAM) synapse capable of spike-timing dependent plasticity was discussed in 2016 by the Politecnico di Milano, IU.NET, and Micron. It had an area improvement over the previous 2T1R structure [23,24]. This synapse circuit used a one transistor one RRAM (1T1R) structure as a tunable connection between a presynaptic neuron (PRE) and a postsynaptic neuron (POST). The RRAM synapse passively transmitted spikes and also updated its weight according to a spike-timing-dependent plasticity (STDP) protocol. The STDP characteristics were characterized and modeled. A two-layer neuromorphic network was shown with pattern learning and recognition with deterministic and stochastic switching.

The RRAM device was a silicon-doped HfO_2 layer with a TiN bottom electrode and Ti top electrode. 1T1R devices were used for pulsed demonstration of driving the top electrode and gate nodes using a waveform generator. The RRAM device had a typical bipolar IV curve. It was demonstrated as a 1T1R synapse connected to PRE and POST as Figure 5.21 (a). Typical spike signals at V_g and V_{te} are shown in Figure 5.21 (b).

A V_g spike from PRE induces a current, which is integrated by POST and eventually leads the synapse to fire. At fire, V_{te} induces potentiation for $\Delta t > 0$ or depression for $\Delta t < 0$, resulting in STDP.

The STDP characteristics, in this case the change of conductance as a function of Δt, are illustrated in Figure 5.22 and were found from both experiments and calculations. The range of conductances are shown from 25 to 500 kohm. The change of conductance was found for increasing initial resistance R_0.

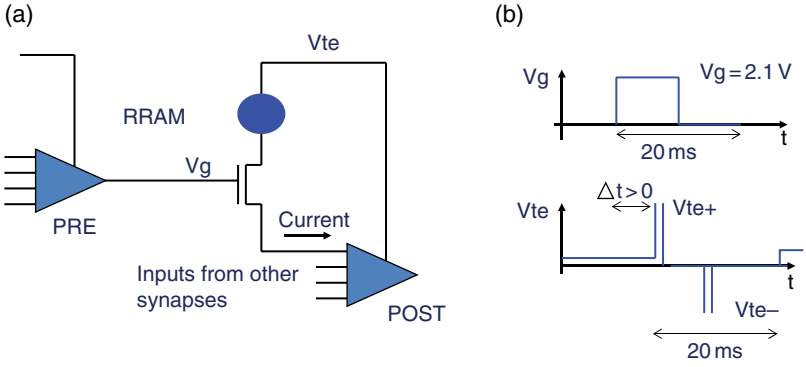


Figure 5.21 1T1R synapse: (a) circuit connected to PRE and POST and (b) typical spike signals. Based on S. Ambrogio *et al.* (Politecnico di Milano, Micron Tech.), *IEEE Trans. on Electron Devices*, April 2016, VLSI, June 2016 [23, 24].

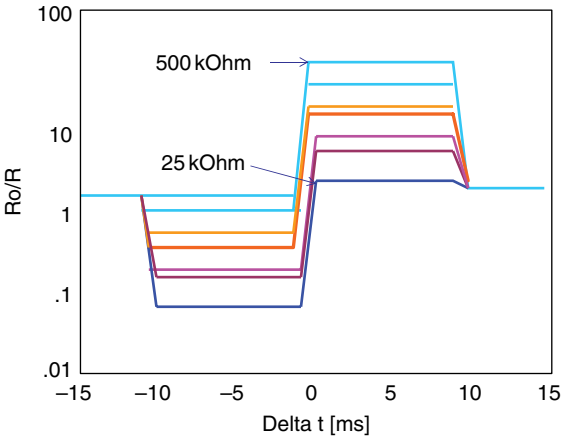


Figure 5.22 STDP characteristics: change of conductance versus Δt. Based on S. Ambrogio *et al.* (Politecnico di Milano, Micron Tech.), *IEEE Trans. on Electron Devices*, April 2016, VLSI, June 2016 [23, 24].

A two-layer neuromorphic network was demonstrated being used with a visual input pattern disturbed by a typical random noise pattern with a typical output pattern and noise sequence appearing at POST. Learning was shown after a given number of cycles.

The 1T1R RRAM synapse capable of spike-timing-dependent plasticity was shown to be an area improvement over the earlier 2T1R structure. STDP functionality was shown along with robust learning by deterministic and stochastic synapses. A 10^3 reduction in power consumption was shown as a result of using a burst operation. The RRAM was based on a silicon-doped HfO_2 layer with a TiN bottom electrode and Ti top electrode.

Learning and classification of handwritten digits from the MNIST database was simulated. A fully connected two-layer perceptron network for simultaneous multipattern learning of 10 digits was used with a second layer consisting of 10 neurons interconnected by an inhibitory synapse permitting differentiating learning of multiple digits. The first layer had 784 neurons and the second had 10. The MNIST handwriting digit classification showed 86% efficiency.

An RRAM-based neuromorphic computing device with a 1T1R crosspoint array was described in 2015 by Tsinghua University [25]. The 1T1R array is shown in Figure 5.23. The 1T1R crossbar array minimized the sneak paths of the 1R and 1R1S array and was shown to have lower power dissipation than a 1R RRAM neuromorphic computing device.

Both the two-terminal 1R and 1S1R crossbar arrays when implemented in a synapse network have good performance but suffer from both sneak path current and program disturb. These limit the computing accuracy of the neural network and also limit the size of the array. In this study, a 1T1R cell is used as the element for an optimized synapse array for neuromorphic computing. A differential pair formed a synapse. Pattern classification was done by simulation using this architecture and the impact of RRAM cell variation on the system was analyzed.

In a column the top electrode of the RRAM is connected to the BL and the source of the transistor is connected to the SL. In a single row, all cells have their transistor gate connected to the same word-line. This organization decreases the sneak current significantly both suppressing disturbance and decreasing power consumption.

In the proposed neural network, a single layer was used based on a 10×6 1T1R synapse array. An RRAM model with asymmetric bipolar switching performance was used. Six columns were divided into three classes in accordance with three output neurons.

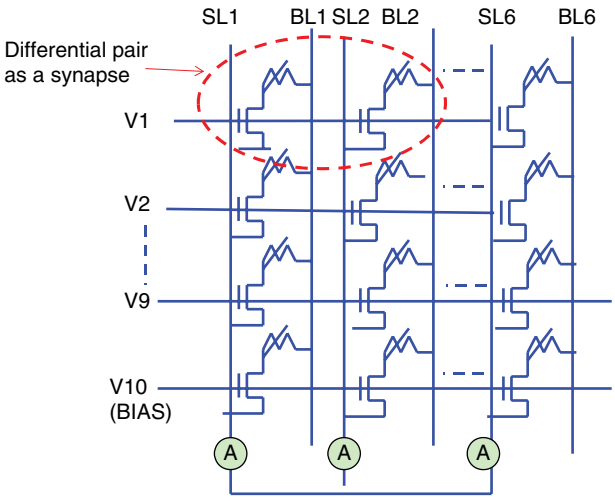


Figure 5.23 Circuit diagram using RRAM 1T1R crossbar array, which eliminates sneak paths. A differential pair is used as a synapse as illustrated. (Based on P. Yao *et al.* (Tsinghua University, Tsinghua National Laboratory), NVMTS, October 2015 [25].

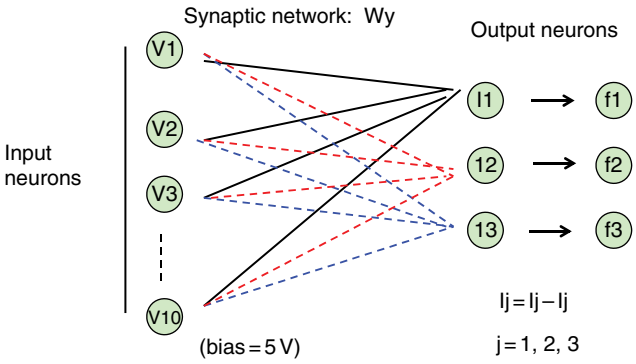


Figure 5.24 Schematic of single-layer bipolar neuron network. Based on P. Yao *et al.* (Tsinghua University, Tsinghua National Laboratory), NVMTS, October 2015 [25] (permission of IEEE).

Every two columns in sequence together connect to a post-neuron. The difference current is the current that flows to the first output neuron. The schematic of the single-layer bipolar neuron network is shown in Figure 5.24.

A total of 30 synaptic weights is realized as two synapse cells representing a synaptic connection. The 10×6 1T1R synaptic array connects with three post-neurons and 10 inputs, which include a constant bias input. Ten training patterns are input as a 3×3 pixel black and white image and the input signal is represented by a gate voltage of 5 or 0 V, corresponding to a black or white pixel. All the 1T1R are HRS initially (dark block). After the pattern is applied, the whole array is at read mode simultaneously. To adjust each cell's weight a parallel programming operation strategy is used column by column where the SET and RESET process are separated. After 15 interations, the final distribution in each column, shown in Figure 5.25, is the input letters "z", "v" and "n".

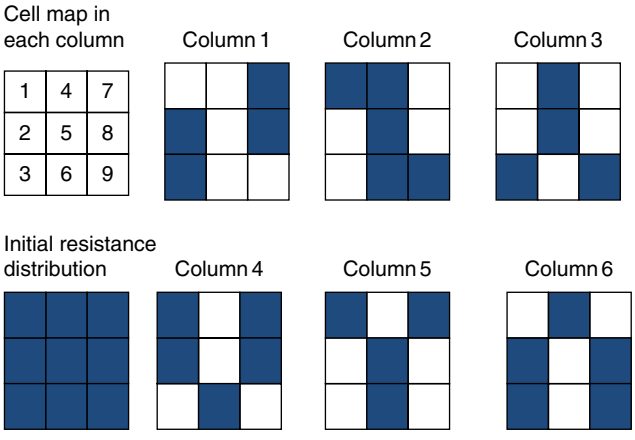


Figure 5.25 The final resistance distribution of each column. Based on P. Yao *et al.* (Tsinghua University, Tsinghua National Laboratory), NVMTS, October 2015 [25].

Compared with the same function done with a 1R-based RRAM crossbar array, the new 1T1R synapse architecture operates more easily and converges more rapidly. The 1T1R structure has low power consumption and the parallel operating mode decreases power dissipation further. The resulting power dissipation was about half that of a 1R RRAM array.

5.5.3 Memory Synapse and Neuron Systems

An oxide-based analog synapse device and an insulator–metal transition (IMT) oscillator neuron device were described for a neuromorphic system in 2016 by POSTECH [26]. The redox reaction was controlled at the metal/$PrCaMnO_3$ (PCMO) interface of the synapse device in order to control synapse characteristics such as switching uniformity, disturbance, retention, and multilevel data storage under identical pulse conditions. For the neuron device, an Mo electrode was shown to have the best data retention characteristics. Using the IMT characteristics of NbO_2 film, an IMT oscillator was developed for the neuron. Accurate pattern recognition was found using the Mo/PCMO synapse array and the NbO_2 oscillator neuron.

The required parameters of the synapse device defined by the ITRS roadmap for semiconductors of 2013 included: energy consumption of <10 fJ, operating speed of > 10 Hz, endurance of 10^9, 32 states (5-bits), and data retention of 10 years. Also required were: analog conductance change, constant rate of change in conductance, symmetric conductance increase and decrease, and a high density 3D fabricated array.

For the Mo/PCMO device, it was found that the conductance is proportional to the device area so scaling could reduce the switching power. The potentiation and depression characteristics of the Mo/PCMO devices under identical pulse with varying voltage amplitude of pulse were shown to exhibit a gradual conductance change. The change of conductance with increasing voltage amplitude is illustrated in Figure 5.26.

The IMT oscillator neuron using NbO_2 had threshold shifting I–V characteristics as shown in the I–V curve in Figure 5.27.

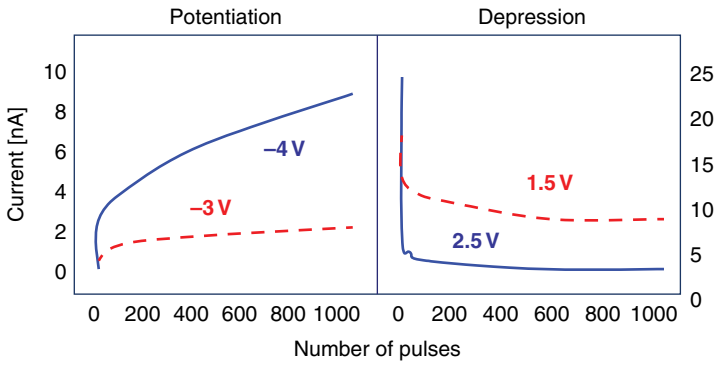


Figure 5.26 Illustration of potentiation and depression characteristics of Mo/PCMO. Based on K. Moon (POSTECH), VLSI-TSA, April 2016 [26].

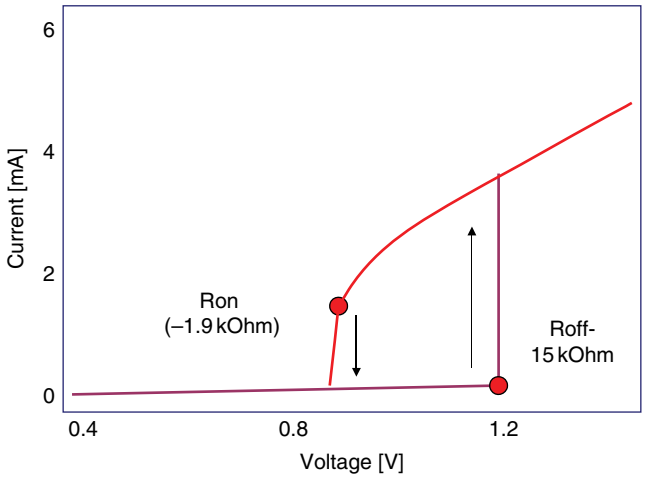


Figure 5.27 NbO_2 IMT oscillator neuron threshold shifting *I–V* characteristics. Based on K. Moon (POSTECH), VLSI-TSA, April 2016 [26].

When the IMT oscillator neuron was observed above the critical threshold voltage of about 1.3 V the oscillator characteristics were as shown in Figure 5.28. The pattern recognition capability of the Mo/PCMO synapse device with an NbO_2 IMT oscillator neuron was tested with a slightly noisy input pattern and the original pattern could be recovered.

Spike-timing dependent plasticity (STDP) in conductive bridge RAM (CBRAM) devices with silicon neurons was discussed in 2016 by Arizona State University [27]. STDP is a neural process that lets biological neural networks learn by strengthening or weakening synaptic connections between neurons. This research showed the potential for using the on-state resistance of a CBRAM according to the biological STDP rule for neuromorphic applications. The CBRAM devices were integrated with CMOS spiking

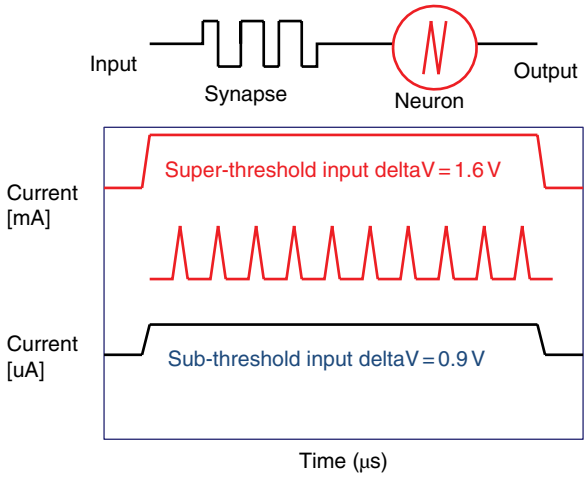


Figure 5.28 Oscillator characteristics of IMT neuron with varying bias conditions. Oscillations above the 1.3 V critical threshold voltage are shown. Based on K. Moon (POSTECH), VLSI-TSA, April 2016 [26].

neuron circuitry using back end of the line (BEOL) post-processing. Different initial resistance values and spike durations were used.

To reduce the large power consumption and low density of traditional CMOS-based implementations, a hybrid neuromorphic architecture was used, which consisted of analog CMOS spike-generating circuits communicating in a massively parallel architecture using nonvolatile resistive memory devices that can mimic biological synapses. Synaptic plasticity is the ability of biological synaptic connections to be strengthened or weakened as a response to spiking activities of neurons. STDP is a neuronal learning mechanism that involves the dependence of change in synaptic connection strength based on the relative timing difference in arrival of voltage spikes generated by the neurons at the synapses.

Before CBRAMs could be used as electronic synapses, it was necessary to show that the conductance of the CBRAMs could be tuned for use with the STDP rule. On-chip STDP behavior was shown in CBRAM devices made in the back end of the line (BEOL) on top of a chip containing a CMOS "integrate and fire" neuron circuit. Experimental evidence of true analog STDP behavior in CBRAM was shown that is similar to the biological STDP learning rule. A CBRAM compact model was used to show STDP behavior and how this behavior could be used for applications such as associative memory.

The CBRAM devices were made as a vertical stack of Ag/GeSe on top of metal pads on the chip containing the CMOS neuron circuits. Two CMOS neurons on the chip were connected using a BEOL CBRAM device. Spikes generated at the nodes of the two neurons were applied across the CBRAM device. To demonstrate STDP learning, a relative timing difference was created between the pre- and post-neuron spikes. The test setup to demonstrate STDP with CMOS neurons connected using a single CBRAM synapse fabricated on chip is shown in Figure 5.29.

The chart in Figure 5.30 illustrates the experimentally measured STDP behavior with an on-chip CBRAM device showing the change in the CBRAM conductance versus spike timing difference for different initial LRS values and a neuron spike width of 7 μs. Experimental results from the die showed an STDP-based analog resistance change behavior in CBRAM devices that follows the biologic STDP learning function.

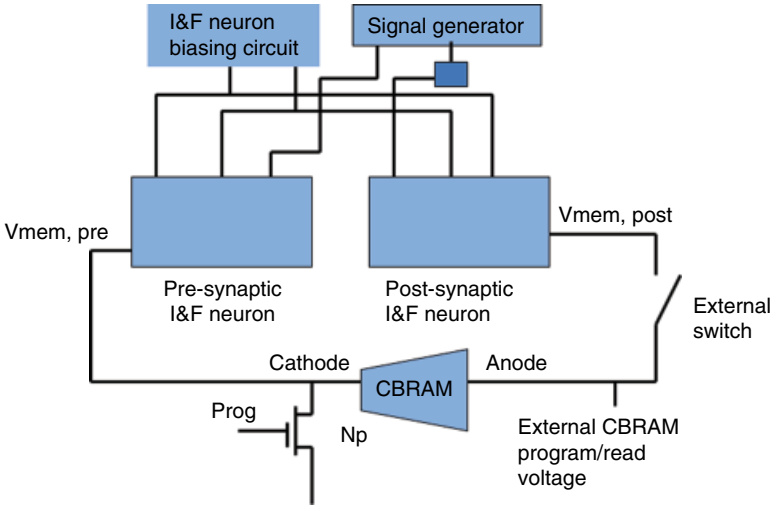
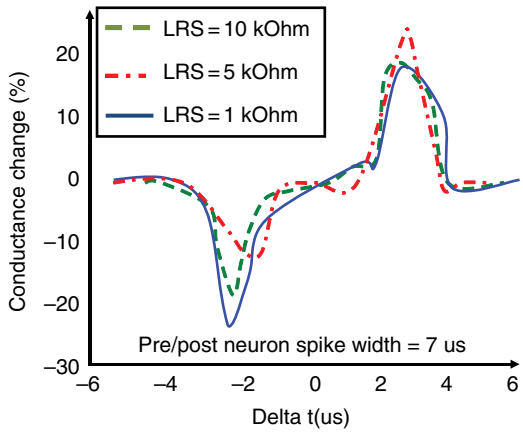


Figure 5.29 STDP demonstration with CMOS neurons connected using a single CBRAM synapse. Based on D. Mahalanabic *et al.* (Arizona State University), ISCAS, May 2016 [27].

Figure 5.30 Change in CBRAM conductance versus spike timing difference with on-chip CBRAM. Based on D. Mahalanabic *et al.* (Arizona State University), ISCAS, May 2016 [27].



5.5.4 Novel RRAM Synapse Applications

The use of RRAMs in neuromorphic systems was discussed in 2014 by CEA-LETI [5]. They showed large-scale energy efficient neuromorphic systems based on RRAM used as stochastic binary synapses. Prototype applications were examined, such as a complex of RRAM memories in neuromorphic circuits for saving energy and increasing performance. They examined visual and auditory pattern extraction using feedforward spiking neural networks. In attempting to imitate the brain, they noted that there are 100 billion neurons in our brain and each of these neurons can have as many as 10 thousand connections to other neurons.

Various neuromorphic systems were studied. An unsupervised neuromorphic system was developed based on GeS_2/Ag CBRAM devices, which were programmed as binary synapses with a stochastic STDP learning rule optimized by a genetic evolution

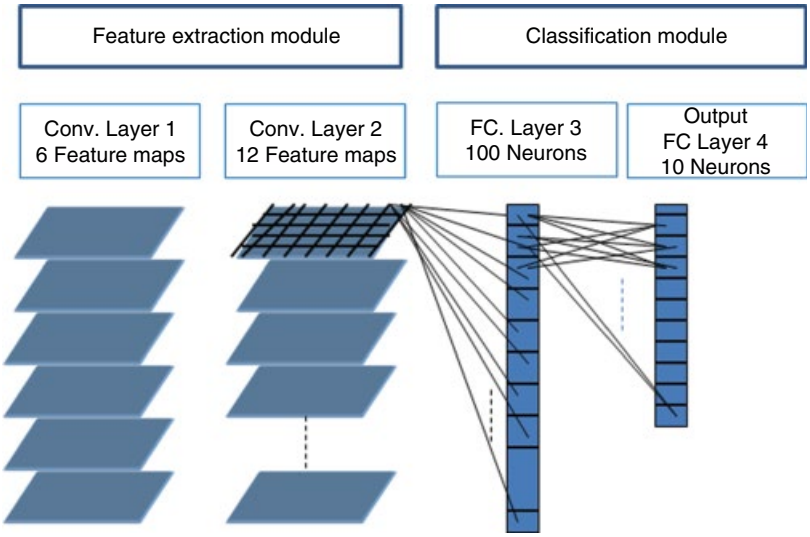


Figure 5.31 Architecture for handwritten digits recognition. Based on B. DeSalvo *et al.* (CEA LETI), IMW, May 2015 [5].

algorithm. Two applications were shown including: read time auditory pattern extraction and visual pattern extractions. Highly accurate audio pattern sensitivity was shown. Figure 5.31 shows the architecture for handwritten digits recognition.

A video detection rate greater than 95% and low synaptic power dissipation in the learning mode were shown. In another implementation, a spike-based hardware convolutional neural network using HfO_2 type OxRAM was developed. Multiple binary OxRAM devices in parallel configurations were used to obtain multilevel synapses. A visual pattern recognition application, the identification of handwritten patterns as digits, was described. The architecture was inspired from the mammalian visual cortex, which is composed of two cascaded convolutional layers to associate the extracted feature maps to the 10 digit categories.

A TiO_x-based RRAM for neuromorphic synapse applications with 64 levels of conductance states due to an optimized interface between the metal electrodes and the TiO_x film was discussed in 2016 by POSTECH [28]. To compensate for the change in switching power with increasing pulse number, the use of fixed voltage pulses in potentiation and of fixed current pulses during depression was tried. By adopting a hybrid pulse scheme, the symmetry of conductance change was shown to be significantly improved under both potentiation and depression conditions.

The improved conductance levels and the symmetry of conductance change were directly related to enhance pattern recognition accuracy, which was confirmed by a neural network simulation. A comparison of the voltage and current pulse modes during depression and evolution of the current pulse programming power with pulse number is shown in Figure 5.32. The power increases gradually.

An Mo/TiO_x-based interface RRAM was developed using a hybrid pulse mode for synaptic applications. An interface switching device with 64 conductance levels was obtained by optimizing the redox reaction at the Mo/TiO_x stack. The hybrid voltage and current pulse method significantly improved the symmetry of the conductance

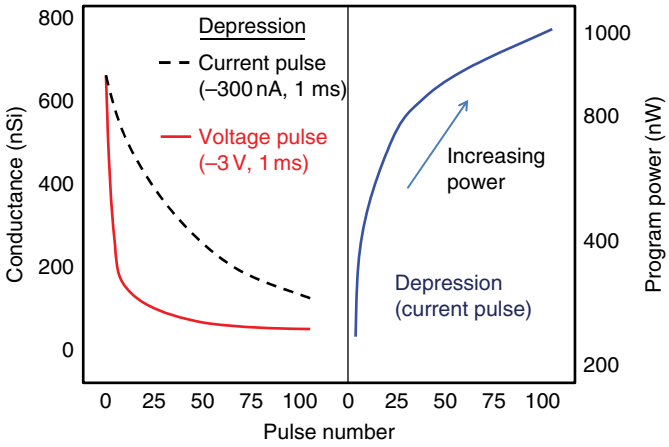


Figure 5.32 Comparison of voltage and current pulse modes for current pulse program power showing conductance in siemens. Based on J. Park *et al.* (Pohang University), *IEEE Electron Device Letters*, December 2016 [28].

change in both potentiation and depression, resulting in an improved pattern recognition accuracy.

Short and long term synaptic plasticity using OxRAM with noisy input data was discussed in 2016 by various French laboratories [29]. A new circuit architecture and a read/program strategy to emulate short and long term plasticity rules using OxRAM cells were discussed. The intrinsic OxRAM device switching probability at ultralow power is used to implement short term plasticity (STP) and long term plasticity (LTP) learning rules. The computational power showed that STP provided for reliable signal detection with noisy input data.

Biological synapses show both short term plasticity (STP) and long term plasticity (LTP). LTP results in stable changes of the synapses according to the timing of pre- and post-synaptic spike events. STP can induce dramatic changes in synaptic weights, which enhances the discrimination ability of single neurons even though LTP characteristics have been the primary research objective of neuromorphic research. The co-implementation of both LTP and STP with the same OxRAM technology is shown.

The emulation of LTP using Ti/HfO_2 OxRAMs cells was demonstrated. In this study, the same technology is used for emulating STP. The programming strategy is to reproduce STP using an OxRAM-based synapse. The method involves having a synaptic weight decreased at each presynaptic spike, which periodically increased with every time interval in the absence of spikes.

The principal circuit that is used to reproduce both STP and LTP rules using nonvolatile OxRAM cells is shown in Figure 5.33. The conductance multiplication during the read operation is done using a buffer, $Y(t)$, which modulates the read voltage for the LTP synapse Sij between the pre-synaptic neuron (i) and the post-synaptic neuron (j).

Long and short term plasticity learning rules are derived for reliable unsupervised learning in spiking neural networks with highly noisy input data. It was shown that the STP permits decreasing the false positives by 40% while maintaining a detection rate of 89% for a neural detection application and a low signal-to-noise ratio of 3.7.

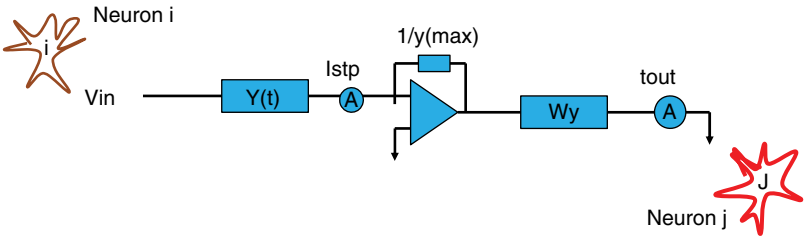


Figure 5.33 Principal circuit used to reproduce STP and LTP rules for an OxRAM cell. The circuit imitates the synapse between neuron (i) and neuron (j). Buffer *Y*(*t*) modifies the synapse for either STP or LTP rules. Based on T. Werner *et al.* (CEA, LETI, MINATEC, CEA, LIST, INSERM), IEDM, December 2016 [29].

5.6 Neural Network Systems Using Ferroelectric RAM Technology

5.6.1 Neural Network Circuits Using Ferroelectric Memory (FeMEM) Synapses

In June of 2013 Panasonic discussed on-chip pattern recognition in a neural network circuit using an NVM [30]. The synapse chip of the neural network consisted of a stack of CMOS circuits and three terminal ferroelectric memristors. The analog and nonvolatile conductance range of the 3 T-FeMEM was used as a synaptic weight to learn the matrix patterns. The original pattern can be recognized from even an incomplete pattern. A schematic cross-section of the CMOS circuit and 3 T FeMEM is shown Figure 5.34.

Neural network circuits can effectively provide low power parallel flexible systems for learning and cognitive recognition. A synapse in which the connection weight is modified by learning controls the electrical signals between neurons. This means that continuous conductance modification is required for a biological synapse. In this case, multiple three-terminal ferroelectric memristors are used as synapses for learning since a gate electrode can control the conductance between the other two electrodes for concurrent learning. The synaptic weight change of a 3 T-FeMEM is shown in Figure 5.35 as a function of the relative timing of the pre- and post-spikes. The spike width of V_{post} was 5 μs. It is similar to the synaptic weight change of a biological synapse.

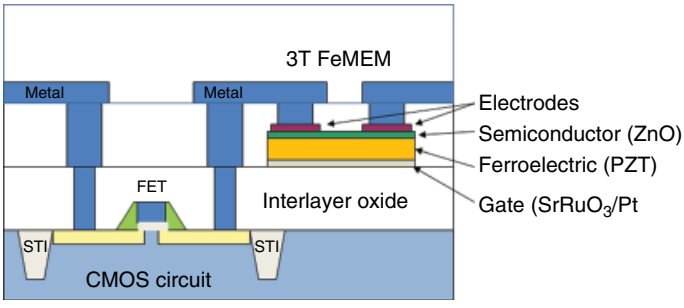


Figure 5.34 Schematic cross-section of CMOS circuit and 3T FeMEM. Based on Y. Kaneko *et al.* (Panasonic), VLSI Technology Symposium, June 2013 [30].

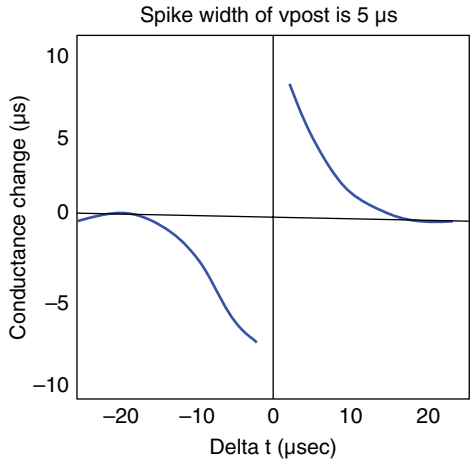


Figure 5.35 Synaptic weight change of 3T-FeMEM versus relative timing of pre/post-spikes with spike width of V_{post} at 5 μs. Based on Y. Kaneko *et al.* (Panasonic), VLSI Technology Symposium, June 2013 [30].

5.6.2 Using the FeMEM in Neural Network Circuits

A neural network circuit was developed that included 9 neurons and 144 synapses using the CMOS circuits and 3 T-FeMEMs. Learning and pattern recall were demonstrated on this neural network chip.

Fabrication of the 3 T-FeMEMs was described in 2014 by Panasonic along with the procedure for integrating the 3 T-FeMEMS into a conventional CMOS circuit [31]. The 3 T-FeMEM circuit had a back-gate thin film transistor structure with a stack of ZnO/ $Pb(Zr,Ti)O_3$ (PZT)/$SrRuO_3$ oxide on a Pt (111) layer. The Pt layer permitted an oriented ferroelectric PZT film. PZT film polarization control was done using the gate voltage, which modulated the channel conductance nonvolatility.

The synapse function, which consisted of continuous control of the analog conductance of the memory, was done by applying pulse gate voltages. In the "integrate and fire" neuron network, the neurons communicate with each other using voltage pulses. An input voltage pulse from another neuron generates a current pulse with amplitude that is proportional to the synapse conductance. These current pulses are then integrated in a leaky integrator. When the integrated voltage reaches a threshold, the pulses are output to other neurons through their synapses and are fed back to the former synapses. All neurons operate simultaneously, resulting in parallel computation.

The two-neuron circuit connected by a synapse composed of the 3 T-FeMEM and a selector is shown in Figure 5.36. A neuron outputs a single rectangle pulse and a triple rectangle pulse to the following synapse. The single rectangle pulse is also fed back to the initial synapse.

Recall was demonstrated by inputting a gray pattern. The recalled pattern approached the learned pattern as the recall proceeded. It was concluded that the synapse chip with 3 T-FeMEM is suitable for use as an artificial electronic synapse.

In August of 2014, Panasonic discussed ferroelectric artificial synapses, which were used in a spiking neural network for recognition of a multishaded image [32]. This pattern recognition is based on a spiking neural network model using 3 T-FeMEMs as synapses. This synapse chip is formed by stacking CMOS circuits and 3 T-FeMEMs. The conductance of the 3 T-FeMEM is gradually changed by varying the amplitude of

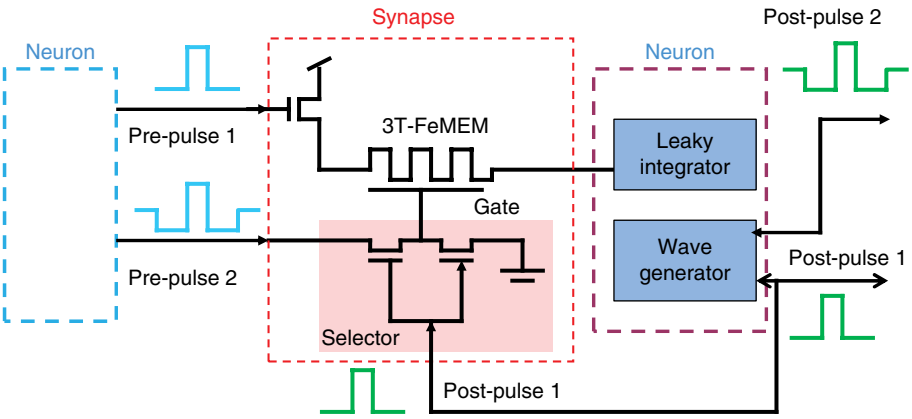


Figure 5.36 Two neuron circuit connected by a synapse containing a 3T-FeMEM and a selector. The first neuron outputs to the synapse a single rectangle pulse to the 3T-FeMEM and a triple rectangle pulse to the selector. The single rectangle pulse is fed back to the synapse as post-pulse 1. Based on Y. Nishitani *et al.* (Panasonic), DRC, June 2014 [31] (permission of IEEE).

the applied voltage pulse. The analog and nonvolatile conductance change of the 3 T-FeMEM is used as a synaptic weight. When an incomplete multishaded pattern is input to the neural network circuit, it automatically completes and recalls a previously learned pattern.

The 3 T-FeRAM is used as a synapse to demonstrate on-chip pattern recognition in a neural network circuit with multiple neurons based on a spiking neural network model. The three-terminal structure offers the advantage of a simple learning scheme. The three-terminal device is capable of realizing signal transmission and synapse learning concurrently in synapse connected neuron pairs.

5.6.3 Ferroelectric Tunnel Junctions in Neuromorphic Circuits

In February of 2014, a ferroelectric tunnel junction (FTJ) was used to simulate a spike-timing dependent plasticity circuit for use in neuromorphic circuits [33]. Memristive behavior has been observed in ferroelectric tunnel junctions. In a conventional FTJ with an ultrathin ferroelectric barrier of <5 nm, tunnel resistance can be changed between ON and OFF states by switching the ferroelectric polarization. It has been shown that the change of resistance can occur continuously rather than abruptly, which defines a ferroelectric tunnel memristor (FTM). An FTM shows large OFF/ON resistance ratios and high operation speed, which is promising for use in multilevel memories and synapsis-like circuits.

The ferroelectric tunnel memristor has resistance that can be continuously tuned by controlling the growth of domain in the ferroelectric tunnel barrier. It has an OFF/ON resistance ratio >100, high operating speed of 10 ns, multilevel storage, and neuromorphic circuits. A compact model was developed for a voltage controlled FTM based on experimental results and ferroelectric switching dynamics. Resistance measurements were taken as a function of the amplitude and duration of programming voltage pulses. Good agreement was found between simulation results and experimental measurement. An FTM-based spike timing dependent plasticity circuit was simulated to show

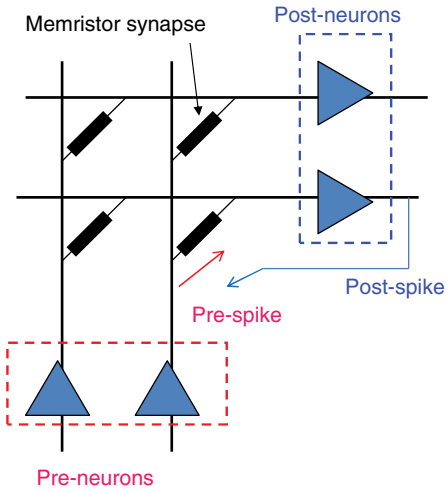


Figure 5.37 Typical neuromorphic network for an STDP application using ferroelectric tunnel memories for synapsis. Based on Z. Wang *et al.* (University of Paris-Sud, Beihang University), *Applied Physics Letters*, February 2014 [33].

its use in neuromorphic circuits. If a programming voltage was applied to the ferroelectric film with a fully up-polarized domain, the down-polarized domains nucleate and propagate and the tunnel resistance can be calculated by the parallel resistance model.

A typical neuromorphic network for an STDP application is shown in Figure 5.37, based on a 2×2 crossbar. The spikes produced by neurons can tune the memristor conductance, which is the synaptic weight. The change of the weight is approximately an exponential function of the time difference between the pre-spike and post-spike.

For LTP, a positive voltage pulse whose width decays with the time frame is generated when the pre-neuron spikes and a negative pulse of constant width is triggered between the top electrode and the bottom electrode when the post-neuron spikes. For LTD, a positive voltage pulse of decaying width is produced between the top electron and the bottom electrode when the post-neuron spikes. Meanwhile, a positive pulse of constant width occurs at the MOS gate when the pre-neuron spikes.

A compact model of the ferroelectric tunnel memristor was based on ferroelectric switching dynamics and experimental results. The model was programmed in Verilog A language for the Cadence platform and Spectre simulator.

Synaptic plasticity can be mimicked by electronic synaptic devices. In January of 2016, ETRI discussed synaptic plasticity in ferroelectric thin film transistors where the channel conductance was gradually modulated by varying the amounts of aligned ferroelectric dipoles [34]. By using ferroelectric thin films as the gate insulator for thin film transistors (TFT), channel conductance can be defined as the synaptic plasticity, and is gradually modulated by variations in the amounts of aligned ferroelectric dipoles. Poly(vinylidene fluoride-trifluoroethylene) [P(VDF-TrFE)]-poly (methyl methacrylate) (PMMA) blended films were chosen and their switching kinetics were investigated. The switching time for ferroelectric polarization was influenced by the amplitude of the applied electric field and the ratio of ferroelectric beta-phases in P(VDF-TrFE)-PMMA films. The switching time of the P(VDF-TrFE) increases when decreasing the pulse amplitude or the ratio of ferroelectric beta-phases by including PMMA. The activation electric field was also found to increase with the blended amount of PMMA.

Synapse TFTs were made using the P(VDF-TrFE)-PMMA as gate insulator and In-Ga-Zn-O for active channels. The drain currents of the synapse TFTs gradually increased when voltage pulse signals of a given duration were applied repeatedly. This suggested that the synaptic weights could be modulated by the number of external pulse signals and that the proposed synapse TFT could be used to mimic the operation of biosynapses.

5.7 Early Neuromorphic Computers Using Phase Change Memory

A 2-PCM synapse for use in a large scale neuromorphic STDP system was discussed in 2012 by CEA-LETI and CEA-LIST [35]. The neural network architecture used the gradual crystallization behavior of a PCM device to emulate both synaptic potentiation and synaptic depression. They used a simplified biological learning rule where long term potentiation and long term depression can be produced with a single invariant crystallizing pulse.

The architecture was simulated on an event-based simulator using a behavioral model for the PCM devices, which was validated with electrical characterization. The system was comprised of about 2 million synapses and directly learns from event-based dynamic vision sensors. It was tested with real data and was able to extract complex and overlapping temporally correlated features such as car trajectories on a freeway. Complete trajectories could be learned with a detection rate above 90%. The synaptic programming power consumption during learning was found to be as low as 100 nW for scaled PCM technology.

An energy efficient 2-PCM synapse using PCM elements as the synapses in large scale neuromorphic systems was discussed in August of 2012 by CEA-LIST, CEA-LETI, and MINATEC [36]. The spiking neural network architecture used the gradual crystallization behavior of the PCM to emulate synaptic potentiation and synaptic depression. A simplified rule was used where long term potentiation and depression can both be products with a single invariant crystallizing pulse. Simulations were run using a behavior model for the PCM devices.

The system of 2 M synapses learned from event-based dynamic vision sensors and was able to extract complex features such as car trajectories on a roadway. Power consumption was estimated to be as low as 100 nW. Robustness to variability was noted. Previously similar implementations have been published with memristor-type devices. PCM is, however, considered a promising embedded memory technology for large scale integration in CMOS.

A hardware implementation of brain-inspired architecture consisting of 2-D crossbar arrays of phase change synaptic devices was discussed in December of 2013 by Stanford University, the University of Pennsylvania, and IBM [37]. These devices were shown to achieve associative learning and perform pattern recognition. Device and array-level studies were done using a 10×10 array of phase change synaptic devices. These devices exhibited pattern recognition that was robust against synaptic resistance variations. Large variations can be tolerated by increasing the number of training iterations. Measurements showed that an increase in initial variation from 9 to 60% causes required training iterations to increase from 1 to 11.

It was demonstrated that brain-like pattern recognition could be implemented using 2-D crossbar arrays of phase change synaptic devices. Individual synaptic device characteristics were presented first for phase change devices in a 10 × 10 array. An example of a single PCM cell DC switching is shown Figure 5.38.

The array level operation and programming method and algorithms used for learning were explained. Finally, analysis on resistance variation and its effect on array level learning and pattern completion were discussed. Gradual programming to achieve analog storage was done by applying multiple pulses to the same PCM cell, as illustrated in Figure 5.39. By adjusting pulse amplitudes and widths, it is possible to do more gradual switching.

The 10 × 10 crossbar array had 100 synaptic devices and 10 recurrently connected neurons. The computer supplied integrated and fire neurons and PCM cells were used as synaptic devices between neurons. The input terminal of each neuron was connected to a bit-line (BL) node and the output terminal of each neuron was connected to a word-line (WL) node. In learning, synaptic weights are updated so that synapses between coactive neurons that fire in the same 100 μs time window get stronger; otherwise the synaptic weight does not change.

It was shown that the synaptic networks could implement a robust pattern recognition through brain-like learning. Test patterns were shown capable of being stored and recalled associatively. Initial resistance variations could be tolerated by adding more training, which consumed more energy.

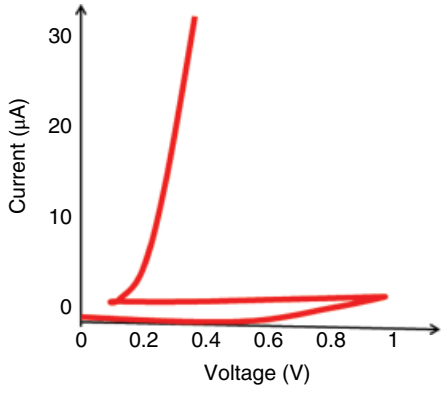


Figure 5.38 Single PCM cell DC switching. Based on S.B. Eryilmaz *et al.* (Stanford University, University of Pennsylvania, IBM), IEEE IEDM, December 2013 [37].

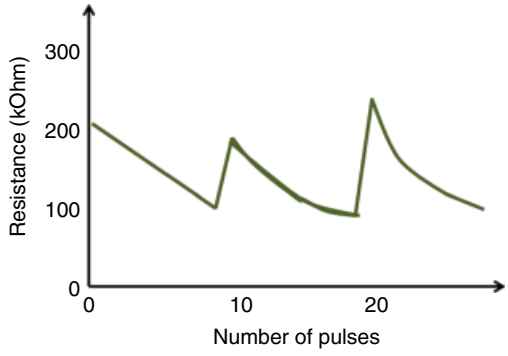


Figure 5.39 Illustration of gradual SET characteristics of a single cell. Based on S.B. Eryilmaz *et al.* (Stanford University, University of Pennsylvania, IBM), IEEE IEDM, December 2013 [37].

A 64 K-bit cell phase change memory (PCM) synaptic array with on-chip neuron circuits for continuous learning was discussed in 2015 by IBM [38]. The neuromorphic core with a 64 k cell PCM synaptic array had a learning capability. The 256 configurable on-chip neuron circuits performed leaky integrate-and-fire (LIF) and a synaptic weight update based on spike-timing dependent plasticity (STDP). A 2 T-1R PCM unit cell separates LIF and synaptic learning paths, minimizing the neuron circuit size. The circuit for the STDP learning algorithm along with a 2 T-1R structure enables both LID and STDP learning to operate asynchronously and simultaneously within the array. This avoids power consumption associated with timing methods as well as other complications. A hardware demonstration of learning with large capacity due to a large array size and analog synaptic weights was shown.

In the LIF mode circuit, shown in Figure 5.40 (a), when the pre-synaptic neuron fires, the axon driver enters its LIF mode. The axon driver generates a short LIF pulse on the LIF WE to turn on LIF TR. This discharges the neuron circuit capacitor through the LIF transistor and PCM cell. The neuron circuit fires if the voltage on the capacitor drops below a threshold voltage and then after a short delay triggers the STDP bit line. The leak resistor continuously restore V_{cap} to the resting potential.

In the STDP mode, shown in Figure 5.40 (b), after the AD turns off the LIF transistor, it continues to enable the STDP transistor with a slowly varying STDP pulse on the STDP word-line. If the neuron circuit generates an STDP bit line pulse while the STDP transistor is on, the PCM cell is programmed. The size of resistance change is determined by the STDP word-line voltage governed by the relative timing between the STDP WL pulse and the BL pulse.

Figure 5.41 illustrates the RESET current in μA versus SET resistance in ohms for a PCM. The area of applicability to storage class memory and for synaptic memory is indicated on the chart. For a synaptic memory device with power <0.8 pJ, the RESET current must be lower than 40 μA and the SET resistance around 10^6 ohms.

In September of 2015, IBM and the Universities of Oxford and Exeter discussed using PCM with FET access devices [39]. A detailed study of accumulation-based computation was done and efficient factorization using PCM cells was shown with the intent of this leading to massively parallelized computations. It was shown that CMOS-integrated PCM cells can be used to perform arithmetic operations using an accumulator-based computational method. Advanced arithmetic capabilities were shown. Intermediate states between pulses were shown to be accessible for up to an hour, which demonstrates that a phase-change accumulator can realize a form of non-von Neumann computing in which the intermediate states of a calculation may be stored simultaneously by the same cell that carries out the computation. Scaling would reduce the power consumption without affecting the computational properties.

The progress in phase change memory in terms of brain-inspired neuromorphic systems implemented with PCM was discussed in 2016 by IBM and Macronix [40]. The PCM field was surveyed over the previous five years. Two paths towards higher density were discussed. One was through 3D integration using PCM and 3D capable access devices. The other was through using multiple bits per cell while managing resistance drive caused by structural relaxation of the amorphous phase. Brain-inspired neuromorphic systems using PCM were also described.

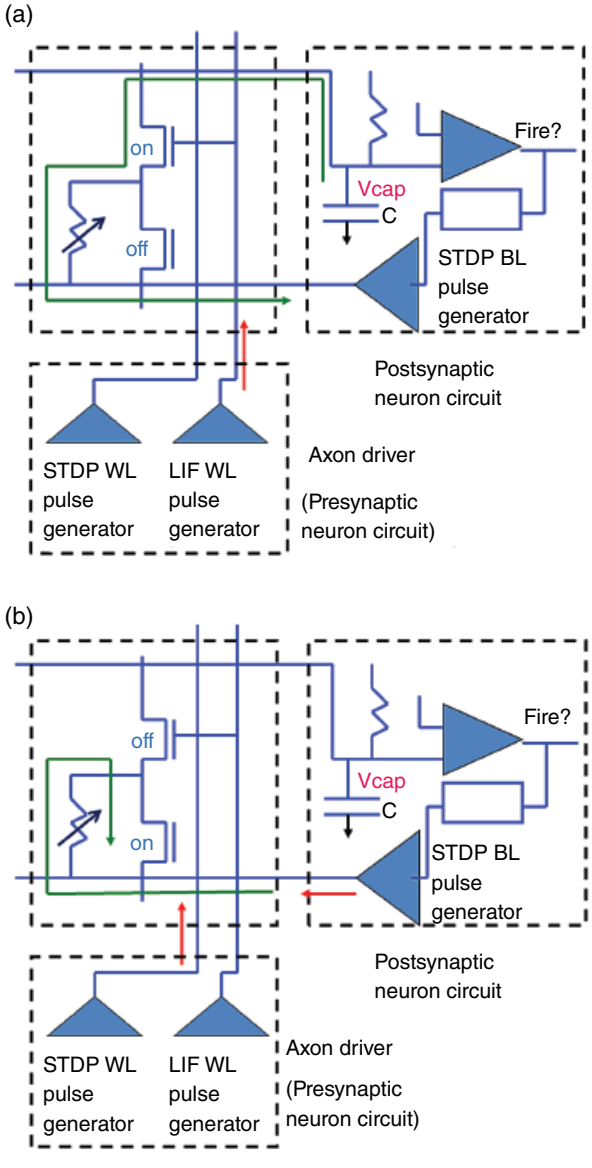


Figure 5.40 Illustration of neuromorphic circuits in (a) LIF mode and (b) STDP mode. Based on S. Kim *et al.* (IBM), IEDM, December 2015 [38] (permission of IEEE).

5.8 Resistive RAMs in Neuromorphic System Design and Application

5.8.1 Design for Synaptic Devices for Neuromorphic Computing

In June of 2014, ASU, Stanford, and the University of Pennsylvania discussed design considerations for synaptic devices for neuromorphic computing [41]. Two terminal

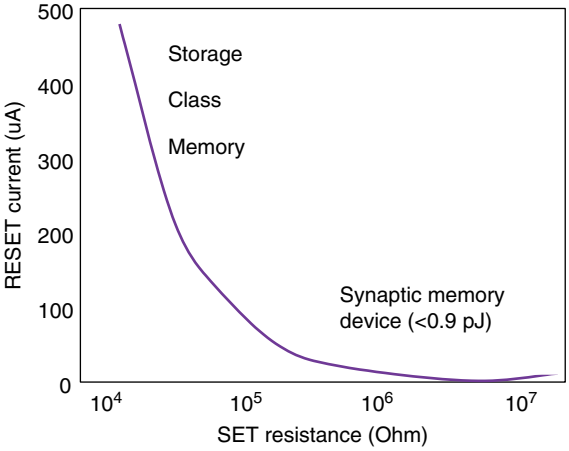


Figure 5.41 Illustration of the RESET current versus SET resistance for a PCM. Based on S. Kim *et al.* (IBM), IEDM, December 2015 [38].

RRAMs with electrically triggered resistance modulation were proposed as synaptic devices for neuromorphic computing. These included phase change memory (PCM), resistive RAM, and CB-RAM. This study considered design considerations of synaptic devices for plasticity and learning. An ideal synaptic device should have characteristics like: dimension, energy consumption, operation frequency, and dynamic range that are scalable to biological systems with comparable complexity.

The HfOx RRAM is considered a good candidate for large scale integration of a neural network since synaptic devices with scalability to sub-10 nm are preferred and an HfO_x RRAM cell has been demonstrated as 10 nm × 10 nm. The energy consumption per synaptic event is around 1–10 fJ and some HfO_x RRAM have already shown sub-pJ energy consumption. Brain-like systems can operate as slow as 1–10 Hz, which gives the 10 ns operating RRAM a lot of flexibility to design trade-offs for energy consumption.

Synaptic plasticity characteristics in biological synapses show analog-like behavior with many synaptic weight states. In general, multilevel states bring advantages to networking. The RRAM has level programmable RESET. A 10 year data retention is required along with about 10^9 endurance cycles. A disadvantage of the RRAM and the CB-RAM is variability in the resistance modulation, which will require further material and device engineering.

Using a RRAM crosspoint array for dictionary learning was discussed in 2014 by Arizona State University [42]. The array had a neurophysics-inspired parallel architecture, which performed matrix product and dictionary update in a fully parallel fashion at a speed that is independent of the matrix dimension. The system was implemented in 65 nm CMOS technology using an RRAM memory cell for high speed unsupervised dictionary learning. It performed more than 3000 times faster than state-of-the-art software, which enabled real-time feature extraction on a single chip.

The resistive crosspoint array structure was used since the crosspoint mimics the structural map of a neural system. At each crosspoint, the conductance of a memory cell represents the synapse weight, as shown the graph in Figure 5.42 of conductance change versus voltage pulse width. RRAM is used due to its nonvolatility, high integration density, and low power consumption. Analogous to a synapse device, the conductance of an RRAM cell is increased (decreased) by a positive (negative) voltage pulse.

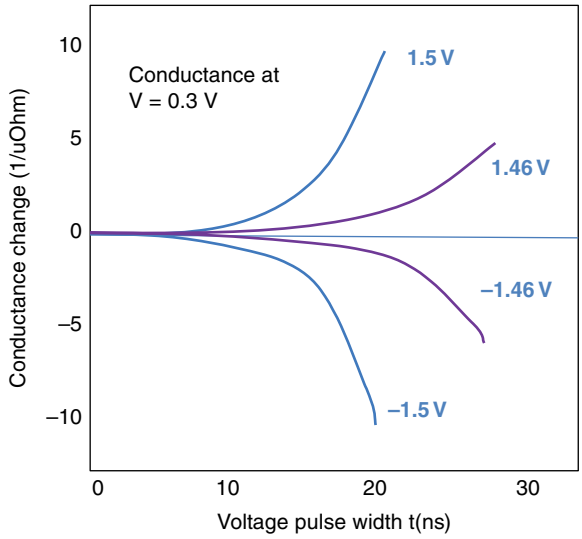


Figure 5.42 Conductance change as a function of voltage pulse width "*t*". Based on D. Kadetotad *et al.* (Arizona State University), Biomedical Circuits and Systems Conference, October 2014 [42].

Suppressing intrinsic variations during the synaptic training process in metal-oxide synaptic RRAMs was discussed in October of 2014 by Peking University [43]. These variations can affect the accuracy of the neuromorphic computations. The influence was investigated of resistance variations on metal-oxide RRAM-based neuromorphic computation systems. Atomic DMC (Dynamic Monte-Carlo) simulations were performed to study the random migration of oxygen vacancies that are responsible for the origin of the variations. A new methodology was suggested to improve the robustness of the metal-oxide RRAM-based neuromorphic computation system.

Neuromorphic computing, which has features of parallelism, adaptivity to complex input information, and error tolerance, was discussed by Stanford University in 2015 [44]. The synapse in a neuromorphic system requires high integration density and low energy consumption, which are properties of oxide-based RRAM. An optimized oxide-based RRAM synaptic device was designed and applied to the performance of a neuromorphic visual system. The intrinsic variation of the RRAM synaptic devices was suppressed to achieve the high recognition accuracy and efficiency required in a neuromorphic visual system.

5.8.2 Using RRAMs in Various Neuromorphic Computing Applications

Neuromorphic computation is used more effectively than von Neumann systems in applications such as image and auditory processing. In December of 2015, Peking University and Sandisk discussed an optimized learning method for grayscale image recognition in an RRAM-based analog neuromorphic system [45]. An analog neuromorphic visual system is developed using a fabricated RRAM array as a synapse network. A new training method is used to optimize the performance of the system by using segmented synaptic behavior. This segmented behavior is due to the dynamics of the conductive filament. Initially the CF thins near the top electrode and resistance increases slowly. When the CF is just ruptured the electric field intensity increases due to the small tunneling gap, which results in faster resistance change. As the gap enlarges,

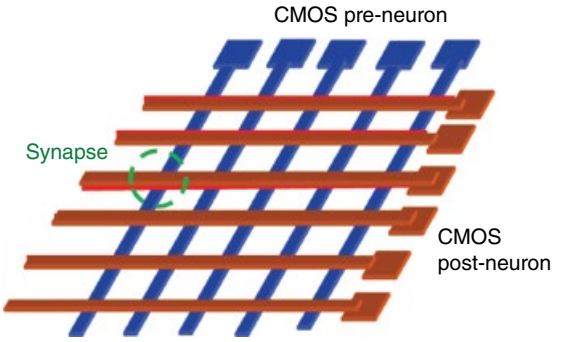


Figure 5.43 Illustration of neuromorphic system of CMOS neurons and RRAM synapses. Based on Z. Chen *et al.* (Peking University, Sandisk), IEDM, December 2015 [45].

the resistance slowly increases with the resistance increase slowing and then finally saturating.

The experimental results using the segmented method indicated that learning accuracy was improved from 77.83 to 91.32% over a nonsegmented system, energy consumption was decreased by more than two orders of magnitude, and learning efficiency improved over traditional learning schemes.

Figure 5.43 shows a neuromorphic system consisting of CMOS neurons and RRAM synapses. In this system the RRAM devices serving as synapses connect the CMOS neurons at the different layers. The conductance of the RRAM synapse represents the strength of the connection between two neurons. This can be modified during the training process by using SET and RESET pulses to change the conductivity of the RRAM.

In December of 2016, Arizona State University and Tsinghua University discussed using a binary neural network with RRAM synaptic devices for classification and on-line training [46]. A large scale neural network was made for use in image recognition. The goal was to find a methodology to binarize the neural network parameters with the intent to reduce the precision of weights and neurons to 1-bit for classification and less than 8-bit for on-line training. An experimental binary neural network was demonstrated on Tsinghua's 16 Mb RRAM macro chip made in 130 nm CMOS. The system performance on the MNIST handwritten digital dataset achieved 96.5% accuracy for both classification and on-line training and nearly 97% accuracy by ideal software implementation.

5.8.3 Large RRAM Array Design for Neuromorphic Computing

Various issues with designing large RRAM arrays in neuromorphic computing applications were discussed in 2016 by IBM, POSTECH, and EPFL [47]. In these neuromorphic schemes, device conductance serves as a synaptic weight allowing the multiply-accumulated operation in these algorithms to be performed using the weight data. An illustration of neuromorphic non-Von Neumann computing is shown in Figure 5.44. Here neurons activate each other through networks of programmable synaptic weights, which can be implemented using dense crossbar arrays of resistive RAMS with variable resistance as 1T1R cells that emulate synapses coupling "neuron"-controlled selector device pairs.

This paper explored the use of NVM and other analog-type devices for artificial neural networks trained with a back-propagation algorithm. The advantage of phase change

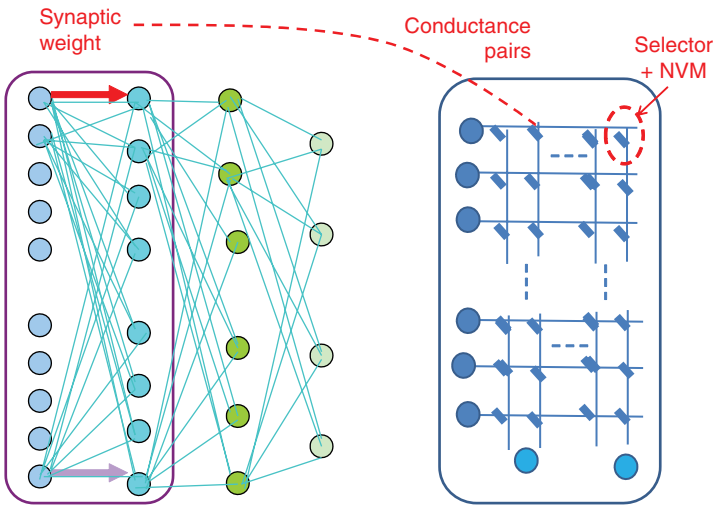


Figure 5.44 An illustration of neuromorphic non-von Neumann computing. Based on A. Fumarola *et al.* (IBM, Pohang University, EPFL-Lausanne), ICRC, October 2016 [47].

memory synapses over the speed and power of GPU-based training had been previously shown and this work assessed the impact of undesirable time-varying conductance change and the trade-offs in designing peripheral circuitry for artificial neural networks.

The benefit of performing computations at the location of the data is in both power and speed.

Another benefit is in the security of data when it does not need to be stored in the cloud but is processed locally. The computations of training using NVM-based synapses must be competitive with that using CPUs or GPUs.

Any on-chip accelerator should be applicable toward networks of different types, such as fully connected deep neural nets or convolutional NN or for networks of different shapes. It was shown that for speed and power of on-chip machine learning, non-volatile memory-based synapses can potentially offer lower power and faster training than GPU-based training for both large and small networks. A critical factor is the design of area-efficient read/write circuitry so that many copies of this circuitry are able to operate in parallel.

Using 2 PCM devices per synapse, a three-layer perceptron with 164 885 synapses was trained with back-propagation on a subset of the MNIST database of handwritten digits using a modified weight update rule compatible with NVM + selector crossbar arrays. It was shown that this weight update modification does not degrade the high accuracies that a three-layer network inherently delivers when trained in software. Nonlinearity and asymmetry in the PCM conductance response limited both training and test accuracy. The asymmetry between the gentle conductance increases of PCM partial SET and the abruptness of PCM RESET was lessened by an occasional RESET strategy, which could be both infrequent and inaccurate. Tolerance results showed that all NVM-type ANN could be expected to be highly resilient to random effects such as variability, yield, and stochasticity, but very sensitive to gradient effects that act to steer all synaptic weights. These observations were extended to assess the impact of

undesigned, time-varying conductance change including drift and leakage of analog CMOS capacitors. A feature of PCM is resistance drift, which is continued relaxation of the material in the amorphous state. The drift is toward an increase in resistance. A potential solution for this issue is a periodic RESET.

PCMO is a nonfilamentary resistance change material in which conductive response was also studied. Resistive switching in PCMO devices is caused by the slow and gradual drift of oxygen ions and vacancies in the polycrystalline PCMO layer. PCMO NVMs show gradual SET and RESET characteristics, but, unlike PCM, there is no need to stop training and perform an occasional RESET operation. RRAM devices based on PCMO can be considered as potentially, providing both a gradual conductance increase and conductance decrease.

5.8.4 Advantages of RRAM over SRAM Crossbar Arrays in Matrix Multiplication

In October of 2015, Sandia National Labs discussed using RRAM crossbar arrays to process large amounts of data for brain-type computing [48]. Analog computation on a crossbar provides fundamental energy scaling advantages over digital memories. The two key computational kernels that are more efficient on a crossbar include: performing a parallel read or a vector matrix multiply and performing a parallel write where every weight is programmed based on the outer product of the row and column inputs. These two computational kernels comprise the basis of neuromorphic algorithms.

For a parallel analog matrix multiply each column of the crossbar performs a vector dot product for a column. Inputs are represented by a voltage or length of a voltage pulse. Weights are represented by the RRAM conductances. The multiplication is done by summing currents. The minimum energy to read the crossbar is determined by the thermal noise of each RRAM.

For a larger crossbar or a higher voltage due to access devices, the energy will be capacitance limited. Even for a capacitance limited read energy, however, the device will still be more energy efficient than a matrix stored in an SRAM array since SRAMs need to be read out one row or word-line at a time and the associated circuitry will need to be charged N times for N rows. In an analog RRAM crossbar, everything can be done in parallel so that the circuitry needs only to be charged once. The conclusion was that the energy to perform a parallel read or parallel write on an analog $N \times N$ RRAM crossbar array scales as $O(N^2)$ while a digital implementation scales as $O(N^3)$.

5.9 Neuromorphic Memories Using Polymer and Flexible Memories

Synaptic behavior in flexible IZO transistors on plastic substrates was discussed in November of 2013 by Nanjing University and the Chinese Academy of Science [49]. Flexible IZO transistors were used for synapses in neural computers. Synapses are the connections between neurons and give neurons computational abilities. In this study, indium–zinc oxide (IZO) was used for flexible synaptic EDL transistors made on a plastic substrate using a self-assembly method at room temperature. The gate dielectric was

made of proton conducting phosphorus-doped nanocrystal SiO_2 electrolyte. Excitatory post-synaptic current, paired pulse facilitation, and long term memory are implemented in the flexible artificial synapses. These IZO flexible electronic synapses have potential for building neuromorphic systems. A schematic diagram of the self-assembled flexible IZO-based transistors is shown in Figure 5.45.

These flexible EDL transistors show an excitatory synaptic current, which is analogous to that found in biological synapses. Pre-synaptic spike duration time-dependent excitory post-synaptic currents are observed in such synaptic transistors because of the slow response of protons in the electrolyte film. Long term plasticity is also represented in the flexible IZO-based synaptic transistor. Flexible artificial synapses such as these are precursors in building neuromorphic systems. These flexible electronic synapses have potential applications for building neuromorphic circuits.

Organic copolymer RRAM used as a synapse in a neuromorphic system was discussed in December of 2013 by CNR-IMEM [50]. An organic RRAM/memristor was used to represent a synapse in a neuromorphic system. The analogy with synapses was demonstrated by showing the artificial circuit to mimic part of the nervous system of a low level animal (a pond snail). A stochastic network made of block copolymers, polyaniline, and gold nanoparticles was described. Similarities of the network and the primitive nervous system were observed and various training algorithms were applied; features similar to imprinting/learning were observed.

The working principle of the device was based on the difference in conductivity of polyaniline in the oxidized and reduced states. A thin polyaniline layer was used in the organic RRAM device with a line of solid electrolyte composed of Li^+ salt-doped polyethylene oxide where all redox reactions occur. The variation of the conductivity is accompanied by the migration of Li^+ ions between the polyaniline and the polyethylene oxide layers. The device must increase its conductivity with the frequency of its involvement in the formation of signal transfer pathways for learning to occur.

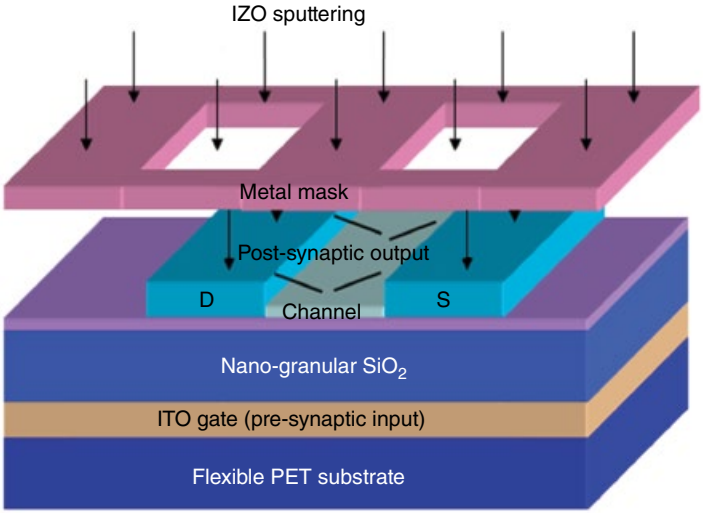


Figure 5.45 Schematic diagram of self-assembled flexible IZO transistors. Based on J. Zhou *et al.* (Nanjing University, CAS), *IEEE Electron Device Letters*, November 2013 [49].

To construct a structure similar to the human brain, a 3D organization of neurons is needed rather than a planar structure. Structuring and supporting random 3D architectures is difficult and the approach used was a bottom-up self-organization method using block-copolymers with phase separation in the formed layers. The main component of the network was a synthesized copolymer with alternating parts of solid electrolyte (polyethylene oxide) and an insulator that acted as a doping agent to polyaniline. The copolymer was mixed with polyaniline and gold nanoparticles, which acted as threshold elements forming Schottky barriers due to the difference in work function between gold and polyaniline. Future work is planned to combine networks of organic devices with living biological objects such as microentities of slime mold.

Polymer RRAM used for neuromorphic applications was discussed in June of 2014 by East China University of Science and Technology, the Chinese Academy of Sciences, and Fudan University [51]. Polymer materials are attractive for implementing RRAM devices due to their low cost, easy processing, mechanical flexibility, and tunable electronic performance and compatibility with CMOS technology. The polymer RRAM is nonvolatile, fast, has low power consumption, a large ON/OFF ratio, high endurance, and long retention. The analog RRAM has the ability to emulate the fundamental synaptic functions of plasticity, which are spike timing-dependent plasticity (STDP), spike-rate dependent plasticity (SRDP), and learning behaviors.

Organic memristors in neuromorphic networks used as analogs of synapses were discussed in November of 2015 by the University of Parma [52]. Organic memristive device-based circuits were explored for biology inspired processing of information. The memristive devices were used in neuromorphic networks and mimicked some of the properties of the human nervous system. The memristive devices were used as electronic analogs of biological synapses.

The synaptic behavior of RRAM made with Ag nanoparticles on flexible substrate was discussed in July of 2016 by Fudan University [53]. The switching behavior of TaN/Al_2O_3:Ag:AnO//ITO RRAMS made on flexible substrates was considered. A schematic cross-section of the proposed device is shown in Figure 5.46.

Intended applications were flexible nonvolatile memory for wearable electronics applications and neuromorphic computers. Embedded Ag nanoparticles give improved device yield and reduce variability in resistance from over 160 to 30%. They also give

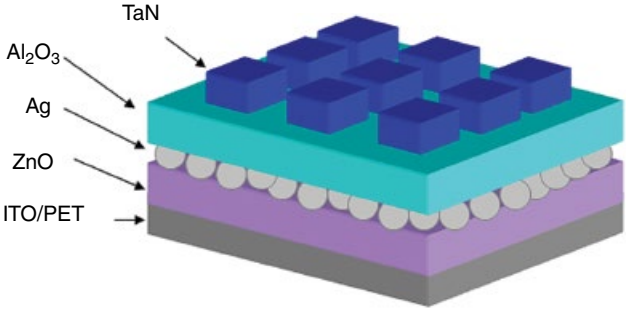


Figure 5.46 Schematic cross-section of TaN/Al_2O_3:Ag:AnO/ITO RRAM on flexible substrate. Based on D.T. Wang *et al.* (Fudan University), *IEEE Electron Device Letters*, July 2016 [53].

reduced variability in SET voltage from 35 to 18% and for RESET voltages from 40 to 11% respectively.

Synaptic behavior was shown in this flexible device, which gave it potential for use in neuromorphic circuit applications. Its low power requirements make it closer to emulating biological systems. An effective method was developed to improve flexible RRAM characteristics by coating Ag nanoparticles. The fabricated flexible RRAM showed a high ON/OFF ratio, low resistance fluctuations, high speed, low SET/RESET voltage, low power, and good retention compared to previously reported flexible RRAM devices. Synaptic behavior was shown, which can be used to emulate synapses.

The pulse response of the devices with Ag nanoparticles is plotted in Figure 5.47. Figure 5.47 (a) shows 50 potentiation sweeps of 0.5 V for 150 ms and (b) shows 200 depression sweeps of –0.5 V for 150 ms. After each pulse, the device was measured at 100 mV read pulse and the read current was plotted. It was concluded that embedded Ag nanoparticles in RRAM could be used to mimic the synapse when integrated in a flexible system.

Organic neuromorphic chips for an efficient human–machine interface were discussed in September of 2016 by the University of Minnesota and the Gwangju Institute of S&T [54]. Their results were expected to incorporate organic electronics into large area neuromorphic systems with complex logic circuits and transparent multifunctional interfaces for direct optical and cellular stimulation. Even though these circuits do not compete with inorganic electronics in performance, the ability of the organic

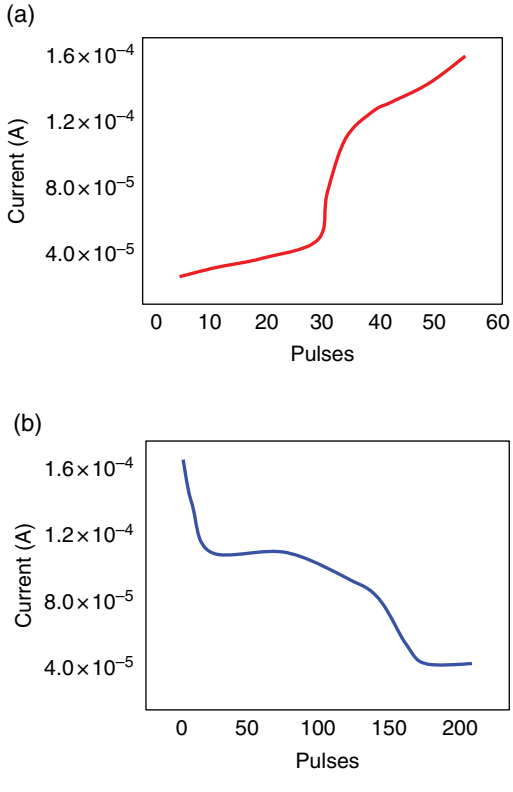


Figure 5.47 RRAM pulse response with Ag nanoparticle: (a) potentiation sweep and (b) depression sweep. Based on D.T. Wang *et al.* (Fudan University), *IEEE Electron Device Letters*, July 2016 [53].

neuromorphic technology to mimic functions of biocompatibility were considered promising for large area applications and for direct biointerfacing hybrid systems. An organic synapse concept is proposed where organic-based neuromorphic devices based on widely available materials can be a route to advance memory circuits and functional neuro-inspired electronics. The organic FET-based synaptic devices were constructed on top of a flexible transparent polyelthlene terephthalate substrate with a stack of ITO/PEDOT PMMA. An ultrathin metallic film with a tunneling layer is formed using ambient oxidation and used for the transistor structure for short-term charge retention. The device had a self-formed floating gate that was used as a memory.

Organic NVM devices with more than 500 distinct conductance states were discussed in February of 2017 by Stanford University, the University of Groningen, Sandia National Labs, and the University de Sao Paulo [55]. The use of a nonvolatile organic electrochemical device was considered as a low voltage artificial synapse for neuromorphic computing. The electrochemical neuromorphic organic device (ENODe) operates with a fundamentally different mechanism from existing RRAMs. ENODe switches at low voltage and low energy of <10 pJ for 10^3 μm^2 area devices. Over 500 distinct, nonvolatile conductance states are exhibited within a 1 V range. High classification accuracy was achieved when implemented in neural network simulations. The plastic ENODes were made on flexible substrates enabling the integration of neuromorphic functions on stretchable electronics substrates.

References

1 Yu, S. (2014) Orientation classification by a winner-take-all network with oxide RRAM based synaptic devices (ASU), ISCAS, pp.1058, June 2014.

2 Woo, J. *et al.* (2016) Multilevel conductance switching of a HfO_2 RRAM array induced by controlled filament for neuromorphic applications (POSTECH), SNW, June 2016.

3 Woo, J. *et al.* (2016) Improved synaptic behavior under identical pulses using AlO_x/HfO_2 bilayer RRAM array for neuromorphic systems (Pohang University). *IEEE Electron Device Letters*, 37 (8), August 2016.

4 Woo, J. *et al.* (2016) Optimized programming scheme enabling linear potentiation in filamentary HfO_2 RRAM synapse for neuromorphic systems (Pohang University), *IEEE Trans. Electron Devices*, 63 (12), December 2016.

5 DeSalvo, B. *et al.* (2015) From memory in our brain to emerging resistive memories in neuromorphic systems (CEA LETI), International Memory Workshop, May 2015.

6 Zhang, W. (2015) An electronic synapse device based on solid electrolyte resistive random access memory (Xidian University, NCKU, National Sun Yet-sen University, Xiamen University, NCTU), *IEEE Electron Device Letters*, 36 (8), August 2015.

7 Clermidy, F. *et al.* (2014) Advanced technologies for brain-inspired computing (CEA-LETI, CEA-LIST, INRIA), ASP-DAC, January 2014.

8 Jang, J., Park, S., Jeong, Y. and Hwang, H. (2014) ReRAM-based synaptic device for neuromorphic computing (Pohang University, Gwangju Institute of Science and Technology), ISCAS, pp. 1054, June 2014.

9 Jang, J.W. *et al.* (2015) Optimization of conductance change in $Pr_{1-x}Ca_xMnO_3$-based synaptic devices for neuromorphic systems (Pohang University of Science and Technology, Gwangju Institute of Science and Technology, IBM). *IEEE Electron Device Letters*, 36 (5), May 2015.

10 Moon, K. *et al.* (2016) Analog synapse device with 5-b MLC and improved Data Retention for neuromorphic system (Pohang University of Science and Technology, Gwangju Institute of Science and Technology, IMEC-Leuven). *IEEE Electron Device Letters*, 37 (8), August 2016.
11 Chou, T. *et al.* (2015) Neuromorphic pattern learning using HBM electronic synapse with excitatory and inhibitory plasticity (National Chiao Tung University), VLSI-TSA, April 2015.
12 Lee, M.-S. *et al.* Implementation of short-term plasticity and long-term potentiation in a synapse using Si-based type of charge-trap memory. *IEEE Trans. on Electron Devices*, 62 (2), February 2015.
13 Sharma, A. *et al.* (2015) Novel CMOS-compatible a-Si based oscillator and threshold switch (Carnegie Mellon University), ESS-DERC, September 2015.
14 Wang, T. *et al.* (2014) 3D synaptic architecture with ultralow sub-10 fJ energy per spike for neuromorphic computation (National Chiao Tung University), IEDM, December 2014.
15 Gaba, S., Sheridan, P., Du, C. and Lu, W. (2014) 3-D vertical dual-layer oxide memristive devices (University of Michigan). *IEEE Trans. on Electron Devices*, 61 (7), 2581, July 2014.
16 Piccolboni, G. *et al.* Investigation of the potentialities of vertical resistive RAM (VRRAM) for neuromorphic applications (CEA, LETI, Minatec, IM2NP, IMEP-LAHC), IEDM, December 2015.
17 Piccolboni, G. *et al.* Vertical CBRAM (V-CBRAM): from experimental data to design perspectives (CEA, LETI, Minatec), IMW, May 2016.
18 Li, H. *et al.* (2016) Four-layer 3D vertical RRAM integrated with FinFET as a versatile computing unit for brain-inspired cognitive information processing (Stanford University, NNDL Narlabs), VLSI, June 2016.
19 Adam, G.C. *et al.* (2016) Highly-uniform multi-layer ReRAM crossbar circuits (University of California Santa Barbara), ESSDERC, September 2016.
20 Gupta, I. *et al.* (2015) A cell classifier for RRAM process development (University of Southampton, Imperial College, London). *IEEE Trans. on Circuits and Systems – II Express Briefs*, 62 (7), July 2015.
21 Gao, B. *et al.* (2016) Oxide-based analog synapse: physical modeling, experimental characterization, and optimization (Tsinghua University, Peking University, Southern University of S&T), IEDM, December 2016.
22 Ielmini, D. *et al.* (2016) Neuromorphic computing with hybrid memristive/CMOS synapses for real-time learning (Politecnico di Milano, IU NET), ISCAS, May 2016.
23 Ambrogio, S. *et al.* (2016) Neuromorphic learning and recognition with one-transistor–one-resistor synapses and bistable metal oxide RRAM (Politecnico di Milano, Micron Technology). *IEEE Trans. on Electron Devices*, 63 (4), April 2016.
24 SAmbrogio, S. *et al.* (2016) Novel RRAM-enabled 1T1R synapse capable of low-power STDP via burst-mode communication and real-time unsupervised machine learning (Politecnico di Milano, Micron Technology), VLSI, June 2016.
25 Yao, P. *et al.* (2015) The effect of variation on neuromorphic network based on 1T1R memristor array (Tsinghua University, Tshinghua National Laboratory), NVMTS, October 2015.
26 Moon, K. (2016) ReRAM-based analog synapse and IMT neuron device for neuromorphic system (POSTECH), VLSI-TSA, April, 2016.

27 Mahalanabic, D. *et al.* (2016) Demonstration of spike timing dependent plasticity in CBRAM devices with silicon neurons (Arizona State University), ISCAS, May 2016.

28 Park, J. *et al.* (2016) TiO_x-based RRAM synapse with 64-levels of conductance and symmetric conductance change by adopting a hybrid pulse scheme for neuromorphic computing (Pohang University). *IEEE Electron Device Letters*, 37 (12), December 2016.

29 Werner, T. *et al*. (2016) Experimental demonstration of short and long term synaptic plasticity using OxRAM multi k-bit arrays for reliable detection in highly noisy input data (CEA, LETI, MINATEC, CEA, LIST, INSERM), IEDM, December 2016.

30 Kaneko, Y., Nishitani, Y., Ueda, M. and Tsujimura, A. (2013) Neural network based on a three-terminal ferroelectric memristor to enable on-chip pattern recognition (Panasonic), VLSI Technology Symposium, June 2013.

31 Nishitani, Y., Kaneko, Y. and Ueda, M. (2014) Artificial synapses using ferroelectric memristors embedded with CMOS circuit for image recognition (Panasonic), Device Research Conference, pp. 297, June 22, 2014.

32 Kaneko, Y., Nishitani, Y. and Ueda, M. (2014) Ferroelectric artificial synapses for recognition of a multishaded image (ATRL Panasonic). *IEEE Trans. on Electron Devices*, 61 (8), August 2014.

33 Wang, Z. *et al.* (2014) Compact modeling of ferroelectric tunnel memristor and its use for neuromorphic simulation (University of Paris-Sud, Beihang University). *Applied Physics Letters*, 104, 053505, February 2014.

34 Kim, E., Kim, K. and Yoon, S. (2016) Investigation of the ferroelectric switching behavior of P(VDF-TrFE)-PMMA blended films for synaptic device applications. *Journal of Physics D: Applied Physics*, January 2016.

35 Suri, M. *et al.* (2012) Interface engineering of PCM for improved synaptic performance in neuromorphic systems (CEA-LETI, CEA-LIST, IEMN), IMW, May 20, 2012.

36 Bichler, O., Suri, M., Querlioz, D. *et al.* (2012) Visual pattern extraction using energy-efficient "2-PCM Synapse" neuromorphic architecture (CEA-LIST, CEA-LETI, MINATEC). *IEEE Trans. on Electron Devices*, 59 (8), 2206, August 2012.

37 Eryilmaz, S.B. *et al.* (2013) Experimental demonstration of array-level learning with phase change synaptic devices (Stanford University, University of Pennsylvania, IBM), IEEE IEDM, December 2013.

38 Kim, S. *et al.* (2015) NVM neuromorphic core with 64 k-cell (256-by-256) phase change memory synaptic array with on-chip neuron circuits for continuous in-situ learning (IBM), IEDM, December 2015.

39 Hosseini, P. *et al.* (2015) Accumulation-based computing using phase-change memories with FET access devices (University of Oxford, IBM, University of Exeter), *IEEE Electron Device Letters*, 36 (9), September 2015.

40 Burr, G.W. *et al.* (2016) Recent progress in phase-change memory technology (IBM, Macronix). *IEEE Journal of Emerging and Selected Topics in Circuits and Systems*, 6 (2), June 2016.

41 Yu, S., Kuzum, D. and Wong, H. (2014) Design considerations of synaptic device for neuromorphic computing (ASU, Stanford, University of Pennsylvania), ISCAS, pp.1062, June 2014.

42 Kadetotad, D. *et al.* (2014) Neurophysics-inspired parallel architecture with resistive crosspoint array for dictionary learning (Arizona State University), Biomedical Circuits and Systems Conference, October 22, 2014.

43 Gao, B. *et al.* (2014) Origin and suppressing methodology of intrinsic variations in metal-oxide RRAM based synaptic devices (Peking University), ICSICT, October 28, 2014.

44 Kang, J.F. *et al.* (2015) RRAM based synaptic devices for neuromorphic visual systems (Stanford University), IEEE International Conference on Digital Signal Processing (DSP), July 21, 2015.

45 Chen, Z. *et al.* (2015) Optimized learning scheme for grayscale image recognition in a RRAM based analog neuromorphic system (Peking University, Sandisk), IEDM, December 2015.

46 Yu, S. *et al.* (2016) Binary neural network with 16 Mb RRAM macro chip for classification and online training (Arizona State University, Tsinghua University), IEDM, December 2016.

47 Fumarola, A. *et al.* (2016) Accelerating machine learning with non-volatile memory: exploring device and circuit tradeoffs (IBM, Pohang University, EPFL-Lausanne), ICRC, October 2016.

48 Agerwal, S. (2015) The energy scaling advantages of RRAM crossbars (Sandia Labs), European Sensory Science Society (E3S), October 2015.

49 Zhou, J., Wan, C., Zhu, L. and Wan, Q. (2013) Synaptic behaviors mimicked in flexible oxide-based transistors on plastic substrates (Nanjing University, CAS). *IEEE Electron Device Letters*, 34 (11), 1433, November 2013.

50 Erokhin, V. (2013) Organic memristive devices: architecture, properties and applications in neuromorphic networks (CNR-IMEM), ICECS, December 8, 2013.

51 Chen, Y. *et al.* (2014) Polymer memristor for information storage and neuromorphic applications (Chinese Academy of Science, Fudan University, East China University of Science and Technology), Materials Horizons, June 2014.

52 Erikhin, V. (2015) Organic memristive devices based circuits for bio-inspired memorizing and processing of the information (University of Parma), MEMRISYS, November 8, 2015.

53 Wang, D.T. *et al.* (2016) Resistive switching and synaptic behaviors of $TaN/Al_2O_3/ZnO/ITO$ flexible devices with embedded Ag nanoparticles (Fudan University). *IEEE Electron Device Letters*, 37 (7), July 2016.

54 Kim, C.H., Sung, S. and Yoon, M.H. (2016) Synaptic organic transistors with a vacuum-deposited charge-trapping nanosheet (Gwangju Institute of Science and Technology), Science Reports, September 2016.

55 van de Burgt, Y. *et al.* (2017) A non-volatile organic electrochemical device as a low-voltage artificial synapse for neuromorphic computing (Eindhoven University of Technology, Stanford, University of Groningen, Sandia National Laboratory, University of Sao Paulo), Nature Materials, February 2017.

6

Big Data Search Engines and Deep Computers

6.1 Overview of Big Data Search Engines and Deep Computers

The Internet of Things (IoT) collects a large amount of data, much of which is stored in large server farms called collectively "The Cloud". These data are only useful if they can be accessed intelligently and efficiently. An example of a basic data search engine is a ternary content addressable memory (TCAM). This is a memory that can compare input search data with stored data and return the location of matched data. TCAMs are used in both MPU designs and in communication chips such as those for IP-routing.

Deep computers and artificial intelligence will also be needed to intelligently process the data in The Cloud. In addition to neuromorphic computers, which were discussed in an earlier chapter, extreme learning machine architectures and artificial intelligence are being developed. Deep learning can be implemented using a restricted Boltzman machine, which consists of a generative stochastic artificial neural network. Deep learning involves extracting complex information from high-dimensional data without using significant amounts of manual engineering. It is useful for image recognition, speech recognition, and natural language understanding.

6.2 Content Addressable Memories Made Using Various Emerging Nonvolatile Memories

A content addressable memory, or associative memory, searches its entire contents in a single clock cycle. It is commonly used for high speed search applications. A ternary CAM (TCAM) can store and query data using three different inputs 0, 1, and an extra input, which allows the TCAM to perform broader searches based on pattern matching.

Historically TCAMs have been implemented using static RAMs. In advanced technologies, the subthreshold leakage of SRAMs result in high power consumption. Several of the emerging memories, such as MTJ MRAMs, FeFETS, and RRAMs, have ultralow standby power and can also be used to make area and power efficient search engines and TCAMS. Each, however, has different trade-offs in other characteristics.

Memories for the Intelligent Internet of Things, First Edition. Betty Prince and David Prince.
© 2018 John Wiley & Sons Ltd. Published 2018 by John Wiley & Sons Ltd.

6.2.1 Ternary CAMs Using Resistive RAMS

RRAMs can be used in TCAM for big data processing to replace SRAMs. A conventional TCAM cell uses 16 transistors. In 2014, NTHU, NCTU, NDL, and ITRI proposed an RC-filtered stress-decoupled four-transistor and two-resistor (4T2R) nonvolatile TCAM [1]. The device could be used to suppress match-line leakage current from match cells, to reduce parasitic load, and to decouple nonvolatile memory (NVM) stress from word length A 128 × 32 nonvolatile TCAM (nvTCAM) macro was made using an HfO_2-based RRAM in 180 nm CMOS technology. In 2015, a 3T1R nvTCAM, which used a multilevel cell (MLC) RRAM was discussed by NTHU, ITRI, and NCTU. The device had a less than 1 ns search delay (T_{sd}) [2]. A conventional 16 T SRAM TCAM cell is shown in Figure 6.1 (a) and the 3T1R RRAM TCAM cell using an MLC RRAM is shown in (b).

The 2 × 64 × 64b 3T1R nvTCAM macro was made using a back-end-of-the-line (BEOL) fast write HfO_2-based RRAM added to a 90 nm CMOS logic process. This device had a 2.27 times cell size reduction compared with SRAM TCAM in the same technology. The challenges of building this nvTCAM included: area and wire routing for the BEOL RRAM, the need for a higher write current for a long data retention time, and the large resistance ratio, which resulted in a large NVM driver area.

A register-based timing extraction method was used to derive T_{sd} from the test chip search time by excluding the path delay time of the test chip as well as the load board. The measured worst case T_{sd} of a 3T1R nvTCAM macro at minimal VDD of 1 V is 0.96 ns. The nvTCAM search engine is shown in Figure 6.2.

The larger current ratio and minimal number of devices stacked between power rails enabled operation at a low 0.48 V VDD with T_{sd} = 11.3 ns. The RRAM write speed for SET and RESET was less than 5 ns. Memory capacity was 2 blocks × 64 rows × 64 bits. Cell size was 0.8704 μm^2. Search speed was 0.96 ns at VDD = 1 V at 25 °C. V_{ddmin} was 0.48 V at 25 °C. Supply voltage was 0.5–1 V for search and 1.8 V for write.

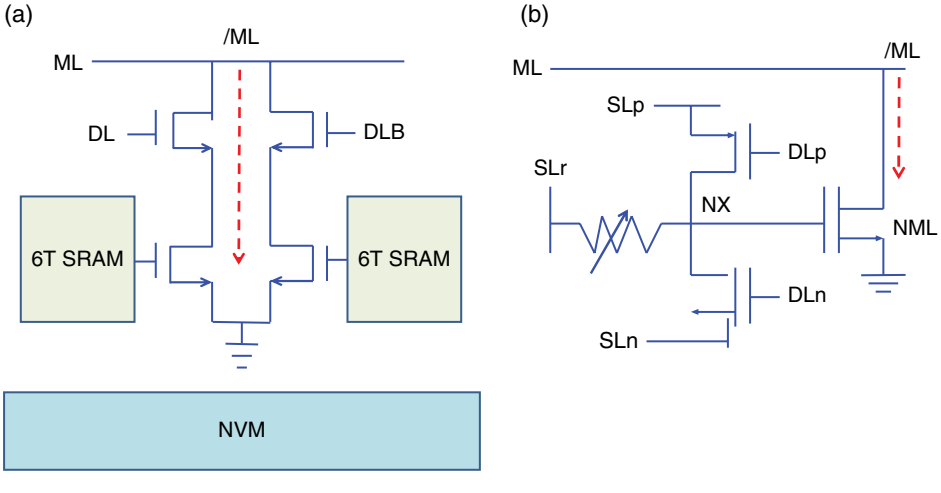


Figure 6.1 TCAM cells: (a) 16-transistor SRAM cell plus NVM and (b) bidirectional voltage divider control for a 3T1R nonvolatile TCAM using MLC RRAM. Based on L. Huang *et al.* (NTHU, NCTU, NDL, ITRI), VLSI Circuit Symposium, June 2014 [1].

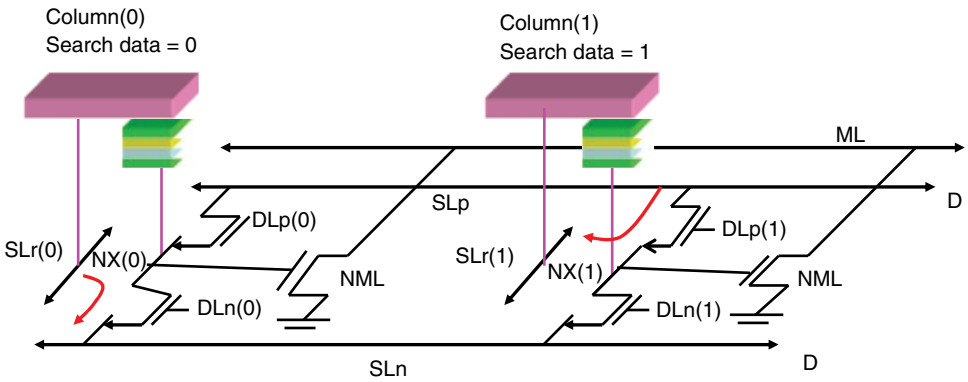


Figure 6.2 Circuit diagram of nvTCAM search engine using RRAM for data storage. Based on M.-F. Chang *et al.* (NTHU, ITRU, NCTU), ISSCC, March 2015 [2].

A nonvolatile ternary content addressable memory (nvTCAM), which was used in network and big data processing, was discussed in 2016 by NTJU, TSMC, and NCTU [3]. The 256 bit wordlength nvCAM was based on an RRAM and was designed to reduce the cell area, search energy, and standby power over that of a SRAM-based TCAM. These advantages were particularly expected in applications with long idle times and frequent-search-few-write operations. A small 2.5 transistor and 1 resistive RAM (2.5T1R) cell was made in a 65 nm CMOS logic process. This cell reduced area and metal line capacitance. A 64×256 bit 2.5T1R nvTCAM macro was made using the 2.5T1R cell and the TSMC logic-process-contact-RRAM. A DFF-based timing extraction method was used to derive T_{sd} by excluding the path delay of the PCB. The measured worst case T_{sd} (1b mismatch time) with $V_{dd} = 1$ V was 1 ns.

6.2.2 CAMs Made Using Magnetic Memory

In February of 2015, Aligarh Muslim University discussed a TCAM made with a five-layer MTJ configuration called a "PentaMTJ" [4]. The structure of the perpendicular magnetic anisotropy (PMA)-based fivelayer TCAM MTJ is shown in Figure 6.3. It is composed of two pinned layers, one free layer, and the three metallic contacts.

It is known that PMA is responsible for a lower threshold current due to the absence of the easy-plane anisotropy term. This TCAM MTJ gives guaranteed disturbance-free read since the net torque acting during sensing always acts to retain the previous state and increases the tolerance to process variations due to its differential nature. A block diagram of a TCAM using the "PentaMTJ" cell is shown in Figure 6.4. This architecture is made up of three parts: a pre-charge sense amplifier (PCSA) for sensing the difference between two states of resistance, a logic circuit, and a TCAM writing cell with reference resistance.

Using the PentaMTJ cell in the TCAM instead of a conventional MTJ provides a cell that is less affected by process variation. The pair of PentaMTJ cells increases the sense margin and robustness. This TCAM has low power consumption compared with existing MTJ-based CAMs since it uses a pre-charge sense amplifier (PCSA) for sensing. It is immune to masking error since the masking is done by putting the transistor in the OFF state rather than by storing the same data in the two MTJs.

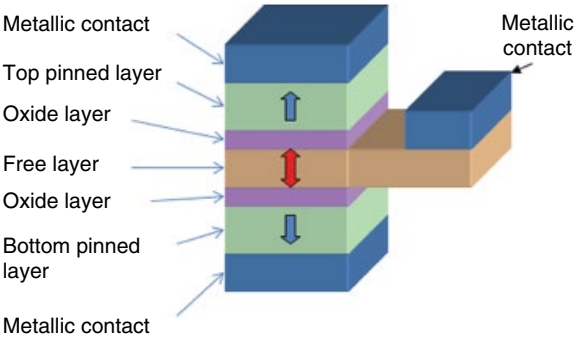


Figure 6.3 Perpendicular magnetic anisotropy (PMA)-based five-layer TCAM MTJ with two pinned layers and one free layer. Based on M. Gupta and M. Hasan (Aligarh Muslim University), *IEEE Trans. on Magnetics*, February 2015 [4].

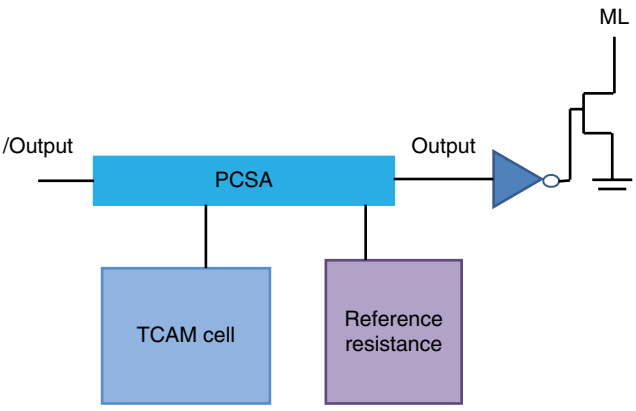


Figure 6.4 Block diagram of TCAM using PentaMTJ cell, a pre-charge sense amplifier (PCSA) for sensing the difference between two states of resistance, a reference resistance, and a logic circuit. Based on M. Gupta and M. Hasan (Aligarh Muslim University), *IEEE Trans. on Magnetics*, February 2015 [4].

A spin memory device that could be used for matching and self-reference function in a non-volatile content addressable memory (CAM) was discussed by Avalanche Technology in November of 2015 [5]. This spintronic memory used a spin–torque interaction combined with conventional spin–torque generated by magnetization polarization. The nvCAM macro was used along with a DRAM macro on a chip for data intensive computing applications such as data mining, search engines, scientific computing, and video processing. A block schematic of the spin CAM chip architecture is shown in Figure 6.5. This spin memory device had matching and self-reference functionality for use in content addressable applications. It included a spin content addressable memory (CAM) structure consisting of a DRAM cell with an embedded dual spin hall MTJ CAM device. This study investigated how to implement a spintronic memory computing device for parallel match-in-space and content addressable applications using SOT MRAM. A spin orbit CAM device and circuits were proposed

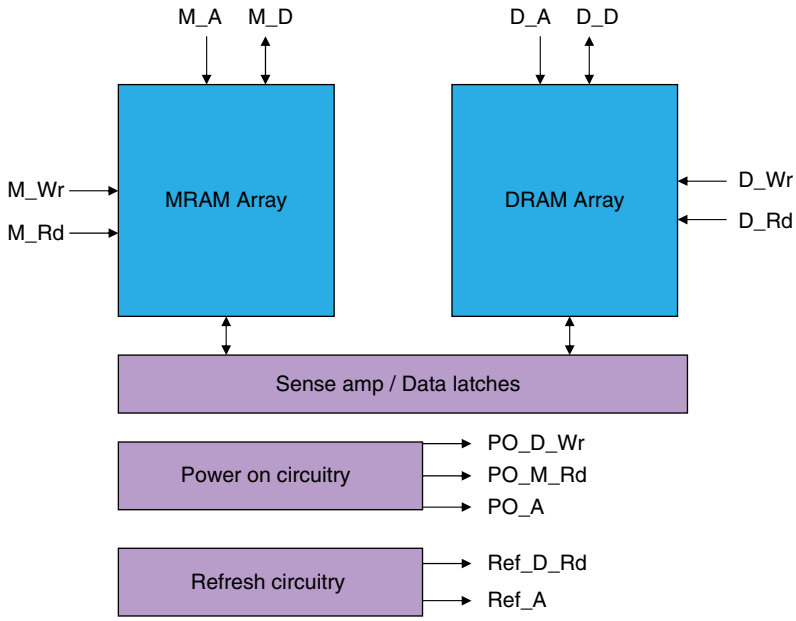


Figure 6.5 Block schematic of spin CAM architecture of a DRAM cell with embedded MRAM. Based on X. Wang *et al.* (Avalanche Technology), *IEEE Trans. on Magnetics*, November 2015 [5].

that are capable of multibit match-in-space functionality and single bit nondestructive self-reference functionality.

A low power, compact, and massive parallel comparison is provided by this spin-orbit memory device. The power consumption could be significantly reduced due to reduced leakage current over that of an SRAM CAM and the massive parallel array could increase system operating speed significantly.

The spin CAM architecture proposed combined DRAM with embedded MRAM. In this architecture of DRAM with embedded MRAM, the DRAM is used to reduce the cell size, while the MRAM is used to store bit content and to refresh the DRAM. During device operation, there is no need to explicitly read on-chip DRAM. DRAM external refresh will not interrupt normal operations. A circuit with an MRAM refreshing DRAM is shown in Figure 6.6.

A multibit parallel match-in-space array circuit structure using a multibit marching and single bit self-reference device was also discussed. This circuit used a spin hall metal layer. It had two writing current paths, the first to switch the free layer magnetization using spin torque, while the second current path switched the comparison reference layer magnetization through the spin torque generated by spin–orbit interactions. Two separated write paths are able to independently switch the magnetization of either the free layer or the comparison reference layer.

The main benefits of the spin–orbit memory device are low power, compact area, and massive parallel comparison. The 1T1R spin–orbit CAM structure reduces the cell area and the massive parallel comparison array increases the system operating speed.

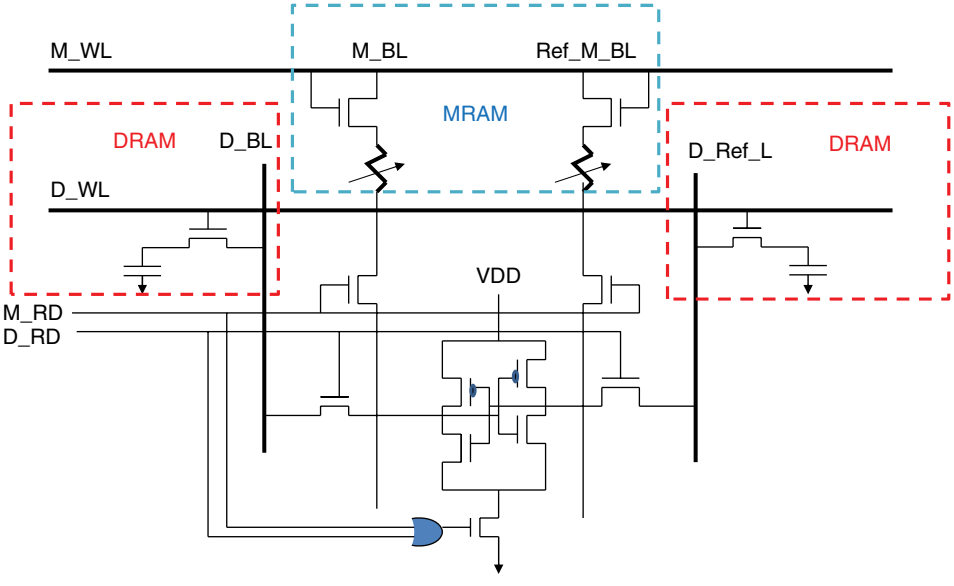


Figure 6.6 Schematic CAM circuit showing an MRAM refreshing DRAM circuit. Based on X. Wang *et al.* (Avalanche Technology), *IEEE Trans. on Magnetics*, November 2015 [5].

6.2.3 CAMs Using Other Emerging Memories

The TCAM is a special type of memory that can compare input search data with stored data and retrieve the matched data. In August of 2014, the University of Pittsburgh discussed TCAMs made of various eNVM technologies [6]. These technologies included tunneling junction (MTJ), resistive RAMs, and ferroelectric FET. It was found that all three types of emerging memories could achieve nearly zero standby power.

6.3 Components of Large Search Engines and Artificial Neural Networks

6.3.1 Using RRAMs in Look-Up Tables in Large Search Engines

Look-up tables (LUT) are necessary in large search engines. In November of 2014, ITRI, National Taipei University, NTHU, and Min Shin University discussed a nonvolatile look-up table that used RRAM for the reconfigurable logic [7]. The circuit used RRAM cells with normally off and instant-on functions for suppressing standby current. The proposed RRAM-based two-input nonvolatile look-up table (nvLUT) circuit decreased the number of transistors by 79% and the area of the nvLUT by 90.4% compared with the area of a SRAM-MRAM hybrid LUT. The areas of two-input RRAM nvLUTs were found to be smaller than MRAM-based two-input nvLUTs. Due to low current switching and high resistance ratios, the RRAM-based nvLUT had 24% less power consumption than SRAM-MRAM hybrid LUTs. Functionality of the adder of the three-input RRAM nvLUT was confirmed using an HfO_x-based RRAM made in 180 nm CMOS logic with a 900 ps delay time.

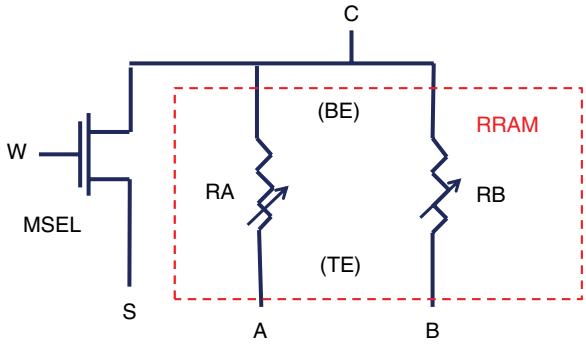


Figure 6.7 Circuit schematic of the 1T2R memory unit with 1T and 2RRAMs which act as a complementary resistive switching (CRS) device. Based on W.-P. Lin (ITRI, National Taipei University, NTHU, Min Shin University), IEEE A-SSCC, November 2014 [7].

A TiN/TiO_x/HfO_x/TiN stack was used for the RRAM. SET operation, which is the low resistance state, used a positive 1.5 V and RESET, which is the high resistance state, used a negative one (−1.1 V). Write time of the HfO_x-based RRAM was as fast as 5 ns. The architecture of the proposed 1T2R memory unit is shown in Figure 6.7. It consisted of an NMOS select transistor and two HfO_x RRAMs. The RRAM devices RA and RB work together as a complementary resistive switching (CRS) device.

A resistive RAM crossbar array can be used for matrix vector multiplication. In 2015, Tsinghua University and Arizona State University discussed using an HfO_x RRAM crossbar array for matrix-vector multiplication [8]. If a 2D filament model of the HfO_x-based RRAM is considered, the conductance is exponentially dependent on the tunneling gap distance d, as illustrated in Figure 6.8.

When a large voltage is applied on the electrodes, d will change due to the electric field and temperature-enhanced oxygen ion migration. The resistivity of the RRAM will switch between the highest resistance state R_{off} and the lowest resistance state R_{on}. In theory an RRAM can achieve any resistance between R_{on} and R_{off}. This study discussed choosing the resistivity of the RRAM device.

The I–V relationship can be expressed as

$$I = I_0 \exp\left(-\frac{d}{d_0}\right)\sinh\left(\frac{V}{V_0}\right).$$

where d is the average tunnel gap distance and I_0 and V_0 are determined experimentally.

The RRAM crossbar array can perform analog matrix vector multiplication. The relationship between the input voltage and the output voltage vector can be expressed as

$$\begin{bmatrix} V_{0,1} \\ \vdots \\ V_{0,M} \end{bmatrix} = \begin{bmatrix} c_{1,1} & \cdots & c_{1,N} \\ \vdots & \ddots & \vdots \\ c_{M,1} & \cdots & c_{M,N} \end{bmatrix} \begin{bmatrix} V_{i,1} \\ \vdots \\ V_{i,N} \end{bmatrix}$$

The matrix parameter $C_{k,j}$ can be represented by the conductivity of the RRAM device and the load resistor as

$$c_{k,j} = \frac{g_{k,j}}{g_s + \sum_{l=1}^{N} g_{k,l}}$$

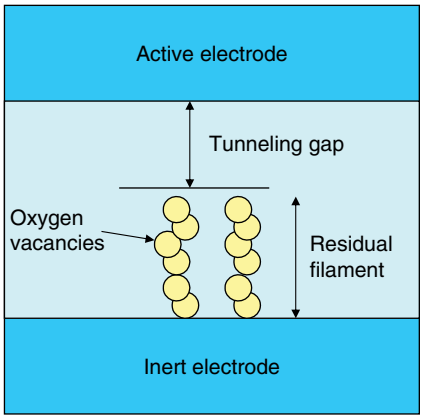


Figure 6.8 Schematic cross-section of 2D filament model of HfO_x-based RRAM. Based on P. Gu *et al.* (Tsinghua University, Arizona State University), ASP-DAC, January 2015 [8].

For a more linear *I–V* relationship for the RRAM device, both the RRAM resistance state, represented by the tunneling gap distance *d* and the applied voltage *V*, should be confined.

Hybrid RRAM and triple level cell (TLC) NAND Flash can be used for data storage in RAID servers in large server farms. In February of 2014, Chuo University and the University of Tokyo discussed hybrid storage of RRAM and three-level cell (TLC) NAND Flash with RAID-5/6 [9]. RAID-5/6 is commonly used in Cloud storage and data warehouses due to having a lower parity overhead than RAID-1. The proposed architecture gave significant improvements in reliability and performance of RRAM and TLC NAND. Data to RRAM was encoded and then written to RRAM.

RRAM crossbar arrays can be used in large memory systems as a potential scaling replacement for DRAM. In February of 2015, the Oakridge National Labs discussed RRAM for its superior characteristics such as high density, low write energy, and high endurance [10]. This study did a comprehensive review of RRAM memory systems using stacked crossbar architecture with peripheral circuits below the multiple layers of RRAM cells. The crossbar architecture was found to have constraints on operating voltages, write latency and array size.

Architectural optimizations were obtained from the relationship between the RRAM write voltage and switching latency. Trade-offs involving voltage drop, write latency, and data pattern were discussed and microarchitectural enhancements were analyzed, including double-sided ground biasing and multiphase RESET operations to improve write performance. A simple compression-based data encoding method was used to reduce the latency. Since the compressibility of a block varies based on its content, write latency is not uniform across blocks. To reduce the impact of slow writes on performance, a new scheduling policy was explored that made writing decisions based on latency and activity of a bank. Experimental results showed that the stacked crossbar architecture improved the performance of a system with RRAM main memory by about 44% over a baseline DRAM system.

6.3.2 Using STT MRAM in Large Artificial Neural Networks

STT-MRAM acting as a stochastic memristive device was discussed in 2014 by the University of Paris-Sud, CNRS, CEA, and LIST [11]. The device was used to implement a "synaptic" function. Basic concepts of STT-MRAM cell behavior were considered for

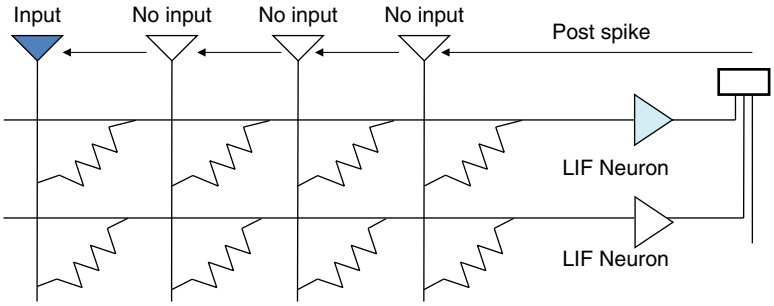


Figure 6.9 Schematic circuit diagram of crossbar architecture for learning system using STT-MRAM synapses and leaky integrate and fire (LIF) neurons. Based on A.F. Vincent *et al.* (University of Paris-Sud, CNRS, CEA, LIST), ISCAS, June 2014 [11].

the potential of implementing synapses capable of learning. A simulation of the synapse was used on an application of "car counting" to show the potential of the technology. A Monte Carlo simulation was used to show the robustness of the technology to device variations.

The ability of a positive current to switch the MTJ in one direction and a negative current to switch it in the opposite direction was taken as similar to a binary bipolar memristor. Using usual mathematical descriptions of an STT-MTJ MRAM, an expression similar to a memristor was derived. An STT-MRAM synapse and a leaky integrate and fire (LIF) neuron learning system is shown in Figure 6.9.

System level simulation of stochastic MTJ synapses were shown using a method previously proposed for conductive bridge RAM. In this system, input neurons present short low voltage spikes that generate current through the crossbar. The conductance of the MTJ acted as a synaptic weight converting voltage into current. The current was integrated by output neurons until one of them generated a spike, which then inhibited the other output neurons.

In April of 2015, the University of Paris-Sud, CEA, and Beihang University discussed further on an STT MRAM used as a stochastic memristive synapse [12]. The STT-MTJ MRAM was used for implementing learning capable synapses. Low, intermediate and high current programming regimes were identified and compared.

An example of the function for system-level simulations for a vehicle counting application showed the potential of using the technology for learning systems. Monte Carlo simulations were done to show robustness to device variations. The simulations also compared system operation with different programming regimes of the STT-MTJs. The intermediate current regime was compared to the high and low current regimes and was shown to permit minimization of energy consumption while retaining robustness to device variations. The conclusion was that applications using STT-MTJ as synapses in robust, low power, cognitive systems were possible.

Artificial neural networks (ANNs) using spin synapses and neurons were discussed in 2016 by Beihang University, the University of Paris-Sud, and Tsinghua University [13]. Spintronic devices using MTJ as a binary device have been explored previously for neuromorphic computing with low power dissipation. This study used a compound spintronic device consisting of multiple vertically stacked MTJs, which behaved as a synaptic device.

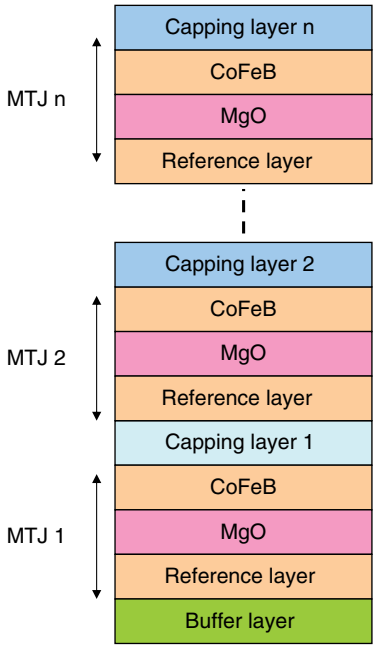


Figure 6.10 Schematic of the compound spintronic device showing "*n*" multilevel states. Based on D. Zhang *et al.* (Beihang University, University of Paris-Sud, Tsinghua University), *IEEE Trans. on Biomedical Circuits and Systems*, August 2016 [13].

It was shown that the compound spintronic device could achieve designable and stable multiple resistance states for use as synapses by doing interfacial and materials engineering of its components. A compound spintronic neuron circuit was developed based on a spintronic device that enabled a multistep transfer function. An artificial neural network was constructed using both the spin synaptic and spin neuron circuits. A test of the NMIST database for handwritten digital recognition was done to evaluate the performance of the ANN. The MNIST database was constructed from NIST databases that contained binary images of handwritten digits.

A schematic of the proposed compound spintronic device is shown in Figure 6.10. The capping layer materials and thickness and MgO tunneling barrier thickness can be manipulated separately to get designed multilevel states. Ideally, N vertically stacked MTJs can exhibit $2N$ states. It was demonstrated that four distinct states could be obtained; then it was inferred that by vertical stacking more MTJs more resistance states could be achieved. Ideally, $2N$ resistance states can be achieved by stacking N MTJs.

A neuron is the basic computational element in an ANN. It generates an output signal depending on the magnitude of the summation of its synaptic weighted input signals. The neuron's output signal is then transmitted by way of the axon as the input of its fan-out neurons. A single MTJ device has been shown to be able to emulate the step transfer function, which corresponds to the switching between its two stable states by a resultant synaptic current in ANNs. For a nonstep transfer function, a compound spintronic neuron that also uses multiple MTJs can function as a single neuron. Figure 6.11 shows a proposed compound spintronic neuron intended to implement a multiple step transfer function.

The performance of the artificial neural network was studied by performing system-level simulations with off-line training on the NMIST database for handwritten digital

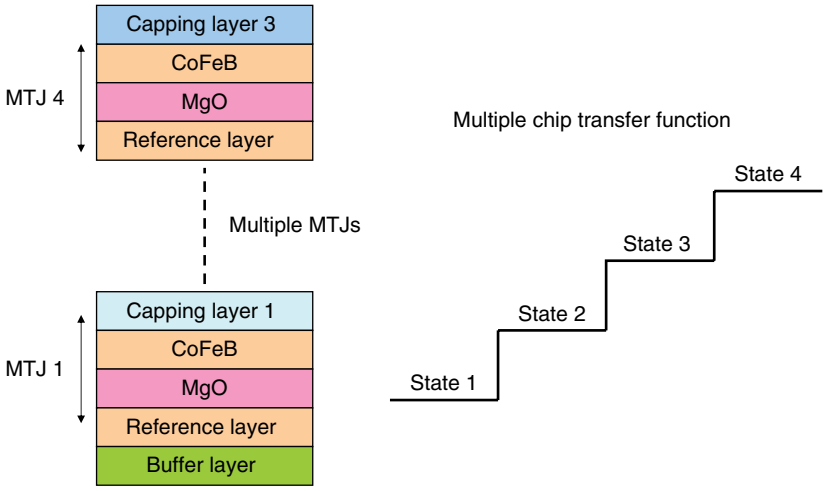


Figure 6.11 Illustration of a compound spintronic neuron using multiple MTJs to implement a multiple step transfer function. Based on D. Zhang *et al.* (Beihang University, University of Paris-Sud, Tsinghua University), *IEEE Trans. on Biomedical Circuits and Systems*, August 2016 [13].

recognition. It was shown that the proposed compound spintronic synapse (CSS) and compound spintronic neuron (CSN) could be used to build a next-generation neural computation platform.

6.4 Memory Issues in Deep Learning Systems

6.4.1 Issues with Partitioning SRAM and RRAM Synaptic Arrays

Memory arrays are being used for on-chip acceleration of weighted sum and update in neuromorphic machine learning. Since the learning algorithms operate on a large weight matrix size, an efficient mapping of the matrix may require partitioning it into a number of subarrays.

In May of 2016, Arizona State University discussed a circuit-level macrosimulator intended to partition a 512×512 weight matrix into RAM-based accelerators using both SRAM and RRAM [14]. For an SRAM, increased partitioning and finer granularity of the array leads to decreased read and write latency and decreased dynamic read and write energy. This is due to the increased parallelism obtained at the cost of larger area and leakage power. For an RRAM accelerator, however, increased partitioning increased the read latency and energy only to a certain partition point due to the overhead of multiple intermediate stages of adders and registers.

In the neural network, the weight is represented by a synapse element. The connections through synapses between neurons form a weight matrix. The key operations of learning algorithms are the weighted sum and weight update. Hardware acceleration of neurocomputing consists of efficiently mapping the learning algorithms on-chip to improve computing speed and reduce power consumption. Early large neuromorphic

hardware platforms have been based on CMOS SRAM-based synaptic arrays. An SRAM accelerator is illustrated in Figure 6.12.

6 T SRAM cells are grouped as one weight element. The weighted sum (read) and weighted update (write) operations are performed on a row-by-row basis. The adder and resister are added to help accumulate the partial weighted sum of each row.

More recently, RRAM crossbar arrays have been used for weighted sum and update in learning algorithms. Since the RRAM can have multiple levels per bit it can be used for storing analog weights at a high density. Scaling up the array size is an issue and requires efficient partitioning.

An RRAM accelerator with a pseudo-crossbar array core (with the crossbar array rotated 90 degrees) is shown in Figure 6.13.

The weight update (write) is through the BL and SL switch matrix with one row activated. Weighted sum operation (read) is through the BL switch matrix (input vector) with all rows activated. If a limited number of read circuits are shared, the weighted sum operation may need to be time-multiplexed.

It was suggested that the large array can be partitioned into $N \times N$ small arrays in a hierarchical manner to speed up the write operation since the weight elements in different subarrays can be updated in parallel. For a read, the vector and matrix were distributed into the subarray partitions and computed in parallel. The results from all subarrays were collected and summed. Multistage adders and registers were used to get the final weighted sum. The resulting read and write latency per operation for the SRAM and RRAM accelerators with different numbers of partitions in a 256 kb (512 × 512) array are shown in Figure 6.14.

The improvement in read latency of the RRAM array is limited due to added adders and registers, which are included in the time-multiplexing. The read and write energy consumption per operation for the SRAM and RRAM accelerators with different numbers of partitions in a 256 kb (512 × 512) array are shown in the figure. While the SRAM continued to improve in read energy consumption with an increased number of

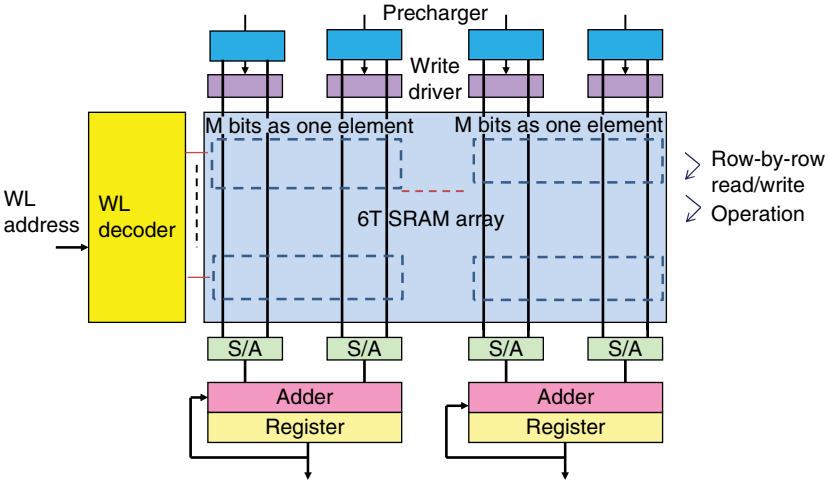


Figure 6.12 Block diagram of SRAM accelerator using CMOS SRAM-based synaptic arrays. Based on P. Chen and S. Yu (Arizona State University), ISCAS, May 2016 [14].

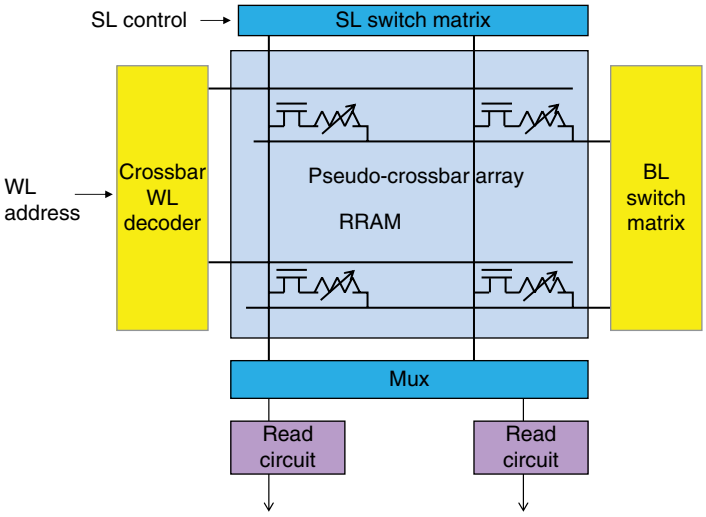


Figure 6.13 RRAM accelerator with pseudo-crossbar array core. Based on P. Chen and S. Yu (Arizona State University), ISCAS, May 2016 [14].

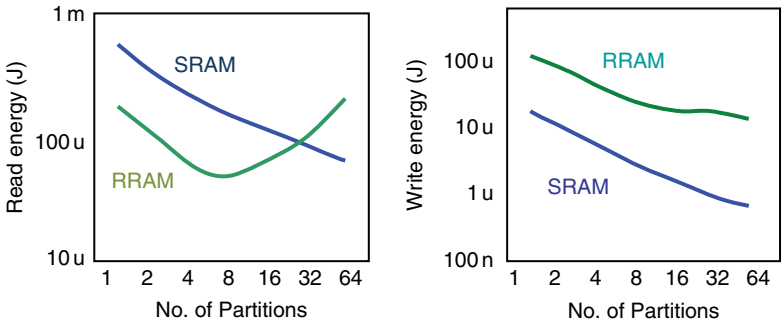


Figure 6.14 Illustration of read and write energy per operation for SRAM and RRAM. Based on P. Chen and S. Yu (Arizona State University), ISCAS, May 2016 [14].

partitions, the RRAM array showed a minimum read energy consumption at about 8.8 partitions, which corresponds to an array size of 64×64. Both arrays continued to improve in write energy with increasing partitioning. The leakage power consumption of the RRAM accelerator was less than that of the SRAM at a small number of partitions, but increased with partitioning to nearly that of the SRAM accelerator at the (8×8) array size corresponding to 64 partitions.

6.4.2 Issues of RRAM Variability for Extreme Learning Machine Architectures

The intrinsic variability of filamentary resistive memory for extreme learning machine architectures was discussed in 2015 by the Indian Institute of Technology of New Delhi

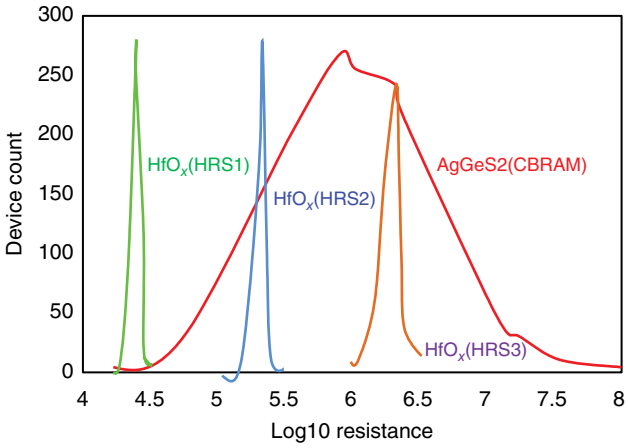


Figure 6.15 Illustration of extracted HRS/R_{off} lognormal distributions for filamentary RRAM devices. Based on M. Suri and V. Parmar (IIT New Delhi), *IEEE Trans. on Nanotechnology*, November 2015 [15].

[15]. This paper showed the unavoidable device variability of RRAMs for designing efficient low power low footprint extreme learning machine (ELM) architectures. The system discussed used the uncontrollable off-state resistance (Roff/HRS) spreads for several RRAM types, as illustrated in Figure 6.15.

The RRAM-ELM architecture proposed was demonstrated in two applications, a diabetes diagnosis test and a SinC curve-fitting regression. The ELM discussed was a single-layer feed forward neural net consisting of: hidden layer synapses with randomly assigned weights, a hidden neuron layer with an infinitely differentiable activation function, and an output layer with synaptic weights determined by a learning rule. The training block architecture proposed is shown in Figure 6.16. It was shown that HRS distributions can be used for input-stage random synaptic weights and random neuron biases.

6.4.3 Issues with RRAM Memories in Restricted Boltzman Machines

Deep learning can be implemented using a restricted Boltzman machine (RBM), which consists of a generative stochastic artificial neural network capable of learning a probability distribution over its set of inputs. RBMs are typically used as building blocks for deep belief networks. In October of 2015, IEF-CNRS and IEMN-CNRS in France discussed the development of a hybrid RRAM-CMOS RBM architecture in which HfO_2-based RRAM devices are used to implement synapses, internal neuron state storage, and stochastic neuron activation function [16]. The RBM architecture was simulated for classification and reconstruction of handwritten digits on a reduced MNIST data set of 6000 images. The required size of the RRAM matrix in the simulated application was on the order of 0.4 Mb. Peak classification accuracy was 92% and an average accuracy of 89% was obtained over 100 training periods. The average number of RRAM switching events was 14 million/period. The I–V characteristics of the OxRAM (oxide-based RRAM) device are illustrated in Figure 6.17.

This device was a crosspoint metal–insulator–metal (MIM) junction with a 10 nm thick HfO_x oxide sandwiched between the Ti/Pt bottom electrode and the Ti/TiN top electrode. The cycle-to-cycle ON/OFF state resistance variability was used for its

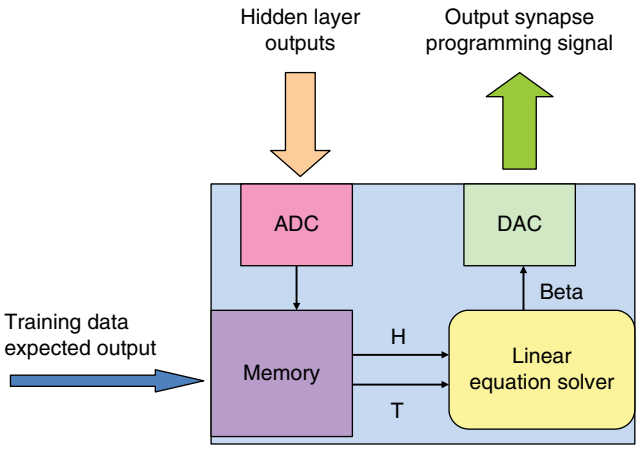
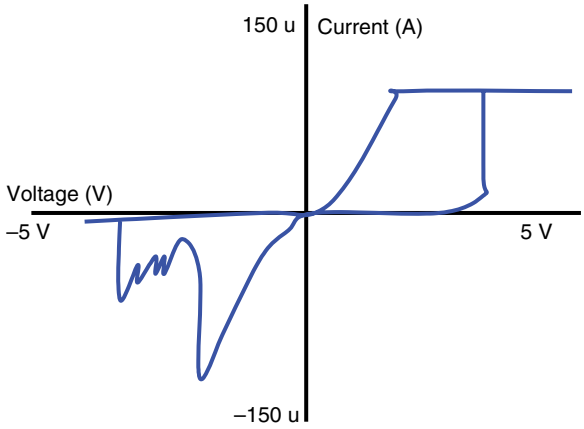


Figure 6.16 Illustration of training block in RRAM extreme learning machine (ELM) architecture, where H is the hidden layer neuron output and T is the expected output. Both are stored inside the training block. Based on M. Suri and V. Parmar (IIT New Delhi), *IEEE Trans. on Nanotechnology*, November 2015 [15] (permission of IEEE).

Figure 6.17 Illustration of *I–V* characteristics of an oxide-based RRAM device. Based on M. Suri *et al,* (IIT, IEF-CNRS, IEMN-CNRS), 15th NVMTS, October 2015 [16].



intrinsic effect for implementing stochasticity for the RBM architecture. The variability spread increases with the device active area dimension. This effect is used for implementing the stochasticity necessary for an RBM architecture. By engineering the RRAM device dimensions or by tuning read voltage the RBM stochasticity can be controlled indirectly.

In December of 2016, Stanford University, University of California Irvine, University of California San Diego, IBM, and Macronix discussed training a restricted RRAM-based Boltzmann machine for use in unsupervised learning in deep networks as a probabilistic graphical model [17]. Currently large scale deep learning machines require thousands of processors, large amounts of memory, and gigajoules of energy. This paper reports using resistive memories for implementing and training a restricted Boltzmann machine (RBM). An RBM is a generative stochastic artificial neural network (ANN)

that can learn a probability distribution over its set of inputs. The RBM implemented was a generative probabilistic graphical model that was a component for unsupervised learning in deep networks.

A 45 synapse RBM was made with 90 PCM elements trained with a variant for the contrastive divergence algorithm implementing Hebbian and anti-Hebbian weight updates. The implementation shows a twofold to tenfold reduction in error rate in a missing pixel pattern completion task trained over 30 epochs, compared with an untrained case. The measured programming energy was 6.1 nJ per epoch using the phase change memory (PCM), which is a factor of 150 times lower than conventional processor-memory systems.

Deep learning involves extracting complex and useful structures from within high-dimensional data without requiring significant amounts of manual engineering. It is particularly suitable for tasks such as image recognition, speech recognition, natural language understanding, and predicting the effects of mutations in DNA, among others.

Training large-scale deep networks of about 10^9 synapses (the human brain has 10^{15} synapses) in conventional von Neumann hardware consumes more than 10 GJ of energy with an important component of this energy being the physical separation of processing and memory. It is estimated that 40% of the energy consumed in general purpose computers is due to the off-chip memory hierarchy. GPUs do not appear to solve the problem since up to 50% of dynamic power and 30% of overall power are consumed by off-chip memory. On-chip SRAM also does not solve the problem since it is area inefficient and cannot scale with system size.

To extract useful information for large amounts of data requires efficient data mining and deep learning algorithms with energy efficiency being crucial. Scaling these systems up with energy efficiency means developing new learning algorithms and hardware architectures that use fine grained on-chip integration of memory with processing. Since the number of synapses in a neural network significantly exceeds the number of neurons, it is essential to solve practical problems of power, device density, and wiring of synapses. Conventional synaptic weights in conventional processors and neuromorphic processors are currently implemented in SRAM or DRAM. High density DRAM processing has separated significantly from conventional CMOS processing requiring high capacity DRAMs to be separate chips, even if connected by TSV chip stacking, which limits via density and increases in power consumption and limits bandwidth. SRAM occupies too much area. Resistive RAMs tend to have analog characteristics that can be used in electronic synapses and many have good scalability. A resistive RAM can implement variations of learning rules in a single chip.

Supervised learning refers to finding the right model with labeled data while unsupervised learning refers to fitting the right model to the underlying probability distribution of data while discovering useful features. Unsupervised learning had been a significant factor in the recently renewed interest in deep learning, which is crucial to using huge amounts of unlabeled data and is similar to the unsupervised nature of learning in humans.

A restricted Boltzmann machine (RBM) is a two-layer probabilistic generative model that efficiently represents the underlying probability distribution of data in a distributive manner, where one layer consists of visible neurons that are associated with observations and another layer consists of hidden neurons. Unsupervised learning using RBM is an important element of deep neural networks for successful generalization where huge

amounts of labeled data are not available. A contrastive divergence (CD) algorithm is a common technique for training RBM. A demonstration was given of CD learning in an RBM with electronic synapses using phase change memory. The RBM was able to learn 3×3 images and retrieve a missing pixel with more than 80% probability of success when five patterns are stored. With fewer patterns the probability of success goes up. Device engineering improved the success.

A proof of concept implementation of a probabilistic graphical model of an RBM was done using 45 synapses implemented with 90 PCM elements trained with CD learning. Synaptic operations consumed 6.1 nJ/epoch compared with an estimated 910 nJ/epoch for a state-of-the-art conventional processor. Fluctuations in the learning progress were found when conductance change control was not sufficient to overcome cycle-to-cycle variations. An opportunity is shown for using emerging NVM devices for probabilistic computing.

6.4.4 Large Neural Networks Using Memory Synapses

In April of 2015, IBM discussed implementing a large-scale artificial neural network (ANN) using two PCM devices to encode the weight of each of 164 885 synapses and studied the associated reliability issues [18]. Artificial neural networks were used to solve specific computational problems in pattern recognition, natural language processing, and similar problems where the computation is parallel and redundant due to large degrees of freedom and has some tolerance to faulty devices. These ANNs were densely connected arrays of simple "neuron" processors in which the computation data was coded by the strength of the interneuron synapse connections.

The PCM conductances were programmed using a crossbar-compatible pulse method and the network was trained to recognize a 5000 example subset of the MNIST handwritten digit database. An 82.2% accuracy was achieved during training along with an 82.9% generalization accuracy on unseen test examples. A simulation of the network performance was developed that includes a statistical model of the PCM response. This permitted estimation of the tolerance of the network to device variation, defects, and conductance response.

Neuromorphic chips have been shown that use CMOS hardware to implement pretrained neural networks. The synaptic weights for each 256 neuron core are stored digitally in banks of SRAM after off-line training in order to establish the values of these weights. NVM devices can potentially be used to represent synaptic weight to provide higher connectivity among a large number of neurons due to their small size and high density. Crossbar NVM arrays can provide on-line learning by firing pulses at each NVM element to update its conductance and therefore affect the synaptic weight. An algorithm can cause each neuron circuit to decide what pulses to fire based on its internal state and all NVM devices in a network layer could be updated in parallel if they are accessed through a nonlinear selector device such as a mixed ionic electronic conductor (MIEC) access device. It has been shown that RRAM or PCM can be programmed using a spike timing-dependent plasticity (STDP) algorithm based on a learning method found in biological neurons, but these networks have been small.

This study described a three-layer perceptron with back-propagation learning made of 916 neurons and 164 885 synaptic connections. The synaptic weights take on both positive and negative values. The weights are represented as conductance differences

between a pair of PCM devices. The network was trained to recognize handwritten digits from the NMIST database. A detailed simulation was done of this network. A three-layer perception network is shown in Figure 6.18, in which the neurons activate each other through layers of synaptic connections with programmable weights.

A layer of connections are implemented using crossbar arrays of NVM elements with selector devices. Synaptic weights are represented by the conductance difference between pairs of NVM and the outputs of one layer of neurons is summed in parallel into the inputs of the next layer.

A three-layer perceptron network with neurons activate each other through layers of synaptic connections with programmable weights. It was shown that a three-layer perceptron ANN, with 164 885 synaptic connections implemented with pairs of PCM memory devices to encode synaptic weights as a difference in device conductance, could be successfully trained to recognize handwritten digits with 82% accuracy.

A 2.9 TOPS/W deep convolutional neural network (DCNN) SoC in 28 nm fully depleted SOI (FD-SOI) was discussed in 2017 by STMicroelectronics. It was intended for an intelligent embedded system [19]. An energy-efficient DCNN processor is discussed with the following attributes: an energy-efficient set of DCNN hardware convolutional accelerators supporting kernel compression, an on-chip reconfigurable data transfer fabric to improve data reuse and reduce on-and-off chip memory traffic, a power efficient array of digital signal processors (DSPs) to support complete real-world

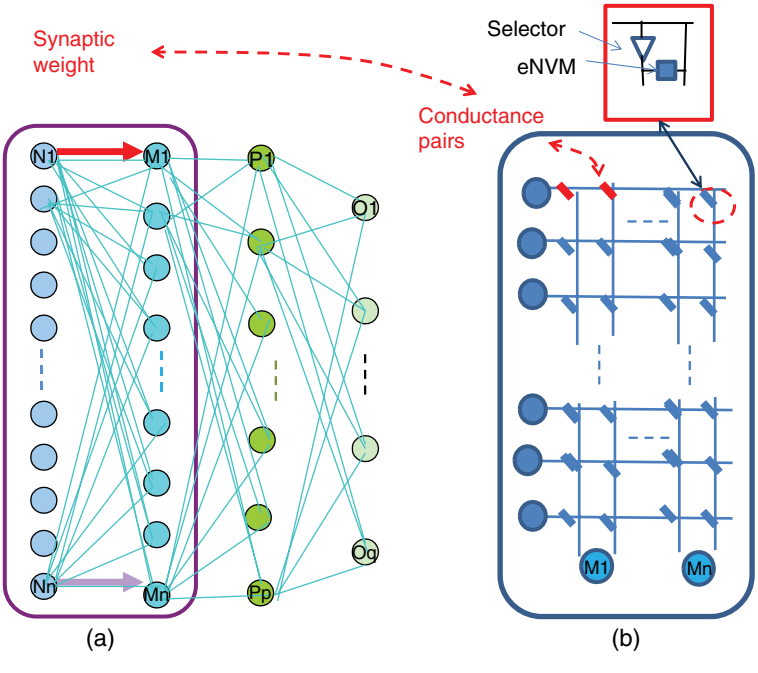


Figure 6.18 Illustration of: (a) three-layer perceptron network, with neurons represented as circles and lines as synaptic weights, and (b) synaptic weights, represented as conductance differences between pairs of nonvolatile memory (NVM) devices. Based on R.M. Shelby *et al.* (IBM), IRPS, April 2015 [18] (permission of IEEE).

computer vision applications, an ARM-based host subsystem with peripherals, a range of high speed IO interfaces for imaging and other types of sensors, and a chip-to-chip multilink to pair multiple devices together.

The SoC test chip was a system demonstrator integrating an ARM Cortex MCU with 128 KB of memory, peripherals, eight DSP clusters, a reconfigurable dataflow accelerator fabric connecting high speed camera interfaces with sensor processing pipelines, and other peripherals including eight convolutional accelerators. The chip has four 1 MB SRAM banks, a dedicated bus port, and fine-grained power gating. Four chips can be connected using high speed serial links to support larger networks without sacrificing throughput. The SoC uses a configurable accelerator framework based on unidirectional links that transport datastreams using a configurable switch. Intermediate results can be stored in on-chip memory. Thirty-six 16b MAC units perform up to 36 MAC operations per clock cycle. A convolutional accelerator configuration per DCNN layer is defined manually.

Each 32b DSP provides specific instructions. The dual-MAC operation loop executes in a single cycle while an independent 2D DMA channel permits overlap of data transfers. The DSPs represent a small fraction of the total DCNN computation but are thought to be more capable of future algorithmic evolution. The prototype chip was made in 28 nm fully depleted silicon on insulator (FD-SOI) technology with an embedded single supply SRAM with low power features and adaptive circuitry to support a voltage range from 1.1 to 0.575 V.

The low power features included that use globally asynchronous and locally synchronous clocking reduced the dynamic power and skew sensitivity due to on-chip variations at lower voltages. Fine-grained power gating and multiple sleep modes for memories decreased the dynamic power and the leakage power. Operation was at 200 MHz with a 0.575 V power supply at 25 °C. Average power consumption was 41 mW with eight chained convolution accelerators (CA) representing a peak efficiency of 2.9 terra-operations-per-second per watt (TOPS/W).

The supply voltage ranged from 0.575 to 1.1 V with a 1.8 V I/O. On-chip RAM was 4 × 1 MB, 8 × 192 KB, and 128 KB. Peak convolution accelerators (CAs) performance was 676 giga-operations-per-second (GOPS). An illustration of the SoC top-level block diagram is shown in Figure 6.19. The device was proven to be effective for advanced real-world, power constrained, embedded applications such as intelligent IoT devices and sensors.

6.5 Deep Neural Nets for IoT

6.5.1 Types of Deep Neural Nets for IoT

An 8.1 TOPS/W reconfigurable convolutional neural network (CNN)–recurrent neural network (RNN) was discussed in 2017 by KAIST [20]. CNNs are used to support vision recognition and processing and RNNs are used to recognize time-varying entities and support generative models. Time-varying visual entities, such as action and gesture can be recognized by combining CNNs and RNNs. The computational requirements of CNNs are different from those of RNNs.

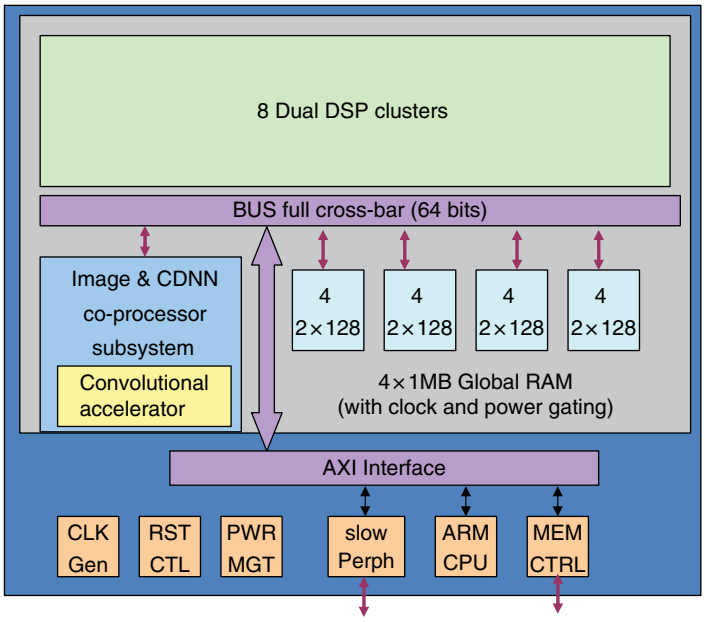


Figure 6.19 Top-level block diagram of a deep convolutional neural network SoC test chip. Based on G. Desoli *et al.* (STMicro), ISSCC, February 2017 [19].

For a CNN convolution layer (CL), the same kernel is used multiple times, as shown in Figure 6.20. Convolutional layers require a massive amount of computation for a relatively small number of filter weights.

For a fully connected layer (FCL) in a CNN, all connections have different weights, as shown in Figure 6.21 (a). In a recurrent neural net (RNN), all connections also have different weights, as shown in Figure 6.21 (b). FCLs and RNNs require a relatively small amount of computation with a huge number of filter weights.

This means that different processors are used for convolutional layers (CL) than for fully connected CNN layers and RNNs. When CLs are accelerated with FCL and RL dedicated SoCs they cannot exploit reusability. When FCLs and RLs are accelerated with SoCs specialized for CLs, they suffer from high memory transaction costs, low PE utilization and a mismatch of the computational patterns. Most studies have used different SoCs for the two types of layers. This work studies a combined CNN-RNN processor, which is highly reconfigurable with high energy efficiency.

This reconfigurable CNN-RNN processor has three main features: (1) a reconfigurable heterogeneous architecture with a CL processor and an FC-RL processor to support general purpose DNNs, (2) an LUT-based reconfigurable multiplier optimized for the dynamic fixed point with on-chip adaptation to exploit maximum efficiency from kernel reuse in the CP, and (3) a quantization table-based matrix multiplication to reduce off-chip memory access and remove duplicated multiplications. The architecture of the deep neural processing unit (DNPU) is shown in Figure 6.22. The part is made using 65 nm CMOS with a 16 mm^2 die area. Energy efficiency is 8.1 TOPS/W.

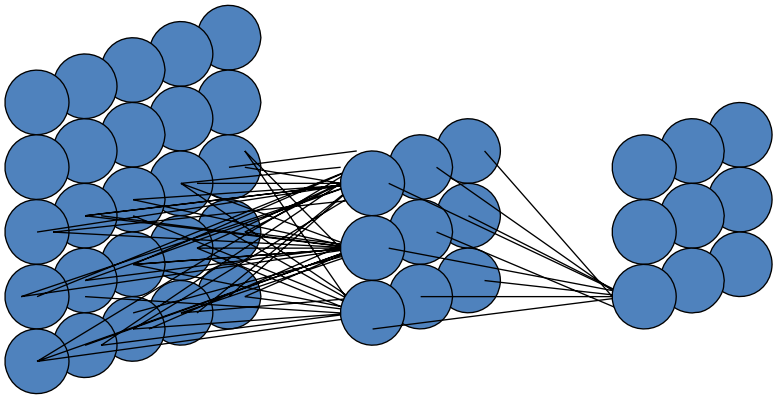


Figure 6.20 Illustration of a convolutional neural net (CNN) convolution layer. Based on D. Shin *et al.* (KAIST), ISSCC, February 2017 [20].

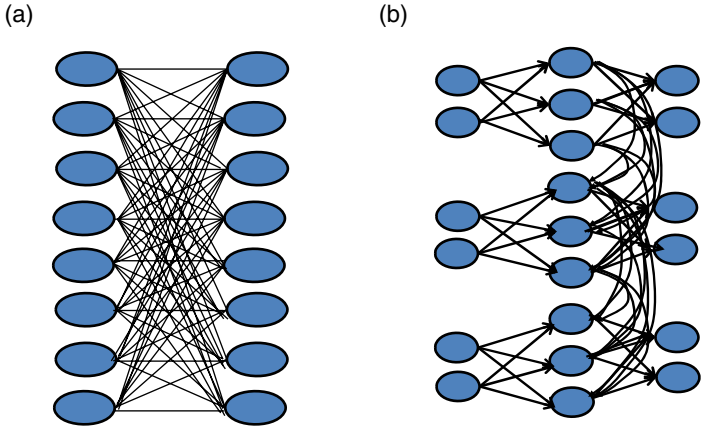


Figure 6.21 Illustration of convolutional neural ets: (a) CNN fully connected layers (FCL) and (b) recurrent neural nets (RNN). Based on D. Shin *et al.* (KAIST), ISSCC, February 2017 [20].

6.5.2 Deep Neural Nets for Noisy Data

In 2017, Harvard University discussed a 28 nm SoC with a 1.2 GHz sparse deep neural net (DNN) engine for IoT [21]. The SoC is intended for Internet of Things devices to enable the interpretation of the complex and noisy real world data resulting from sensor-rich systems. Energy efficiency is important. Programmable fully connected deep neural network accelerators have flexible support for general classification talks with high accuracy. Earlier machine learning accelerators have focused either on computer vision CNNs, which consume high power at around 280 mW, or spiking neural networks with low accuracy of about 84% on MNIST.

A 28 nm SoC with a programmable fully connected (FC)-DNN supported a range of general classification tasks with high accuracy. It offered hardware support for data

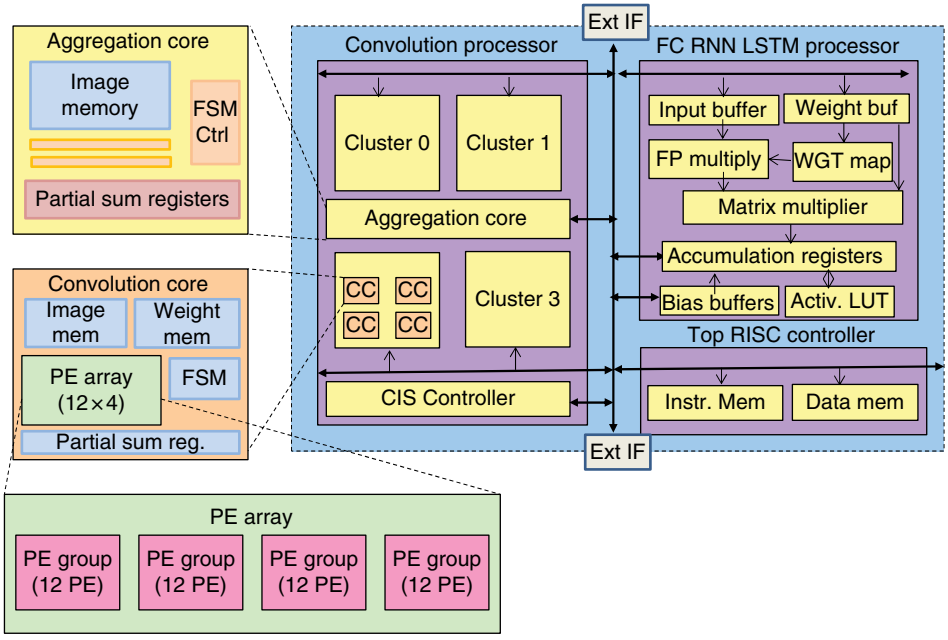


Figure 6.22 Block diagram of deep neural processing unit (DNPU). Based on D. Shin *et al.* (KAIST), ISSCC, February 2017 [20].

sparsity by eliminating unnecessary computations that reduced energy. It improved algorithmic error tolerance for weights and datapath computations. It improved circuit level timing violation tolerance and had timing violation detection to reduce energy using V_{dd} scaling. It had high classification accuracy of 98.36% on MNIST. The accelerator had maximum throughput at 1.2 GHz.

The SoC used an RISC cluster. The DNN engine connected through an asynchronous bridge. A four-way banked on-chip memory stored the weights for the DNN model up to 1 MB and had low latency access to the DNN engine. The DNN engine was a five-stage SIMD-type programmable sparse matrix vector machine that could process arbitrary DNNs. A sequencer dynamically scheduled operations for the DNN with up to 8 FC layers with up to 1024 nodes per layer. The eight-way MAC datapath processed eight concurrent neuron computations at a time with 8b or 16b weight precision. A small 512B SRAM buffer kept the list of active node indexes in the previous layer from which W-MEM addresses are generated.

For MNIST the average loads, operations and cycles were reduced by over 75%, which improved energy and throughput. For the memory, switching activity was reduced so bit-flips were reduced. A bit-masking technique was used to mask individual bit errors in the weight word. This permitted the accelerator to tolerate SRAM read timing violation rates at 98.36% accuracy. Timing violation tolerance was improved to 98.36% accuracy over the MNIST test set, which supported VDD reduction to 0.715 V. A system block diagram of the 28 nm SoC with DNN engine is shown in Figure 6.23. Overall energy was reduced by more than nine times using 8 bit weights.

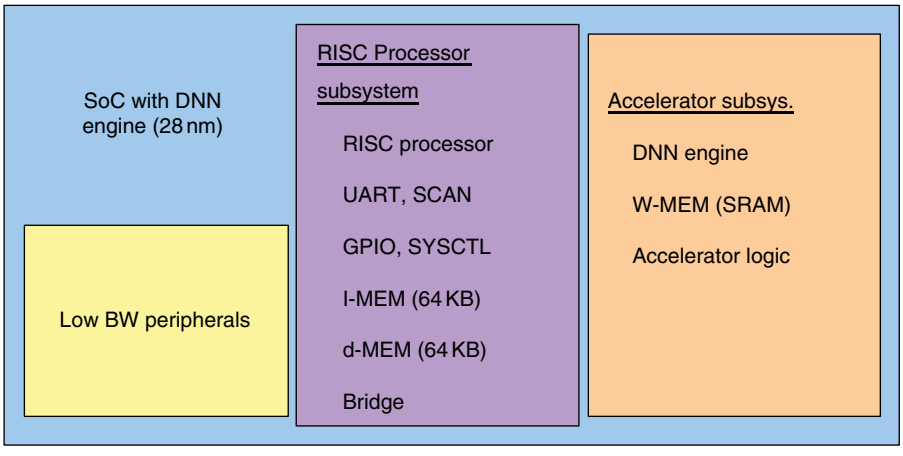


Figure 6.23 System block diagram of 28 nm SoC with deep neural net (DNN) engine for interpretation of noisy real world data from IoT devices. Based on P. Whatmough *et al.* (Harvard), ISSCC, February 2017 [21]

6.5.3 Deep Neural Nets for Speech and Vision Recognition

A scalable speech recognizer using hardware accelerated automatic speech recognition (ASR) with deep NN models and voice-activated power gating for use in personal assistants was discussed in 2017 by MIT and Analog Devices [22]. A wakeup mechanism such as voice activity detection (VAD) is needed to power-gate the system. ICs for ASR and VAD are described. The power-gated speech recognizer concept is shown in Figure 6.24. The front end, accoustical models, and neural net search are power-gated by the VAD detection.

While most recent work on NN hardware has targeted convolutional networks (CNNs) for computer vision, ICs for keyword detection with DNNs have had power as low as 3.3 mW. This study uses a DNN for low power ASR. The feedforward DNN accelerator uses an SIMD, as shown in the block diagram in Figure 6.25.

The ASR/VAD test chip performed all stages of ASR transcription from audio samples to test using voltages from 0.6 V at 10.2 MHz to 1.2 V at 86.8 MHz. The VAD was functional from 0.5 V at 1.68 MHz to 0.9 V at 47.8 MHz. The core was partitioned into five voltage areas. The SRAM could operate at 0.25 to 0.2 V above the logic supply up to 1.2 V. Latch-based clock gates were inserted at 76 locations resulting in a 30–40% reduction in core power at full load. Search efficiency varied from 2.5 to 6.3 nJ per hypothesis. A variety of ASR tasks with vocabularies up to 145 words could be run in real-time. Core power scaled by 45 times.

A 0.26–10 TOPS/W subword parallel convolutional neural net (CNN) processor in 28 nm FD-SOI technology was discussed in 2017 by KU Leuven [23]. Convolutional neural nets are classification algorithms for visual recognition. They are expensive in terms of energy due to the amount of data movement required and the billions of convolution computations.

This study presents the concept of hierarchical recognition processing using an energy scalable convolutional network processor with efficiencies up to 10 TOPS/W.

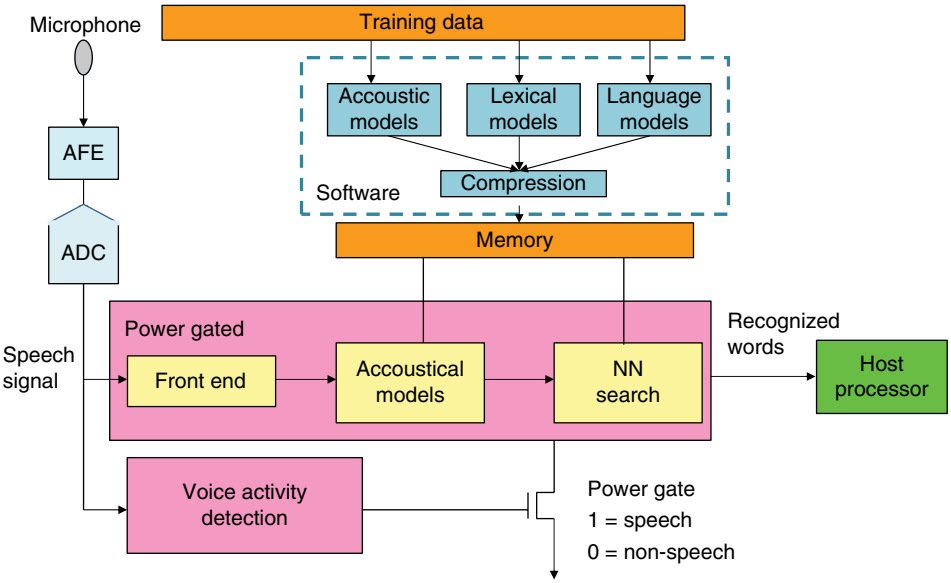


Figure 6.24 Block diagram of power-gated automatic speech recognition deep neural net chip. Based on M. Price *et al.* (MIT, Analog Devices), ISSCC, February 2017 [22].

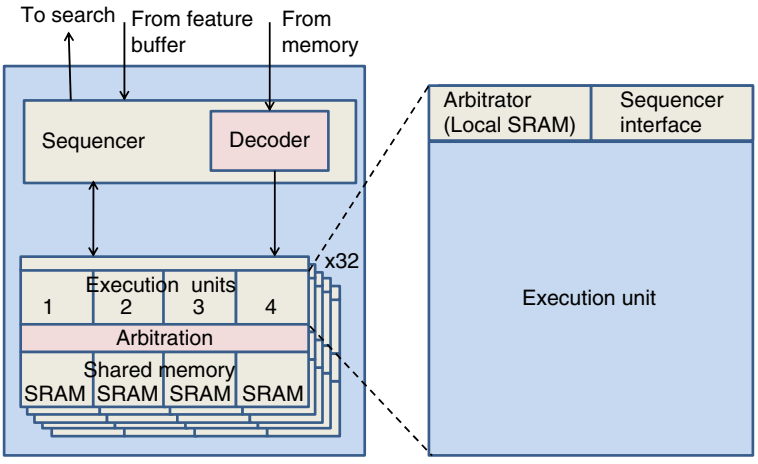


Figure 6.25 Block diagram of feedforward DNN accelerator using an SIMD neural network evaluator. Based on M. Price *et al.* (MIT, Analog Devices), ISSCC, February 2017 [22].

The processor enables always-on visual recognition in wearable devices. A block diagram of the top-level architecture of the processor is shown in Figure 6.26.

This convolutional network processor allowed use of a wide range of topologies. It had a 16b SIMD RISC instruction set with custom instructions. It was equipped with 2D for convolution, 1D-SIMD arrays, and a scalar unit. An on-chip memory used 64 2kB single-port SRAM macros storing a maximum of 64 K × N words. Three blocks could be read or written in parallel. A six-stage pipelined processor executed

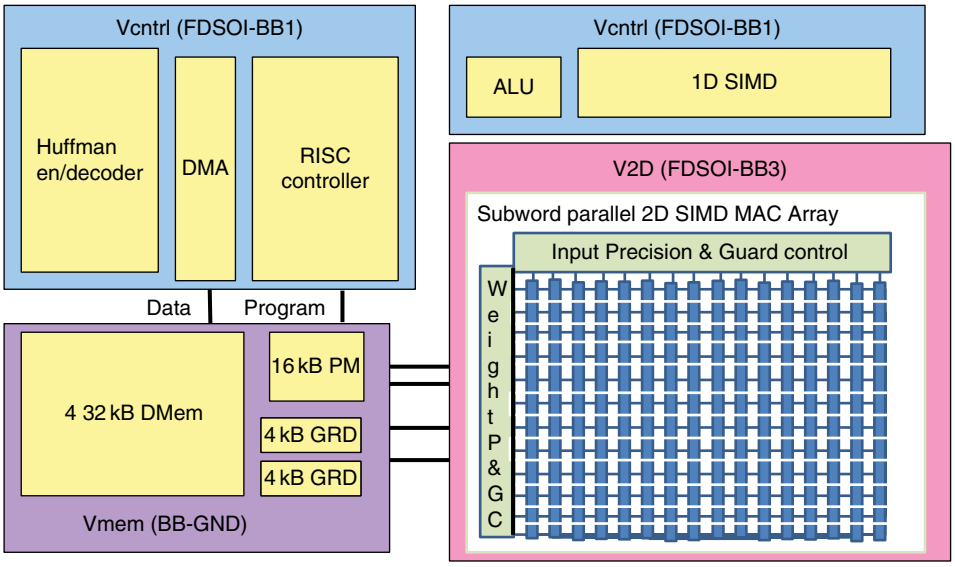


Figure 6.26 Block diagram of top-level architecture of a CNN processor with efficiencies up to 10 TOPS/W. Based on B. Moons *et al.* (KU Leuven), ISSCC, February 2017 [23].

convolutions. A 256b FIFO reduced memory bandwidth by reusing and shifting features along with the x-axis. All intermediate output values were stored in accumulation registers so there was no data transfer between MACs and no frequent write-back to SRAM. This 1.87 mm^2 chip ran at 200 MHz at 1 V at room temperature. Energy was 6.2 μJ/f at an average 6.5 mW, which shows the potential for always-on recognition through hierarchical processing.

A 0.52 mW convolutional neural net face recognition processor and a CIS face detector intended for the next generation of wearable devices were discussed in 2017 by KAIST [24]. A face recognition system was developed as a life-cycle analyzer constantly recording the people encountered together with time and place information. This face recognition had an always-on capability, which permitted it to be used for user authentication for secure access to personal systems. A convolutional neural network (CNN) was used for high accuracy and to enhance device intelligence. In order for this system to be used with personal devices required that the CNN processors consume low enough power to permit more than 10 hours of operation with a 190 mA h coin battery.

A low power CNN face recognition processor integrated with an always-on Haar-like face detector is proposed for smart wearable devices. Three features are used for low power: an analog–digital hybrid Haar-like face detector for face detection, an ultralow power CNN processor with wire I/O local distributed memory, and a separable filter approximating convolutional layers and a transpose-read SRAM for low power CNN processing. The system had two chips: a face image sensor and the CNNP. The first provides always-on face detection. The second completes the facial recognition. The overall block architecture with the face image sensor and the CNN processor is shown in Figure 6.27.

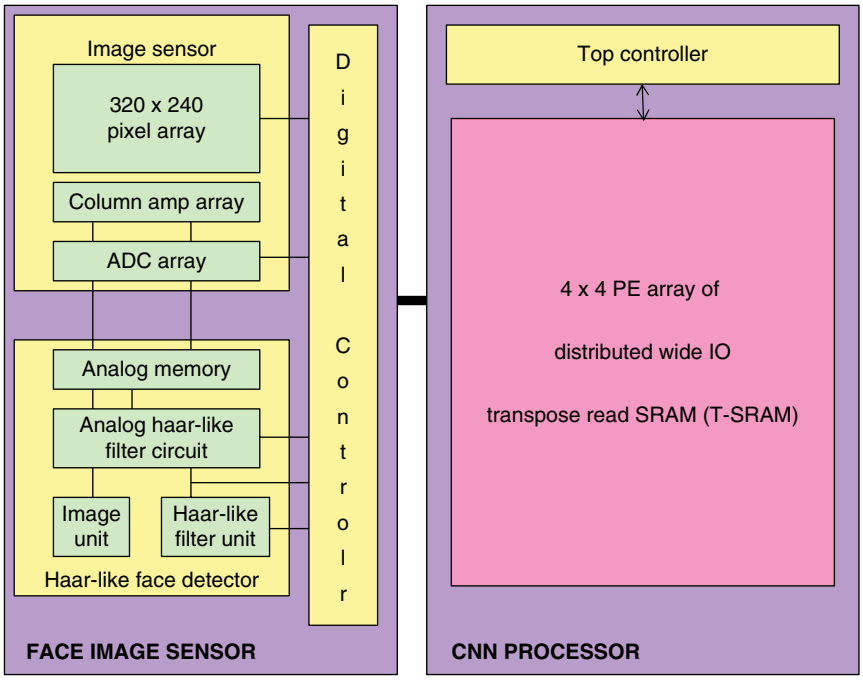


Figure 6.27 Overall block architecture with face image sensor and CNN processor. Based on K. Bong *et al.* (KAIST), ISSCC, February 2017 [24].

The overall energy consumption of the facial recognition device was reduced by 39%. The CNN processor could operate at 0.46 to 1.0 V supply with 5 to 100 MHZ clock frequency. The peak power at 0.46 V was 5.3 MW and at 1.0 V was 211 mW. The minimum energy point was 0.46 V. Energy efficiency at MEP is 1.06 nJ/cycle. The proposed system dissipates 0.62 mW on average at a 1 fps frame rate.

A 288 μW programmable deep processor with nonuniform memory hierarchy for mobile intelligence was discussed in 2017 by the University of Michigan and CubeWorks [25]. Recently there has been increased interest in deep learning for mobile IoT. Deep learning has been a powerful tool for applications such as speech and face recognition and object detection.

These are applications that are often needed at the edge of the Internet in situations involving mobile IoT. By having the deep learning at the edge of the Internet, local data can be used and only data for meaningful events need to be forwarded to the Cloud. This has the potential for improving the security of data since it can be secured locally. Many mobile applications are always on, so low power is critical. Rather than focusing on high performance reconfigurable processors optimized for large scale deep neural networks that can consume as much as 50 mW and more, bandwidth and power can be traded off in edge devices. Off-chip weight storage in DRAM can be avoided since it adds power consumption when data are manipulated in the Cloud due to the intensive off-chip data movement.

A low power programmable deep learning accelerator (DLA) was proposed with all weights stored on-chip for mobile intelligence. Low power of less than 300 μW was achieved in four ways: (1) four processing elements were located in the 279KB weight storage memory, which minimized data movement overhead; (2) a non-uniform memory hierarchy offered a trade-off between small low power memory banks for frequently used data and larger high density banks for the high power infrequently accessed data and the synaptic weights for the deep learning algorithms can be deterministically scheduled at compilation time using optimal memory assignments, which avoids the need for SRAM caches with their power and area overhead; (3) a 0.6 V 8 T custom memory was specifically designed for DNNs with a sequential access mode, band-by-bank drowsy mode control, power-gating for peripheral circuits, and voltage clamping for data retention; (4) flexible, compact memory storage was achieved using reconfigurable fixed point bit precision from 6 to 32b for neurons and weights. The 40 nm technology deep learning processor included: DLA, RISC processor, and MBus interface, which enabled integration into a complete sensor system. The DLA power was 288 μW at 374 GOPS/W. A top-level diagram of the DLA is shown Figure 6.28.

In the DLA, a nonuniform memory access (NUMA) architecture is used to provide balance between the memory area and access energy. The processor element (PE) NUMA memory floor plan uses signal gating circuits. The SRAM arrays use a dynamic drowsy mode operation since only a few banks are actively accessed in a specific

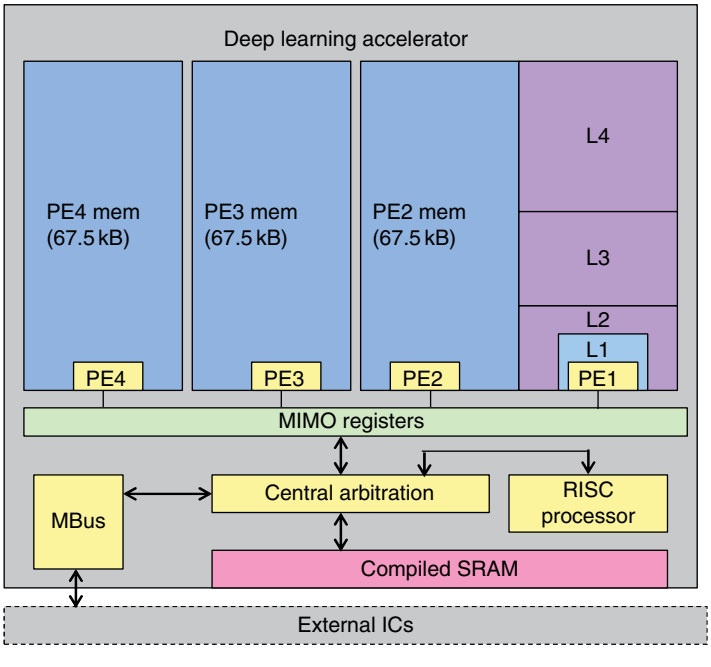


Figure 6.28 Top-level diagram of the deep learning accelerator (DLA). Based on S. Bang *et al.* (University of Michigan, CubeWorks), ISSCC, February 2017 [25].

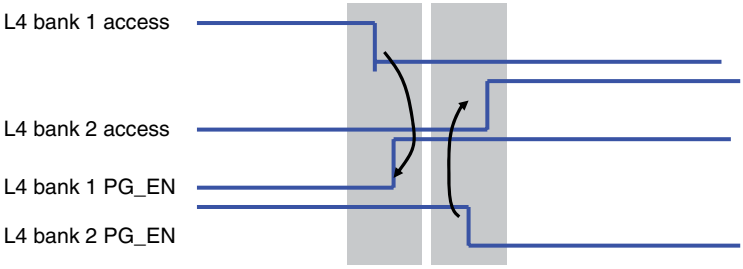


Figure 6.29 SRAM synaptic drowsy mode. Based on S. Bang *et al.* (University of Michigan, CubeWorks), ISSCC, February 2017 [25].

PE while the others are idle most of the operating time. A timing diagram for the SRAM dynamic drowsy mode is shown in Figure 6.29.

The 40 nm CMOS test chip measured data access power in L1 at 60% less than in L4. Memory drowsy mode operation reduced leakage by 54% due to peripheral circuits consuming less leakage power. The peak efficiency was 374 GOPS/W at 288 μA at 0.65 V and 3.9 MHz. An application for keyword spotting and face detection was ported on to the DLA. Both DNN classifications fit into the 270kB on-chip memory with <7 ms latency.

6.5.4 Deep Neural Nets for Other Applications

A 135 mW fully integrated data processor for next-generation DNA sequencing (NGS) or mapping was discussed in 2017 by National Taiwan University and NCTU [26]. With traditional methods, DNA sequencing is very time consuming, but DNA mapping can be partitioned into suffix array (SA) sorting and backward searching. Dedicated hardware has been proposed for low complexity backward searching but SA sorting has never been implemented. This study discusses a fully integrated NGS data processor that accomplished both SA sorting and backward searching.

A distributed sort algorithm is used to reduce sorting complexity. Throughput for SA sorting is maximized using 2 K insertion sorting elements. Overflow and splitter caches are used to reduce computation latency. This NGS processor achieves several orders of magnitude improvement in energy and throughput compared to high end generic processors. A block diagram of the architecture of the NGS processor is shown in Figure 6.30. Operations in the NGS processor included: splitter reordering, suffix sorting, suffix grouping, parameter analyzing, backward searching, splitter cache, and overflow cache.

The 40 nm NGS processor had a chip area of 7.84 mm^2 compared to over 300 mm^2 for various high end processors also used for DNA analysis. External memory was 1GB and local memory was 384 KB. Operating frequency was 200 MHz compared with over 1000 MHz for comparable alternative processors and power was 135 mW compared to over 4 orders of magnitude for alternative processors. Features of the NGS processor compared favorably with the use of a comparable technology 8-core processor and 448-core graphics processor.

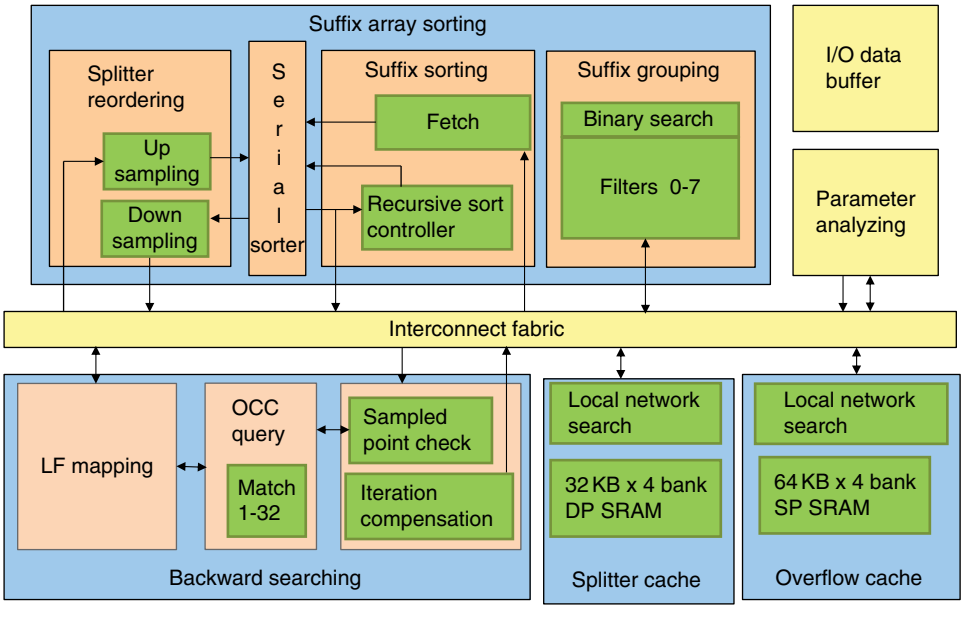


Figure 6.30 Block diagram showing architecture of NGS processor. Based on Y.-C. Wu *et al.* (National Taiwan University, NCTU), ISSCC, February 2017 [26].

References

1 Huang, L. *et al.* (2014) ReRAM-based 4T2R nonvolatile TCAM with 7× NVM-stress reduction, and 4× improvement in speed-word-length-capacity for normally-off instant-on filter-based search engines used in big-data processing (National Tsing Hua University, National Chiao Tung University, NDL, ITRI), VLSI Circuits Symposium, pp. 99, June 2014.

2 Chang, M.-F. *et al.* (2015) A 3T1R nonvolatile TCAM using MLC ReRAM with sub-1 ns search time (NTHU, ITRU, NCTU), ISSCC, March 2015.

3 Lin, C.C. *et al.* (2016) A 256b-wordlength ReRAM-based TCAM with 1 ns search-time and 14× improvement in word length-energy efficiency-density product using 2.5T1R cell (NTHU, TSMC, NCTU), ISSCC, February 2016.

4 Gupta, M. and Hasan, M. (2015) Design of high-speed energy-efficient masking error immune PentaMTJ-based TCAM (Aligarh Muslim Univ.). *IEEE Trans. on Magnetics*, 51 (2), February 2015

5 Wang, X. *et al.* (2015) Spin-orbitronics memory device with matching and self-reference functionality (Avalanche Technology). *IEEE Trans. on Magnetics*, 51 (11), 1401504, November 2015.

6 Bayram, I. and Chen, Y. (2014) NV-TCAM: Alternative interests and practices in NVM designs (University of Pittsburgh), NVM-SA, August 2014.

7 Lin, W.-P. (2014) A nonvolatile look-up table using ReRAM for reconfigurable logic (ITRI, National Taipei University, NTHU, Min Shin University), IEEE Asian Solid-State Circuits Conference, November 10, 2014.

8 Gu, P. *et al.* (2015) Technological exploration of RRAM crossbar array for matrix-vector multiplication (Tsinghua University, Arizona State University), ASP-DAC, January 2015.
9 Tanakamaru, S., Yamazawa, H., Tlkutomi, T. *et al.* (2014) Hybrid storage of ReRAM/TLC NAND Flash with RAID-5/6 for Cloud data centers (Chuo University, University of Tokyo), IEEE ISSCC, February 2014.
10 Xu, C. *et al.* (2015) Overcoming the challenges of crossbar resistive memory architectures (Oakridge National Laboratory), HPCA, February 2015.
11 Vincent, A.F. *et al.* (2014) Spin-transfer torque magnetic memory as a stochastic memristive synapse (University of Paris-Sud, CNRS, CEA, LIST), ISCAS, pp. 1074, June 1, 2014.
12 Vincent, A. *et al.* (2015) Spin-transfer torque magnetic memory as a stochastic memristive synapse for neuromorphic systems (University of Paris-Sud, CEA, Beihang University). *IEEE Trans. on Biomedical Circuits and Systems*, 9 (2), April 2015.
13 Zhang, D. *et al.* (2016) All spin artificial neural networks based on compound spintronic synapse and neuron (Beihang University, University of Paris-Sud, Tsinghua University). *IEEE Trans. on Biomedical Circuits and Systems*, 10 (4), August 2016.
14 Chen, P. and Yu, S. (2016) Partition SRAM and RRAM based synaptic arrays for neuro-inspired computing (Arizona State University), ISCAS, May 2016.
15 Suri, M. and Parmar, V. (2015) Exploiting intrinsic variability of filamentary resistive memory for extreme learning machine architectures (IIT New Delhi). *IEEE Trans. on Nanotechnology*, 14 (6), November 2015.
16 Suri, M. *et al.* (2015) Neuromorphic hybrid RRAM-CMOS RBM architecture (IIT, IEF-CNRS, IEMN-CNRS), 15th NVMTS, October 2015.
17 Eryilmaz, S. *et al.* (2016) Training a probabilistic graphical model with resistive switching electronic synapses (Stanford, University of California at Irvine, University of California at San Diego, IBM, Macronix). *IEEE Trans. on Electron Devices*, 63 (12), December 2016.
18 Shelby, R.M. *et al.* (2015) Non-volatile memory as hardware synapse in neuromorphic computing: a first look at reliability issues (IBM), IRPS, April 2015.
19 Desoli, G. *et al.* (2017) A 2.9TOPS/W deep convolutional neural network SoC in FD-SOI 28 nm for intelligent embedded systems (STMicro), ISSCC, February 2017.
20 Shin, D. *et al.* (2017) DNPU: An 8.1TOPS/W reconfigurable CNN-RNN processor for general-purpose deep neural networks (KAIST), ISSCC, February 2017.
21 Whatmough, P. *et al.* (2017) A 28 nm SoC with a 1.2 GHz 568 nJ/prediction sparse deep-neural-network engine with >0.1 timing error rate tolerance for IoT applications (Harvard), ISSCC, February 2017.
22 Price, M., Glass, J. and Chandrakasan, A. (2017) A scalable speech recognizer with deep-neural-network acoustic models and voice-activated power gating (MIT, Analog Devices), ISSCC, February 2017.
23 Moons, B. *et al.* (2017) ENVISION: A 0.26-to-10 TOPS/W subword-parallel dynamic-voltage-accuracy-frequency-scalable convolutional neural network processor in 28 nm FDSOI (KU Leuven), ISSCC, February 2017.
24 Bong, K. *et al.* (2017) A 0.62 mW ultra-low-power convolutional-neural-network face-recognition processor and a CIS integrated with always-on Haar-like face detector (KAIST), ISSCC, February 2017.

25 Bang, S. *et al.* (2017) A 288 μW programmable deep-learning processor with 270 KB on-hhip weight storage using non-uniform memory hierarchy for mobile intelligence, ISSCC, February 2017.

26 Wu, Y.-C. Hung, J.-H. and Yang, C.H. (2017) A 135 mW fully integrated data processor for next-generation sequencing (National Taiwan University, National Chiao Tung University), ISSCC, February 2017.

7

Memory in Security Issues for IoT

7.1 Introduction to Memory in Security Issues for IoT

On-chip security systems are critical for IoT applications such as wearable devices and portable medical instruments. Any IoT application may store sensitive user information that needs to be protected. All nonvolatile memories such as the EEPROM, Flash, RRAM, FeRAM, and STT MRAM have security challenges, which involve persistent data that can be tampered with such as a password or cryptographic key. The vulnerability is due to the fundamental dependency of these memories on environmental parameters such as magnetic fields and temperature, which can be used to tamper with data. TAGs are generated for memory data authentication in embedded processor systems. Data integrity can be protected by attaching an identifying tag to individual data that becomes invalid if the data is altered. Face-detection is a critical feature of many security systems. Several acceleration techniques exist for enabling this capability on a CPU-FPGA platform.

Many nonvolatile memories have stochastic switching mechanisms and intrinsic variability that open a range of potential applications in hardware security. RRAMS, FeRAMs, and MRAMs have been studied for use as physical unclonable functions (PUF) in security systems.

7.2 Memories Used as Physical Unclonable Functions (PUFs)

Memory can be used in security implementation by providing a physical unclonable function (PUF).

A PUF is a physical entity that is embodied in a physical structure and is easy to evaluate but is hard to predict. The PUF generates a secured key by using the physical nature of an electronic system since the physical structure of every electronic system is unique due to inherent differences occurring during manufacturing, even on the same manufacturing line. PUFs have been developed based on CMOS SRAMs, CMOS RRAM, spintronic MRAMs, phase change memories, and Flash memories. RRAM and MRAM PUFS have good Hamming distance due to inherent variations in the technology.

Memories for the Intelligent Internet of Things, First Edition. Betty Prince and David Prince.
© 2018 John Wiley & Sons Ltd. Published 2018 by John Wiley & Sons Ltd.

7.2.1 Using RRAM for a Physical Unclonable Function

RRAMs have a stochastic switching mechanism and intrinsic variability, which pose challenges for use as a memory. However, these features open a range of potential applications in hardware security.

RRAMs can be used for physical unclonable functions (PUFs) for hardware security applications. For RRAMs, variation of one of the device properties is frequently used in a PUF. The performance of the RRAM PUF depends on the reliability of the RRAM and may degrade. In February of 2015, Global Foundries proposed a PUF based on RRAM resistance variability, which had the advantage of being intrinsic in the physical mechanisms with reconfigurability [1]. It is critical for RRAM resistance to have truly random variation for the PUF to be unique and unclonable.

A PUF uses physical randomness and manufacturing variability to provide security primitives. This robustness is affected by temperature and voltage dependence of the resistance and also by retention characteristics.

A large separation between the interchip and intrachip Hamming distance is a measure of uniqueness and reliability. An RRAM based on TaO_x was used for the PUF. A simulated resistance distribution in the LRS and HRS is shown in Figure 7.1. The LRS follows a normal distribution while the HRS follows a lognormal distribution.

A challenge applied on a PUF generates a response that involves converting physical variability in PUF hardware into digital bit-string randomness. The ability to reconfigure the PUF by reprogramming the RRAM arrays helps overcome reliability degradation. The uniqueness of the PUF is its ability to distinguish a particular chip among a group of chips of the same type. It can be measured by the Hamming distance (HD) between responses of a pair of PUFs for a given challenge. This is the interchip HD. Uniqueness is expected to be 50%. The intrachip HD should be 0%. An example of a PUF implementation based on RRAM resistance variability is shown in Figure 7.2.

In May of 2015, Arizona State University discussed an optimized design for a physical unclonable function (PUF) primitive based on an RRAM crosspoint array [2]. The stochastic, or random, effect of the PUF comes from the resistance variation of the RRAM

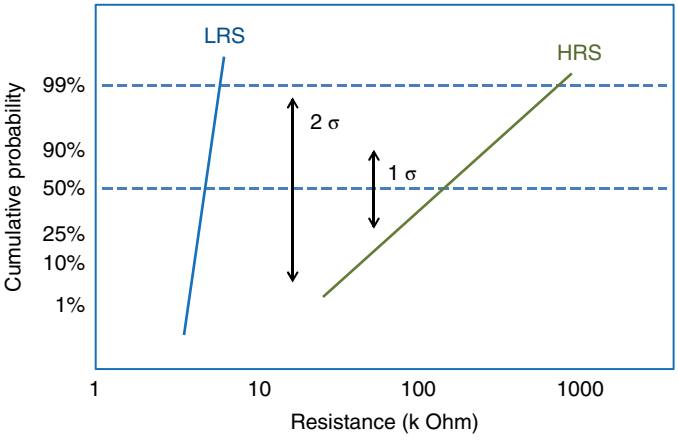


Figure 7.1 Simulated resistance distribution in LRS and HRS of TaO_x-based RRAM PUF. Based on A. Chen (Global Foundries), *IEEE Electron Device Letters*, February 2015 [1].

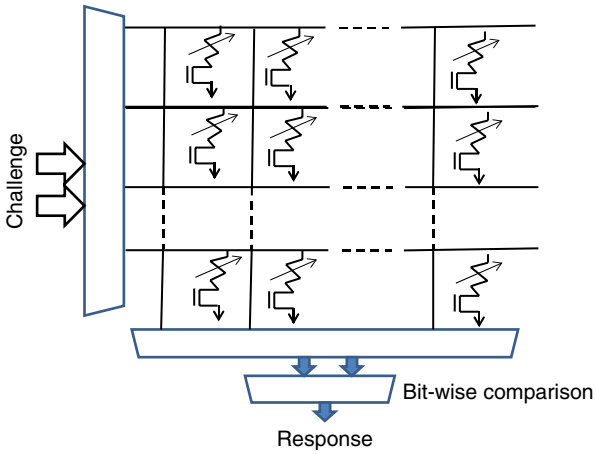


Figure 7.2 Circuit diagram of PUF implementation based on RRAM resistance variability. Based on A. Chen (Global Foundries), *IEEE Electron Device Letters*, February 2015 [1].

cells in the array. A four-cell selection method was used to create the large number of challenge–response pairs that are needed to achieve a high security level. To check the PUF performance uniqueness and reliability, an RRAM crosspoint array was made and the device parameters were calibrated.

It was shown that the RRAM PUF works correctly across a wide temperature range without degradation, although the IR drop resulting from interconnect resistance in the array can limit the performance. To reduce the effect of the IR drop, a reverse scaling rule on feature size was proposed for the RRAM PUF. This resulted in an increase in area of the chip, but improved the performance significantly. Compared to a 45 nm SRAM PUF, an RRAM PUF array size with $F = 200$ nm has a 45% smaller area while maintaining better robustness against invasive side-channel attacks.

The stochastic switching mechanism of the RRAM and the intrinsic variability of its resistance are issues for the RRAM memory application. They are, however, an opportunity to design a physical unclonable function (PUF) for hardware security.

In June of 2015, the Politecnico di Milano discussed scalable low power random number generator (RNG) blocks [3]. These blocks are needed for encryption in modern communications systems. A true RNG must be based on a random physical phenomenon such as individual fluctuations in random telegraph noise (RTN) or random trapping/detrapping phenomena in dielectrics.

This study based the RNG on a SET variability in an RRAM. The RNG relies on a single RRAM device that is repeatedly programmed at a constant voltage close to the nominal SET voltage. Because of the statistical variability of the SET voltage, the SET transition only occurs in about 50% of the applied pulses. This results in a bimodal distribution of analog resistance, which is finally converted into a "0" and "1" distribution of output voltage values through digital regeneration with a CMOS inverter.

In December of 2015, TD Research and GlobalFoundries discussed RRAM-based PUF for hardware security applications [4]. The performance of an RRAM PUF depends on the RRAM device characteristics. The reliability of the RRAM PUF may degrade as the retention is lost, with read instability or with thermal variation. PUF uniqueness is

maintained as long as the randomness in RRAM resistance distribution is preserved. Figure 7.2, shown earlier, is an example of a schematic of a physical unclonable function based on a 1T1R RRAM array. Challenges are addresses of a pair of *n*-bit memory data. Responses are generated by bit-wise comparison of resistance or read current between the two *n*-bit data. All cells are programmed to the HRS. Resistance comparison may be affected by the sense amplifier resolution. The challenge–response pairs generated convert stochastic RRAM resistance variation into random bit streams that can be used as security primitives.

The impact of RRAM device characteristics on the performance of RRAM-based PUF was systematically analyzed. While the magnitude of the effects may vary, the general trend and parametric dependence are expected to remain valid. A reliable PUF requires the RRAM to have truly random variation but also to have sufficiently low noise and long retention.

In December of 2015, Arizona State University discussed using a 1Kb RRAM array as a PUF [5]. The randomness of the RRAM PUF comes from the stochastic switching mechanism and from the intrinsic variability of the RRAM devices. A 1Kbit HfO_2-based 1T1R RRAM was used as the PUF. The experimental results showed that the selection of the reference current affects the uniqueness. It was found that increasing the number of dummy cells improved the uniqueness of the RRAM. The reliability of the RRAM PUF was determined by the RRAM data retention.

A new design was suggested in which the sum of the read currents from multiple RRAM cells was used for generating one response bit. This was found to statistically minimize the risk of early lifetime failure. It was shown that with eight cells per bit, the retention time was more than 50 hours at 150 °C or 10 years at 69 °C, as shown in Figure 7.3, which plots retention time versus $1/KT$ with an activation energy (E_a) of 1.15 eV, where K is the Boltzman constant and T is the temperature in degrees kelvin. It was

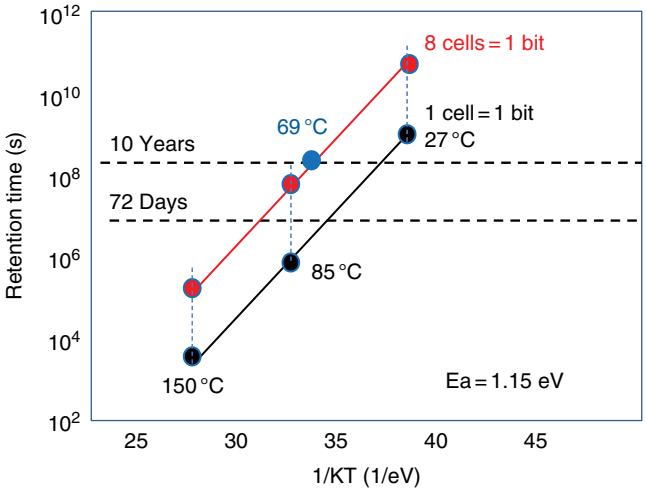


Figure 7.3 Retention time versus $1/KT$ for a single cell PUF response and retention time for a PUF response bit represented by different numbers of cells along with retention degradation over time for eight cells per bit along with an extrapolated retention time at 85 and 27 °C using E_a = 1.15 eV. Based on Rui Liu *et al.* (Arizona State University, Tsinghua University), *IEEE Electron Device Letters*, December 2015 [5].

found that the RRAM PUF is a viable technology for a hardware security primitive with an inter-Hamming distance of 49.8% and an intra-Hamming distance of 0%.

In May of 2016, the Politechnico di Milano and Micron Technology discussed physical unbiased generation of random numbers using two coupled RRAM devices [6]. The two-resistance method permits overcoming the requirement for probability tracking, where the operating voltage must be tuned to adjust the generation probabilities of 0 and 1. Random number generation (RNG) is required for generating random encryption keys for secure transmission protocols. Compact RNG circuits are useful if capable of generating true random numbers with high entropy quality and high throughput. In order to generate true random numbers, it is necessary to identify a convenient on-chip entropy source and to design a circuit corresponding to it for generating the random bit stream with high throughput and stability.

Several sources of random fluctuation are available in an RRAM. Random telegraph noise (RTN) appears as a random fluctuation of current or resistance between two levels, but is difficult to control and RTN has been shown unstable in the RRAM. The approach selected replies on stochastic switching in two coupled RRAM devices driven by the voltage pulse and having different parallel or series configurations.

Unbiased RNG relies either on a comparison between the resistance levels after independent switching or on alternative switching in series/parallel RRAM. All of these concepts adopt two coupled RRAM devices to give 50% probability of zeros and ones by self-compensation methods. The methods used included: a parallel-reset configuration based on HRS resistance variation, a serial-reset configuration based on V_{reset} variation, and a parallel-set configuration based on V_{set} variation. Figure 7.4 (a) illustrates the parallel-RESET circuit with two RRAM cells "P" and "Q" connected in parallel. P and Q both represent 1T1RRAM structures. Figure 7.4 (b) shows the sequence of applied signals starting from P to Q in HRS. The cells are first independently SET using a positive voltage across P and Q and the output. They are then RESET using a negative voltage

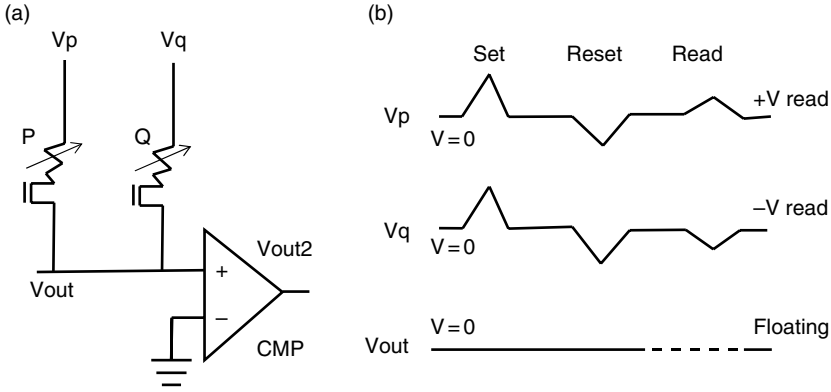


Figure 7.4 Circuit with two RRAM cells "P" and "Q" connected in parallel where both cells have a 1T+1R configuration. (a) The circuit diagram of the parallel RESET circuit with the two RRAM cells connected in parallel and the output feeding into a comparator. (b) Timing diagrams showing the sequence of applied signals on P and Q. The cells are first SET using a positive voltage from P and Q to the output. They are then RESET using a negative voltage across P and Q to the output for RESET. P and Q are then READ using the voltage divider configuration. Based on S. Balatti *et al.* (Politechnico di Milano, Micron Tech.), *IEEE Trans. on Electron Devices*, May 2016 [6].

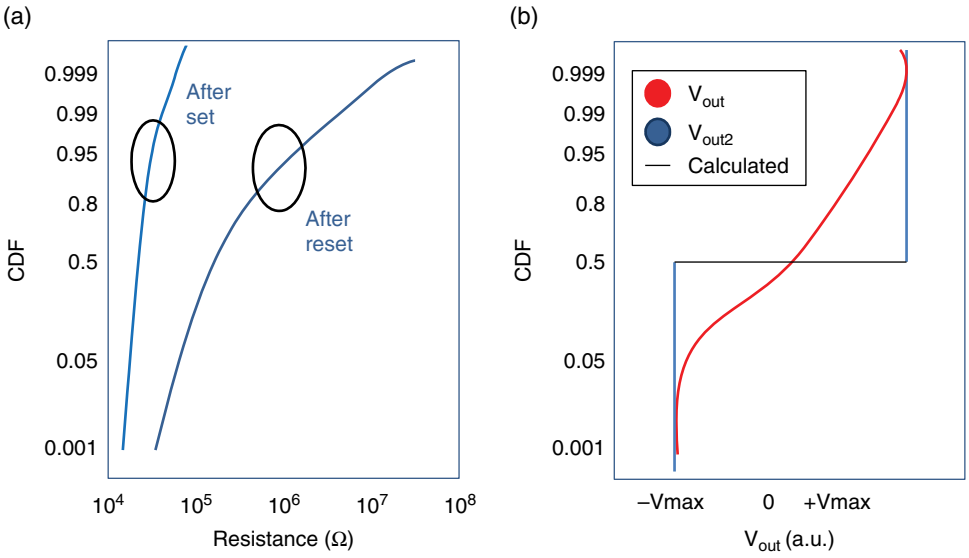


Figure 7.5 (a) Cumulative distribution function (CDF) of resistance after SET and after RESET for cells P and Q and (b) corresponding distribution of V_{out} and V_{out2} with and without comparator. Based on S. Balatti *et al.* (Politechnico di Milano, Micron Tech.), *IEEE Trans. on Electron Devices*, May 2016 [6].

across P and Q and the output. P and Q are then READ using the voltage divider configuration. The analog comparator is used to digitally regenerate V_{out}.

The cumulative distributions of resistance after SET and after RESET for cells P and Q are shown in Figure 7.5 (a) and the corresponding distributions of V_{out} and V_{out2} with and without a comparator are shown in Figure 7.5 (b). In all cases, the bimodal distributions of high and low voltage were found with low to high transition at 50%. RNG was improved by adding digital regeneration. The results of randomness tests for the parallel RESET method with and without the von Neumann correction support the two RRAM methods for future on-chip RNG.

In May of 2016, Arizona State University discussed the design of a reliable RRAM-based PUF for security primitives [7]. A PUF must be highly reliable, particularly when it is being used with cryptographic hash modules to generate keys. For high reliability, normally error correction codes are used based on data input. This, however, tends to increase the hardware overhead of the PUF, which increases cost and also reduces the entropy of the system and results in both hardware and software security issues. The system that was designed used a compact and reliable PUF architecture based on resistive RAM (RRAM). In this design, the sum of the read out currents of several RRAM cells is used to generate one response bit. A design of an RRAM crossbar array where each PUF bit is represented by two RRAM cells is shown in Figure 7.6.

This technique tends to statistically minimize any early lifetime failure due to RRAM retention degradation at high temperature or under voltage stress. A device model was used that was calibrated with the HfO_x RRAM data. The model showed that, with 8 cells per bit, it was possible to ensure less than 10^{-6} intra-Hamming distances, which equates to 99.9999% reliability for a lifetime greater than 10 years at 125 °C. The RRAM PUF was embedded into the secure hash algorithm (SHA) 256. The SHA-256 algorithm generates a nearly unique 256-bit hash, which is a one-way function that cannot be decrypted back.

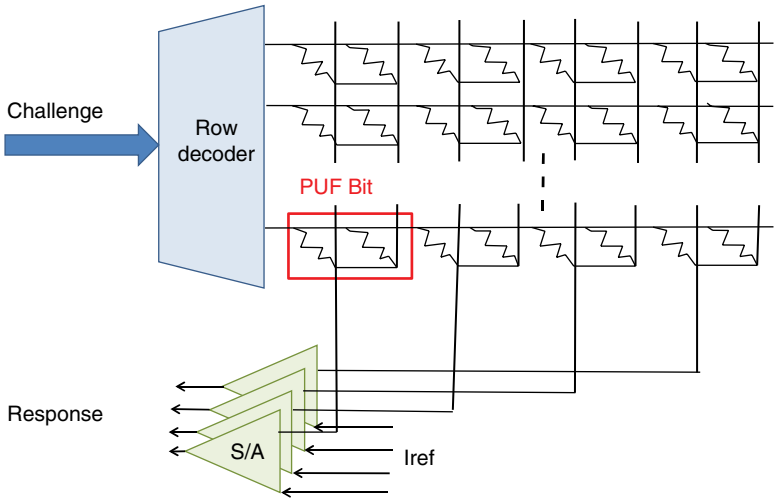


Figure 7.6 Illustration of an RRAM crossbar array with each PUF bit represented by two RRAM cells. Based on A. Shrivastava *et al.* (Arizona State University), ISCAS, May 2016 [7].

The hardware overhead of the proposed RRAM PUF-based architecture was significantly lower than one that uses a traditional RRAM PUF with ECC. Eliminating ECC by adding redundancy into the RRAM array significantly increased the security level of the PUF.

In June of 2016, Panasonic discussed using a 180 nm RRAM-based PUF in an FPGA with a bit error rate less than 0.5% after 10 years at 125 °C for 40 nm embedded applications [8]. By using an RRAM PUF with an RRAM memory core, the footprint is minimized and security is improved. A new ID generating method and reproducing algorithm for RRAM PUF was described. Detailed resistance value information was obtained using a timing sense amplifier and a resistance to time-count converter.

The RRAM cell was an It/Ta_2O_5/TaO_x/TaN stack that switched using a conductive filament consisting of oxygen vacancies. The position of these vacancies and the CF diameter were random for each bit so that variation of resistance values showed a log-normal distribution. The resistance value of each bit in the LRS had an unclonable characteristic. A three-step algorithm ID generating method was proposed that determined the median of the lognormal distribution as a threshold.

An RRAM used as a PUF for reliable security primitives for authentication and in-place secret key generation was discussed in July and October of 2016 by the University of South Florida and Pennsylvania State University [9,10]. PUFs are cost effective and reliable security primitives widely used in authentication and in-place secret key generation applications. RRAM offers easy integration with CMOS due to minimal changes in the process technology and also show resistance variability characteristics due to inherent defects in the conducting filament formed in the metal oxide layer. RRAM-based PUF designs use either the probabilistic switching of the RRAM or the resistance variability during forming, set, and reset processes.

A strong arbiter PUF was proposed based on a 1T1R metal oxide RRAM bit cell such as used in conventional RRAM memory arrays. A conductive filament in these RRAMs is formed of oxygen vacancies in the dielectric between two metal electrodes during the SET operation. The filament is dissolved back to the metal electrode during the RESET

operation. An illustration of the MO-RRAM SET and RESET operation is shown in Figure 7.7 and the RRAM memory *I–V* curve is shown in Figure 7.8.

A conventional voltage sense amplifier was used to generate the response. The PUF was simulated using 65 nm predictive technology models for CMOS and a Verilog-A model for an HfO_2-based RRAM. The proposed PUF architecture showed a mean intradie Hamming distance of 0.13% and interdie HD of 52.3%.

Optimizing an RRAM-based PUF using a new differential read out method was discussed by Tsinghua and Arizona State Universities with ISRC, Durham, NC, in January of 2017 [11]. A 1K-bit 1T1RRAM array was demonstrated. A differential read out method was used to reduce the effect of resistance window degradation. The reliability

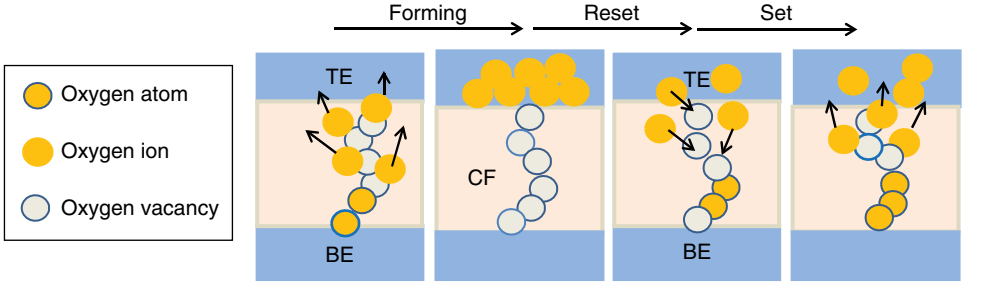


Figure 7.7 Illustration of the MO-RRAM SET and RESET operation showing the initial forming of the conductive filament, the RESET operation with dissolution of the filament and the SET operation with reform of the filament in the gap left during RESET. Based on R. Govindaraj and S. Ghosh (University of South Florida, Pennsylvania State University), ICCD, October 2016 [10].

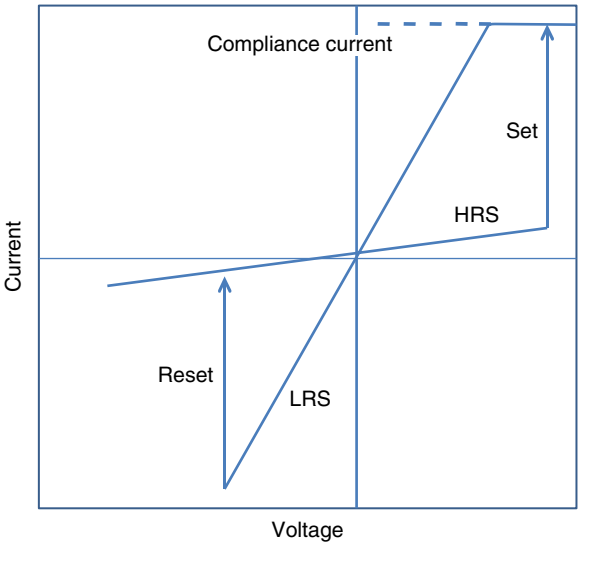


Figure 7.8 RRAM memory *I–V* curve showing the compliance current level that limits SET current for the low resistance state (LRS) and the bipolarity of applying the opposite voltage for RESET to the high resistance state (HRS). Based on R. Govindaraj and S. Ghosh (University of South Florida, Pennsylvania State University), ICCD, October 2016 [10].

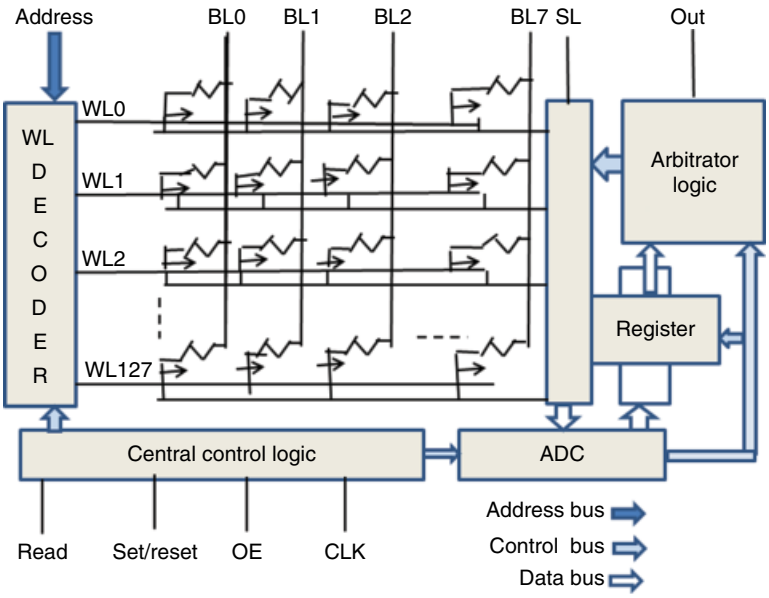


Figure 7.9 Schematic circuit diagram of the proposed PUF based on an RRAM array. (Based on Y. Pang *et al.* (Tsinghua University), *IEEE Electron Device Letters*, February 2017 [11] (permission of IEEE).

of the RRAM PUF was optimized using an enhanced reliability design and using oxide stack engineering. The optimized RRAM PUF showed nearly ideal uniqueness with the interchip Hamming distance being close to the ideal 50% and the intrachip Hamming distance close to 0%. This showed that the RRAM PUF has potential for robust security solutions in IoT applications. A schematic circuit diagram of the proposed PUF based on an RRAM array is shown in Figure 7.9.

7.2.2 Using MRAMs as Physical Unclonable Functions

MRAMs can be used for physical unclonable functions. As long as the magnetic function technology is being widely used in various logic and memory platforms, it can also be used in the security function for authentication of ICs. The inherent variations in the MRAM can be used to generate secret keys and signatures without physically storing values in the MRAM.

In June of 2014, Nanyang Technical University and Purdue University discussed an STT-MRAM-based physical unclonable function (PUF) [12]. A PUF based on the unpredictable disorder of physical devices is used to address security issues related to cryptographic key generation. A PUF uses the disorder of device properties resulting from the fabrication variations to generate a unique and unpredictable output when interrogated by an input challenge. Responses generated from the PUF by the same challenge must be repeatable and robust.

A memory-based PUF based on STT-MRAM is suggested as a key generation primitive for embedded computing systems. The resistance of STT-MRAM cells, which are initialized to the same state, are compared and response bits are generated by using the

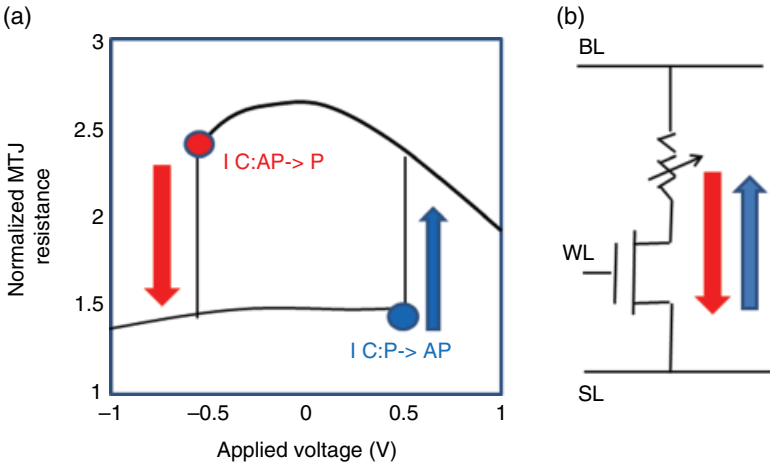


Figure 7.10 Magnetic tunnel junction (MTJ): (a) Resistance–voltage characteristics of MTJ. Switching from AP > P when I > I_c (AP > P) and from P > AP when I > I_c:(P > AP). (b) 1T1MTJ cell circuit schematic. Based on L. Zhang, *et al.* (Nanyang TU, Purdue University), ISCAS, June 2014 [12].

inherent random mismatches between them. A new M-PUF using a hybrid STT-CMOS technology was found capable of producing reliable responses for secure key generation with marginal design overheads.

An MTJ device is a structure in which a layer of MgO dielectric is sandwiched between two paramagnetic layers, one of which is free to rotate and the other of which is pinned. When the magnetic layers are oriented in the same direction the resistance through the MgO is low and current flows. When the magnetic layers are oriented in the opposite direction the resistance is higher and the current does not flow. A memory hysteresis curve of resistance through the MTJ versus applied voltage is shown in Figure 7.10 (a). The direction of current is indicated in the 1T1MTJ cell in Figure 7.10 (b).

An automatic write-back method was used to regenerate stable responses securely under different operating conditions by sensing the resistance difference between two STT-MRAM cells in complementary states, which reduces the STT-PUF's reliance on ECC.

To improve the robustness of the response bits regeneration, an automatic write-back (AWB) method is suggested, which does not compromise the resilience of the STT-PUF against possible attacks. Simulations have shown that the STT-PUF can produce raw response bits with a uniqueness of 50.1% and entropy of 0.985 bits per cell. The worst case bit-error rate (BER) under varying operating conditions is 6.6×10^{-6}.

In August of 2014, the University of South Florida discussed an MRAM-based physically unclonable function (PUF) using the uncontrollable geometric variation in MRAM cells [13]. The geometric variations were found to generate a preferred ground state in each MRAM cell in the PUF. Security properties of the PUF were analyzed using simulation and fabrication. An intradistance of 0.0225 and an interdistance of 0.47 were found for this PUF.

Physically unclonable functions (PUFs) in hardware security devices have advantages over software-based secure ID and key storage. A PUF was proposed for MTJs in November of 2014 by Toshiba along with a method for extracting the PUF based on intrinsic properties of spin transfer switching (STS) [14]. The STS characteristics of MTJs with perpendicular layers were measured and analyzed to determine how to

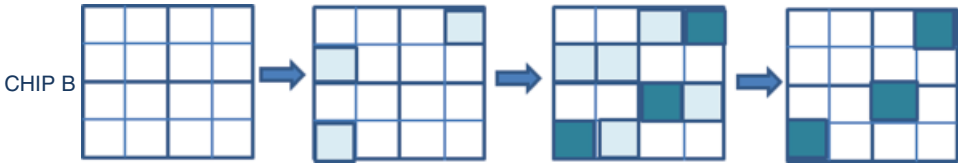


Figure 7.11 Schematic view of PUF extraction from collected MTJ voltage data categorized into three alternatives representing the parallel state (white), 50% parallel or anti-parellel state (gray) and anti-parallel state (black). Based on J. Das *et al.* (University of South Florida), IEEE-NANO, August 18, 2014 [13].

extract PUF properties. A PUF in an integrated circuit is a function that uses inherently unique device variations and is therefore difficult to clone.

A new PUF was proposed for MTJ involving the magnitude of the switching current depending on the process variations. The time-dependent behavior of the MTJ switching is an important aspect of the MTJ-based PUF. When a challenge signal is received from external devices, the PUF generates a response bit stream based on unique IC properties. This study shows the basic principles and methods for extracting a PUF from the STS characteristics in MTJs. A schematic view of the PUF principles is shown in Figure 7.11. For repeated V(PUF) corresponding to voltages for PUF extraction, the collected MTJ data are categorized into three alternatives: less than 50% (white)(parallel state), near 50% gray (parallel or anti-parallel state), and more than 50% (black) (anti-parallel state). The resulting black bits show the desired PUF being obtained.

In February of 2015, the University of South Florida discussed a geometry-based magnetic physical unclonable function (PUF) [15]. In this case the intrinsic geometric variations in the MRAM cell are used to build a PUF. A dedicated hardware device can be used for processing signals with the PUF, as shown in Figure 7.12. In this case X is the hardware device that processes the signals between the PUF and the circuit for authentication. Various attributes of the MRAM cells based on random geometric variations are used as PUF attributes. The fabricated PUF cells generate entropy up to 0.99, which is comparable to competing PUFs in other technologies. The footprint of this PUF cell is smaller than most silicon-based PUFs.

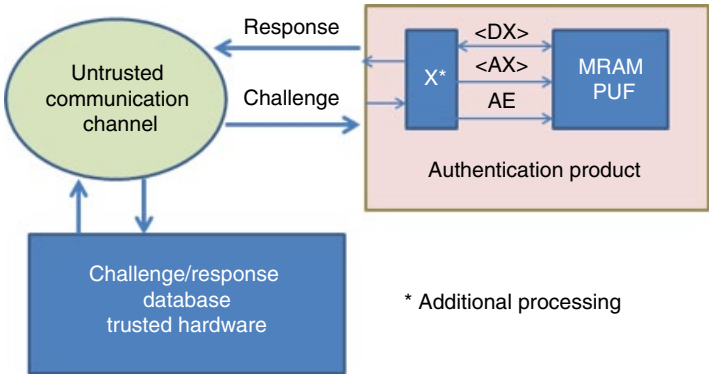


Figure 7.12 Concept of MRAM PUF authentication protocol. In this system, X is the hardware device that processes the signals between the PUF and the circuit for authentication. Based on J. Das *et al.* (University of South Florida), *IEEE Trans. on Nanotechnology*, February 2015 [15].

In March of 2015, the Politecnico di Torino discussed physical unclonable functions (PUFs) as emerging cryptographic primitives for use in implementing low-cost device authentication and secure secret key generation [16]. This study proposed a new PUF design based on STT-MRAM memory. The high variability affecting the electrical resistance of the MTJ device in anti-parallel magnetization was used to create the PUF. The proposed solution was shown to be robust, unclonable, and unpredictable.

In April of 2015, Nanyang Technical University and Purdue University discussed optimizing the function of memory-based PUFs in generating device specific random security keys [17]. Existing memory-based PUFs must buffer data in the memory before it is used to produce random bits. This results in both additional area and additional energy consumption. A new memory-based PUF is proposed that uses nonvolatility and random variability to produce random bits. The random bit generation process does not disturb data bits already stored in memory. An example was designed using STT-MRAM, which showed the desired results using the proposed method.

In April of 2015, Nanyang Technical University discussed an STT-MRAM based PUF with multiple response-bits per cell [18]. This PUF used the characteristics of STT-MRAM stabilized using a new automatic write-back technique. The alterability of the MTJ state was used to expand the response-bit capacity per cell. The analysis showed the advantage of the method for reliability enhancement. The bit error rate (BER) went from 10:1 to 10:6 for worst case conditions. Response per bit capacity per cell improved from 1 bit to 1:48 bits. In comparison with conventional memory PUFs, this approach is better in chip area use and in energy for producing a response bit.

In May of 2015, the University of South Florida discussed a geometry-based magnetic MRAM PUF [19]. RAMs can also be used as random number generators. Random numbers are used in Monte Carlo simulations, and various security applications such as cryptographic key generation. Spin dice can be formed as random number generators using conventional STT MRAM, to generate an *m*-bit random number by concatenating *m* spin dies. Stochastic programming of the STT MRAM by current-driven STT are used as the die. The final state is sensed by passing a current through the MTJ.

In June of 2014, Purdue University proposed using a complementary polarizer STT MTJ structure for implementing an on-chip spin dice [20]. This consisted of a single free layer, a tunneling oxide layer, and two complementary polarized pinned layers. The complementary polarizer spin dice (CPSD) is shown in Figure 7.13.

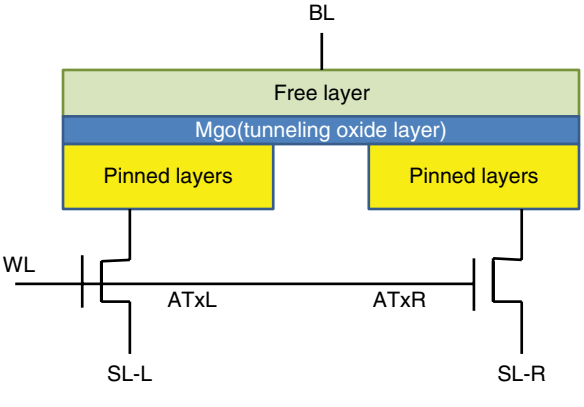


Figure 7.13 Schematic circuit diagram pf complementary polarizer spin dice (CPSD) consisting of a single free layer, a tunneling oxide layer, and two complementary polarized pinned layers. Based on J. Das *et al.* (University of South Florida), *IEEE Trans. on Nanotechnology*, May 2015 [19].

The resistance between BL and SLL and between BL and SLR depends on the magnetization of the fixed layer with respect to the pinned layers (PLs). Since the PLs connected to SLL and SLR are polarized in a complementary manner, the fixed layer magnetization is always parallel to one of them and anti-parallel to the other. "Rolling the Dice" is done by passing current from the BL to the SLR. There is a positive feedback loop that stabilizes the fixed layer magnetization during sensing and preserves the randomness. The primary contribution of power consumed is the power to reset and stochastically program the spin dice activation energy (E_a). E_a determines the frequency of random switching events in the MTJ due to thermal fluctuations since a lower E_a increases the frequency of random switching events. If the frequency of random switching events is high enough, the state may be randomized using thermal fluctuations alone, which eliminates the requirement for reset and programming. This would permit an energy efficient on-chip random number generator.

7.2.3 Using Flash Memory as a Physical Unclonable Function

In March of 2015, ETRI and KAIST discussed PUFs using Flash memory for IC authentication [21]. Inherent fluctuations in the characteristics of Flash memory devices can be used to produce different outputs, which can be used to identify particular devices. For example, a difference in programming and erasing efficiency comes from a widely distributed threshold voltage. Statistical fluctuations in the threshold voltage are an advantage for PUF applications. Various characteristics of PUFs are studied using fabricated Flash memories. Simulations were done to support the experimental results and to show that the unpredictability is induced by variations in the gate dielectric thickness.

PUF characteristics of gate all around (GAA) SONOS Flash memories were investigated. V_{read} was set to distinguish between two binary states under ordinary memory operation. These were the erased state and the programmed state. The experiment generated a PUF output and V_{read} was replaced by V_{puf}, which is the statistical median V_{th} of the programmed state. If $V_{th} > V_{puf}$, a logic state of "1" is extracted and if $V_{th} < V_{puf}$, a logic state of "0" is extracted. Since for even memory cells made in the same process and equipment, each memory cell has a different program/erase (P/E) efficiency, there are inevitably induced variations of V_{th}. This is illustrated in the schematic diagram in Figure 7.14, showing (a) ordinary operation of Flash memory and (b) the proposed memory operation for PUF circuitry.

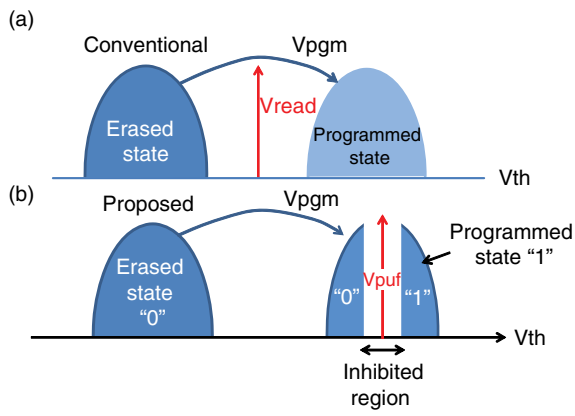


Figure 7.14 Diagrams of V_{read} versus V_{th} for the GAA SONOS Flash memories showing: (a) erase and programmed state distributions and (b) erased versus V_{puf} as the statistical median V_{th} of the programmed state. Based on M.S. Kim *et al.* (ETRI, KAIST), *IEEE Trans. on Nanotechnology*, March 2015 [21].

V_{th} is considered a good metric for judging physically unclonable outputs since it is easily and reliably measurable in Flash memory devices. The main characteristics of PUFs are: unclonability, uncontrollability, unpredictability, and robustness. These characteristics were each successfully tested for. The characteristics being used for the PUF is the thicknesses of the thin gate dielectrics in a SONOS memory structure. These inherent process variations are uncontrollable, unclonable, and unpredictable. They permit ΔV_{th} to vary by a factor of 1.5.

The V_{th} is therefore favorable for creating a PUF.

7.2.4 Other Memories used as Physical Unclonable Functions

In April of 2014, Nanyang Technical University discussed developing reconfigurable PUFs (RPUFs) using process parameter variability and programming sensitivity of phase change memory (PCM) [22]. One of the PCM-based RPUF methods extracted its challenge–response pairs from the measurable differences of the PCM cell resistances programmed by randomly varying pulses. An imprecisely controlled regulator was used to protect the privacy of the challenge–response pair in the case where the configuration state of the RPUF becomes known. The other PCM-based RPUF developed produces the random response by counting the number of programming pulses required to make the cell resistance converge to a predetermined target value. The merging of the challenge–response pair reconfiguration and evaluation reduces the vulnerability of the devices to malicious prediction attached by limiting the number of accessible challenge response pairs between two consecutive reconfigurations to one. Both methods of creating PUFs were evaluated using 180 nm PCM chips.

7.3 On-Chip Memory-Based Security Systems

7.3.1 Introduction to On-Chip Security Systems

On-chip security systems are critical for IoT applications. These systems are frequently implemented in a single chip. Retail intelligence systems can filter and analyze video surveillance data both for security and for marketing purposes. Small neurotechnology fingerprint recognition kits can run on low power microcontrollers and extract a fingerprint template using less than 128 kB of memory. Biometrics are routinely used to tie user identity to their mobile systems. Access control tablets use a combination of biometrics, keypads, intercoms, and cameras to identify access seekers. Facial recognition software is reaching higher levels of accuracy in smaller footprints. Automatic license plate recognition software captures high definition images of plates using a wide field of view and can capture multiple plates in a single image in high resolution. Research into the field of security is accelerating as we seek to define better information in smaller chips using less power.

7.3.2 Physically Secure Key and TAG Storage

An RRAM chip can be used for physically secure key storage. In April of 2014, SMIC and Fudan University discussed a RRAM chip for physically secure key storage. The chip had security features for resisting fully invasive attacks [23]. These features include

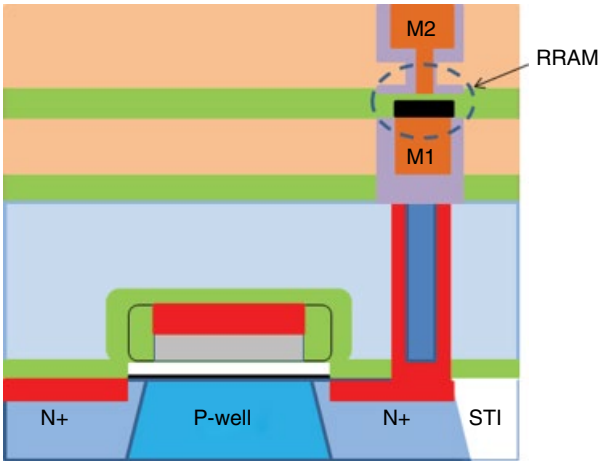


Figure 7.15 Schematic cross-section of the 1T1R CB-RAM cell formed in the metal layers above the transistor. Based on Y. Xie *et al.* (Fudan University and SMIC), *IEEE Trans. on Circuits and Systems*, April 2016 [23].

deprocessing and microscopy observation, resisting side-channel attacks by providing symmetrical power and timing read signals, resisting malicious writing by a reduced write protection method with feedback, and resisting data interception attack across a pin boundary by using on-chip integration with the logic platform. The chip was made in a 130 nm conventional CMOS logic process and used as the key storage for a demonstrative information security platform with a MIPS-based cryptoprocessor. The RRAM used was a 1T1R conductive bridge CB-RAM using CuxSiyO. A schematic cross-section of the 1T 1RRAM cell is shown in Figure 7.15.

Experiments were done of reverse engineering. These showed fully invasive attack resistant features. Since the key storage uses a conductive filament, which is randomly placed in the dielectric between the top and bottom electrodes, the small size of the filament makes it difficult to see so that the chip successfully resists fully invasive attacks.

Experiments emulating side-channel attacks showed no difference between 0 and 1. Side-channel attacks observe and analyze the leakage information such as power timing and electromagnetic signals of running systems to infer secret information. A 2T2R bit structure was used with one cell in the HRS and the other in the LRS. For the read cycle, the bias current generator produces equal read currents for IL and IR to the bitline pair simultaneously. The internal control signals are carefully delayed. The average power of the 2T2R bit and read method shows no obvious difference for a 0 and a 1. The sense amplifier's output was registered when not latched to provide the same timing for reading a 0 and a 1. The write circuits are protected from malicious writes by providing nonreversible state changes. Only the changes from HH to HL or LH are allowed and the reverse directions are prohibited. This protects the programmed RRAM bit from malicious writes. It was also shown by statistical analysis that no statistical correlation exists between the two sequences for "1" and "0" and that a power analysis attack cannot distinguish a 0 or a 1 by the read power signals.

It was experimentally shown that the information security platform could correctly encrypt and decrypt using the RRAM key storage. The chip has an advantage of a small

area and a lower write voltage and current (i.e., lower power) along with security features compared to embedded key storage using an anti-fuse.

In January of 2016, the University of New South Wales and University of Notre Dame discussed improving TAG generation for memory data authentication in embedded processor systems [24]. Data integrity can be protected by attaching an identifying tag to individual data. If the data is maliciously altered, the related tag becomes invalid and the attack will be detected.

A method of improving an existing tag design for memory data was discussed. The existing tag design is for memory data in embedded processor systems, where data that are stored in the memory or transferred over the bus can be tampered with and need to be authenticated before use. The design is cost effective and can counter a data integrity attack using random values; however, the design was found to be vulnerable if the fake data are not randomly selected. If the values chosen as fake data are from a reduced set then the data alteration might go undetected. This issue is discussed and a low cost enhancement to ensure full range distribution of the tag values is proposed. This system uses nonlinear randomization operations. The improved design achieves a full range of random distributions of tag values for all data, which increases the security level of the design against a chosen value attack. The improved randomness is achieved using bit flip and nonlinear Galous field multiplication operations. The bit-level flip functions, even with small blocks, greatly improved the tag's randomness for each individual data, providing similar security against the chosen value attack as against a random attack.

In November of 2016, Pennsylvania State University discussed the security challenges of emerging nonvolatile memories such as STT MRAM, which involve persistent data that can allow an adversary to retrieve sensitive information such as a password or cryptographic keys [25]. The vulnerability is due to the fundamental dependency of these memory technologies on environmental parameters such as magnetic fields and temperature, which can be used to tamper with data.

For example, for an STT MRAM, the polarization of the fixed layer is robust, but the polarization of the free layer could be toggled using either spin polarized current or a magnetic field. An AC magnetic field is of more concern since both storage polarities will be affected under an AC magnetic field. During a write operation, the AC field will either assist or suppress the threat. If the AC frequency is faster than the write time, it could affect writing both data polarities.

The relative polarity of the layers could also be affected by temperature, which can affect the pinning on the fixed layer. As the temperature rises the polarity could flip. The threat is to data privacy as sensitive data could endure unauthorized access. While a magnetic or heat shield can be built around the device, the cost and weight in various applications may not be suitable. Such applications include IoT and embedded cache in a processor. High and asymmetric write current and long and asymmetric write latency can serve as side channels, which can expose information such as the number of 1s and 0s in a memory word. An example of the read and write latency of an 8 MB STT MRAM cache with process variation is illustrated in Figure 7.16. The write latency of an STT MRAM is a function of the thermal stability factor which depends on the retention time. A longer read and write latency provides more opportunity for side channel attacks on data privacy. The STT MRAM is also susceptible to process variation, which increases the thermal stability of bits randomly particularly for larger arrays.

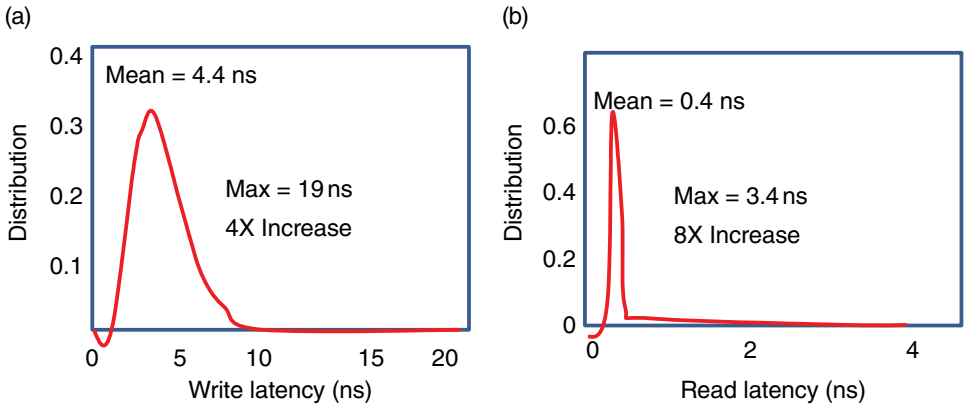


Figure 7.16 8MB STT MRAM cache with process variations: (a) write latency distribution and (b) read latency distribution. Long read/write latency gives a wide attack window for an adversary. Based on S. Ghosh *et al.* (Pennsylvania State University), ICCAD, November 2016 [25].

The stability of the MTJ free layer is a function of its volume so it is possible to enhance the robustness of the MTJ against tampering by increasing the size. This may improve stability in a low frequency field, but not provide protection in a high frequency field.

A sensor can be used to detect a magnetic field or thermal attack so corrective steps can be taken. Such a sensor can be a small replica of the STT MRAM array but designed to fail early as an alert signal. Such a sensor can capture the angle and frequency of the magnetic attack` and also the high temperature thermal attack. Sensors can be distributed through the array. A combination of sensors and ECC can be used for further insurance against attack. A variable strength ECC was suggested.

7.3.3 Face and Feature Detection in Security Systems

Face detection is a critical feature of many security systems. While there has been extensive work on face detection, detecting a large number of faces with different scales and variations, such as pose, expression, or illumination, makes this a computationally expensive classification algorithm. In May of 2016, Arizona State University discussed a number of acceleration techniques for enabling this capability on a CPU-FPGA platform [26]. A state-of-the-art face detection algorithm was used that uses a large number of simple classifiers.

The algorithm is first mapped using the integrated OpenCl environment for FPGA. A nested architecture is used to speed up the memory access and the computing iterations. This multilayer architecture distributes parallel computing cores with the memory. The core size and number of cores are then further optimized to achieve read-time face detection. The architecture of the face detection model used is shown in Figure 7.17. It consists of scaling, channel generation with 10 channels, integral channel computation with 10 integral channels, classifier output computation with 2000 weak classifiers times five templates, and nonmaximum suppression states to suppress redundancies.

This model uses the rigid templates-based detector, which achieves high precision and recall rates compared to other models. It is accelerated in the CPU/FPGA to achieve

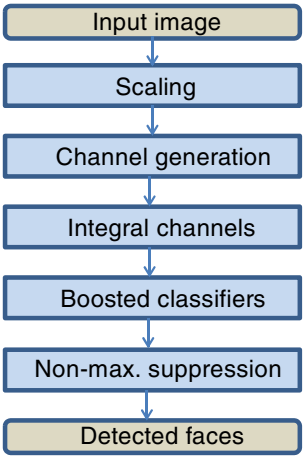


Figure 7.17 Architecture of a state-of-the-art face detection model. It includes: scaling, channel generation, integral channel computation, classifier output computation, and nonmaximum suppression states to suppress redundancies. Based on A. Mohanty (Arizona State University), ISCAS, May 2016 [26].

real time detection rates without accuracy degradation. The classifier stage is composed of five trained models each with 2000 weak classifiers. The time taken by the CPU for computation steps in the face detection algorithm by the weak classifiers computation consumes as much as 91.5% of the total computation time. FPGAs are thought to be energy efficient hardware accelerators and the classifier acceleration was done on the FPGA. The noncritical stages were done on the host CPU. A conventional face detection model is used. The input features used were simple rectangular pooling regions. The software was OpenCL.

7.3.4 Security in Embedded Systems

Security in embedded systems is a critical issue. Such systems include many IoT applications such as wearable devices, portable medical instruments, and gaming devices. Security is an issue due to the presence of private and critical information about users in these systems. The main challenge to security in embedded systems is resource constraints, which include limited computation capability, battery life, and data storage. A major susceptibility of these systems is the use of unsafe languages such as C and C++. Runtime attacks, such as buffer overflow-based stack smashing and code reuse attacks, are common. In November of 2016, Nanyang Technological University and Hong Kong University of Science and Technology discussed runtime attacks on memory such as buffer overflow stack smashing and code reuse attacks in embedded systems [27]. An approach was discussed of enforcing fine grained control flow integrity at a basic block level to defend against such attacks. The goal is to verify the target address of the control flow instruction, which can be modified in an attack. There are two stages to such an attack. These are off-line profiling of the program intended to extract control flow information and runtime control flow checking to use the extracted information. An architectural design of a control flow checker was used, which monitors the program execution during runtime. The concept was implemented in an FPGA. The results showed that this technique was effective against runtime attacks with 100% verification accuracy. There was less than 1% performance overhead when implemented on an FPGA and dynamic power consumption was 78 mW.

References

1. Chen, A. (2015) Utilizing the variability of resistive random access memory to implement reconfigurable physical unclonable functions (Global Foundries), *IEEE Electron Device Letters*, 36 (2), February 2015.
2. Chen, P.Y. *et al.* (2015) Exploiting resistive cross-point array for compact design of physical unclonable function (Arizona State University), IEEE HOST, May 5, 2015.
3. Balatti, S. *et al.* (2015) True random number generation by variability of resistive switching in oxide-based devices (Politec. di Milano). *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, 5 (2), June 2015.
4. Chen, A. (2015) Comprehensive assessment of RRAM-based PUF for hardware security applications (TD Research, GlobalFoundries), IEDM, December 2015.
5. Rui Liu *et al.* (2015) Experimental characterization of physical unclonable function based on 1kb resistive random access memory arrays (Arizona State University, Tsinghua University). *IEEE Electron Device Letters*, 36 (12), December 2015.
6. Balatti, S. *et al.* (2016) Physical unbiased generation of random numbers with coupled resistive switching devices (Politechnico di Milano, Micron Technology). *IEEE Trans. on Electron Devices*, 63 (5), May 2016.
7. Shrivastava, A. *et al.* (2016) Design of a reliable RRAM-based PUF for compact hardware security primitives (Arizona State University), ISCAS, May 2016.
8. Yoshimoto, Y. *et al.* (2016) A ReRAM-based physically unclonable function with bit error rate <0.5% after 10 years at 125 °C for 40 nm embedded application (Panasonic), VLSI Technology Symposium, June 2016.
9. Govindaraj, R. and Ghosh, S. (2016) A strong arbiter PUF using resistive RAM (University of South Florida), SAMOS, July 2016.
10. Govindaraj, R. and Ghosh, S. (2016) A strong arbiter PUF using resistive RAM within 1T1R memory architecture (University of South Florida, Pennsylvania State University), ICCD, October 2016.
11. Pang, Y. *et al.* (2017) Optimization of RRAM-based physical unclonable function with a novel differential read-out method (Tsinghua University). *IEEE Electron Device Letters*, 38 (2), February 2017.
12. Zhang, L. *et al.* (2014) Highly reliable memory-based physical unclonable function using spin-transfer torque MRAM (Nanyang Technical University, Purdue University), ISCAS, pp. 2169, June 1, 2014.
13. Das, J. *et al.* (2014) A novel geometry based MRAM PUF (University of South Florida), IEEE-NANO, August 18, 2014.
14. Marukame, T. *et al.* (2014) Extracting physically unclonable function from spin transfer switching characteristics in magnetic tunnel junctions (Toshiba). *IEEE Trans. on Magnetics*, 50 (11), November 2014.
15. Das, J. *et al.* (2015) MRAM PUF: A novel geometry based magnetic PUF with integrated CMOS (University of South Florida). *IEEE Trans. on Nanotechnology*, 14 (3), February 2015.
16. Vatajelu, E.I. *et al.* (2015) STT MRAM-based PUFs (Politecnico di Torino), DATE, March 9, 2015.
17. Zhang, L. *et al.* (2015) Optimizing emerging non-volatile memories for dual-mode applications: Data storage and key generator (Nanyang Technical University, Purdue University). *IEEE Trans. on CAD of IC and Systems*, April 28, 2015.

18 Zhang, L. *et al.* (2015) Highly reliable spin-transfer torque magnetic RAM based Physical unclonable function with multi-response-bits per cell (Nanyang Technical University). *IEEE Trans. on Information Forensics and Security*, April 10, 2015.

19 Das, J. *et al.* (2015) MRAM PUF: A novel geometry based magnetic PUF with integrated CMOS (University of South Florida). *IEEE Trans. on Nanotechnology*, 14 (3), May 2015.

20 Fong, X. *et al.* (2014) Generating true random numbers using on-chip complementary polarizer spin-transfer torque magnetic tunnel junctions (Purdue University), DRC, June 22, 2014.

21 Kim, M.S. *et al.* (2015) Investigation of physically unclonable functions using Flash Memory for integrated circuit authentication (ETRI, KAIST). *IEEE Trans. on Nanotechnology*, 14 (2), March 2015.

22 Zhang, L. (2014) Exploiting process variations and programming sensitivity of phase change memory for reconfigurable physical unclonable functions (Nanyang Technical University, Singapore). *IEEE Trans. on Information Forensics and Security*, 9 (6), April 4, 2014.

23 Xie, Y. *et al.* (2016) A logic resistive memory chip for embedded key storage with physical security (Fudan University and SMIC) *IEEE Trans. on Circuits and Systems*, 63 (4), April 2016.

24 Liu, T. *et al.* (2016) Improving TAG generation for memory data authentication in embedded processor systems (University of New South Wales, University of Notre Dame), ASP-DAC, January 2016.

25 Ghosh, S. *et al.* (2016) Security and privacy threats to on-chip non-volatile memories and countermeasures (Pennsylvania State University), ICCAD, November 2016.

26 Mohanty, A. (2016) High-performance face detection with CPU-FPGA acceleration (Arizona State University), ISCAS, May 2016.

27 Das, S., Zhang, W. and Liu, Y. (2016) A fine-grained control flow integrity approach against runtime memory attacks for embedded systems (Nanyang Technological University, Hong Kong University of Science and Technology), DAS, November 2016.

Index

a

activation energy (*Ea*) **315**
adaptive learning 225
ADAS (advanced driver assistance systems) 61–66, 143
AES (advanced encryption standard) **29**, 47, 98
a-IGZO (amorphous-InGaZnO) 173, **190**, 192
a-InGaZnO (amorphous InGaZnO) 172, 189
AMI (advanced metering infrastructure) **11**
analog 222
analog synapse 228, 231
ANN (artificial neural networks) 261, 279–280, 285, 287–288
ANT 48, 59
antifuse 127, **128**
artificial intelligence 3, 271
artificial neural net 271, 276, 279
associative learning 254
associative memory *see* CAM (content addressable memory)
ATD (address transition detection) 135
atom switch 91
attack window 319
automotive 61, 118, 122, 125, 138–140
  body processors 8, 70
  ECU (engine control unit) 141
  infotainment 69
  power train 139
  specifications 122, 136–140
  under-the-hood 62, 66, 68
autonomous cars 8, 64
AWB (automatic write-back) 312, 314

b

back propagation 227
bar code 1, 41, 178 *see also* UPC (Universal Product Code)
battery operated 27–29
BBHEI (band-to-band hot electron injection) 105–107, **135**, 140
beacon 3, 43, 75
bending 198, 210
bending radius 173
BEOL (back-end-of-the-line) 160, 186
Big Data 78
binary neural net 260
binary synapse 247
biological synapse 225, 229
BIOS (basic input/output system) 93
bit-by-bit write 111
bit-error rate 312, 314
BL (bitline) 103
Bluetooth 2, 42–45, 52, 57–59
BSN (body sensor node) **49**, 50
BTBT (band-to-band tunneling) 126
BTBT HHE (band-to-band tunneling hot hole erase) 137
BTBT HHI (band-to-band hot hole injection) 143, 147

Memories for the Intelligent Internet of Things, First Edition. Betty Prince and David Prince.
© 2018 John Wiley & Sons Ltd. Published 2018 by John Wiley & Sons Ltd.

c

CAM (content addressable memory) **271**
  TCAM (ternary CAM) **271**, **272**
CAN 24, 66, 70
CB-RAM (conductive bridge RAM) **38**, 49–51, 59–60, 193, 224–225, 233–234, 245, 317
  CAS (copper atom switch) 181, 184–185, 187
CD (contrastive divergence) 287
CESL (contact etch stop liner) 99
CF (conductive filament) 221–222, 309
CG (control gate) 104, 106
charge trapping flash **19**, 65, 68, 91, 133, 169
  BE-SONOS **151**
  MANOS (metal-Al_2O_3-nitride-SiO_2-silicon) 155
  MONOS (metal-oxide-nitride-oxide-silicon) **68**–69, 76, 133, 136
  Schottky barrier charge trapping cell 152
  SG-MONOS (split gate MONOS) **124**, 133, 137, 143
  SONOS (silicon-oxide-nitride-oxide-silicon) 133–135, 145–146, 315
checkpointing 25
CHEI (channel hot electron injection) 93, 113, 120, 135, 147
classification 260
cloud 177, 271, 278, 296
CMP (chemical mechanical polishing) 126, **141**, 235
CNT (carbon nanotube) memory 200
coin cell batteries 6, 23, 28
communications protocols *see* Bluetooth; NFC (near field communications); RFID (radio frequency identification); SPI; USB (universal serial bus); Zigbee
compliance current 224
conductance 227
  change 259
copolymer 169
CPSD (complementary polarizer spin dice) 314
crosspoint (bar) array 182, 235, 242, 244, 253, 255, 258, 262, 277–279, 282, 288, 308–309
  pseudo-crossbar array 283
CRS (complementary resistive switching) 277
cryptographic
  hard modules 308
  key 303, 318
  primitives 314
  security 178
cryptoprocessor 60–61, 77, 317

d

data integrity 303
  attack 318
datalogging 20, 51, 61
data server 78
DCNN (deep convolutional neural network) 288–290
DDC (deep depleted channel) **130**
deep computers 271
deep learning **271**, 281, 284–286, 296–297
depression **221**, 224–225, 228–229, 239, 244, 248, 254
DIBL (drain induced barrier lowering) 104
dictionary learning 258
dielectric constant 174
dielectric HiK 102
DLA (deep learning accelerator) 297
DNA sequencing 298
DNN (deep neural nets) 293, 297
DNPU (deep neural processing unit) 290, 292–293
drone **3**
DRSL (data retention storage life) 122
DSP (digital signal processors) 288–290

e

ECC (error correction code) 70, 308–309, 312, 319
ECG (electrocardiograph) 52–54
ECU (electronic control unit) 68
EEPROM (electrically erasable programmable read only memory) 18–19, **41**, 77–78, 104

automotive grade 114–115
poly-to-poly erase 113
single polysilicon 18–19, **91**–92, 96–97, 102–105
source-side-injection 113
split gate 113
eFlash (embedded Flash) memory 18
code flash 69, 115, **137**, 139, 141
corner enhanced poly-to-poly 115
data flash 69, 115, **137**, 139, 141
double poly 18, 56, 130
poly-to-poly erase 112–113, 117
SCE (split gate Flash memory configuration element) **117**–118
single poly 18, 56, 96
split gate **112**, 116–117
stacked gate 112
triple poly 117
EG (erase gate) 124–125
electronic label 175
electronic paper 181
ELM (extreme learning machine) 284–285
embedded memory 21
emerging memory **19**
EMI (electromagnetic interference) 63, 140–141
energy band diagram 230
energy harvesting 6, 17, 23–24, 26, 37, 49
engine control units 8
engine processors 8
EOT (effective oxide thickness) **146**, 151
EPC (electronic product code) **41–42**
error correction codes *see* ECC (error correction code)

f

face detection 298, 303, 319–320
face recognition 226–227, 295–296, 316
FD-SOI (fully depleted silicon on insulator) 62, 288–289, 293
feature detection 319
FE (embedded ferroelectric) capacitors 180
FeFET (ferroelectric field effect transistor) 25, 174
TCAM 276
FeMEM (ferroelectric memory) 250–251
graphene-ferroelectric memory 170–172
imprint 172
organic ferroelectric memory 169–170, 175
P(VDF-TrFE) copolymer 170–172, 174, 180
P(VDF-TrFE-CFE) terpolymer 170–171
PZT 251
FeRAM (ferroelectric RAM) 17, 20, 24–25, 29–32, 34, 42, 49–54, 72, 74, 77, 91, 169–170, 174, 221, 250–251
2T2C FeRAM 173–174
ferroelectric TER (tunnel electroresistance) 181
ferroelectric TFT (thin film memory transistor) 179–180, 253
FG (floating gate) 105
filamentary RRAM 237
FinFET 56, 103, 142–143, 234
fingerprint recognition kit 316
fire event 241
flexible electronics 169, 173–174, 190, 209, 215, 264
flexible substrate 181, 187, 190, 208
flex strength 214
flipchip 207–208
FLOTOX 130
FN (Fowler-Nordheim) tunneling 93, 96, 101, 105, 145
FO-WLP (fan-out wafer level packaging) 210–213
FPGA (field programmable gate array) 103, **185**, 320
FTJ (ferroelectric tunnel junction) 181, 252–253
FTM (ferroelectric tunnel memristor) 252–253
fuse 19, 128

g

GAA (gate-all-around) 152, 207, 315
GPU (graphics processing unit) 261, 286
GQD (graphene quantum dots) 200–201

h

half cell 100
half MOS 100–102
hamming distance 303–304, 307–308, 310
hardware security 305, 307
HBM (high bandwidth memory) 79
HBM (homogeneous barrier modulation) RRAM 229
HHI (hot hole injection) 135
HKMG (Hi-K metal gate) 128, 147
HRS (high resistance state) 198
HTOL (high temperature operating life) 122
HTS (high temperature storage) 213
HV (high voltage) 112
HVCMOS (high voltage CMOS) 106–107
hybrid neuromorphic **222**

i

IGZO (InGaZnO) 172, 180, 189–191, 201
IIoT (Intelligent Internet of Things) 1, **17**
image recognition 286
imprint time (FeRAM) **170**
IMT (insulator-metal transition) 244
IMT oscillator neuron 244–246
infotainment processors 8, 62, 65
inkjet printing 169, 174, 176, 197, 201
integrate and fire 251, 255
intelligent homes **4**
intelligent hospitals **13**
Internet edge 296
IoT (Internet of Things) 1, 146
ISM (industrial, scientific and medical) 46
ITO (indium–tin–oxide) 194–196
IZO (indium–zinc oxide) synaptic transistor 263

k

k–1000 21
K–1024 92
key storage 316, 317
key word spotting 298

l

learning rule
  winner take all 228
LIF (leaky integrate-and-fire) 256–257
LIN interconnected network 65, 70
logic eNVM (logic embedded non-volatile memory) 99
LRS (low resistance state) 198, 224, 228, 234
LTD (long term depression) 225
LTP (long term plasticity) 249
LTP (long term potentiation) 225, 250
LUT (look-up tables) 276, 290
  nvLUT (nonvolatile LUT) 276

m

machine learning 281
magnetic attack 319
main memory 70
malicious writes 317
matrix multiplication 262
MFF (magnetic flip-flops) 36
MCU (microcontroller unit) 17
memory based security systems 316
memory hierarchy 78, 80
memristor 250
MO-RRAM (metal oxide resistive RAM) 221–224 *see also* ReRAM (resistive RAM)
Miller-Abrahams hopping **195**
MIM (metal-insulator-metal) 174, 184, 197, 221
MLC (multi-level cell) 188, 224, 272
MNIST handwritten digits 227, 242, 261, 280–281, 284, 287–288, 292
MONOS *see* charge trapping flash
MO-RAM (metal oxide RRAM) 221, 224–225, 310
MOSFET (MOS field effect transistor) 103
motor control 75, 77
MRAM (magnetic RAM) 17, 35–36, 49, 62, 275
  CAM (content addressable memory) 276
  MTJ (magnetic tunnel junction) 35–36, 273, 276, 279–281, **312**, 313

multilevel states 280
PMA (perpendicular magnetic anisotropy) 274
spin hall (orbit) MRAM 275
spin synapses 279
spintronic device 280
spintronic neuron 280
spintronic PUF 303
stochastic memristive synapse 279
STT (spin-torque-transfer) MRAM 19, 36, 49–51, 279
TCAM MTJ 273, 276
MRAM PUF 303, 311–315
MTJ (magnetic tunnel junction) *see* MRAM (magnetic RAM)
MTP (multitime-programmable) 56, 99, 101–102, 127–128
multi-level characteristics 222, 225
multi-level program 111, 195

n

nanoparticle memory *see* NC (nanocrystal) flash
nanowire 207
NC (nanocrystal) flash 158, 169, 194, 200–205
synaptic organic 264
NDR (negative differential resistance) **230**
neural network 250, 258, 286–287
CNN (convolutional neural network) 261, 290–291, 293, 295–296
DNN (deep neural network) 261, 291, 297
FC-DNN (fully connected) 261, 291
RNN (recurrent neural net) 290
neuromorphic array 232
neuromorphic computer 23, 222, 254
neuromorphic computing 221, 236, 239, 242, 257–258, **259**, 260–261
neuromorphic system 228, 242, 244, 247, 260
neuron 221, 232–233, 241, 244, **280**
neuron network 243, 253
NFC (near field communications) 2, **42**–44, 51–53, 56–57, 117, 177–179
nonfilamentary RRAM 236
nonlinear randomization 318
NOR Flash 66, 73, 120, 144
nvDFF (nonvolatile data flip-flop) 32–33
nvFF (nonvolatile flip-flop) 32–33, 35, 37–40, 180
NVL (nonvolatile logic) 32–35
NVM (nonvolatile memory) 18, 87
NVP (nonvolatile processor) 39–40
nvSRAM (nonvolatile SRAMs) 38, 39, 103

o

ONO (oxygen-silicon-nitride) 106
organic circuits (NAND, NOR, SRAM, FF) 175
organic FeFET 173
organic semiconductor 176
organic synapse 266
OTP (one-time-programmable) 18, 127–128, 160
OxRAM (metal oxide RAM) 29
oxygen vacancies 309
oxygen vacancy migration 237–238

p

pattern recognition 224, 250, 252, 254–255
Pb(Zr,Ti) 25, 26, 251
PCB (printed circuit board) 31
PCM (phase change memory) 17, 79, 91, **225**, 240, 254–256, 261, 286–288
PCMO 245, 262
PCSA (precharge sense amplifier) 274
pentacene 176
perceptron 227, 242, 261, 287–288
network 288
PIP (poly-insulator-poly) 138
PML (product markup language) **41**
PMMA (poly methyl methacrylate) **253**
polyimide substrate 190
polymer memories 169, 262
poly-to-poly erase *see* eFlash (embedded flash) memory
Poole-Frenkel emission 195, 238
PoP (package-on-package) 210, 214
potentiation **221**, 224–225, 228–229, 239, 244, 248, 254
power analysis attack 317

power gating 32, 35–36
PPROM (print-programmable ROM) 176
printed memory 177–178
program disturb 118
  immunity 119
PROM (programmable read only memory) **41**
PSE (polymer solid electrolyte) 184, 186–187
PUF (physical unclonable functions) **303–304**, 305–316
  challenge-response pairs 305
  Flash memory PUF 315–316
  geometry based 313
  primitive 304
  RPUF (reconfigurable PUF) 316
  uniqueness 304
PVD (physical vapor deposition) 202p
PVDF (polyvinylidene flouride) 25–26
P(VDF-TrFE) 170, 173
P(VDF-TrFE)-PMMA **253**, 254
PVT (process, voltage, and temperature conditions) 122
PZT 25 *see also* Pb(Zr,Ti)

q

QR Code **1**, 177–178
Quick Recall 24

r

RAID servers 278
random bit generation 314
random variability 314
RBM (restricted Boltzman machine) 271, 284, **285**, 286–287
redistribution layer 214
remnant polarization (ferroelectric) 171
RF (Radio frequency) 46
RFID (radio frequency identification) 2, **41**–42, 60, 66, 97–99, 169, 175, 209
RF noise (radio frequency noise) 139
RIE (reactive ion etching) 228
RNG (random number generator) 305, 307–308, 314
roll-to-roll 174, 179
ROM (read only memory) **41**
RRAM (resistive RAM) 17, 26, 29–30, 37–38, 59, 91, 193, 195, 221–222, 240, 242, 258, 262, 278, 286
  AZO (Al-doped ZnO) 195
  flexible substrate 187
  graphene oxide 195
  on IGZO 189, 192
  inorganic RRAM 188
  MO-RRAM (metal oxide RRAM) **221**, 222, 235
  nanowire RRAM 193
  organic neuromorphic 263, 264
  parylene-C (polychloro-*para*-xylylene-C) 182 *see also* CB-RAM (conductive bridge RAM)
  PCMO (PrCaMnO3) RRAM 226–228, 244
  PEDOT:PSS polymer 194
  polycrystalline thin film 193
  polymer RRAM 169, 181, 184
  PUF 303, **304**, 305–311
  synaptic device 224, 227
  TCAM 271–272, 276
RTN (random telegraph noise) 305, 307
runtime attacks 320

s

SA (sense amplifier)
  single ended 115
SAN (self aligned nitride) 148–149
search engines 271–272, 276
security 303, 304
security challenges, STT MRAM
  magnetic field 318
  temperature 318
security primitives **304**, 308
SCE (short channel effects) 119
SCE (split gate Flash memory cell configuration element) 117–118
screen print 176
secure key 303, 314
self aligned process 114
sensors 24
server 79
  server farms 271, 278

SG (select gate) 104, 106
SG-MONOS *see* charge trapping flash
SHA (secure hash algorithm) 47, **308**
SHE (secure hardware extension) 70
side channel attacks 317
SILC (Stress induced leakage current) 130, 159
silicon nanowire arrays 193 *see also* RRAM (resistive RAM)
SIMOX (separation by implantation of oxygen) 96
single wire connectivity 48
SiP (single-in-line package) 210
SL (source line) 103
SLA (switching logic array) 118
smart automobiles 6–7
smart card 22, 77, 117, 122, 146
smart city 1–**2**
smart electrical grid **12**
smart energy network **11**
smart factories **13**
smart highways **10**
smart homes 4, 74, 75
  environmental sensors 75
smart hospital **14**
smart labels 177–179
smart meter **10**–11, 71–74
smart office buildings **12**
sneak path current 232, 242
SoC (system on chip)
SOI (silicon-on-insulator) wafers 96, 159, 207, 208
solar cell 26
solid electrolyte RAM 225 *see also* CB-RAM (conductive bridge RAM)
SONOS *see* charge trapping flash
SoP (system on polymer) 209
speech recognition 222, 286, 293
SPI 2, 44, 46, 60, 72
spiking neural nets 239, 251, 254
spin-RAM *see* MRAM (magnetic RAM)
spintronic *see* MRAM (magnetic RAM)
  PUF 303
split gate *see* eFlash (embedded flash) memory, split gate
SS (subthreshold swing) 190
SSHEI (source side hot electron injection) 116–117, 148
stacked gate 112
STDP (spike-timing dependent plasticity) 221, **225**, 239–242, 245–248, 252–257, 287
STI (shallow trench isolation) 119
stochastic switching 305–306
storage class memory 79, 258
STN (strontium titanate nickelate) 198–199
STP (short term plasticity) 249–250
STS (spin transfer switching) 312
STT-MRAM *see* MRAM (magnetic RAM)
STT-PUF 312
supervised learning 286
synapse 221, 243
  analog 222
  binary 222
synaptic connections 245
synaptic devices 257–258
synaptic modeling (RRAM) 236
synaptic network 243
synaptic plasticity 221, 230, **246**, 253, 258
synaptic training 259
synaptic weight 249, **253**, 258, 261, 279, 281–282, 284, 287, 297
  learning rules 221

t

tag generation 318
TCAM (ternary content addressable memory) 23, 78, **271**, 273–274, 276
  nvTCAM (nonvolatile TCAM) 272–274
TDDB (time dependent dielectric breakdown) 139–140, 143
tensile strain 173
TER (tunneling electroresistance) 181
terpolymer 169
testing RRAMs 236
TFT (thin film transistor) 172–173, 179–180
  CdS 174
thermal attack 319
thermal cross-talk **235**
thermal stability factor 318

thin chips 208
by cavity formation 209
by SOI 208
3D vertical 231, 234
transconductance (Gm) 100, 101, 147
transfer process (chip integration)
transparent TFT 191
TPC (transiently powered computers) 24–25
TSV (through silicon via) 23, 127, 214, 226, **231**
tunnel gap variation 222

u

UART (universal asynchronous receiver/transmitter) 42
ULP (ultralow power) 24, 33, 56
UPC (Universal Product Code) *see* bar code
USB (Universal Serial Bus) 24, **47**

v

vacancies (oxygen) **222**, 224
VAD (voice activity detection) 293–294
variable resistance memories 239
Verilog **253**
vision recognition 222, 234, 259, 294
vision sensors 254
VRRAM (vertical pillar RRAM) 232, 234

w

wearable electronics **5**–6, 49, 209, 264
WiFi **2**, 42, 46
WLAN (wireless local area network) 2
WLP (wafer level packaging) 210
WORM (write once read many) memory 176

z

Zigbee 47, **48**, 52, 75