Grids in Very Large Scale Integration Systems

by

Albert Çiprut

Submitted in Partial Fulfillment

of the

Requirements for the Degree
Doctor of Philosophy

Supervised by
Professor Eby G. Friedman

Department of Electrical and Computer Engineering
Arts, Sciences and Engineering
Edmund A. Hajim School of Engineering and Applied Sciences

University of Rochester
Rochester, New York
2019

Dedication

This work is dedicated to my mother, Jacqueline Gabay Çiprut, and in memory of my father, Erol Çiprut.

Table of Contents

Biographical Sketch **ix**

Acknowledgments **xii**

Abstract **xiv**

Contributors and Funding Sources **xvi**

List of Tables **xvii**

List of Figures **xviii**

1 Introduction **1**

1.1 My Thesis 1

1.2 Motivation 2

1.3 The Grid 4

1.4 Outline 8

2 Grids in Integrated Systems **12**

2.1 Density 14

2.2 Regularity 18

2.3 Path Diversity 22

2.4 Computational Efficiency 26

2.5 Summary 28

3 Non-Volatile Resistive Crossbar Arrays **30**

3.1 Non-Volatile Resistive Devices 32

3.1.1 Device Parameters 37

3.1.2 Applications 39

3.2 Non-Volatile Resistive Memory System 41

3.3 Challenges 44

3.3.1 Write Operations 44

3.3.2 Read Operations 46

3.3.3 Selectors 48

3.4 Summary 51

4 On-Chip Power Delivery with Fully Integrated Voltage Regulators **52**

4.1 Low Dropout Voltage Regulator 59

4.2 Fully Integrated On-Chip LDO Regulators 61

4.3 Stability Analysis 63

4.3.1 Effect of Output Capacitance on LDO Stability 66

4.3.2 Improving LDO Performance 69

4.4 Summary 70

5 Modeling Size Limitations of Resistive Crossbar Array With Cell Selectors 73

5.1 Models of Crossbar Array Design Parameters 75

5.1.1 Driver size 76

5.1.2 Voltage degradation across selected cell 79

5.1.3 Read margin 84

5.2 Enhancement of Nonlinearity Factor 87

5.2.1 Driver size 89

5.2.2 Voltage degradation across selected cell 90

5.2.3 Read margin 92

5.3 Design Requirements for Varying Array Size 93

5.3.1 Driver Resistance 94

5.3.2 Voltage Degradation and Device Nonlinearity 95

5.3.3 Read operation 96

5.4 Design Of A Crossbar Array Based On These Models 98

5.5 Summary 100

6 Energy Efficient Write Scheme for Non-Volatile Resistive Crossbar Arrays with Selectors **102**

6.1 Write Operations 106

6.2 Energy Models 108

6.2.1 Energy Efficient Bias Scheme 113

6.2.2 Impact of Nonlinearity Factor 116

6.2.3 Write Pulse Width 117

6.3 Energy Efficienct Hybrid Write Scheme 119

6.3.1 Optimal Choice of Bias Scheme 121

6.3.2 Overhead and Challenges 126

6.4 Summary 128

7 Stability of On-Chip Power Delivery Systems with Multiple Low Dropout Regulators **131**

7.1 Stability of Parallel Connected LDOs 133

7.2 Existing Work 136

7.3 Evaluating the Stability of Multiple LDOs 138

7.3.1 Effect of Number of LDOs on Grid Stability 140

7.3.2 Source of Instability 142

7.3.3 Degradation of Resonant Frequency 146

7.3.4 Condition for Stability 151

7.4 Effect of Design Parameters on Grid Stability 154

7.4.1 LDO Design Parameters 155

7.4.2 Power Grid Parameters 161

7.5 Summary 165

8 Distributed Pass Gates in Power Delivery Systems with Digital Low Dropout Regulators 167

8.1 Distributed Pass Gates 169

8.1.1 Grid Centroid 170

8.1.2 Proposed Heuristic 172

8.2 Power Grid Analysis 176

8.2.1 Distribution Topologies 176

8.2.2 Comparison of Pass Gate Distribution Topologies 178

8.3 Summary 184

9 Conclusions 186

10 Future Work 190

10.1 RowHammer Effect 191

10.2 RowHammer Effect in Nonvolatile Resistive Arrays 192

10.3 Proposed Research Direction 194

10.4 Summary 196

Bibliography **198**

Appendix A Derivation of Switching Energy Consumption **216**

Appendix B Off-chip Power Delivery Network **218**

Biographical Sketch

Albert (Avi) Çiprut received the Bachelor of Science degree in Electronics Engineering from Sabanci University, Istanbul, Turkey in 2013, and the Master of Science degree in Electrical Engineering from the University of Rochester, Rochester, New York in 2016. He was an intern with the Power Team, Google Inc., Mountain View, California in 2016. His current research interests include memory systems based on emerging memory technologies, on-chip power delivery systems, and electronic design automation. The following publications are a result of the work conducted during his doctoral study.

Journal papers

- A. Ciprut and E. G. Friedman, "Energy Efficient Write Scheme for Non-Volatile Resistive Crossbar Arrays with Selectors," *IEEE Transactions on Very Large Scale Integration Systems*, Vol. 26, No. 4, pp. 711–719, April 2018.

- A. Ciprut and E. G. Friedman, "Modeling Size Limitations of Resistive Crossbar Array with Cell Selectors," *IEEE Transactions on Very Large Scale Integration Systems*, Vol. 25, No. 1, pp. 286–293, January 2017.

- A. Ciprut and E. G. Friedman, "Stable On-Chip Power Delivery Systems with Multiple Low Dropout Regulators," *IEEE Transactions on Very Large Scale Integration Systems* (in press).

- A. Ciprut and E. G. Friedman, "Distributed Pass Gates in Power Delivery Systems with Digital Low Dropout Regulators," *IEEE Transactions on Very Large Scale Integration Systems* (in submission).

Conference papers

- A. Ciprut and E. G. Friedman, "Hybrid Write Bias Scheme for Non-Volatile Resistive Crossbar Arrays," *Proceedings of the IEEE International Symposium on Circuits and Systems*, pp. 1–5, May 2018.

- A. Ciprut and E. G. Friedman, "On the Write Energy of Non-Volatile Resistive Crossbar Arrays with Selectors," *Proceedings of the IEEE International Symposium on Quality Electronic Design*, pp. 184–188, March 2018.

- A. Ciprut and E. G. Friedman, "On the Stability of Distributed On-Chip Low Dropout Regulators," *Proceedings of the IEEE International Midwest Symposium on Circuits and Systems*, pp. 217–220, August 2017.

- A. Ciprut and E. G. Friedman, "Design Models of Resistive Crossbar Arrays with Selector Devices," *Proceedings of the IEEE International Symposium on Circuits and Systems*, pp. 1250–1253, May 2016.

Patent

- A. Ciprut and E. G. Friedman, "Energy Efficient Write Scheme for Non-Volatile Resistive Crossbar Arrays with Selectors," 2017 (disclosed).

Acknowledgments

In this section, I would like to thank and express my gratitude towards a number of people who have made my Ph.D. experience worthwhile and supported my efforts in realizing the works embodied in this dissertation. First, I thank my advisor and mentor Professor Eby G. Friedman for giving me the chance and the opportunity to pursue this intellectual journey under his supervision. His teachings significantly influenced my thought process, having a major positive effect both on this dissertation and me as a person. As time goes by, I realize how lucky I am to have worked with him.

I thank Professor Engin Ipek, Professor Zeljko Ignjatovic, and Professor Marc D. Porosoff for serving on my proposal and defense committees. Their valuable feedback has improved this dissertation. I would also like to express my gratitude to Professor Yonathan Shapir for serving as the chairperson in my defense committee.

I thank Gregory Sizikov for providing the opportunity to work with him at Google. This internship experience taught me the value and importance of working with people like Gregory who can retain a continuous positive attitude that motivates and empowers.

I would like to thank my colleagues (in no particular order): Tahereh, Abdu, Gleb, Kanisha, Rassulisha, Boris, Alex, Inna, Ravi, and Mohammad. These people were like a second family to me here in Rochester. In addition, I would like to thank Ruth Ann Williams, for the amazing blueberry cakes as well as the interesting conversations and viewpoints that she has shared with us.

I thank many friends for the great moments I get to keep forever. After all, it's all about the moments. Your friendship has carried this Ph.D./Rochester experience to a whole new level of amazingness and I thank you all for that.

I thank my family, especially my mother and my little brother, for their continuous support throughout my lifetime. Your love was a critical source of fuel that kept me going through this challenging experience. I can't imagine how it would have been without your presence.

Lastly, I want to express my deep gratitude to this country, the United States of America, for opening its doors and providing the opportunity to better myself in ways which would hardly be possible otherwise.

Abstract

With transistor scaling, the process in which high performance, very large scale integrated (VLSI) systems are engineered has changed due to significant interconnect resistance, noise, and coupling effects. One common design solution that has not changed however has been the grid, a structural topology that enhances different characteristics of a VLSI system, such as area, reliability, and design complexity. In this dissertation, the fundamental role of a grid structure in VLSI systems is explored and a set of design challenges in grid-based circuits are addressed; specifically, the design challenges of nonvolatile resistive memory arrays and on-chip power grids with integrated linear voltage regulators.

The dissertation starts by introducing the characteristics of a grid structure and the influence of grids on different aspects of integrated systems, such as power delivery networks, memory systems, digital logic, and automated routing. The following chapters address a set of challenges in grid-based systems, continuing with the computational complexity of designing nonvolatile resistive memories as well as the write

energy of these memory arrays. A set of closed-form expressions that intuitively model the size limitation of nonvolatile resistive memories with cell selectors are described to relax the computational requirements. Furthermore, the write energy of resistive memory arrays is explored, and models estimating the write energy are presented. Based on the insights gained from these models, an energy efficient bias scheme is proposed to reduce the write energy.

Moreover, the stability of on-chip power grids in the presence of multiple linear voltage regulators is evaluated. The decreasing stability of a power grid when increasing the number of regulators is described. The integration challenges of digital linear regulators with resistive power grids is also discussed. A methodology to distribute the pass transistors of a digital linear regulator is proposed to mitigate voltage variations across a grid.

The benefits of a grid topology in complex integrated systems often comes at a cost of various design challenges. This dissertation provides insight into the complex relationship between grids and VLSI systems.

Contributors and Funding Sources

This work was supervised by a dissertation committee consisting of Professor Eby G. Friedman (advisor), Professor Zeljko Ignjatovic, and Professor Engin Ipek of the Department of Electrical and Computer Engineering, and Professor Marc D. Porosoff of the Department of Chemical Engineering. All of the work described in the dissertation was completed by the student with support from Professor Eby G. Friedman.

This work was supported by the U.S.-Israel Binational Science Foundation under Grant No. 2012139, the National Science Foundation under Grant Nos. CCF-1329374, CCF-1526466, CNS-1548078, and CCF-1716091, the Intel Collaborative Research Institute for Computational Intelligence (ICRI-CI), IARPA under Grant No. W911NF-14-C-0089, AIM Photonics under Award No. 059447-007, the Singapore Ministry of Education Tier 2 under Grant No. MOE2014-T2-2-105, and by grants from Cisco Systems, Qualcomm, and OeC.

List of Tables

3.1 Summary of typical device parameters [72, 75] 38

5.1 Parameters for read operation 83

5.2 Design parameters 99

5.3 Varying array sizes to satisfy $V_{cell}/V_{write} = 0.75$ 99

6.1 Summary of parameters for write operation 111

8.1 Mean of distributions 183

8.2 Variance of distributions 183

B.1 Off-chip parasitic impedances 218

List of Figures

1.1 Grid systems are used in a variety of design and engineering problems, (a) architecture, (b) Manhattan gridiron plan, and (c) graphic design of the New York Times. 4

1.2 Repeated cell structures forming a grid in die photographs, (a) Intel 4004, and (b) IBM Power9 processor. 5

1.3 Grid based IC design, (a) DRAM memory array, (b) programmable logic array (PLA), (c) interconnect routing, and (d) mesh-based network for communication within multicore microprocessors. 6

2.1 Grid structures, (a) square grid, and (b) triangular grid. 13

2.2 Top view of a unit memory cell with a cell height of h and width of w. 15

2.3 Placement of memory cells within a grid structure for higher density. 16

2.4 Examples of grid structure in integrated circuits for high density, (a) memory array, and (b) programmable logic array (PLA). 17

2.5 Graph Cartesian product of two generators produces a regular graph, (a) triangle graph and heptagon, (b) path graph and hexagon, and (c) two path graphs producing a grid graph. 19

2.6 Regularity of a network-on-chip (NoC) based on a mesh topology. . . 20

2.7 Redundancy insertion in PLAs to enhance yield. The regular structure of a PLA supports fault tolerant circuits. 21

2.8 Alternate paths between two points within a mesh improve robustness and reduce the effective parasitic impedance between the source and sink. 23

2.9 Power delivery networks, (a) tree structured, and (b) mesh-based topology. 24

2.10 Effective resistance across a grid (between marked points) with respect to grid density, assuming an equal number of rows and columns. . . . 25

2.11 Routing multiple pins, (a) regular grid with uniform lengths, and (b) Hanan grid. 27

3.1 Crossbar array structure. 30

3.2 A 2 x 2 resistive crossbar array. 31

3.3 Switching mechanism of RRAM, (a) set operation, and (b) reset operation [68]. 33

3.4 Switching mechanism of PCM, (a) set operation, and (b) reset operation. 34

3.5 The switching mechanism of MRAM, (a) set operation, and (b) reset operation. 35

3.6 Two types of MRAM switching mechanisms, (a) field assisted switching and (b) spin-transfer-torque (STT). 36

3.7 Smallest memory cell matched to the metal pitch with an area of $4F^2$. 38

3.8 Memory hierarchy of a computer system [43]. 39

3.9 An MRAM based main memory from Everspin Technologies [76]. . . 40

3.10 PCM and RRAM based SSDs, (a) Intel–Micron [77], and (b) Toshiba [78]. 41

3.11 Typical memory system with the peripheral circuitry forming a bank of multiple crossbar arrays [79]. 42

3.12 Leakage currents within a resistive crossbar array due to biased unselected cells. 45

3.13 Variation of interconnect resistance with respect to the metal pitch [82]. 46

3.14 The alternative path through the unselected rows and columns generates sneak path currents, lowering the read margin. 47

3.15 MOSFET as a selector to increase the I_{on}/I_{off} ratio of a memory cell. 48

3.16 A non-volatile resistive cell incorporating a two terminal selector device. 49

3.17 The effect of a selector device, (a) selectorless bipolar RRAM, and (b) bipolar RRAM integrated with a selector [85]. 50

4.1 Regulator to die distance, (a) physically distant high efficiency off-chip regulator, (b) in-package voltage regulator, and (c) multiple point-of-load on-chip voltage regulators producing heterogeneous voltages. . . 53

4.2 Fully integrated on-chip voltage regulators, (a) eight distributed voltage regulators across an IBM DDR3 I/O core [89], and (b) integrated power management on a cellular mobile IC from ST Ericsson [87]. . . 55

4.3 Power efficiency of voltage regulators for different conversion ratios, (a) switched-capacitor converter with 2 volt input voltage [91], (b) switching buck regulator with 1.5 volt input voltage [92], and (c) linear regulator with 1.1 volt input voltage [90]. 57

4.4 The effect of the output current on power conversion ratios, (a) switching buck regulator and linear regulator [95], and (b) switched-capacitor converter [91]. 58

4.5 Linear regulator, (a) simplified representation, and (b) practical circuit. 59

4.6 Linear regulator, (a) with large output capacitor, and (b) with small output capacitor. 62

4.7 Small–signal model of an LDO regulator. 63

4.8 Frequency response of an LDO regulator, (a) large output capacitor, and (b) small output capacitor. 66

4.9 Closed loop frequency response of the output impedance. The ESR of C_{out} is ignored. Decreasing the bandwidth of the error amplifier increases the output impedance, increasing the voltage droop. 68

4.10 Different LDO topologies, (a) slew-rate enhancement using current amplifier [99], and (b) adaptive RC compensation with an adaptive boost technique [104]. 70

5.1 Biasing scheme for a crossbar array when (a) writing to a cell, (b) reading from a cell. 75

5.2 Driver circuit. 76

5.3 Circuit model of crossbar array during a write operation. 80

5.4 Ratio of the voltage drop across the worst case selected cell to the driver voltage during a write operation. 81

5.5 The circuit model of the crossbar array during a read operation where R_{sense} is the input resistance of the sense amplifier and R_{sneak} is the sneak path resistance of the resistive memory cells between the (un)selected column(s) and unselected rows. 82

5.6 Ratio of the voltage drop across the worst case selected cell to the driver voltage during a read operation. 84

5.7 Comparison of the read margin between the analytic model and simulation. 86

5.8 Enhancing cell nonlinearity for (a) write operation with $V/3$ biasing scheme, and (b) read operation with floating biasing scheme. 87

5.9 Ratio of the voltage drop across the worst case selected cell to the driver voltage during a write operation under the $V/3$ biasing scheme. 91

5.10 Ratio of the voltage drop across the worst case selected cell to the driver voltage during a read operation under the floating biasing scheme. 91

5.11 Comparison of the read margin between the model and simulation for the floating biasing scheme. 93

5.12 Analytic model of driver resistance with respect to varying array sizes for $K_r = 10$, $K_{r(write)} = 2 \times 10^3$, and $K_{r(read)} = 10^3$ that satisfies $\frac{V_{driver}}{V_{read}} = \frac{4}{3}$. 94

5.13 Voltage degradation vs. array size where $V_{source} = V_{write}$ (solid lines) and $V_{source} = V_{read}$ (dashed lines). $R_{sense} = 100\ \Omega$. 96

5.14 Read margin with respect to array size based on the parameters listed in Table 5.1 for (a) $R_{int} = 0\ \Omega$, and (b) $R_{int} = 2.5\ \Omega$. The solid lines describes the grounded biasing scheme whereas the dashed lines describes the floating biasing scheme. 98

6.1 Bias schemes for a two bit write operation, (a) $V/2$ bias scheme, and (b) $V/3$ bias scheme. 106

6.2 Energy consumption of a crossbar array with respect to (a) array size, and (b) number of selected cells, assuming $R_{on} = 10^4\ \Omega$, $R_{off} = 10^7\ \Omega$, $K_{V/2} = 20$, $K_{V/3} = 1,000$, and $V_{write} = 2\ V$. 111

6.3 Effect of the number of selected cells on the energy consumption of a crossbar array for the $V/2$ and $V/3$ bias schemes, assuming $K_{V/2} = 20$ and $K_{V/3} = 1,000$. 112

6.4 Effect of the nonlinearity factor on the energy consumption of a crossbar array for the $V/2$ and $V/3$ bias schemes, assuming $n = 4$. 113

6.5 Comparison of the energy consumption in terms of the array size and number of selected cells for the $V/2$ and $V/3$ bias schemes, assuming $K_{V/2} = 20$ and $K_{V/3} = 1,000$. 114

6.6 Energy savings of the $V/3$ bias scheme as compared to the $V/2$ bias scheme assuming the same parameters listed in Fig. 6.5. The solid line is the contour where the energy consumption between the two bias schemes is equal. 115

6.7 Ratio of the nonlinearity factors $K_{V/3}$ to $K_{V/2}$ to maintain equal energy consumption for the $V/2$ and $V/3$ bias schemes in terms of the array size and number of selected cells. 116

6.8 Ratio of the switching energy to the total energy in terms of the array size, $R_{on} = 10^4\ \Omega$, $R_{off} = 10^6\ \Omega$, and $n = 4$. 118

6.9 Writing an eight bit word. Four bits of the new string are the same as the old string; however, only three bits are selected since one bit requires a reset whereas the other three bits require a set operation. . 120

6.10 Steps during the proposed energy efficient write scheme. 121

6.11 Energy improvement in terms of the number of selected cells, assuming $N = 128$, $K_{V/2} = 20$, and $K_{V/3} = 345$. The proposed write operation chooses the most energy efficient bias scheme based on the number of selected cells n with respect to n_{th}. 122

6.12 Energy savings for different array sizes and number of selected cells considering, (a) $K_r = 1000/20$, (b) $K_r = 345/20$, and (c) $K_r = 345/50$. 123

6.13 Number of selected cell in which the energy for both bias schemes are equal with respect to K_r and the array size N. 125

7.1 Linear regulator used to analyze the stability of multiple connected LDOs, (a) conventional low dropout regulator, and (b) Bode plot of a regulator under different load conditions. 133

7.2 Comparison of single LDO to multiple connected LDOs sharing a common power grid, (a) power delivery network with an input voltage of 1.2 volts and an output voltage of 1 volt, (b) transient response to a load varying from 175 mA to 210 mA in 10 ns considering one and 15 LDOs, and (c) transient response to a load varying from 1 mA to 3 mA in 10 ns considering one and 15 LDOs. The parasitic impedances are listed in Table B.1 in the Appendix. The input and output capacitance are, respectively, 1 nF and 50 pF per LDO. The load current as well as the input and output capacitors proportionally increase with the number of regulators. Each regulator therefore operates under the same load conditions and AC characteristics (see Fig. 7.1b). 135

7.3 Model of a power delivery system with parallel connected LDOs. The impedance of the power delivery network observed from the input of the LDOs is represented as a lumped impedance Z, (a) multiple LDOs attached to the same power grid, and (b) the grid impedance split per LDO, and (c) each LDO is separated based on the corresponding grid impedance at the input of the LDO. 139

7.4 Model of a power delivery system with parallel connected LDOs, (a) with an off-chip parasitic impedance, and (b) distribution of the parasitic impedance when the LDOs operate under the same load conditions [158]. The quiescent current of the LDOs is assumed to be negligibly small. 141

7.5 Reduction of parallel connected LDOs operating under the same load conditions [158]. 141

7.6 Transient simulation of multiple connected LDO regulators where the solid line describes the circuit shown in Fig. 7.4a and the dashed line describes the circuit shown in Fig. 7.5, (a) three LDOs, and (b) 15 LDOs. 142

7.7 Small-signal model of the simplified circuit shown in Fig. 7.5 [158]. . 143

7.8 Bode plot of circuit model shown in Fig. 7.5. The effect of an increasing number of parallel LDOs, (a) open loop gain, and (b) phase. With five parallel LDOs, the open loop gain rises above 0 dB beyond the initial unity gain frequency, producing an unstable system. $C = 1$ nF, $L =$ 100 pH, $R = 100$ $\mu\Omega$, $C_{out} = 50$ pF, and $I_{load} = 210$ mA per LDO. . . 145

7.9 Input impedance, (a) considering the LDO regulator, (b) model of the input impedance as an RC circuit, and (c) comparison of the magnitude of the input impedances. 148

7.10 Effect of different number of LDO regulators, (a) resonant frequency, and (b) quality factor, based on the circuit characteristics considered in Fig. 7.8. The model is based on the small-signal circuit shown in Fig. 7.7. 150

7.11 Output impedance of multiple LDOs sharing a common load. 152

7.12 Decreasing phase margin shifts the complex poles of the output impedance $p_{\pm}^{Z_{out}}$ to the RHP. 154

7.13 Open loop transfer characteristics of an LDO assuming the simulation setup shown in Fig. 7.8, (a) a large output capacitance, and (b) a small output capacitance. 156

7.14 Complex poles of the open loop transfer function $H(s)$ become equal to the complex zeros with decreasing output capacitance. 157

7.15 Reduction in the open loop gain of the LDO significantly lowers the UGF. Decreasing the mid-band gain from 40 dB to 32 dB reduces the UGF (considering the first 0 dB crossing) from 100 MHz to 12 MHz. 158

7.16 Effect of the UGF on the output power grid shared by ten LDOs, (a) UGF at 12 MHz, (b) UGF at 25 MHz, and (c) UGF at 100 MHz. An output capacitance of 100 pF and an input capacitance of 1 nF per LDO is considered with the off-chip power grid described in the Appendix. A total load variation from 1.75 A to 2.1 A is assumed (equally divided among the LDOs). 159

7.17 Increasing the corner frequency of $|Z_{in}|$ reduces the interaction between the input and output power grids over a wider range of frequencies, (a) LDO with an additional pull-up transistor to reduce the resistance of the error amplifier R_{ea} observed from the input power network, and (b) different corner frequencies of $|Z_{in}|$ under a fixed output capacitance by increasing the quiescent current I_q of the LDO. 160

7.18 Increasing the corner frequency of Z_{in} reduces the interaction of the output capacitance with the input power grid, improving the stability of the power delivery network. The same power delivery network described in Fig. 7.15 is assumed with a UGF of 100 MHz. 161

7.19 C4 parasitic resistance R_{C4} is increased to reduce the quality factor and improve the stability of the power grid. Transient response of the output grid considering, (a) R_{C4} = 2 mΩ, (b) R_{C4} = 1 mΩ, and (c) R_{C4} = 0.1 mΩ. The power delivery network and parasitic impedances are listed in the Appendix with a load variation of 1.75 A to 2.1 A evenly distributed among the ten LDOs. The resonant frequency is 112 MHz. 163

7.20 Pole movement of the output impedance at the driving point, (a) considering single and multiple LDOs, and (b) several C4 parasitic inductances. 164

7.21 Effect of inductance and number of LDOs on the input capacitance required to prevent a phase shift of more than 45^o at the resonant frequency (considering the circuit characteristics used in Fig. 7.5). . . 165

8.1 A digital low dropout regulator. 168

8.2 Pass gates located at the centroid of a grid to source the distributed load currents. 170

8.3 Centroid of two load currents. 171

8.4 Iterative process for determining the centroid of three load currents, (a) the centroid is initially assigned to load current I_1, (b) a new centroid between I_1 and I_2 is determined and replaces the old centroid, (c) a new centroid is determined between the current centroid and I_3, replacing the old centroid, and (d) the final centroid is replaced by a source connected to the load currents. 173

8.5 Recursive process to determine multiple centroids. 174

8.6 Iterative process for determining the location of the quadrant centroids, (a) power grid divided into four quadrants, and (b) a centroid is placed within each quadrant. A diamond represents an individual centroid. . 175

8.7 A power grid composed of 16 regions with separate centroids. 175

8.8 Different pass gate distribution topologies, (a) top–bottom [96, 150], (b) daisy chain [172], and (c) distribution from [169]. 177

8.9 Power grid analysis assuming a uniform load distribution, (a) proposed centroid-based distribution topology, (b) top–bottom topology [96, 150], (c) daisy chain topology [172], and (d) distribution topology from [169]. 179

8.10 Power grid analysis considering nonuniform load distribution, (a) proposed centroid-based distribution topology, (b) top–bottom topology [96, 150], (c) daisy chain topology [172], and (d) distribution topology from [169]. 181

8.11 Monte Carlo simulations evaluating the difference between the maximum and minimum voltage across a power grid considering four different pass gate distribution topologies. 182

10.1 Noise coupling from a selected row to an unselected adjacent row. . . 193

B.1 Off-chip power delivery network model considered in Sections 7.1 and 7.4 [165]. 219

Chapter 1

Introduction

"The forest is magnificent, yet it contains no perfect trees."

– Gye Fram

1.1 My Thesis

Grids (or meshes) are common geometric structures used throughout the design of very large scale integration (VLSI) systems, such as high performance microprocessors, due to the ordered structure that eases circuit scalability, lowers design complexity, improves performance, increases transistor density, and enhances reliability.

1.2 Motivation

The exponential growth in the complexity and compute power of VLSI digital systems has significantly enhanced human civilization [1]. This increasing growth in compute power supported by algorithmic discoveries have enabled tackling difficult problems that were once thought to be unsolvable [2, 3]. The demand for computational power however has yet to cease, continuously pushing the semiconductor industry to design and manufacture ever more complex VLSI systems [4, 5].

From the very first practical microprocessors, such as the Intel 4004, to the latest and most advanced examples, such as the IBM Power9, there has always been tremendous challenges in the design, development, manufacture, and test of these highly complex systems [6–10]. While the Intel 4004 contained a few thousand transistors in 1971 [6], there was little support from electronic design automation (EDA) tools. Structured and automated placement and routing tools started to appear in the early 1980's, years after the development of the first microprocessors [11]. To develop complex integrated circuits (IC) in the late 1960's and early 1970's, companies utilized a large number of engineers and technicians to manually craft and optimize these circuits [12]. Alternatively, modern microprocessors often contain many tens to hundreds of billions of transistors. These integrated circuits are significantly more expensive to develop and manufacture despite the support provided by advanced EDA tools [13]. This increased design challenge is primarily due to reliability and noise

coupling issues in deeply scaled technology nodes as well as the increased functional complexity of advanced ICs that even the most capable EDA tools have difficulty in tackling [13–16].

A vast number of engineers spend countless hours developing complex ICs based on a variety of principles and guidelines to manage the complexity of designing these VLSI systems. In the early days of IC development, area and speed were the primary criteria shaping these design principles. The successful continuation of transistor scaling relaxed some of these initial design criteria, such as area, and enabled massive computational power which reinforced complex design procedures. Scaling however has also brought challenges that have not yet been considered, such as process (manufacturing) variability and thermal reliability. As a result, the design guidelines and tools have evolved over time [12,17]. Today, the primary critera shaping design principles and guidelines are speed, power, reliability, signal integrity, and security [18, 19]. While the process in which complex ICs are engineered continues to change, one particular topology has been used throughout the development of integrated systems; the grid.

In Section 1.3, the concept of grids and the relationship of grids to VLSI systems are reviewed. Those aspects of VLSI systems that rely on grid structures are described. In Section 1.4, an outline of this dissertation is provided, summarizing the following chapters.

1.3 The Grid

The grid is a set of intersecting horizontal and vertical lines forming an ordered structure (as described in greater detail in Chapter 2). Grids embody a rich set of properties, such as regularity, modularity, scalability, and density. Many applications over a wide range of disciplines use grids to tackle different kinds of problems and challenges, as illustrated in Fig. 1.1 [20–22]. From architecture to city planning to

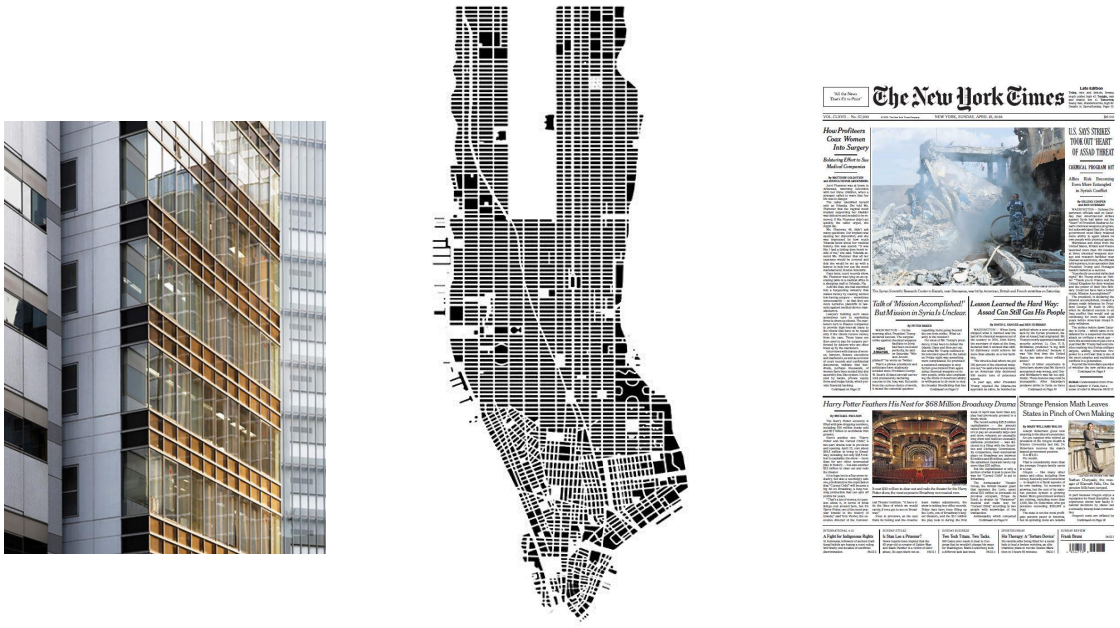

Figure 1.1: Grid systems are used in a variety of design and engineering problems, (a) architecture, (b) Manhattan gridiron plan, and (c) graphic design of the New York Times.

graphic design, grids are used for a variety of reasons. For example, graphic designers adopt grids in newspapers, websites, and other media platforms to organize

content in a structured and compact fashion [23]. In city planning, grids are used to enhance scalability, efficiently use space, and ease transportation [24]. In architecture and construction, grid systems form efficient supporting structures to equally distribute carried loads while benefiting from structural redundancy to improve robustness against failures [25].

Similarly, VLSI systems have exploited grid structures, as shown in Fig. 1.2 [26,27]. Grids have been used in integrated systems since the early days of IC design.

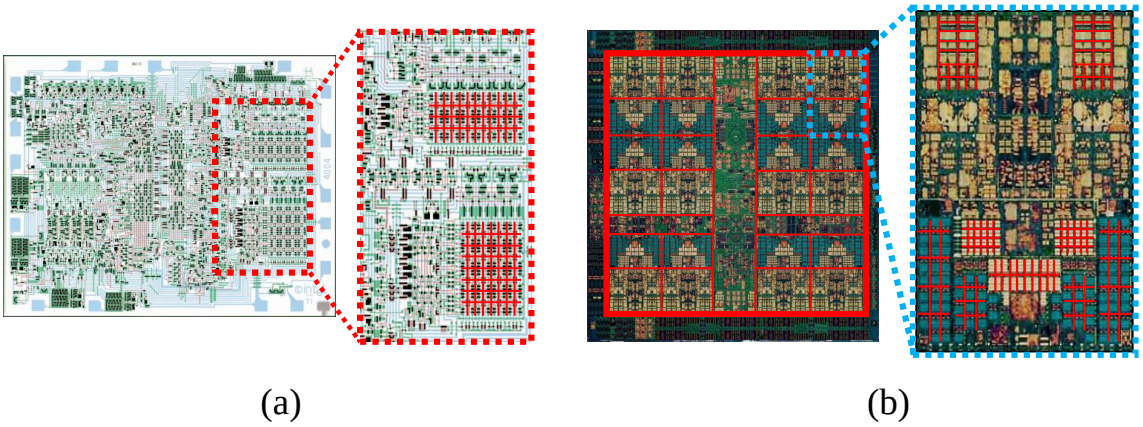

Figure 1.2: Repeated cell structures forming a grid in die photographs, (a) Intel 4004, and (b) IBM Power9 processor.

One of the first microprocessors, the Intel 4004, included an on-chip memory laid out in a grid structure (see Fig. 1.2a). The grid, shown in Fig. 1.2a, is a DRAM memory array formed of three transistor cells [7,28]. The use of grids in memories has led to high density, low cost, and reliable ICs. In contrast, the initial development of logic circuits lacked design guidelines and were implemented in more of a random fashion [29]. With transistor scaling, circuit complexity has increased while density and yield

has become more critical. Early logic circuits were therefore not as reliable and dense as memories [29].

To match the yield and density of memory ICs, structured design methodologies for the logic gates have been developed [30]. The use of grids has been extended to many other parts of an IC other than memory, from logic to interconnect to network communication within multicore systems, as shown in Fig. 1.3 [31–34]. To provide

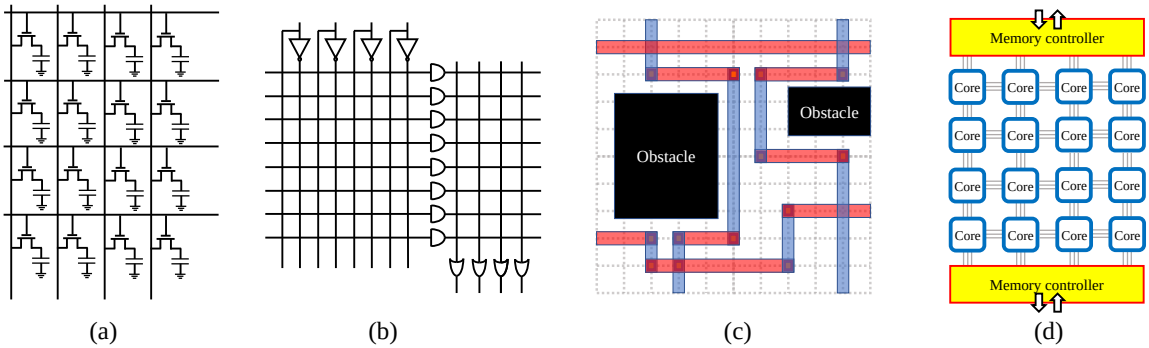

Figure 1.3: Grid based IC design, (a) DRAM memory array, (b) programmable logic array (PLA), (c) interconnect routing, and (d) mesh-based network for communication within multicore microprocessors.

a more efficient structure to logic circuits, programmable logic arrays (PLA) were developed [32]. These circuits enable memory-like, structured logic circuits, forming grid-based functional units. The fixed structure of a PLA however has come at a cost of reduced design flexibility, lowering performance. Initially, this structured design approach was preferred to improve circuit density despite suboptimal performance. With the advent of advanced EDA methodologies however other structured design

techniques such as standard cell-based circuits are usually preferred [31]. Interconnects in VLSI systems consider rectilinear routes where each metal layer is typically directed along a preferred axis (horizontal or vertical direction) [35]. Automated routing tools use a grid topology as a guideline to generate these interconnects, each aligned with the grid (see Fig. 1.3c) [33]. While diagonal routing reduces interconnect length and increases performance, grid-based rectilinear routing is mostly preferred due to the greater computational efficiency [33, 35]. Since a complex VLSI system can contain many billions of nodes, computationally tractable heuristic approaches are preferred over optimal but computationally expensive techniques. In multicore VLSI systems, the cores are typically connected within a mesh structure [34, 36–39]. The regularity of mesh-based multicore networks support scaling the number of cores [37–39]. Scalability of the network becomes critical as the number of on-chip cores increases. This increase in multicores is due to on-chip thermal and power constraints, pushing the development of highly parallel architectures to improve computational efficiency under a fixed power budget [40].

In addition to these aforementioned applications, grids have been considered in a variety of other aspects of a VLSI system such as the on-chip power delivery and clocking networks, and field programmable gate arrays (FPGAs) (see Chapter 2) [18, 41, 42]. The applicability and influence of grids across a wide range of design methodologies in VLSI systems have resulted in the physically structured and ordered

layout found in modern VLSI systems (see Fig. 1.2b). The seminal influence of grids in high performance ICs stem from the need to manage these highly complex systems. Fundamentally, the grid system forms a set of rules, creating a framework to effectively manage complexity, allowing the formation of tractable design solutions. The grid system is therefore an inseparable design element in the development of complex integrated systems.

1.4 Outline

In this dissertation, applications of grid structures to VLSI systems are examined in Chapter 2, beginning with the definition and attributes of grids. The grid offers different properties that can enhance speed, power, reliability, and density in VLSI systems. The influence of grids in the IC design process, such as power delivery, clock distribution networks, memory systems, programmable logic arrays, and automated routing, is discussed.

In Chapter 3, different types of nonvolatile resistive memories are reviewed. The working mechanisms of resistive RAM (RRAM), phase change memory (PCM), and magnetoresistive RAM (MRAM) are explained. These devices exhibit different strengths and weaknesses and are therefore used in different types of memory systems (e.g., cache, main memory, or disk). Similar to charge-based memories, such as DRAM,

SRAM, and flash, nonvolatile resistive memories are physically based on a grid structure. The advantages and challenges of grids (i.e., crossbars) and mitigation techniques, such as including selector devices, are discussed.

In Chapter 4, on-chip power delivery systems with integrated voltage regulators are reviewed. The power delivery network is typically structured as a grid. This network is supported by multiple on-chip voltage regulators in high performance ICs to reduce power consumption via dynamic voltage scaling (DVS). Different types of voltage regulators along with the advantages and disadvantages are described. The working mechanism of low dropout (LDO) voltage regulators is discussed with an emphasis on regulator stability.

In Chapter 5, closed-form expressions that model critical metrics of nonvolatile resistive crossbar arrays are provided. These crossbar arrays are widely considered as a likely memory technology that will replace DRAM and flash memory. The performance and capacity of resistive crossbar arrays are however limited due to sneak paths and the parasitic interconnect resistance within the grid. To reduce computational time and provide intuitive design guidelines, a set of compact closed-form expressions modeling these aforementioned limitations is described.

The modeling approach discussed in Chapter 5 is extended in Chapter 6 to evaluate the write energy consumption of nonvolatile resistive crossbar arrays. The write energy of resistive memory arrays are described, considering data dependencies and

bias schemes. Moreover, a hybrid, energy efficient write scheme utilizing multiple bias schemes is proposed to decrease the write energy based on the number of selected cells.

In Chapter 7, the stability of on-chip power delivery networks with multiple integrated capacitorless LDO regulators is explored. As the number of on-chip LDOs that share a common grid grows, the stability of the power delivery network degrades. This effect is primarily due to the decreasing resonance frequency with increasing number of LDO regulators. The degradation of grid stability is intuitively described, and an expression that relates stability to the number of LDOs is provided.

Moreover, the challenges of integrating digital LDOs within an on-chip power grid is described in Chapter 8. In a power delivery system with distributed pass transistors, the flow of current depends upon the physical location of the pass transistors, which can have a major effect on the power noise. A methodology to distribute the pass gates of a digital LDO is proposed based on grid centroids. The proposed distribution topology is compared to earlier topologies in terms of steady state IR drops within a power grid.

Lastly, a summary and concluding remarks of this dissertation are offered in Chapter 9, and future research directions are described in Chapter 10. The effect of noise coupling among adjacent rows and columns in a resistive memory array is explained.

In deeply scaled technologies where the interconnect pitch is small, capacitive coupling can disturb the memory cells along the unselected lines. This reliability issue, similar to the RowHammer problem, is discussed in greater detail.

Chapter 2

Grids in Integrated Systems

"The grid system is an aid, not a guarantee. It permits a number of possible uses and each designer can look for a solution appropriate to his personal style. But one must learn how to use the grid; it is an art that requires practice."

– Josef Müller-Brockmann, *Grid Systems in Graphic Design*, 1981

Grids are common structures within integrated systems such as memories [43], logic arrays [30], power delivery networks [44], networks-on-chip (NoC) [45], and automated layout [46]. The grid structure provides several benefits to different aspects of VLSI circuits. Depending upon the application, grids can be used to form an orderly structure to improve scalability and robustness, enhance density, or reduce computational time. A grid is a geometric structure formed of similar tiles, such as a square, rectangle, or triangle, tightly packed together, as shown in Fig. 2.1. A tile is a substructure that is copied multiple times to form a larger and regular structure. In

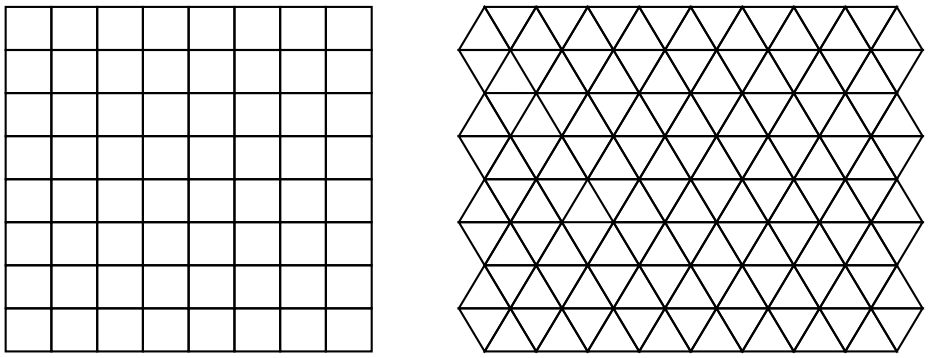

Figure 2.1: Grid structures, (a) square grid, and (b) triangular grid.

this dissertation, a grid is considered to be composed of a tile of squares or rectangles (see Fig. 2.1a) that need not be uniform. Note that the term, grid, is interchangeably used with the word, mesh.

The grid structure offers several properties often favored in complex integrated systems. These traits are density, regularity, and path diversity. Density is critical in VLSI systems due to limited die area. Since area is the primary factor determining cost and fabrication yield [29, 47], ICs are typically crammed with as many transistors as possible to enable a multitude of computational capabilities under a fixed area constraint. Regularity is highly desired in VLSI systems since regular structures are easier to scale, fabricate, design, and test [42, 48]. Regular structures produce more predictable results. Path diversity is the number of paths between two connected points. Structures including redundancy (e.g., redundant communication links or metal interconnects) against a variety of failures, such as faulty ineterconnects or

congested communication networks [47, 49], improve the overall robustness and yield of VLSI systems. In addition, grid structures are utilized by fast algorithms, enabling the computationally efficient automation of the VLSI design process, such as interconnect layout (see Section 2.4).

In this chapter, these properties of a grid useful to the development of VLSI systems are considered. In Section 2.1, the grid as an enabling structure for dense VLSI systems is described. In Section 2.2, a discussion on the regularity of grids and application to high density circuits is provided. In Section 2.3, the path diversity of grids and the relationship to system robustness is explained. In Section 2.4, the role of grids in automated layout is described. In Section 2.5, a summary of this chapter is offered.

2.1 Density

Density has different meanings depending upon the field (e.g., linear algebra [50], graph theory [51]). In this dissertation, density describes the total area occupied by the transistors in terms of the total on-chip area. An integrated circuit is dense if significant on-chip area is utilized by transistors. Density is critical to VLSI systems since the greater the number of transistors, the more functionality that can be

provided. In addition, die area is a major factor affecting cost. Larger dies also typically produce lower yield. Circuits that utilize significant die area at high density are therefore highly favored.

Grids form compact layouts, efficiently using on-chip area. These structures are therefore common in VLSI circuits, particularly those circuits requiring extremely high density such as memory arrays. For illustrative purposes, consider an arbitrary memory cell, as shown in Fig. 2.2. A memory cell is a circuit or device that stores a

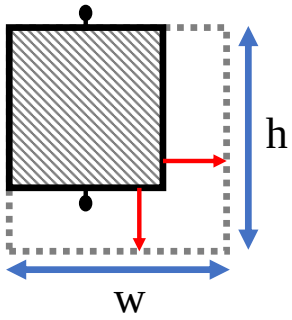

Figure 2.2: Top view of a unit memory cell with a cell height of h and width of w.

single bit of information, either a 0 or a 1. The shaded box represents the memory cell with two pins enabling either a write or read operation from/to the cell. The total cell area, $h{\cdot}w$, is the area occupied by the memory circuit, as constrained by the design rules of a particular technology node[1]. Considering an overall on-chip area of $H{\cdot}W$ (H and W are, respectively, the height and width of the on-chip area), the total number of memory cells in this area cannot be more than $\frac{H{\cdot}W}{h{\cdot}v}$. The layout structure

[1]Every component in an IC has a minimum physical separation from another neighboring component to mitigate manufacturing errors.

with the greatest number of cells naturally forms a grid structure, as illustrated in Fig. 2.3. A grid is therefore the most common structure in those circuits requiring high density.

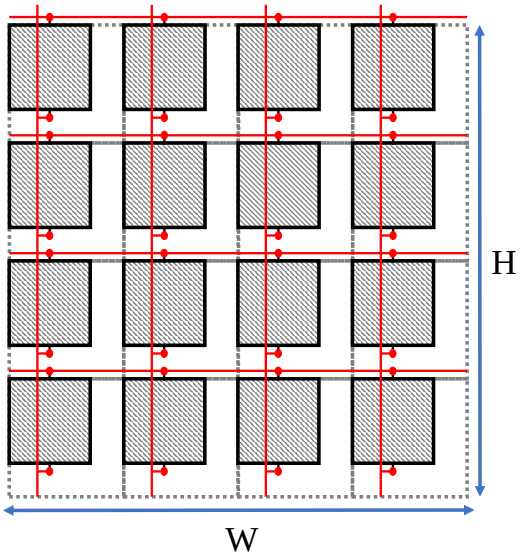

Figure 2.3: Placement of memory cells within a grid structure for higher density.

An exemplary memory array is depicted in Fig. 2.4a [43]. A memory array also requires additional peripheral circuits such as row and column decoders to function properly. Memory systems consist of multiple arrays rather than a single large array since the performance of the memory decreases as the array size grows [52]. A memory IC therefore consists of multiple sub-arrays as well as peripheral circuits to support read and write operations.

The grid structure in memories provides superior circuit density as compared to logic circuits designed in a more random fashion. A similarly structured, grid-based design approach has therefore been applied to logic circuits [30]. One of the early

examples of a logic oriented grid structure was the programmable logic array (PLA), as shown in Fig. 2.4b [29, 32, 53]. The PLA is a one time programmable circuit

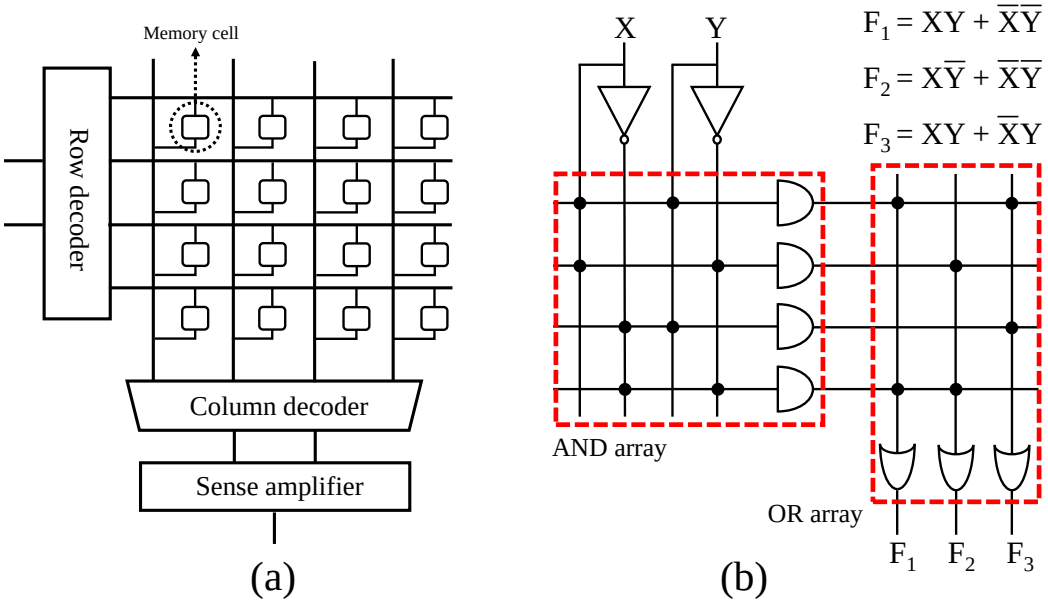

Figure 2.4: Examples of grid structure in integrated circuits for high density, (a) memory array, and (b) programmable logic array (PLA).

that computes multiple Boolean logic functions using two arrays; an AND array and an OR array. These arrays are programmed by placing a short at a crosspoint to connect a horizontal line to a vertical line. The inputs, initially passing through the AND array, are fed to a large AND gate. The OR array transmits the output of the AND gates to the OR gates. As a result, multiple boolean functions in the form of sum-of-products is obtained (see F_1, F_2, and F_3 shown in Fig. 2.4b). The PLA consumes significantly less area as compared to random logic, improving both density and performance.

Density was a particularly major concern in the early days of the semiconductor industry when the transistor channel length was on the order of a few micrometers. PLAs were therefore a highly favorable design approach. Today, however, area is less of a concern due to advanced technology nodes where transistor channel lengths range from a few tens of nanometers to a few nanometers [18]. Moreover, with the advent of advanced layout tools, logic circuits based on standard cells are commonly used, limiting the use of PLAs to a few specialized applications such as finite state machines [54–56]. Since capacity remains critical in memory systems, grid structures are still in common use in memory circuits.

2.2 Regularity

Grids are also used due to the structural regularity. Regular structures are preferred in VLSI systems due to the ease of fabrication and scalability. The definition of regularity in [57] for structural analysis is described here. A regular structure consists of multiple similar substructures applied in a repetitive fashion (e.g., see Fig. 2.1).

A structure represented as a graph $\mathrm{G}(V, E)$, formed of a set of vertices V and edges E, is regular if $\mathrm{G}(V, E)$ is a graph product of two or three subgraphs, also known as generators [57]. The graph Cartesian product is used to illustrate the formation of regular structures from two generators. The graph Cartesian product of two graphs $G_1(V_1, E_1) \square G_2(V_2, E_2)$ is the graph formed of the set of vertices V_1

X $V_2 = \{(x_i, y_j)|x_i \in V_1, y_j \in V_2\}$ with edges between a pair of vertices (x_i, y_j) and $(x_i{}', y_j{}')$ if and only if $(y_j, y_j{}')$ is an edge of G_2 when $x_i = x_i{}'$ and $(x_i, x_i{}')$ is an edge of G_1 when $y_j = y_j{}'$ [51]. The graph Cartesian product copies the graph G_1 for every node of G_2 (and vice versa) [51], as illustrated in Fig. 2.5. Note that a generator

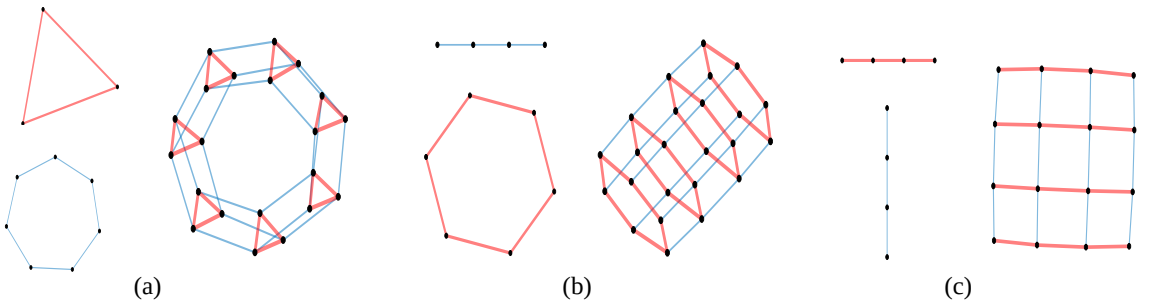

Figure 2.5: Graph Cartesian product of two generators produces a regular graph, (a) triangle graph and heptagon, (b) path graph and hexagon, and (c) two path graphs producing a grid graph.

is replicated for every node of the second generator, thereby forming a repetitive structure that is called regular. Similarly, a grid can be formed of two path graph[2] generators, as shown in Fig. 2.5c. The grid graph is therefore regular and consists of repetitive tiles.

Regularity in circuits and systems has several benefits, one of which is scalability. A regular structure is scalable due to the repetitive nature that supports additional structural elements (e.g., nodes in a graph, or rows or columns in an array). The repetitive substructure provides a guideline for integrating additional tiles into a regular structure (e.g., a new node connects to the head or tail of a path graph

[2]A path graph is a set of nodes connected in series, forming a path-like structure.

to preserve the sequential order). Scaling is critical to VLSI systems since circuits need to accommodate an increasing number of transistors in evolving generations of technology nodes to improve speed, reduce power, increase yield, and/or increase functional and memory capacity. Scalable circuits leverage the increasing number of transistors without significant changes in topology or design methodology. In contrast, circuits that are difficult to scale need to be significantly re-designed to accommodate the additional circuitry, increasing design time and cost.

An exemplary regular and scalable on-chip structure is the mesh-based network-on-chip (NoC), as shown in Fig. 2.6a [45]. With increasing number of transistors,

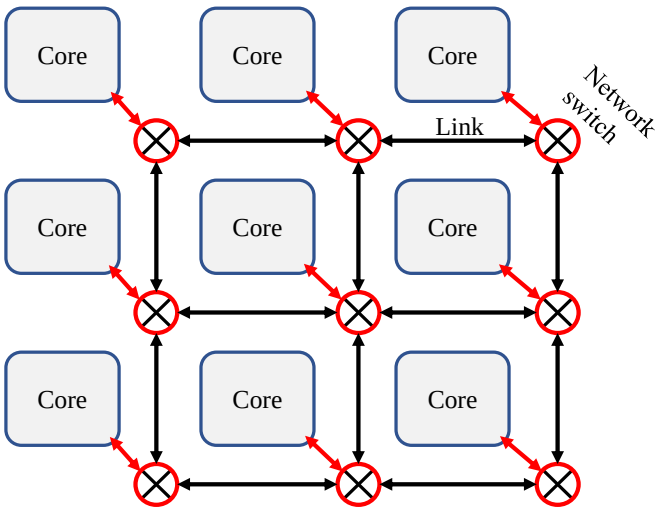

Figure 2.6: Regularity of a network-on-chip (NoC) based on a mesh topology.

multiple cores integrated onto a single die has become feasible [38]. The number of on-chip cores has reached several tens of cores in today's commercial products and continues to improve computational performance under a fixed power budget

[58–61]. A scalable network topology, such as a mesh topology, is therefore preferred to accommodate the increasing number of cores without re-designing the communication network in evolving generations of microprocessors [45].

Furthermore, regularity simplifies the addition of redundancy to improve fault tolerance and yield. VLSI systems typically contain redundant elements to tolerate fabrication failures. For example, an additional column or row can be inserted into

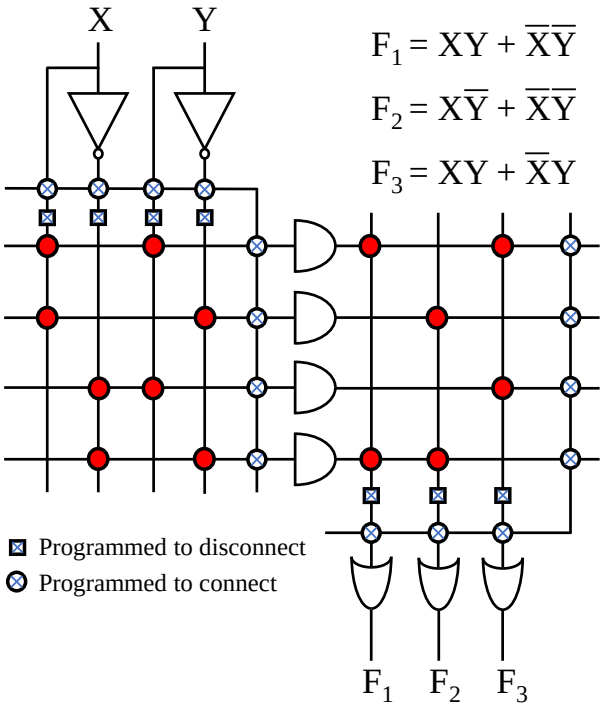

Figure 2.7: Redundancy insertion in PLAs to enhance yield. The regular structure of a PLA supports fault tolerant circuits.

a memory or programmable logic array to replace a faulty column or row [47]. An exemplary PLA circuit with a redundant row and column is shown in Fig. 2.7 [62,63]. The redundant rows and columns (initially disconnected) are connected to the AND

and OR arrays to replace faulty connections after testing the circuit whereas the faulty lines (initally connected) are disconnected from the arrays[3] [47]. Note that redundancy insertion in arrays requires low area due to the regularity of the grid. In contrast, random logic does not exhibit a structured pattern, making the insertion of redundancy more difficult [47]. Random logic therefore requires significantly greater area as compared to a grid-based circuit topology.

2.3 Path Diversity

Path diversity is the number of paths connecting two points. Grid structures are used to reliably communicate within networks. In a grid, a multitude of alternate paths often exist. Different alternative paths supported by a grid is illustrated by the example shown in Fig. 2.8. The high path diversity within dense grids enhances the robustness of connected paths against failures such as electromigration, fabrication defects, and/or congestion. The number of paths between two points on a grid (assuming only upward and rightward directions) separated by n rows and m columns is

$$\binom{m+n}{n} = \frac{(m+n)!}{(n!)(m!)}. \tag{2.1}$$

Path diversity therefore significantly increases with the number of rows and columns. The multitude of paths within a grid enhances several aspects of VLSI systems. One

[3]One possible approach to disconnecting faulty lines is using fusible links. Fusible links, initially shorted, can be blown to form an open circuit [47].

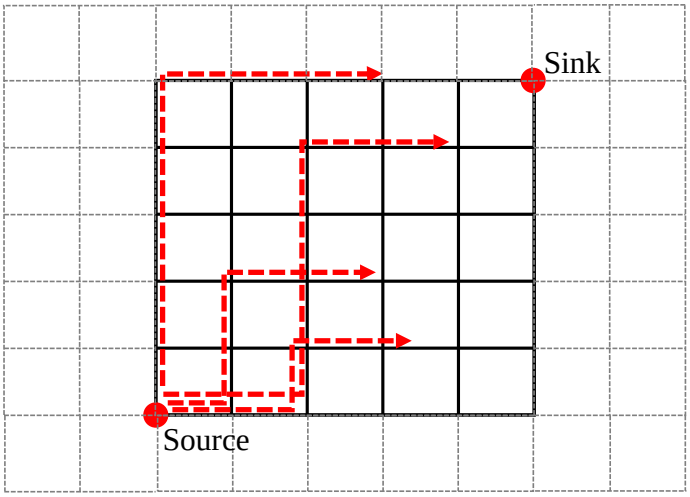

Figure 2.8: Alternate paths between two points within a mesh improve robustness and reduce the effective parasitic impedance between the source and sink.

important example is the power delivery network. A power delivery network consists of metal interconnects spanning a VLSI circuit delivering the current necessary to billions of loads to successfully operate an IC [44].

Two interconnect topologies are commonly used to delivery power; tree and mesh, as shown in Fig. 2.9 [18, 64]. The preferable topology of a power delivery network within an IC depends upon the total current, maximum voltage noise, and available metal layers. A tree structured power network was often used in early microprocessors where the current delivered to an IC is small and the wide metal interconnects exhibited negligible parasitic resistance [26]. Tree structured power networks are however susceptible to failure since a single interconnect fault along a path in a tree can produce an open circuit between the power supply and the circuit, catastrophically

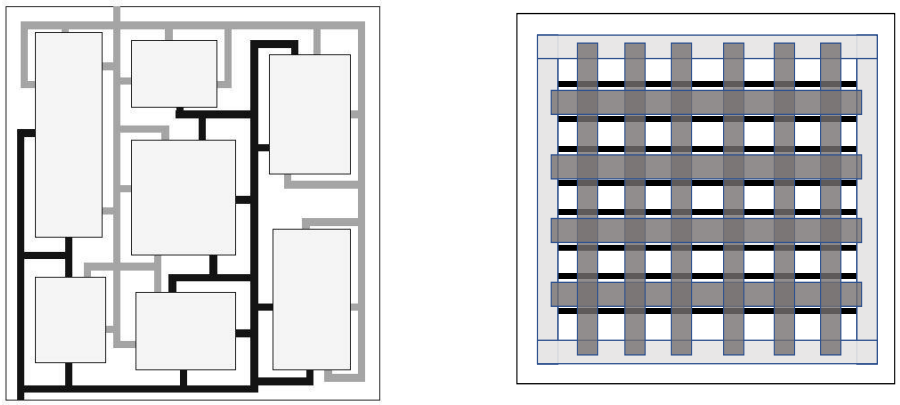

Figure 2.9: Power delivery networks, (a) tree structured, and (b) mesh-based topology.

increasing the output impedance of the network. Tree structured power networks are therefore unsuitable for delivering power in modern high performance VLSI systems where the total current is on the order of hundreds of amperes and the interconnect resistance needs to be below milliohms.

Alternatively, a grid structured power network improves robustness against interconnect failures since alternative paths within a grid can replace the faulty interconnects. Furthermore, the abundance of alternate paths reduces the effective parasitic resistance between the current loads and the source of power, as illustrated in Fig. 2.10. The number of paths significantly increases as additional rows and columns are inserted, greatly lowering the resistance between the source and load. Similarly, the detrimental effects of interconnect failures are mitigated by greater path diversity. The maximum grid density (number of rows or columns, or interconnect

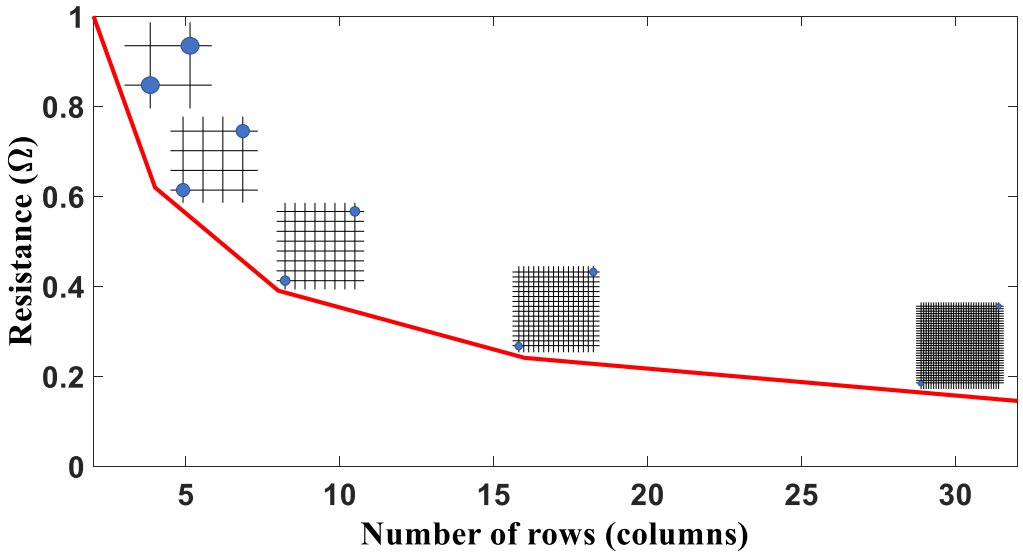

Figure 2.10: Effective resistance across a grid (between marked points) with respect to grid density, assuming an equal number of rows and columns.

width) however depends upon the available metal resources as well as physical design rules. On-chip metal resources are typically limited, requiring careful management and distribution among the data signals, decoupling capacitance, shield lines, and clock networks to satisfy performance and signal integrity requirements. The available metal resources however can be effectively managed with grids. Since the grid is also a regular structure, the metal resources (number of rows or columns and interconnect widths) are easier to manage than a power network formed of an irregular structure. Grids are therefore the dominant interconnect topology to delivery power across complex integrated systems.

2.4 Computational Efficiency

Grids are often used in electronic design automation; specifically, to route an IC. Grid-based routing improves computational performance as opposed to gridless routing [33]. During net routing, a grid is superimposed over the routing region (see Fig. 1.3c). Depending upon the type of routing (global or detailed), the grid edges represent routing tracks where interconnects are allowed [33, 64]. The interconnects are placed on these available tracks connecting different terminals while considering potential obstacles (e.g., signal pins, I/Os) and minimizing cost (such as interconnect length).

A critical metric characterizing routing quality is the average length of the interconnect [46]. Long interconnects produce larger parasitic impedances. Routers therefore typically search for the shortest path between pins (or terminals). Grid-based routing requires less computational time to solve the single source shortest path problem. The run time of the single source shortest path problem for an arbitrary graph with positive edge lengths can be as low as $O(Elog(V))$ (based on the Dijkstra algorithm) [65]. If the graph is a grid of uniform length, the complexity for solving the shortest path problem can decrease to linear time $O(V)$ (based on the Hadlock algorithm) [66].

Multi-pin routing is another problem considered in automated layout tools. The problem of determining the shortest tree connecting multiple terminals is described as

a Steiner problem [67]. Since on-chip metal interconnects are directional (horizontal or vertical depending upon the metal layer), the rectilinear distance (also known as the Manhattan distance or taxi cab metric) between nets is considered as a primary metric. As a result, the shortest routing cost connecting multiple pins is described as the minimum rectilinear Steiner tree (MRST) problem. A Hanan grid is a type

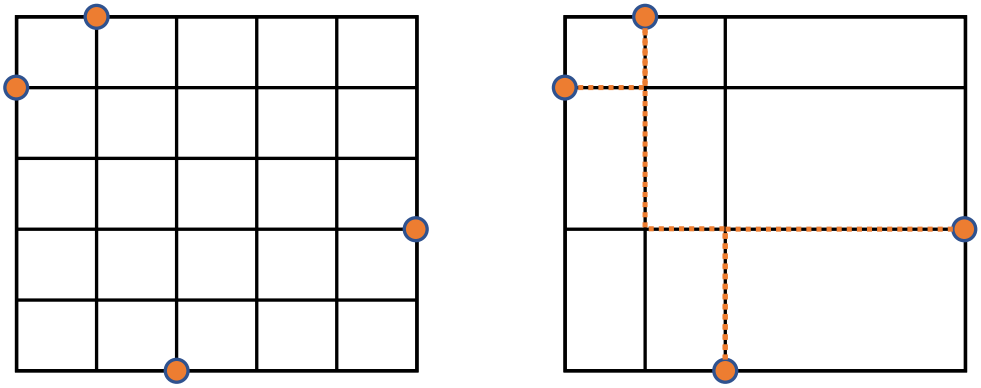

Figure 2.11: Routing multiple pins, (a) regular grid with uniform lengths, and (b) Hanan grid.

of grid that efficiently solves the MRST problem. A Hanan grid is a grid formed by vertical and horizontal lines passing through a set of vertices, as shown in Fig. 2.11. Based on the Hanan theorem, a Hanan grid is guaranteed to contain an MRST that connects these vertices [67]. A uniform grid can therefore be reduced to a Hanan grid to minimize the search space of the available paths connecting a set of terminals [46].

2.5 Summary

In this chapter, the properties of a grid useful in developing VLSI circuits and systems are introduced. Grids are often used in different types of integrated systems due to several structural properties that are highly desirable in VLSI systems. These properties are density, regularity, and path diversity. A grid is an efficient, compact structure to reduce area and is commonly used in a variety of VLSI circuits, such as memories and PLAs, due to the inherent circuit density advantages. These grid-based systems exhibit a structural regularity which is also highly scalable, preserving the system topology and circuit design across evolving technology generations while leveraging the increasing number of transistors. This regularity further improves circuit yield by supporting enhanced fault tolerance at low overhead. Circuit duplication to add redundancy is avoided. Yield is further improved in grid-based power delivery networks by providing a large number of paths to connect the billions of distributed on-chip loads to the off-chip power sources. The path diversity of a grid network reduces the detrimental effects of power interconnect failures. In addition, grid structures enable computationally tractable algorithms for the automated layout of VLSI systems. The compute time for solving the shortest path as well as the MRST problem is greatly reduced, enhancing the efficient design of complex VLSI systems. These benefits stem from the structural properties of a grid which improve a variety of traits. Grid structures are therefore widespread in complex integrated

circuits, supporting many design methodologies to create robust, area efficient, and scalable VLSI systems.

Chapter 3

Non-Volatile Resistive Crossbar Arrays

Crossbar arrays, also known as cross-point arrays, are commonly used structures in memory systems to enhance density. A crossbar array is formed of intersected bundles of vertical and horizontal lines, as shown in Fig. 3.1. A memory cell is placed

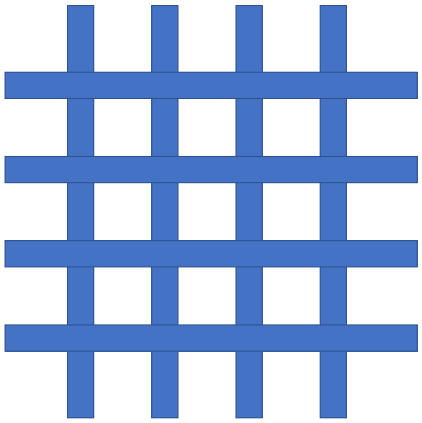

Figure 3.1: Crossbar array structure.

within an intersection and accessed by appropriately biasing the pair of horizontal and vertical lines. These lines are formed of metal used to interconnect circuits on an integrated circuit. Current memory systems such as dynamic random access

memory (DRAM) and flash memory are based on crossbar structures to improve memory capacity. While these memory systems have enabled large storage capacities, the advantage of low cost has started to diminish due to several limitations. One of the primary limitations of these charge based memories is scalability driving up fabrication costs. Over the past ten years, emerging resistive devices have utilized crossbar array topologies to create scalable and high density non-volatile memory systems. Unlike charge based technologies, a resistive memory device stores one or multiple bits of information in the form of resistance. These devices are naturally non-volatile, offering significant benefits in static energy savings.

A resistive crossbar array is illustrated in Fig. 3.2. Similar to vias that connect

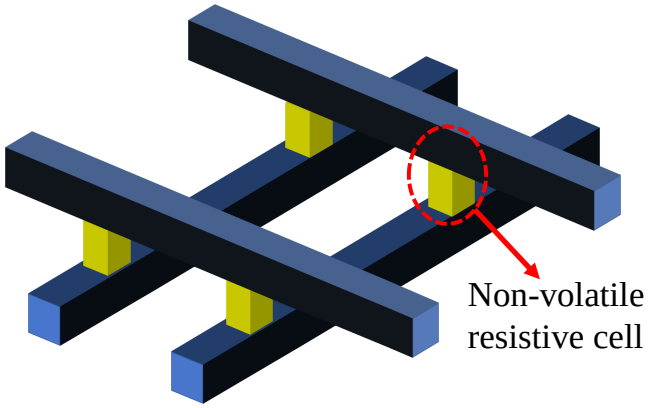

Figure 3.2: A 2 x 2 resistive crossbar array.

different metal layers, the resistive devices are placed within the metal layers and can be stacked, allowing vertical integration. The memory array can therefore be placed above the CMOS circuitry, saving die area. Moreover, depending upon the type of

non-volatile resistive device, the cell area can be significantly decreased and matched to the pitch of the metal lines, thereby forming the smallest cell possible in a given technology node.

In this chapter, a general review of non-volatile resistive crossbar arrays is provided. In Section 3.1, different kinds of resistive devices, primary parameters, and applications are explained. In Section 3.2, memory systems consisting of resistive crossbar arrays and the peripheral circuits are described. In Section 3.3, the challenges of developing resistive crossbar arrays are discussed. In Section 3.4, some conclusions are offered.

3.1 Non-Volatile Resistive Devices

Different kinds of non-volatile resistive devices have been explored for memory applications [68–70]. The most widely considered resistive devices are resistive random access memory (RRAM), phase change memory (PCM), and magnetoresistive random access memory (MRAM) [71, 72]. While random access memory (RAM) pertains to a type of memory system, the MRAM, RRAM, and PCM acronyms have been adopted to refer to these resistive devices. A non-volatile resistive cell is a two terminal device in which the resistance depends upon the state of the device. A non-volatile resistive cell operating in the on-state (the set state) exhibits a lower resistance as compared to the off-state (the reset state). The switching mechanisms

as well as materials in which these devices are made greatly differ for different types of non-volatile resistive devices.

An RRAM is formed from three stacked metal-insulator-metal layers, where the insulator layer is rich in oxygen ions, as shown in Fig. 3.3. The resistance state

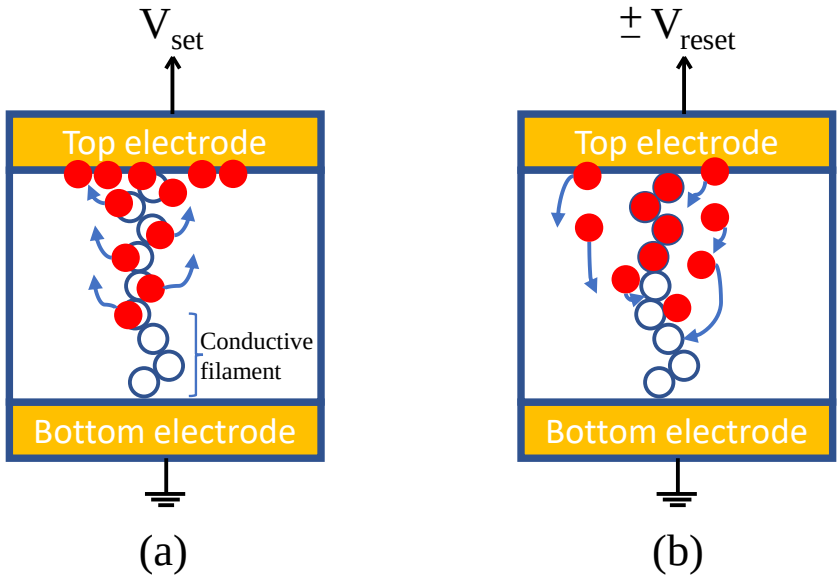

Figure 3.3: Switching mechanism of RRAM, (a) set operation, and (b) reset operation [68].

of an RRAM depends upon the oxygen vacancies within the insulating layer. A specific voltage applied across the device results in reversible breakdown, sweeping the oxygen ions across the metal layers and changing the available energy states within the insulating layer. In the absence of oxygen ions, the increased number of energy states forms a conductive filament inside the insulator, allowing electrons to pass through, lowering the resistance [68]. This effect can be reversed by appropriately setting the voltage across the device to replace the oxygen ions, removing the energy

states and shortening the conductive filaments to increase the resistance. The reset voltage can either be positive or negative depending upon the metal and insulator materials. If a non-volatile resistive device has a positive reset voltage, it is called unipolar. Conversely, if the device has a negative reset voltage, the device is called bipolar.

A PCM is formed of a phase change material sandwiched between two metal contacts with an embedded heater element, as shown in Fig. 3.4 [69]. By applying

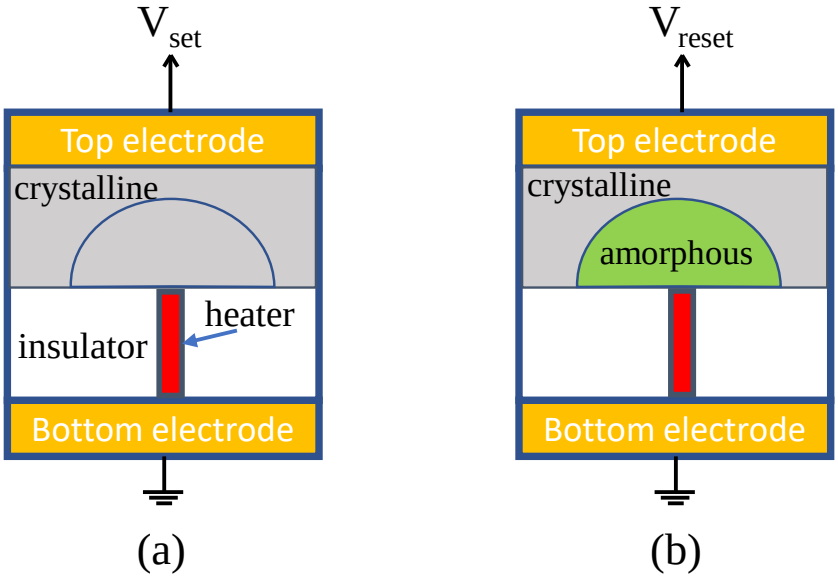

Figure 3.4: Switching mechanism of PCM, (a) set operation, and (b) reset operation.

current through the heater, the temperature of the phase change material is increased. As a result, the atomic structure of the phase change material changes either to crystalline or amorphous. Amorphous structures exhibit a higher resistance than crystalline forms. A reset operation occurs when the phase change material becomes

amorphous, increasing the resistance of the device. Alternatively, a set operation lowers the resistance by changing into a crystalline form. PCM devices are typically unipolar with a higher reset voltage. A reset operation therefore consumes more power than a set operation [69].

An MRAM is formed of a thin insulator or conductor placed between two ferromagnetic layers, as shown in Fig. 3.5 [73]. One ferromagnetic layer has a fixed

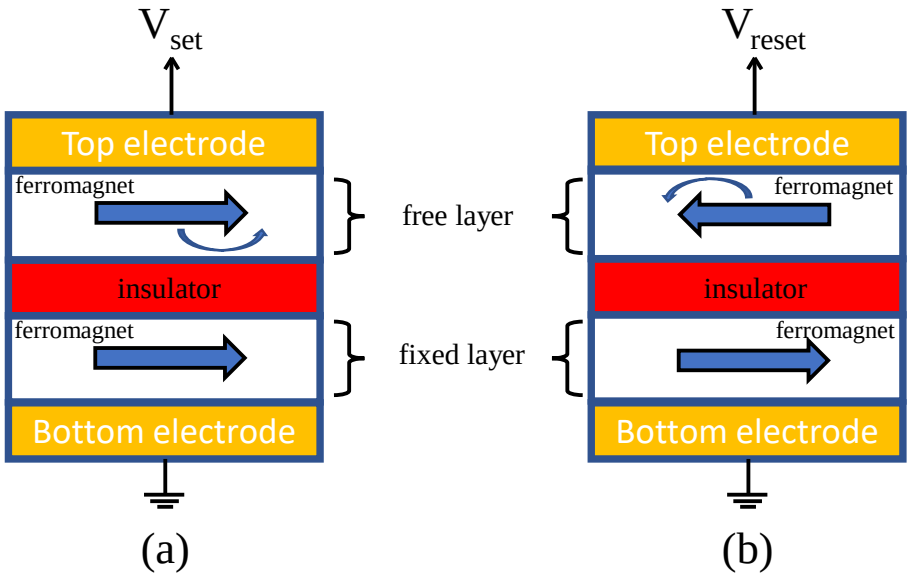

Figure 3.5: The switching mechanism of MRAM, (a) set operation, and (b) reset operation.

magnetic polarity (fixed layer), and the other ferromagnetic layer has a variable polarity (free layer). By changing the magnetic polarity of the free layer, the MRAM can be set either to a high or low resistance state. A ferromagnet sets the spin of the electrons to either up or down depending upon the magnetic polarity of the ferromagnet. A secondary ferromagnet receiving these electrons counteracts the spin of the

electrons if the magnetic polarity of the two ferromagnets is opposite to each other (i.e., antiparallel). Under this condition, the MRAM is in a high resistance state. Conversely, if the two magnetic polarities are aligned (i.e., parallel), the MRAM is in a low resistance state. Different types of MRAMs exist depending upon the material between the two ferromagnets. If the material between the two ferromagnets is an insulator, the device is called a magnetic tunnel junction (MTJ). If the material is a conductor, the device is called a spin-valve device. Moreover, the MRAM device can be switched (to either the on or off resistance state) by either injecting a current pulse with sufficient amplitude and duration or by applying a magnetic field, as shown in Fig. 3.6. The switching mechanism using current injection is called spin-

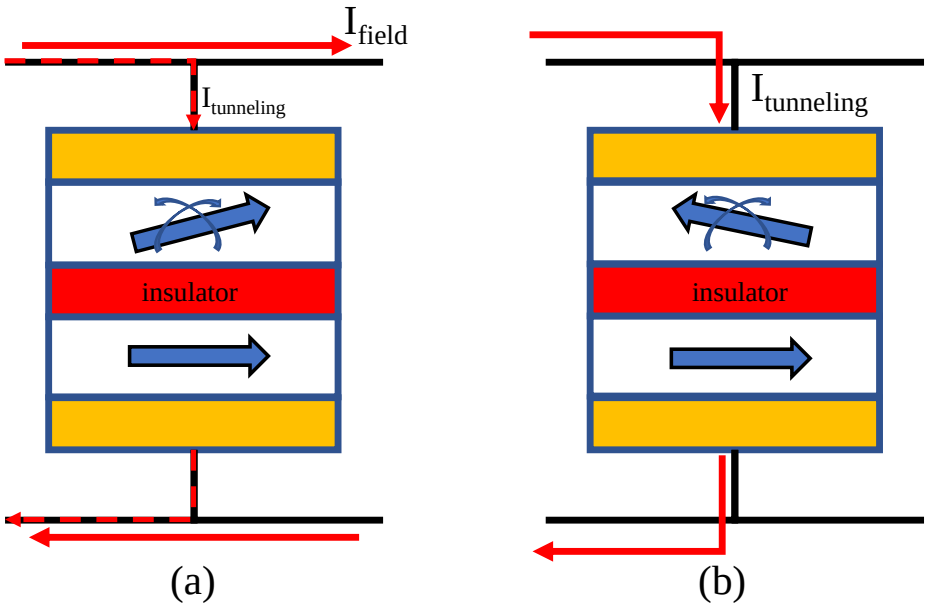

Figure 3.6: Two types of MRAM switching mechanisms, (a) field assisted switching and (b) spin-transfer-torque (STT).

transfer-torque (STT). The switching mechanism using a magnetic field is called field driven. A combination of both mechanisms is called field assisted switching. Field driven write mechanisms typically require larger currents (mA) whereas STT-MRAM can switch at significantly lower currents (μA) [74]. Lower switching currents are preferred due to lower power consumption and a smaller voltage drop across the parasitic interconnect resistance.

3.1.1 Device Parameters

A non-volatile resistive cell can be characterized by different device parameters, such as threshold voltage (V_{th}), on–off resistance (R_{on-off}), endurance, switching speed (t_{sw}), area, and polarity. The endurance is the maximum number of switching events a device can endure before becoming dysfunctional. Due to the different switching mechanisms and materials, different types of devices exhibit different strengths and weaknesses, as listed in Table 3.1. PCM and RRAM have limited endurance, orders of magnitude lower than MRAM. PCM and RRAM are therefore not appropriate for circuits requiring frequent switching. Moreover, PCM and RRAM are significantly slower than MRAM. MRAM however exhibits significantly lower resistance values as compared to PCM and RRAM. The current required to switch an MRAM is therefore significantly higher, requiring larger driver circuitry and interconnects with smaller parasitic resistance. In terms of area, an RRAM and PCM cell can be as small as

Table 3.1: Summary of typical device parameters [72, 75]

Device parameters	**PCM**	**RRAM**	**MRAM**
V_{th}	< 3 V	< 3 V	< 1.5 V
R_{on}	10^4 to 10^6 Ω	10^4 to 10^6 Ω	10^3 Ω
R_{off}/R_{on}	10^2 to 10^3	10^2 to 10^3	< 10
t_{sw}	10 to 100 ns	10 to 100 ns	sub–10 ns
Endurance	10^9	10^6 to 10^{12}	10^{15}
Area	4 to 30 F^2	4 to 12 F^2	6 to 50 F^2
Polarity	unipolar	unipolar and bipolar	bipolar

*F is the minimum feature size of a technology node.

$4F^2$ where F is the minimum feature size of a given technology node. The minimum feature size typically determines the pitch of the lower level metal interconnect. The smallest achievable cell size is illustrated in Fig. 3.7. While MRAM can be similarly

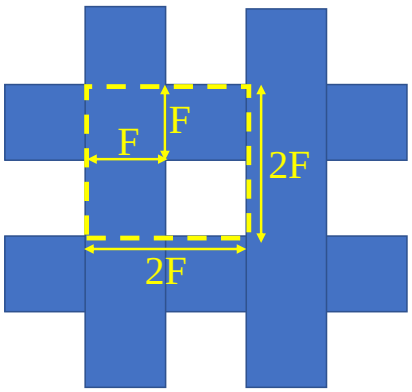

Figure 3.7: Smallest memory cell matched to the metal pitch with an area of $4F^2$.

integrated within the metal lines, the area is typically larger than $4F^2$. Since these devices have different strengths and weaknesses, the choice of device type depends

upon the application. A universal memory in which one device is considered across the entire memory hierarchy has therefore yet to be achieved.

3.1.2 Applications

Non-volatile resistive devices are primarily considered as a solution to the scaling bottleneck imposed by charge based memories such as SRAM, DRAM, and flash. Conventionally, different types of memory systems are used within a compute system, forming a memory hierarchy [43]. The memory hierarchy of a computer system is

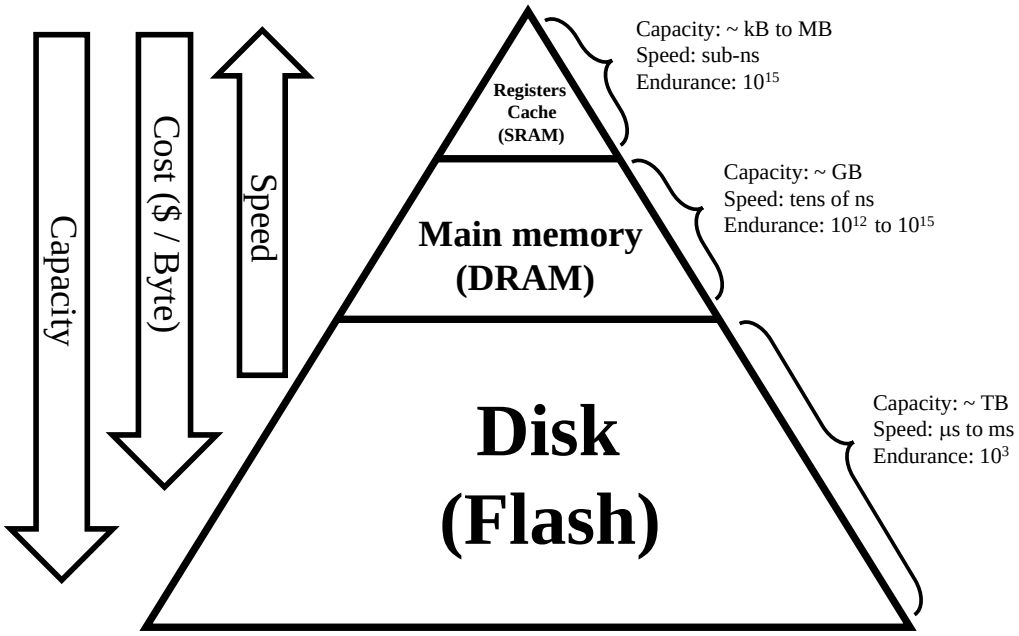

Figure 3.8: Memory hierarchy of a computer system [43].

shown in Fig. 3.8. The memory hierarchy provides a systematic method to distribute the different kinds of memories across a compute platform to provide high performance

as well as high capacity at low cost. Each level within the hierarchy has different requirements of speed, endurance, and capacity. At the top of the hierarchy, the memory is placed within the integrated circuit, requiring frequent switching at high speeds. For this application, MRAM is preferred over RRAM and PCM due to the advantages of higher endurance and fast switching speed. The high switching speeds and unlimited endurance of MRAM have lead to new memory products. An MRAM main memory product from Everspin Technologies is shown in Fig.3.9. This

Figure 3.9: An MRAM based main memory from Everspin Technologies [76].

MRAM based main memory has a maximum capacity of 256 Mb and is DDR3 DRAM compatible with a bandwidth of 1,333 MT/sec (MegaTransfers per second) per pin.

Conversely, at the bottom of the hierarchy, high capacity is the primary requirement with relaxed constraints of speed and endurance. RRAM and PCM are therefore considered as DRAM and flash replacement. Solid state drives (SSD) based on RRAM and PCM have also been developed [77,78]. Transmission electron microscopy (TEM) images of a PCM and RRAM based SSDs are shown in Fig. 3.10. The RRAM based SSD developed by Toshiba has a capacity of 32 Gb [78], whereas the PCM based

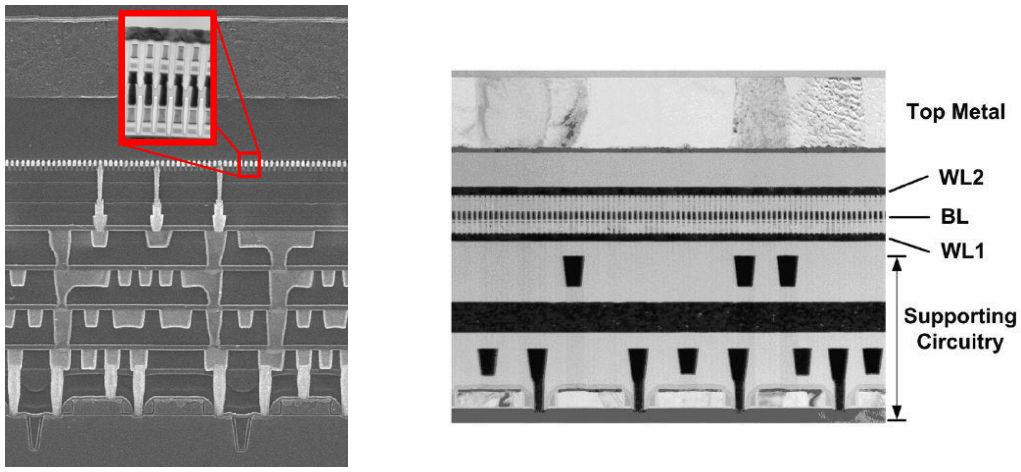

Figure 3.10: PCM and RRAM based SSDs, (a) Intel–Micron [77], and (b) Toshiba [78].

SSD developed by Intel and Micron has a capacity of 128 Gb [77]. Note that the resistive memory is embedded within higher metal layers above the CMOS circuitry, improving the area efficiency. Crossbar arrays are typically located above the peripheral circuitry. The peripheral circuits support the memory array, interfacing with the memory bus, receiving and executing the memory operations.

3.2 Non-Volatile Resistive Memory System

A memory system conventionally contains multiple crossbar arrays rather than one large array to overcome the performance limitations imposed by large arrays (see Section 3.3). A hierarchical distribution is followed, as shown in Fig. 3.11. A bank of

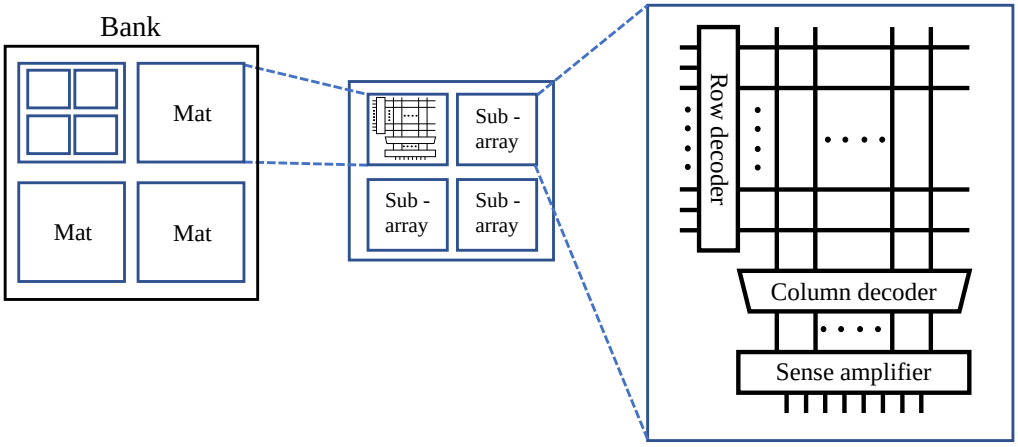

Figure 3.11: Typical memory system with the peripheral circuitry forming a bank of multiple crossbar arrays [79].

memory contains multiple sub-blocks, each called a mat. A mat contains an array of crossbar arrays, interconnected via shared peripheral circuits. A crossbar array within a mat is controlled by a peripheral circuit consisting of row and column decoders, row drivers, and sense amplifiers (see Fig. 3.11). The row and column decoders receive the memory address and access the corresponding cells within the array. Depending upon the request, the decoders select the appropriate cells and enable an electric pulse sufficient to write or read a cell. The rows within a crossbar array are called wordlines and the columns are called bitlines. The row drivers are placed between the wordlines of a crossbar array and the row decoder to supply sufficient current to the crossbar array. Larger crossbar arrays require larger drivers due to greater leakage currents. Moreover, in addition to the column decoder, sense amplifiers interface with the bitlines to read the state of the selected cells.

Current mode sensing, as compared to voltage mode sensing, is typically preferred in resistive crossbar arrays due to faster sensing [80]. By applying a read voltage across a selected cell, the sense amplifier receives a read current which is compared to a threshold current level. If the current is higher than the threshold current, a low resistance state is detected, resulting in a data output of 1. Conversely, if the current is below the threshold current, the sense amplifier produces a 0. The sense amplifiers consist of several transistors and are therefore significantly larger than the metal pitch of a column. To improve the area efficiency, sense amplifiers are typically multiplexed among multiple bitlines. The area efficiency of a resistive crossbar array is the ratio of the physical area occupied by the crossbar array to the area of the peripheral circuits. The area efficiency of resistive memory systems is typically poor due to the small size of a single array. To improve area efficiency, several techniques have been adopted such as multiplexing and placing the array above the peripheral circuitry [78, 81]. In addition, the row and column decoders can be shared between adjacent crossbar arrays, thereby doubling the array size per decoder [81]. While larger arrays increase memory capacity and improve area efficiency, challenges that limit the maximum size of a crossbar array exist.

3.3 Challenges

Non-volatile resistive devices are placed within a crossbar array to form a dense memory. The size of a resistive crossbar array (i.e., number of rows and columns) is however limited. The leakage currents from the unselected cells, sneak paths, and parasitic interconnect resistance prohibit scaling the size of an array and affect the write and read operations differently.

3.3.1 Write Operations

A primary bottleneck during a write operation is the leakage current of the unselected cells, as shown in Fig. 3.12. During a write operation, the selected row and column are biased to access the cell. The unselected rows and columns are also biased to prevent disturbance of the unselected cells along the selected row and selected column. As a result, the unselected cells along the unselected rows and columns receive a partial bias (depending upon the bias scheme, see Chapters 5 and 6) and leak undesired current. In large arrays, this leakage current aggregates, resulting in large currents flowing through narrow metal interconnects. Thus, a voltage drop is produced across the parasitic interconnect resistance, reducing the voltage across the selected cells. The degradation in the voltage across the selected cells reduces

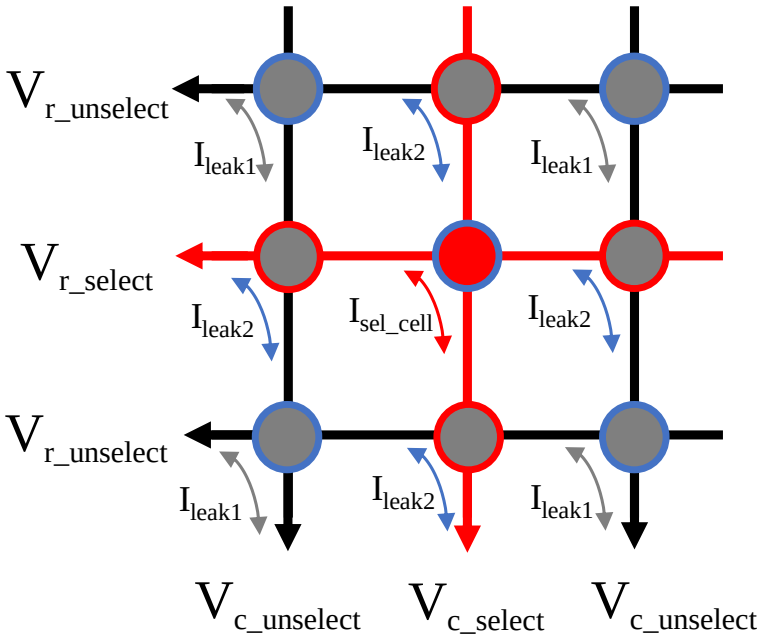

Figure 3.12: Leakage currents within a resistive crossbar array due to biased unselected cells.

the switching speed, potentially resulting in a write failure. Moreover, large leakage currents require larger drivers, increasing the area of the peripheral circuits and dissipating greater power.

The interconnect resistance increases with advancing technology nodes, exacerbating the voltage losses. The parasitic resistance of a crossbar array is shown in Fig. 3.13 for different metal pitches. Note that each cell along a row and column can add as much as tens of ohms of interconnect resistance in advanced technology nodes. Hence, rows and columns that carry hundreds of resistive cells can exhibit an

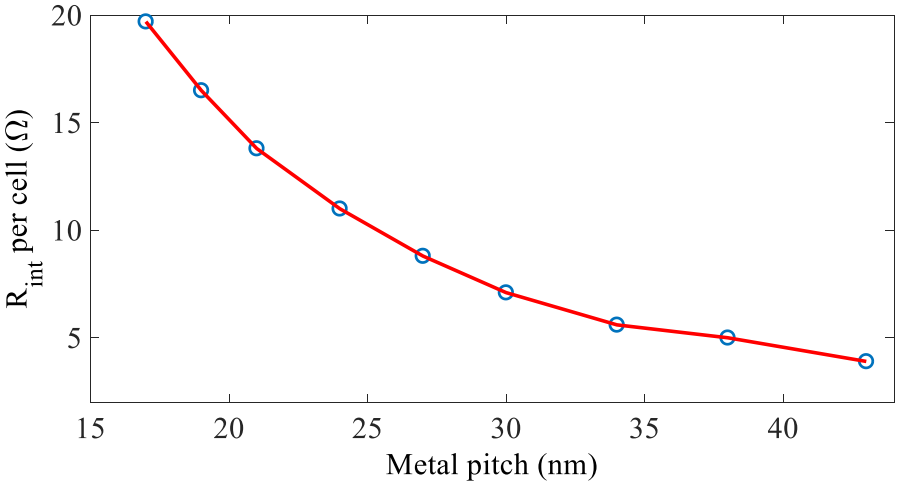

Figure 3.13: Variation of interconnect resistance with respect to the metal pitch [82].

interconnect resistance on the order of thousands of ohms, limiting the capacity of a crossbar array.

3.3.2 Read Operations

During a read operation, all or some of the unselected rows and columns are unbiased, and remain floating, depending upon the bias scheme. As a result, the cells along the unselected rows and columns form alternative paths, also known as sneak paths, between the selected row and the input of the sense amplifier at the selected column. The sneak path currents interfere with the currents through the selected cells, flowing into the same sense amplifier and distorting the information of the resistance states of the selected cells, as illustrated in Fig. 3.14. In large

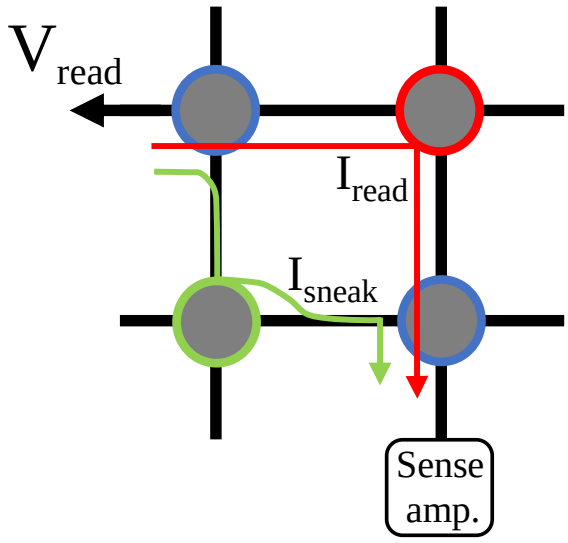

Figure 3.14: The alternative path through the unselected rows and columns generates sneak path currents, lowering the read margin.

arrays, the number of sneak paths grows drastically, increasing the noise added to the current through the selected cell and lowering the read margin. The ability to distinguish between two resistance states within a memory array determine the read margin (see Chapter 5). A high read margin lowers the burden on the sense amplifiers to distinguish different resistance states. If a resistive cell stores more than one bit (i.e., multiple resistance states), reading becomes significantly more challenging due to the decreased noise margin between resistance states. The read margin and the capability of the sense amplifier therefore imposes another limit on the size of a crossbar array. An effective way to increase the read margin and lower the voltage drop across the parasitic interconnect resistance is using a selector device with the non-volatile resistive cells.

3.3.3 Selectors

To reduce the leakage current during a write operation as well as to increase the resistance of the sneak paths during a read, the non-volatile resistive cells can be integrated with an additional selector device. The selector device suppresses the currents under a low voltage bias while supporting higher currents under a high voltage bias. Different three terminal devices such as an MOS transistor and bipolar junction transistor as well as two terminal devices such as a silicon-based diode and metal-insulator-metal tunneling barrier have been considered [71,83]. A one-selector-one-resistor (1S1R) notation is conventionally used to denote a resistive cell integrated with a two terminal selector and a one-transistor-one-resistor (1T1R) is used to denote a resistive cell integrated with a transistor. A resistive device with a MOSFET selector is illustrated in Fig. 3.15. Transistor based selectors provide enhanced iso-

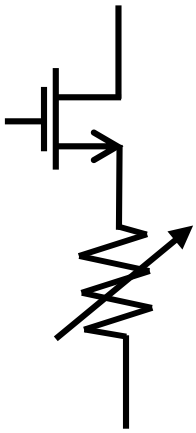

Figure 3.15: MOSFET as a selector to increase the I_{on}/I_{off} ratio of a memory cell.

lation between the selected cells and unselected cells within a crossbar array. This solution however significantly increases cell area and inhibits scalability. Two terminal selectors can however be vertically integrated within a non-volatile resistive cell, preserving the area, as shown in Fig. 3.16. A wide range of two terminal selectors

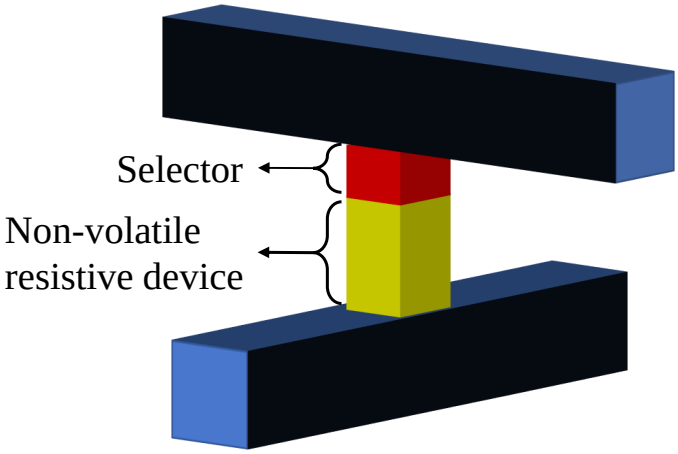

Figure 3.16: A non-volatile resistive cell incorporating a two terminal selector device.

exist which can be classified into two categories, unipolar and bipolar. In addition, depending upon the material and the type of non-volatile resistive cell, the selector can be a silicon based diode, self-rectifying resistive device, or metal-insulator-metal (MIM) with different kinds of tunneling mechanisms depending upon the thickness of the insulator material [84]. MRAM exhibits significantly lower resistance as compared to RRAM and PCM, resulting in greater leakage and sneak path currents. In MRAM based arrays, transistors are therefore preferred over two terminal selectors for enhanced isolation between the unselected cells and the rest of the array. In RRAM

and PCM based crossbar arrays, two terminal selectors are typically sufficient. The effect of a selector device on the I–V characteristics of a non-volatile resistive cell is shown in Fig. 3.17. Note that the selectorless device exhibits high current during the

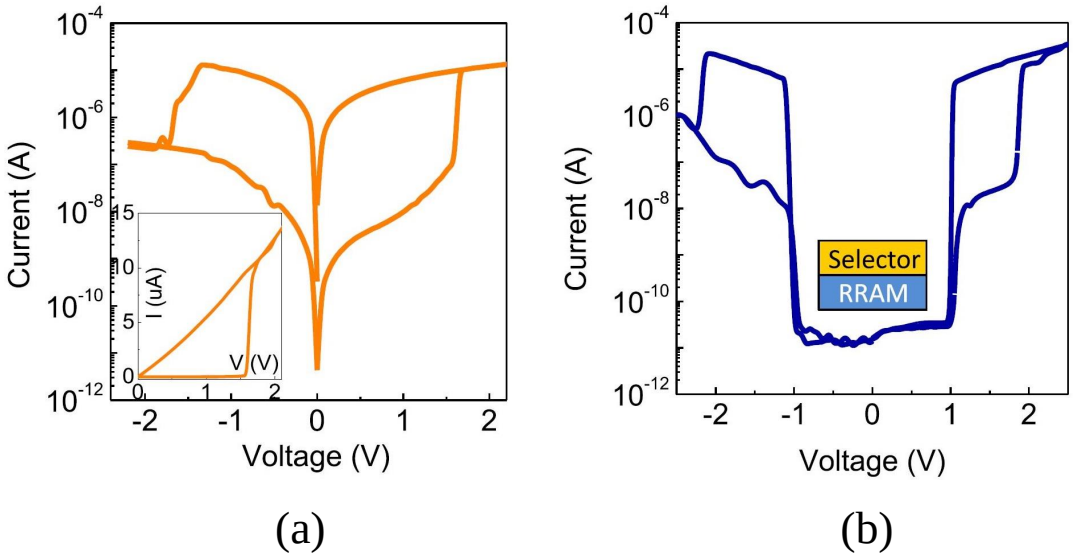

Figure 3.17: The effect of a selector device, (a) selectorless bipolar RRAM, and (b) bipolar RRAM integrated with a selector [85].

on-state for a wide range of voltages whereas the 1S1R device exhibits high current only above the selector threshold voltage. Below the selector threshold voltage, the current is orders of magnitude smaller. A selector therefore significantly suppresses the current leaking through the unselected cells and has a profound effect on the energy consumption, read margin, and voltage losses across the parasitic interconnect resistances in large crossbar arrays.

3.4 Summary

Crossbar arrays based on non-volatile resistive devices such as RRAM, PCM, and MRAM have become a promising technology due to the advantages of small form factor, non-volatility, and scalability. While these devices behave similar to a variable resistor, due to different materials and switching mechanisms, these devices exhibit different strengths and weaknesses. RRAM and PCM based devices exhibit a higher and wider range of resistances; however, at lower switching speeds and endurance as compared to MRAM. MRAM exhibits a lower resistance over a narrower range but higher switching speeds and higher endurance. As a result, different kinds of crossbar arrays are considered within different levels of the memory hierarchy depending upon the type of device. Resistive crossbar arrays leak significant current due to partially biased unselected cells and sneak paths, limiting the size and capacity of a single array. A resistive memory system therefore consists of multiple crossbar arrays rather than one large array. Selector devices are used to lower leakage and sneak path currents to improve performance and increase the capacity of the array. While MRAM crossbar arrays use transistor based selectors, arrays based on PCM and RRAM use two terminal selectors, achieving higher density in a smaller area at the expense of less isolation.

Chapter 4

On-Chip Power Delivery with Fully Integrated Voltage Regulators

Voltage converters are placed on-chip to improve the spatiotemporal granularity of the on-chip power delivery and management system as well as to decrease power consumption and enhance quality of power (QoP). A voltage converter is a circuit that produces current at a desired voltage. A voltage converter that regulates the output voltage using an internal closed-loop feedback mechanism is referred to as a voltage regulator. A voltage regulator regulates the output voltage by monitoring the output and adapting the circuit with respect to the current load variations while maintaining a specific constant voltage.

Benefits of Integrated Voltage Regulators

In recent years, voltage regulators have been placed closer to the die, as shown in Fig. 4.1, due to several advantages [86–88]. Bringing the voltage regulator from the

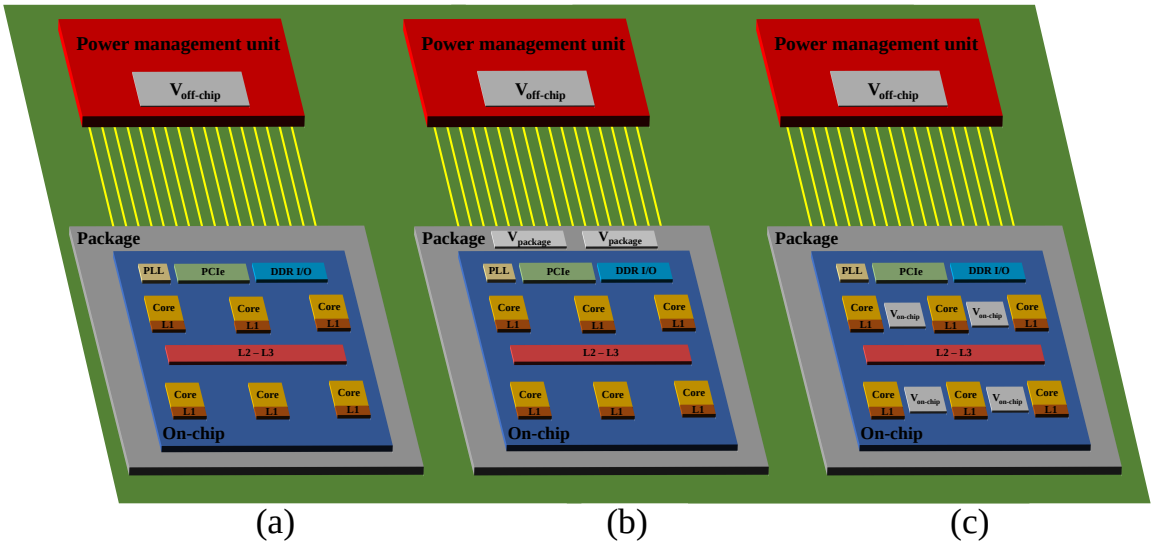

Figure 4.1: Regulator to die distance, (a) physically distant high efficiency off-chip regulator, (b) in-package voltage regulator, and (c) multiple point-of-load on-chip voltage regulators producing heterogeneous voltages.

board to the package to on-chip significantly increases the communication bandwidth of the voltage regulator. As a result, the on-chip voltage supply can be configured to quickly adapt to workload variations at a high speed (sub–microseconds), lowering the power consumption with faster dynamic voltage scaling (DVS). In addition, by distributing multiple on-chip regulators, a separate voltage supply can be dedicated to different sub-systems within an IC. For example, placing separate voltage regulators within each core in a processor can enable per core DVS. This local optimization of voltage sources significantly lowers the power consumed by the idle sub-systems while simultaneously supporting a high voltage core operating under aggressive workload conditions. Different voltages are available at different times to match a diverse variety of workload conditions at different locations within an IC.

Integrating the voltage regulators on-chip also improves the QoP. The QoP of an integrated circuit greatly affects the performance and is critical to the robustness of a system. The QoP refers to the ability of the on-chip power delivery network to maintain a constant, accurate, and noiseless voltage regardless of variations in the load current [18]. A high QoP implies a voltage supply that is highly accurate and resilient to load current variations with small voltage droops. Integrating the voltage regulators on-chip improves the QoP due to several reasons. Bringing the voltage regulators closer to the load decreases the parasitic impedance between the voltage source and the load. Moreover, high bandwidth regulators can more quickly respond to fast load variations, reducing the impedance observed from the load into the power grid [88], while providing a constant voltage for a wide range of load conditions.

The benefits of on-chip voltage regulators have led to commercial ICs with multiple fully integrated regulators. Some example ICs with embedded voltage regulators are shown in Fig. 4.2. In Fig. 4.2a, an IBM DDR3 I/O core contains eight distributed voltage regulators to supply power to the noise sensitive high speed circuits. In Fig. 4.2b, the power management unit is integrated with the cellular system to save board area and reduce cost. Moreover, the Intel Xeon [60] and IBM Power8 [90] processors embed on-chip voltage regulators to decrease the power dissipated by the high performance ICs.

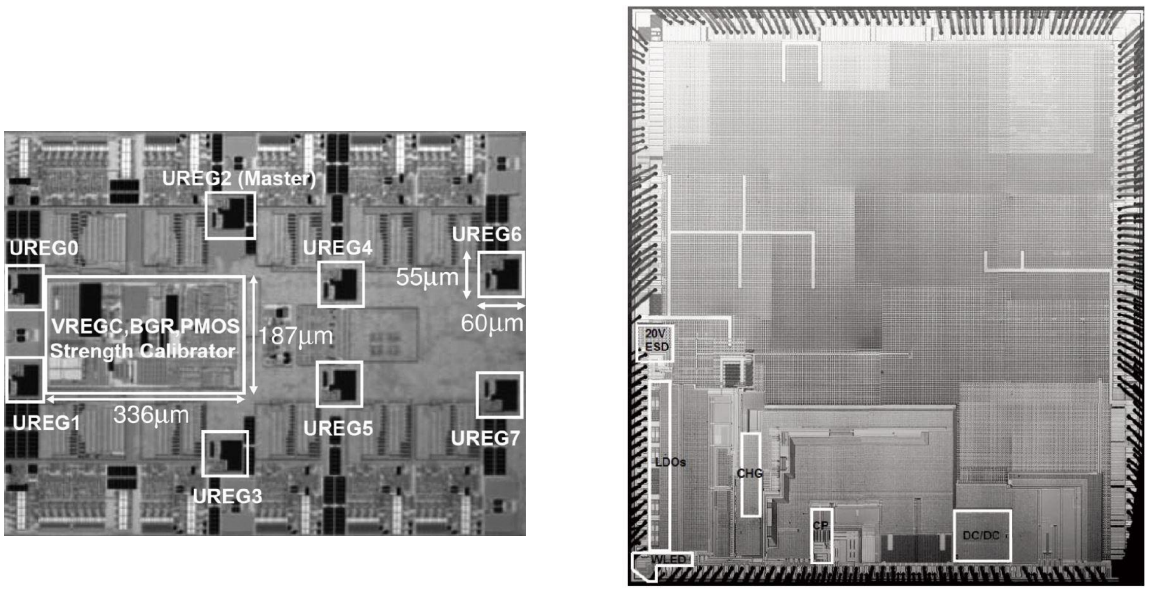

Figure 4.2: Fully integrated on-chip voltage regulators, (a) eight distributed voltage regulators across an IBM DDR3 I/O core [89], and (b) integrated power management on a cellular mobile IC from ST Ericsson [87].

Strengths and Weaknesses of Different Voltage Regulators

On-chip switching, switched-capacitor, and linear voltage regulators are typically used as on-chip regulators. These different types of regulators exhibit different strengths and weaknesses. The choice of regulator type depends upon the application-specific power efficiency requirement, load conditions, and area limitations. Switching and switched-capacitor regulators can down convert ($V_{in} > V_{out}$) as well as step up ($V_{in} < V_{out}$) the input voltage while linear regulators can only down convert the input voltage. A switching converter that steps up the input voltage is called a boost converter. A switching converter that down converts the input voltage is called a

buck converter. Fully integrated on-chip switching and switched-capacitor regulators exhibit moderate to high power efficiency ranging from 70% to the mid 80% and are typically sensitive to the output current and conversion ratios [91, 92]. Several exceptions achieving low 90% efficiencies have been demonstrated [88, 93]. The on-chip switching buck converter described in [88] exhibits high efficiency ($> 90\%$) over a wide range of load conditions. This regulator, integrated on a microprocessor, uses high quality off-chip air core inductors (integrated within the package), enabling high output power while retaining good efficiency without significantly increasing the on-chip area. Fully integrated switching converters with on-chip inductors however occupy significant area to produce high current (> 100 mA) and exhibit significantly lower conversion efficiencies due to the low quality on-chip inductors [94]. In [93], a switched-capacitor regulator with a peak efficiency of 93% is demonstrated. This fully integrated converter, however, occupies significant on-chip area (0.366 mm^2) and only produces up to 1 mA while exhibiting high efficiency for a narrow range of conversion ratios. Alternatively, the power efficiency of linear regulators is high under low conversion ratios and low under high conversion ratios. The efficiency of different types of fully integrated voltage regulators with respect to the conversion ratio is shown in Fig. 4.3. Note that switched-capacitor and switching regulators exhibit higher power efficiency at higher conversion ratios as compared to linear regulators. Moreover, switched-capacitor converters can provide a limited discrete set of

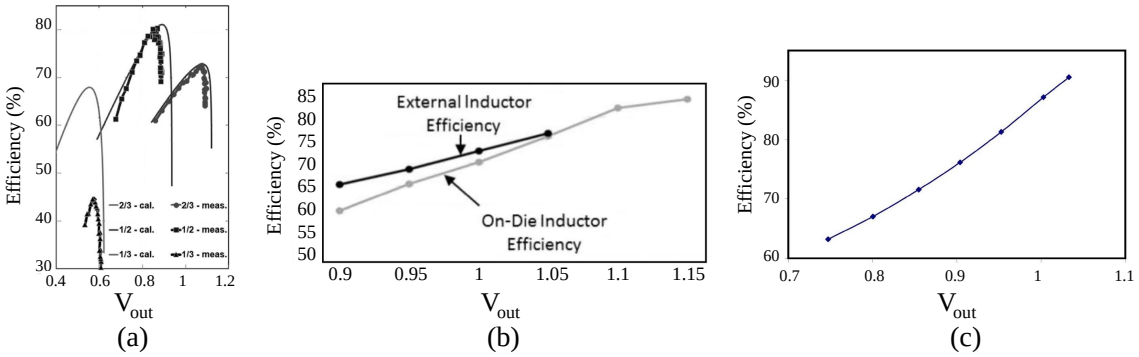

Figure 4.3: Power efficiency of voltage regulators for different conversion ratios, (a) switched-capacitor converter with 2 volt input voltage [91], (b) switching buck regulator with 1.5 volt input voltage [92], and (c) linear regulator with 1.1 volt input voltage [90].

voltages at a reasonable conversion efficiency whereas switching and linear regulators can provide a wider range of voltages. In terms of load dependent variations, linear regulators maintain power efficiency for a wide range of load conditions under a fixed conversion ratio if the quiescent current is significantly lower than the load current. The effect of the output current on the power conversion efficiency for different types of fully integrated voltage regulators is illustrated in Fig. 4.4. Note that switching and switched-capacitor converters are highly sensitive to the output current as compared to linear regulators that maintain the same power efficiency over a wide range of load conditions. In terms of area, linear regulators require significantly less area as compared to switching and switched-capacitor regulators since on-chip linear regulators do not require bulky inductors or capacitors. As a result, a large amount of on-chip current can be regulated without significantly increasing the dedicated physical area. On-chip linear regulators are therefore highly appropriate when multiple

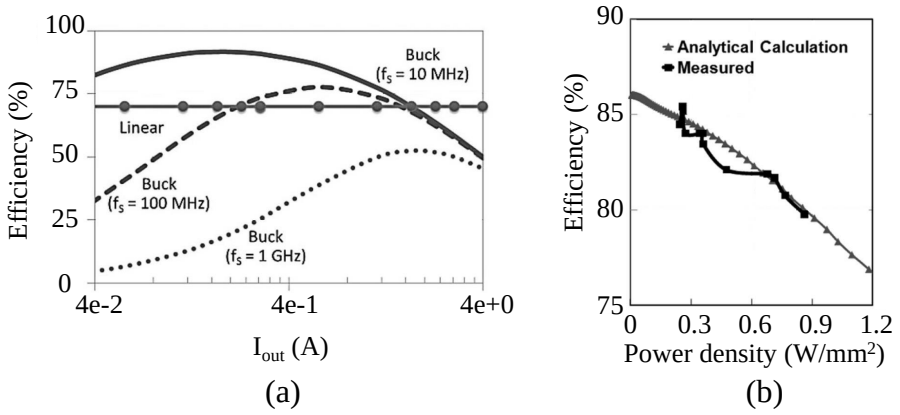

Figure 4.4: The effect of the output current on power conversion ratios, (a) switching buck regulator and linear regulator [95], and (b) switched-capacitor converter [91].

fully integrated on-chip regulators are needed to provide a wide range of voltages with low area. Several industrial microprocessors have been developed with several fully integrated on-chip linear regulators to decrease power consumption. In [96], a four core Intel processor contains four on-chip linear regulators, and in [90], a twelve core IBM processor utilizes 64 on-chip linear regulators to decrease the power by enabling per core DVS.

In this chapter, background on linear regulators is provided. In Section 4.1, the working mechanism of a linear regulator is described. In Section 4.2, on-chip linear regulators and the difference between a fully integrated and a standard linear regulator are reviewed. In Section 4.3, the stability and design challenges of fully integrated on-chip linear regulators are discussed. In Section 4.4, some conclusions are offered.

4.1 Low Dropout Voltage Regulator

The working mechanism of a linear regulator is similar to down converting a voltage using a resistive voltage divider with two resistors, as shown in Fig. 4.5a. Assuming the bottom resistor represents the load condition determined by the work-

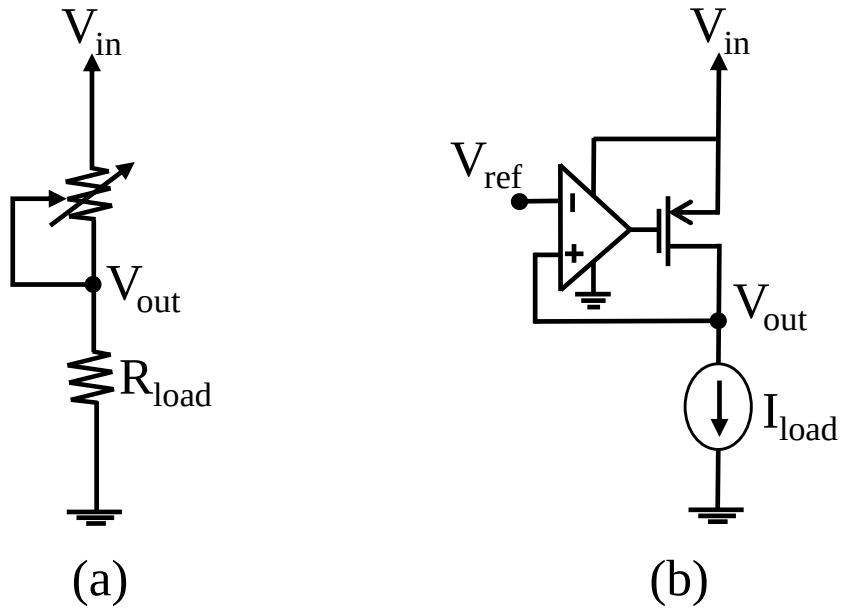

Figure 4.5: Linear regulator, (a) simplified representation, and (b) practical circuit.

load, the upper resistor connected to the input voltage is varied to retain a constant output voltage for a wide range of load conditions (i.e., different resistances, since the workload is dynamic). The upper resistance detects changes in the load current by monitoring the output voltage and responding to ensure the down converted voltage remains roughly constant. A practical version of this regulation mechanism is shown in Fig. 4.5b. The variable upper resistor is realized using a pass transistor. The

linear regulator is described as a low dropout (LDO) regulator if a PMOS (p-type metal oxide semiconductor) pass transistor is used. This naming convention is due to the greater output voltage headroom when using a PMOS device as compared to an NMOS (n-type MOS) device. With an NMOS pass transistor, the drain–to–source voltage, also known as the dropout voltage, needs to be sufficiently high to retain an output voltage at least a threshold voltage below the gate voltage of the pass transistor. This minimum requirement on the dropout voltage limits the power efficiency. Replacing the NMOS pass transistor with a PMOS pass transistor solves this problem since the source terminal is connected to the input of the regulator rather than the output, enabling regulation at lower dropout voltages (hence, a low dropout regulator or LDO). For the rest of this dissertation, an LDO regulator is assumed in the discussions of linear regulators. This discussion can also be applied to linear regulators using NMOS pass transistors.

The pass transistor within the LDO is typically driven by the error amplifier. The error amplifier has two inputs, one input for the reference voltage and the other input for the output voltage. The reference voltage sets the output voltage of the regulator. The regulator reaches a steady state voltage under a fixed workload condition when, ideally, the output voltage is equal to the reference voltage. The reason the reference voltage is not directly connected to the load (as opposed to a linear regulator) is because of the insufficient current drive of the reference generator circuit. Reference

voltage generators are typically connected to high impedance terminals such as the input of an operation amplifier (op-amp). The primary concern of a reference generator is to produce a stable and accurate voltage that is highly resilient to process, temperature, and voltage (PVT) variations. The reference voltage in a regulator is compared to the output voltage using an error amplifier. The error amplifier produces an amplified signal called an error voltage which is the difference between the reference and output voltages multiplied by the amplifier gain. This voltage sets the source–to–gate voltage to ensure the dropout voltage remains constant for different output currents. This working mechanism governs the basic principles of any type of linear regulator. The design techniques and challenges however greatly differ between LDO regulators with and without large output capacitors [97–99].

4.2 Fully Integrated On-Chip LDO Regulators

The primary difference between a fully integrated on-chip linear regulator and a standard off-chip linear regulator is the output capacitor. Voltage regulators are typically supported by an output capacitor to supplement the output current during those fast load transitions that surpass the regulator response time (due to the limited bandwidth of the regulator), as shown in Fig. 4.6. The output capacitor is critical to decrease the voltage droop and lower power noise. While standard linear regulators incorporate a large off-chip output capacitor (on the order of μF), fully integrated

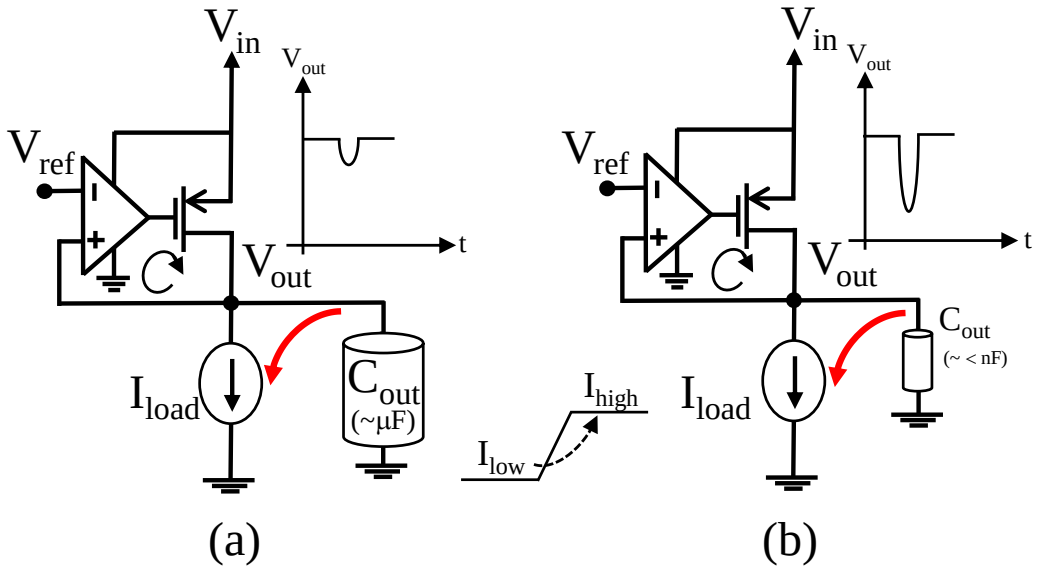

Figure 4.6: Linear regulator, (a) with large output capacitor, and (b) with small output capacitor.

on-chip regulators do not include large output capacitors due to constraints on area. Fully integrated on-chip LDO regulators are therefore also called capacitorless LDO regulators [98,99]. Several fully integrated on-chip LDO regulators have been demonstrated, incorporating large output capacitors either with a dedicated pin to connect to an off-chip capacitor or by utilizing significant die area. In both cases, the physical area is significant and may not be feasible for those ICs requiring multiple on-chip linear regulators. In this discussion, fully integrated on-chip LDO regulators refer to on-chip LDOs with small output capacitors (up to a few hundred picofarads).

The absence of a large output capacitor significantly affects the design and performance of an LDO for two reasons. First, the absence of an output capacitor requires

additional techniques to provide a fast response to high speed load variations to prevent large voltage droops. Second, the stability of the LDO, typically ensured by the large output capacitor, needs to be maintained by a different approach.

4.3 Stability Analysis

To investigate the effects of the output capacitor on LDO stability, a simplified small–signal model of a regulator is considered, as shown in Fig. 4.7 (based on the circuit shown in Fig. 4.5b). The output of the error amplifier and the output of the

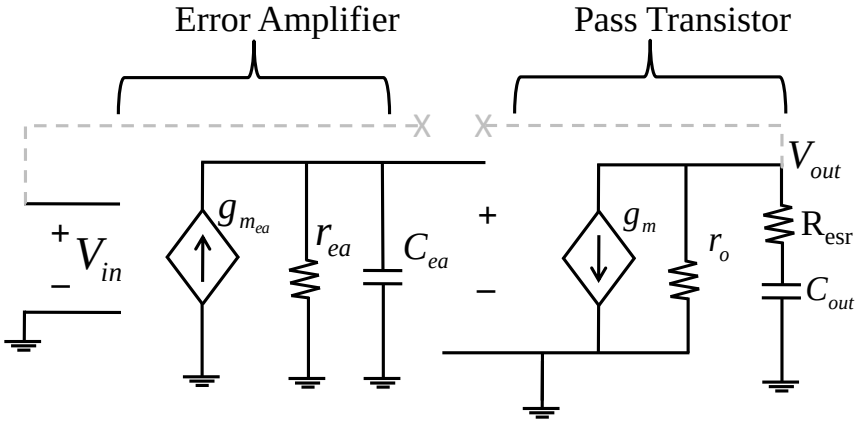

Figure 4.7: Small–signal model of an LDO regulator.

regulator create two poles, w_{p1} and w_{p2}, at an angular frequency, respectively,

$$w_{p1} = \frac{1}{C_{ea} r_{ea}}, \tag{4.1}$$

$$w_{p2} = \frac{1}{C_{out}(r_o + R_{esr})}, \tag{4.2}$$

where C_{ea} and r_{ea} are, respectively, the output capacitance and output resistance of the error amplifier, and C_{out}, r_o, and R_{esr} are, respectively, the output capacitance, output resistance, and equivalent series resistance (ESR) of the output capacitor. In addition, a left-half-plane (LHP) zero is formed due to the parasitic series resistance of the output capacitor R_{esr} at the angular frequency,

$$w_{z1} = \frac{1}{C_{out}R_{esr}}. \tag{4.3}$$

The stability of the open-loop small–signal model is evaluated using phase margin. The phase margin is the difference between 180^o and the total phase shift of the circuit at the unity gain frequency. Analytically, the phase margin is

$$180^o - \mid \angle H(0) - \angle H(f_{0db}) \mid, \tag{4.4}$$

where $\angle H(f)$ is the phase of the transfer function and f_{0db} is the unity gain frequency. The transfer function $H(w)$ of the small-signal model shown in Fig. 4.7 is

$$H(s) = \frac{V_{out}}{V_{in}}(s) = \frac{A_{ol}(1 + \frac{s}{w_{z1}})}{(1 + \frac{s}{w_{p1}})(1 + \frac{s}{w_{p2}})}, \tag{4.5}$$

where A_{ol} is the open-loop gain of the LDO. The higher the phase margin, the more stable the system. Conversely, the lower the phase margin, the closer the system is to unintentionally invert the input. This effect converts negative feedback into positive feedback, causing instability. Specifically, if the phase margin is nonpositive,

the system is unstable. Unstable circuits are dysfunctional and typically exhibit an oscillatory response. The minimum required phase margin depends upon the application, varying between 45^o and 90^o (60^o is common) [100,101]. Circuits sensitive to process and temperature variations are often overdesigned to maintain a minimum phase margin of 65^o [101]. A phase margin of at least 40^o is typically required for voltage regulators [88]. To increase the phase margin, single pole behavior below the unity gain frequency is typically preferred.

The gain and location of the poles and zeros depend upon the DC bias of the regulator set by the load current. Hence, the load condition plays an important role on the stability of the regulator. In particular, the output pole w_{p2} exhibits large variations with respect to the load current due to the output resistance of the pass transistor r_o. Under heavy load conditions (milliamperes), the pass transistor operates in the linear region, exhibiting low output resistance and shifting w_{p2} to a higher frequency. As the load current decreases, the pass transistor moves from the linear region into the saturation region and eventually into the subthreshold region under light load conditions (< 1 mA), shifting w_{p2} to a significantly lower frequency. w_{p1} also varies depending upon the load condition since the DC bias changes. The dominant pole however is a weak function of the load current and does not vary as much as w_{p2}.

4.3.1 Effect of Output Capacitance on LDO Stability

Regulators with a large output capacitor ($C_{out} >> C_{ea}$) exhibit a dominant pole formed by the output capacitor (i.e., $w_{p2} > w_{p1}$). In addition, due to the large output capacitance, the equivalent series resistor (ESR) of the output capacitor generates a LHP zero located within the LDO bandwidth [102]. The LHP zero improves the stability of the regulator by canceling the non-dominant pole, reducing the phase shift, as shown in Fig. 7.1a. Off-chip linear regulators with a large output capacitor therefore typically require a small ESR to maintain a stable system [102].

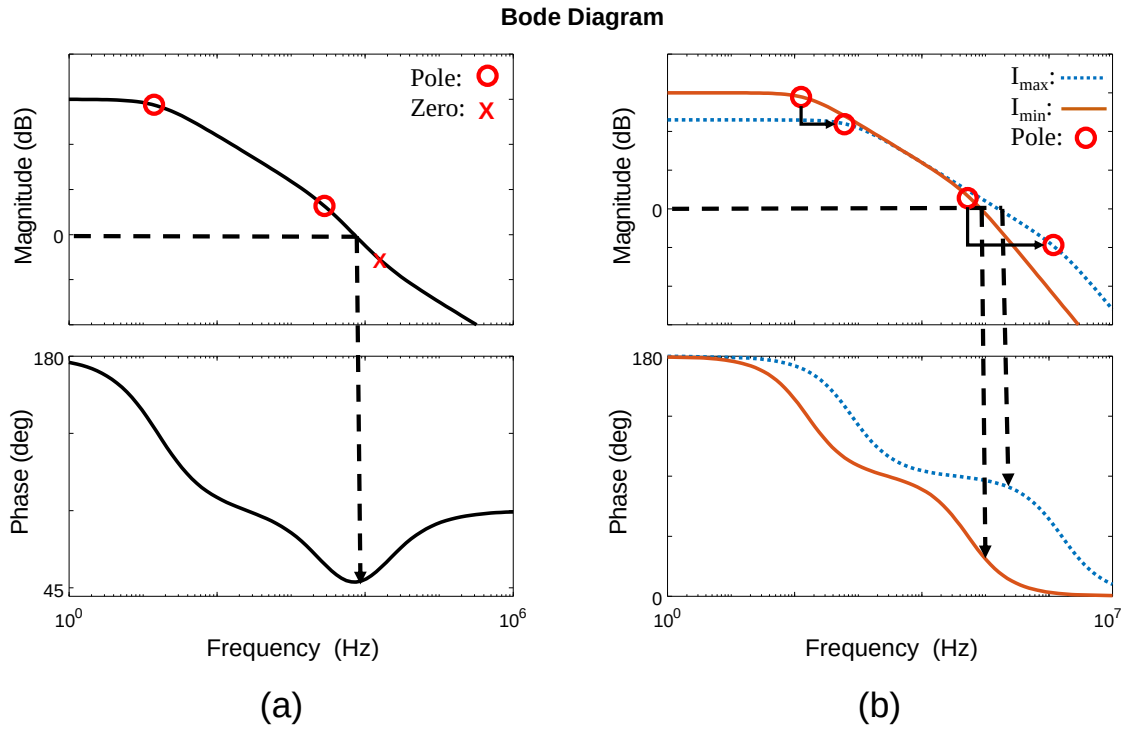

Figure 4.8: Frequency response of an LDO regulator, (a) large output capacitor, and (b) small output capacitor.

Conversely, fully integrated capacitorless on-chip LDOs exhibit a dominant pole generated by the error amplifier and a non-dominant pole generated by the output capacitance. The frequency response of a capacitorless LDO is shown in Fig. 7.1b. The stability of a capacitorless LDO depends significantly upon the load condition since the output pole is the non-dominant pole. The ESR generated zero shifts to a higher frequency, typically beyond the LDO bandwidth due to the small output capacitance. The zero therefore has a negligible effect on the stability of the system. Since the non-dominant pole shifts to a higher frequency under heavy load conditions (i.e., high output current), the separation between the two poles, w_{p1} and w_{p2}, increases, thereby also increasing the phase margin. Under light load conditions, however, the two poles are close in frequency, jeopardizing the stability of the system. Hence, the worst case stability of a capacitorless LDO regulator occurs under light load conditions (see Fig. 7.1). Note that the open loop gain also affects the stability. A low gain LDO is more stable since the unity gain frequency is lower. Sufficient gain however (typically above 40 to 50 dB) is necessary to reduce the steady state error between the output voltage and the reference voltage.

To improve the stability under light load conditions, a pole splitting approach is often used [98, 99]. Simply lowering the bandwidth of the error amplifier to decrease w_{p1} shrinks the overall LDO bandwidth. While this approach can increase the pole splitting and therefore the phase margin, the speed of the LDO significantly decreases,

increasing the voltage droop. To quantify this effect, the output impedance of the LDO regulator is considered. The output impedance is

$$Z_{out}(s) \approx \frac{r_o(1+\frac{s}{w_{p1}})}{A_{ol}+(1+\frac{s}{w_{p1}})(1+\frac{s}{w_{p2}})}. \tag{4.6}$$

The output impedance initially increases due to the more narrow bandwidth of the error amplifier and eventually decreases due to the output capacitance of the LDO regulator, as shown in Fig. 4.9. Note that decreasing the bandwidth of the LDO

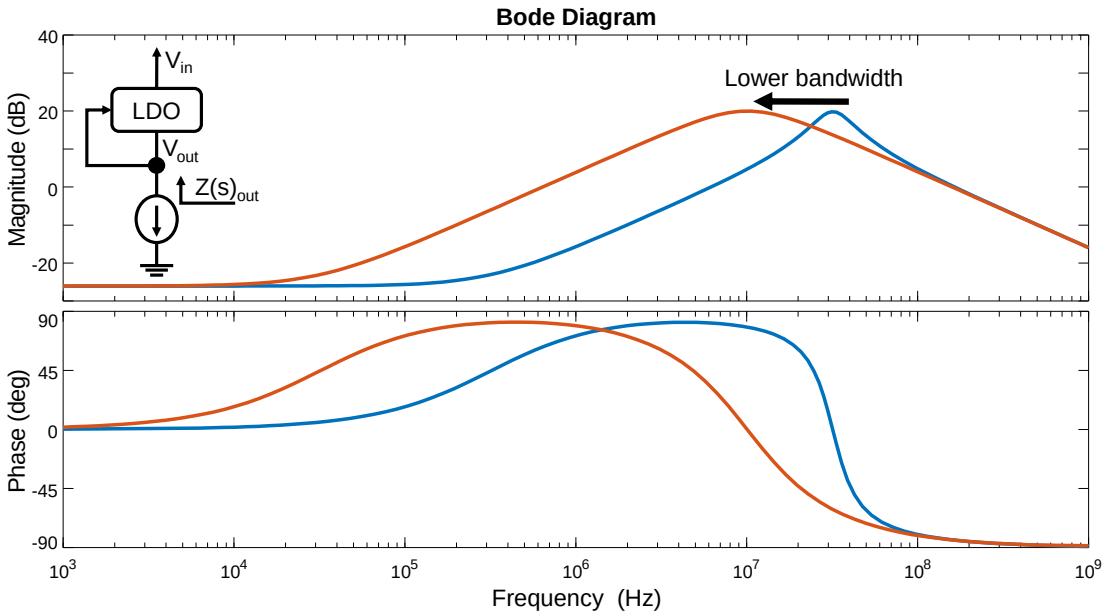

Figure 4.9: Closed loop frequency response of the output impedance. The ESR of C_{out} is ignored. Decreasing the bandwidth of the error amplifier increases the output impedance, increasing the voltage droop.

(i.e., decreasing w_{p1}) results in a higher output impedance over a wider range of frequencies as compared to a higher bandwidth. The voltage droop followed by a

transient load variation is therefore larger for low bandwidth LDOs. As a result, alternative approaches are required to improve the phase margin without significantly increasing the voltage droop.

4.3.2 Improving LDO Performance

Alternate solutions have been offered to enable pole splitting while maintaining a low voltage droop [98]. For example, in [99] and [103], a slew rate enhancement approach is adopted to reduce the charge–discharge time of the pass transistor gate capacitance while splitting the poles. The proposed LDO topology uses a current amplifier in series with a capacitor between the input and output of the pass transistor (i.e., the gate and source terminals), as shown in Fig. 4.10a. The amplified capacitor current caused by the instantaneous change in the LDO output increases the charge–discharge rate of the gate capacitance, lowering the voltage droop. In [104], a digitally controlled active compensation network is proposed, as shown in Fig. 4.10b. By varying the resistance and capacitance of an RC compensation network connected between the input and output of the pass transistor, a variable LHP zero is generated. The variable zero cancels the non-dominant pole w_{p2}, increasing the phase margin. To lower the voltage droop, an adaptive boosting technique is used. By tracking the variations in the load conditions, the quiescent current of the error amplifier is temporarily increased (i.e., boosted) to lower the voltage droop during fast load

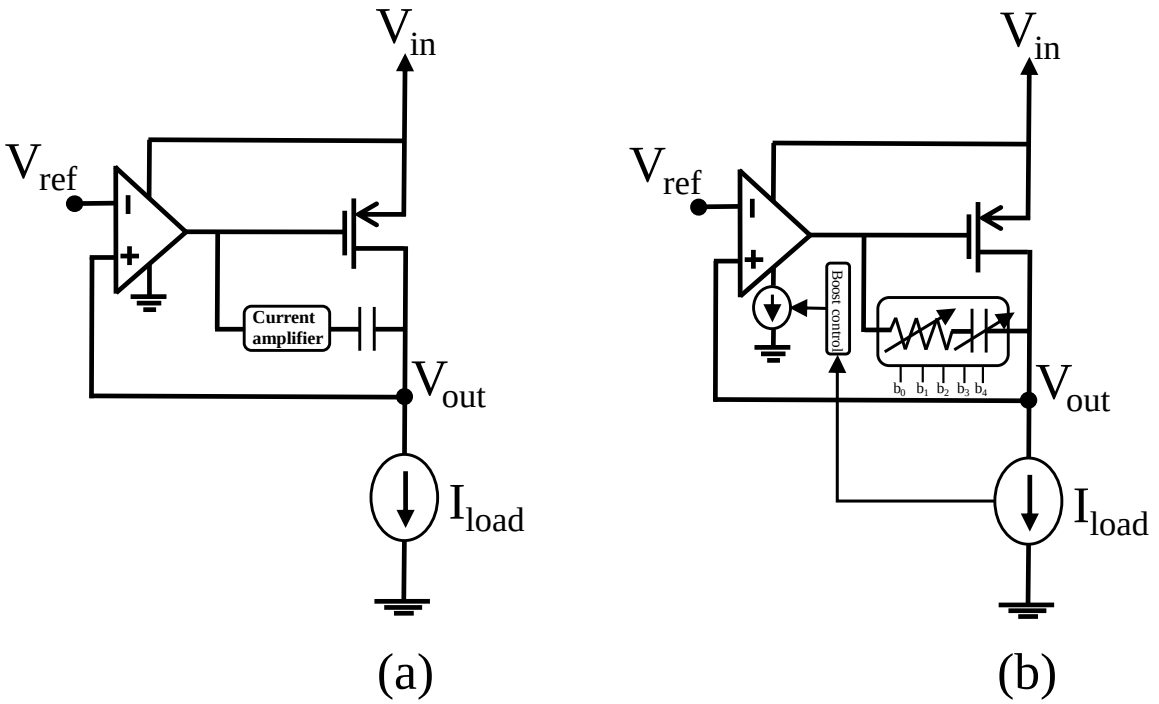

Figure 4.10: Different LDO topologies, (a) slew-rate enhancement using current amplifier [99], and (b) adaptive RC compensation with an adaptive boost technique [104].

transitions. Many other approaches have been proposed to improve the stability and response time of fully integrated LDOs [89, 98, 105–108]. Ultimately, the choice of LDO depends upon the strengths and weaknesses offered by the circuit topology which is determined by the requirements of the application [98].

4.4 Summary

In this chapter, a general overview of the advantages of embedding on-chip voltage regulators is provided. Fully integrated on-chip voltage regulators enable power

management at higher speeds, supporting fast DVS. Moreover, embedding multiple regulators enables local optimization of the voltage supply, providing a higher voltage to the high performance cores while reducing the voltage of the idle or low performance subsystems. Integrated voltage regulators can therefore be used to significantly decrease power while improving the QoP.

Switching, switched–capacitor, and linear regulators are embedded on-chip. Linear regulators, among the different types of regulators, require the least area since no bulky inductors or large capacitors are necessary. A fully integrated on-chip LDO typically has an output capacitor on the order of a few hundred pF whereas a standard off-chip LDO has a large output capacitor on the order of a few μF. The significant reduction in output capacitance greatly changes the design of an LDO since stability is no longer ensured. Fully integrated LDOs therefore exhibit a dominant pole due to the error amplifier as opposed to the output capacitance. Moreover, the LHP zero generated by the ESR is at too high frequency due to the small output capacitance. As a result, the stability of an LDO is highly sensitive to the load, exhibiting the worst case stability under light load conditions. A pole splitting approach is typically adopted to increase the difference between the dominant and secondary pole frequencies. Simply lowering the LDO bandwidth to increase the phase margin increases the voltage droop since the regulator response time is significantly less. As an alternative, several solutions have been offered such as slew rate enhancement using

current amplifiers or adaptive RC compensation networks with adaptive boost. These techniques enable a fast output to input feedback path, decreasing the voltage droop while enabling sufficient pole splitting to increase the phase margin.

Chapter 5

Modeling Size Limitations of Resistive Crossbar Array With Cell Selectors

Resistive crossbar arrays were developed before the invention of emerging memory technologies such as MRAM, RRAM, and phase change memory (PCM) [109, 110]. With the recent development of RRAM devices [111], resistive crossbar arrays, for use in memory, have gained increasing popularity due to the advantages of $4F^2$ density and non-volatility. Existing analyses of resistive crossbar arrays show that the array size is limited by the degradation in read margin and voltage loss across the cells due to parasitic interconnect resistances, sneak path leakage currents, and on-off resistance ratios [112–115]. These analyses have been primarily simulation based. In [116], a matrix based theoretical solution is presented for solving the voltages and currents of each cell within a crossbar array. This study however does not provide intuitive models to support the design of resistive crossbar arrays due to the complexity of

large arrays. Moreover, large matrix sizes are computationally complex. Therefore, simple analytic models that can intuitively characterize the limitations imposed on resistive crossbar arrays and project device and circuit requirements for large scale arrays would be useful [117].

In this chapter, three challenges in designing resistive crossbar array are considered; the driver size, voltage degradation across the selected cell, and the read margin. For each of these issues, models have been developed which provide intuition into the design of resistive crossbar arrays while also clarifying device requirements and limitations on the array size as interconnects continue to scale. These models are valid for both unipolar and bipolar memory elements. Moreover, different biasing schemes for writing and reading are compared to clarify possible advantages and design tradeoffs.

In Section 5.1, models of the driver size, voltage degradation across the selected cell, and read margin are described and compared to simulation. In Section 5.2, these models are considered under different biasing schemes to enhance nonlinearity and to mitigate size limitations. In Section 5.3, projected device requirements for large arrays are discussed. In Section 5.4, an example of the application of the proposed models to the crossbar array design process is demonstrated. In Section 5.5, some conclusions are offered.

5.1 Models of Crossbar Array Design Parameters

Expressions that model three primary design parameters of resistive crossbar arrays, the driver size, voltage degradation across the selected cell, and read margin are introduced in this section. For simplicity, an equal number of rows and columns are assumed under worst case conditions. For the write operation, the $\frac{V}{2}$ biasing scheme [118] is considered. For the read operation, the scheme in which a read voltage is applied to the selected row while connecting the remaining portion of the rows to ground and the columns to sense amplifiers [119] is considered, as shown in Fig. 5.1. In the following sub-sections, the driver resistance, voltage degradation across the selected cell, and read margin are discussed.

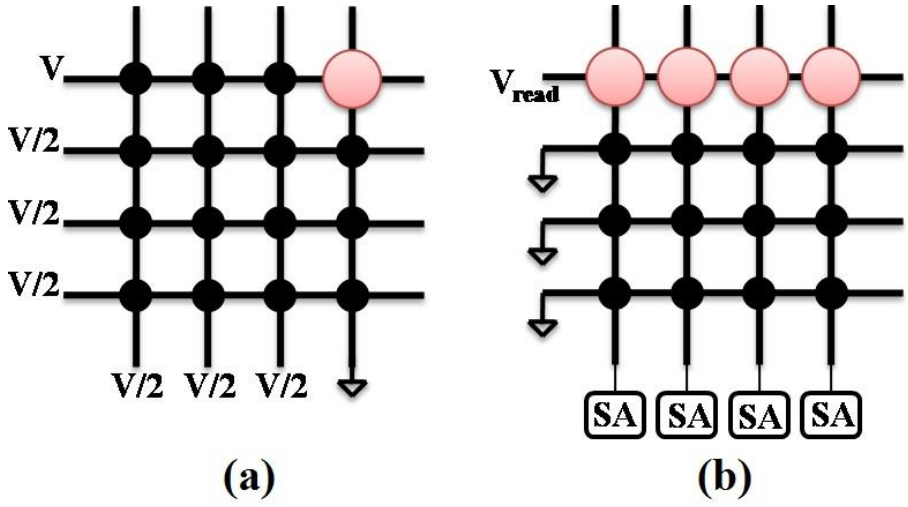

Figure 5.1: Biasing scheme for a crossbar array when (a) writing to a cell, (b) reading from a cell.

5.1.1 Driver size

An important advantage of a crossbar structure in memory systems is physical density. Resistive crossbar arrays however require large peripheral circuits due to the high current required to drive large arrays of closely packed devices. The physical area of a crossbar array is ultimately determined by the cell size and the peripheral circuitry, as well as the drivers.

The driver resistance is the output resistance of the driving circuit, illustrated in Fig. 5.2. This output resistance depends upon the input resistance of the selected row as well as the voltage drop across the selected cell. Although the lower bound

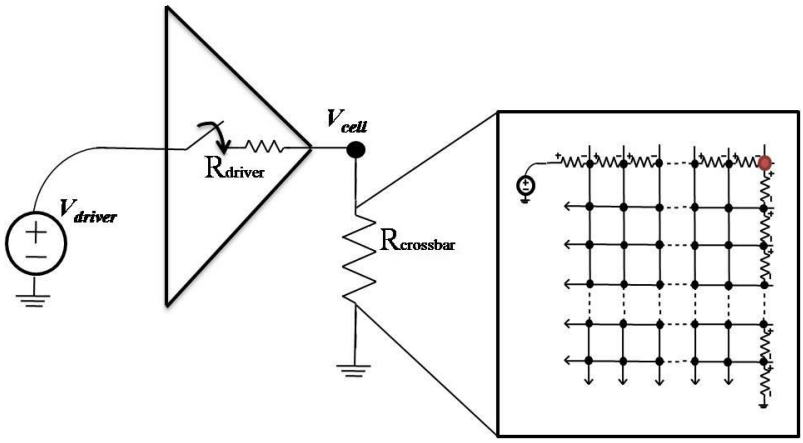

Figure 5.2: Driver circuit.

on the resistance of a single memory element could reach tens of kilo-ohms in an RRAM crossbar structure, the effective resistance between the driver at a selected row and the sense amplifier at a selected column(s) drastically decreases with larger

array size. Since the effective resistance is also dependent on the number of selected cells, the driver resistance varies depending upon whether a read or write operation is executed. In this analysis, the interconnect resistance and input resistance of the sense amplifier are considered to be negligible.

For a write operation, the worst case condition occurs when the selected cell is initially in the on-state and switches to the off-state. Since selector devices are in series with the resistive memory elements, a nonlinear relationship between the cell voltage and current exists. Hence, the resistance of each cell varies nonlinearly with the voltage across the cell. This nonlinearity is described by the nonlinearity factor. For a worst case analysis, in which the highest current is required by the crossbar array, half selected cells are assumed to be in the on-state. Based on these assumptions, the following expression for the driver resistance at the selected row is

$$R_{driver(write)} = \frac{R_{ON}(\frac{V_{driver}}{V_{cell}} - 1)}{\frac{N-1}{K_r} + 1}, \tag{5.1}$$

where R_{ON} is the resistance of a memory cell (the selector and resistive memory element) in the on-state, V_{driver} is the driver output voltage when the driver resistance is zero, V_{cell} is the voltage drop across the selected cell, N is the array size (number

of rows or columns), and K_r is the nonlinearity factor,

$$K_r = \frac{I_{cell}(V_{write})}{I_{cell}(V_{write}/2)} = 2 \times \frac{R_{ON|V_{write}/2}}{R_{ON}}, \tag{5.2}$$

where $R_{ON|V_{write}/2}$ is the on-state cell resistance when the voltage across the cell (V_{cell}) equals half of the write voltage. K_r is the ratio of the current flowing through the selected cell to the current flowing through the half selected cell. The nonlinearity factor characterizes to what extent the current flowing into the unselected columns compares to the current flowing into the selected column.

For the case where multiple devices are selected, as in the case of a read operation, the constraint on the driver resistance becomes more stringent. During a single read operation, all of the cells on the selected row are selected. Considering the worst case condition when all of the selected cells are in the on-state, the driver resistance is

$$R_{driver(read)} = \frac{R_{ON}(\frac{V_{driver}}{V_{cell}} - 1)}{N}. \tag{5.3}$$

The driver resistance during a read operation is independent of the selector devices, and inversely proportional to the size of the crossbar array.

5.1.2 Voltage degradation across selected cell

An important limitation on the size of a resistive crossbar array is the interconnect resistance. With interconnect scaling, the resistance per cell has drastically increased, reaching 2.5 Ω for the 22 nm node [120]. It is therefore crucial to consider the effects of parasitic resistance when executing an operation. The worst case selected cell is farthest from the driver on the selected row, and farthest from ground on the selected column. For low nonlinearity factors, since the difference in resistance of a half selected cell in the on- and off-state remains significant, voltage degradation is data pattern dependent. To consider the worst case voltage degradation, all half selected cells and the selected cell are assumed to be in the on-state. The indicated cell shown in Fig. 5.1(a) is an example of a worst case cell for a 4 x 4 crossbar array during a write operation.

For writing, a circuit model of a crossbar array that includes the interconnect resistance along the selected row and column is considered. Furthermore, in this model, it is assumed that equal current flows through the half selected cells between the selected row and the unselected columns, as shown in Fig. 5.3. Based on this assumption, the voltage across the worst case selected cell is

$$\frac{V_{cell}}{V_{write}} = \frac{1}{\frac{NR_{int}}{R_{ON}}(\frac{N-1}{K_r}+2)+1}, \tag{5.4}$$

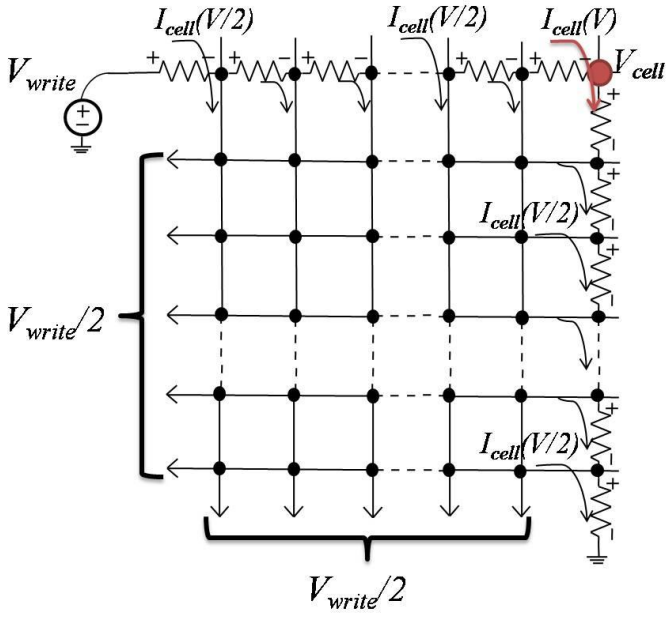

Figure 5.3: Circuit model of crossbar array during a write operation.

where R_{int} is the interconnect resistance per cell. As illustrated in Fig. 5.4, (5.4) agrees with SPICE, exhibiting a maximum error of 6.5% for voltage ratios above 0.5. Increasing interconnect resistance per cell decreases the voltage across the selected cell due to IR losses. This degradation becomes more severe and nonlinear as the array size scales. This behavior is due to increased current flow into the selected row and column with increasing number of rows and columns. Since the number of half selected cells increases with larger array size, the total current flowing into both the selected row and column increases. Larger array sizes therefore exacerbate the voltage degradation across the selected cell due to increased current flow and interconnect resistance.

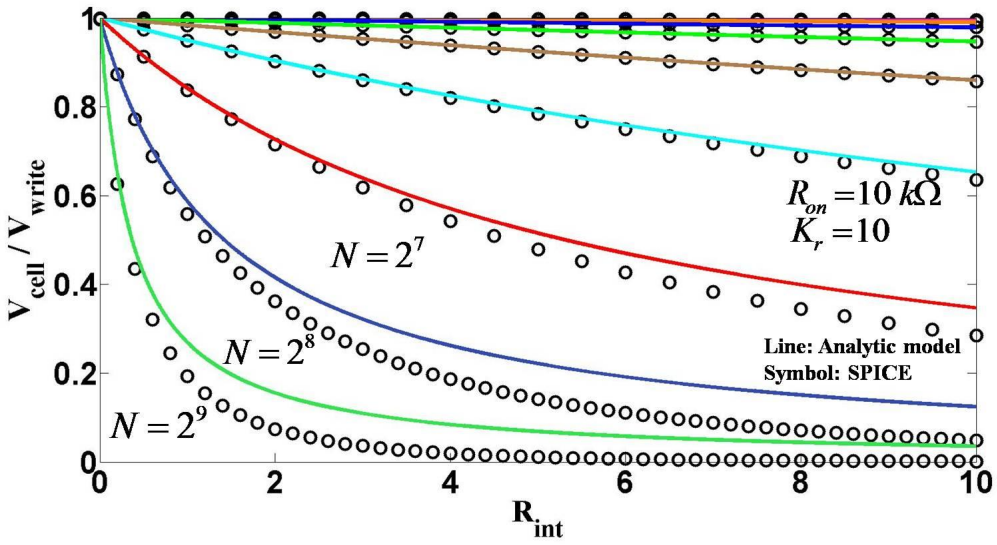

Figure 5.4: Ratio of the voltage drop across the worst case selected cell to the driver voltage during a write operation.

For reading, a circuit model of a crossbar array is shown in Fig. 5.5. The worst case cell for the read case is farthest from the driver on the selected row and farthest from the sense amplifiers on the selected columns. Since all of the cells in the same row are selected, any voltage degradation is data pattern dependent. The worst case condition occurs when all of the cells on the selected row are on, including R_{cell}. Based on the circuit model shown in Fig. 5.5, the ratio of the worst case cell voltage to the read voltage is

$$\frac{V_{cell}}{V_{read}} = \frac{1}{(1+\frac{N^2 R_{int}}{\alpha R_{sel(L)}})(1+\frac{1}{R_{ON}(\frac{1}{R_{sense}}+\frac{N-1}{R_{sneak}})})}, \tag{5.5}$$

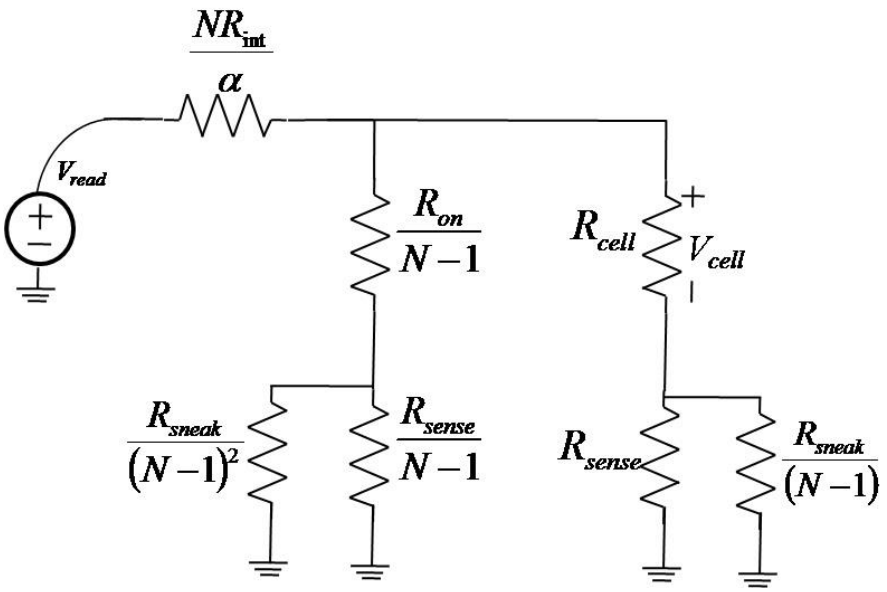

Figure 5.5: The circuit model of the crossbar array during a read operation where R_{sense} is the input resistance of the sense amplifier and R_{sneak} is the sneak path resistance of the resistive memory cells between the (un)selected column(s) and unselected rows.

where R_{sense} is the input resistance of the sense amplifier, R_{sneak} is the resistance of the cells between the (un)selected column(s) and unselected rows, α is a fitting parameter, and $R_{sel(L)}$ is

$$R_{sel(L)} = R_{ON} + (\frac{R_{sneak}}{N-1} || R_{sense}). \tag{5.6}$$

Expression (5.5) agrees with SPICE, exhibiting a maximum error of 6.6% for voltage ratios above 0.25, as illustrated in Fig. 5.6 (based on the parameter values of R_{ON}, α, R_{sense}, and R_{sneak} listed in Table 5.1). Similar to the degradation in cell voltage during a write operation, a larger interconnect resistance increases IR losses which is

further exacerbated with a larger array size. The degradation is more severe during a read operation since selection of a single row causes a full read voltage to drop across all of the cells in that row. All of the cells in the selected rows are therefore selected as opposed to selecting a single cell during a write operation.

Note that the value of R_{sneak} listed in Table 5.1 depends upon the voltage drop across the sense amplifier. It is assumed that the voltage drop is below the threshold voltage of the cell selector. The input resistance R_{sense} needs to be sufficiently low to maintain a low voltage at the sensing end of the columns which is ideally grounded. This low input resistance requirement imposes a serious challenge on the design of the sense amplifiers.

Table 5.1: Parameters for read operation

Parameters	Values
R_{ON}	10 kΩ
α	1.5
R_{sense}	100 Ω
R_{sneak}	$\frac{K_r}{2} \times R_{ON}$
R_{OFF}	10 MΩ
K_r	2×10^3
R_{tran}	R_{ON}

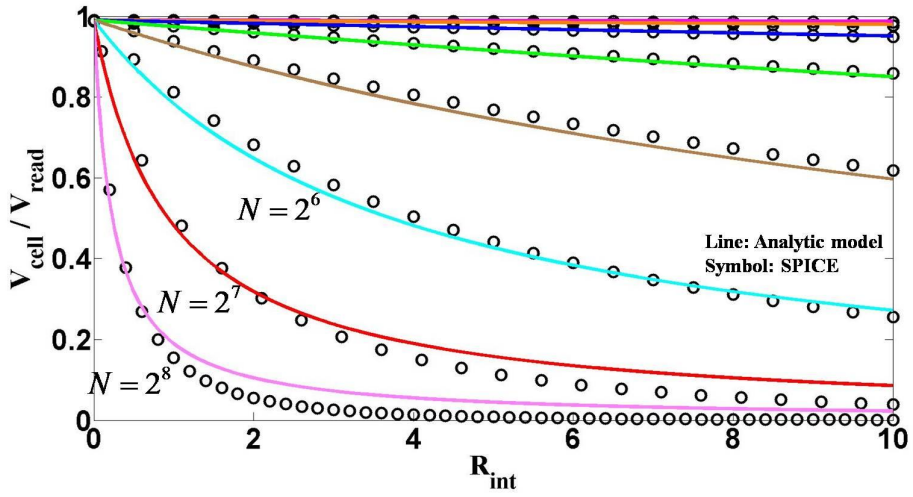

Figure 5.6: Ratio of the voltage drop across the worst case selected cell to the driver voltage during a read operation.

5.1.3 Read margin

An important figure of merit that determines the ability of a sense amplifier to distinguish between two states is the read margin. The read margin is

$$ReadMargin = \frac{(I_{sense(L)} - I_{sense(H)})R_{tran}}{V_{read}}, \tag{5.7}$$

where R_{tran} is the transresistance of the sense amplifier which is matched to R_{ON}, $I_{sense(L)}$ is the current flowing into the sense amplifier when the target cell is on, and $I_{sense(H)}$ is the current flowing into the sense amplifier when the target cell is off. The worst case read margin occurs when reading an on-state when all of the cells along the selected row are on, and when reading an off-state when all of the cells along the selected row are off. In the worst case condition, the selected row is farthest from

the sense amplifiers [see Fig. 5.1(b)]. Based on these worst case conditions and the circuit model shown in Fig. 5.5, $I_{sense(L)}$ and $I_{sense(H)}$ are described, respectively, as

$$I_{sense(L)} = \frac{V_{read}}{R_{ON}R_{sense}(\frac{1}{R_{sense}} + \frac{1}{R_{ON}} + \frac{N-1}{R_{sneak}})(1 + \frac{N^2 R_{int}}{\alpha R_{sel(L)}})} \tag{5.8}$$

$$I_{sense(H)} = \frac{V_{read}}{R_{OFF}R_{sense}(\frac{1}{R_{sense}} + \frac{1}{R_{OFF}} + \frac{N-1}{R_{sneak}})(1 + \frac{N^2 R_{int}}{\alpha R_{sel(H)}})} \tag{5.9}$$

where R_{OFF} is the resistance of a memory cell in the off-state, and $R_{sel(H)}$ is

$$R_{sel(H)} = R_{OFF} + \left(\frac{R_{sneak}}{(N-1)} || R_{sense}\right). \tag{5.10}$$

Expression (5.7), based on the expressions of $I_{sense(L)}$ and $I_{sense(H)}$ in, respectively, (5.8) and (5.9), agrees with SPICE, exhibiting a maximum error of 6.6% for read margins above 0.25 based on the parameter values listed in Table 5.1, as illustrated in Fig. 5.7.

Note the degradation in voltage across the cell with increasing array size (or interconnect resistance) which can fall below the threshold voltage of the selector. The selector resistance can dominate the overall memory cell resistance, making the on- and off-states indistinguishable. It is therefore crucial to consider the threshold voltage of the selector when estimating the read margin or voltage drop across a cell.

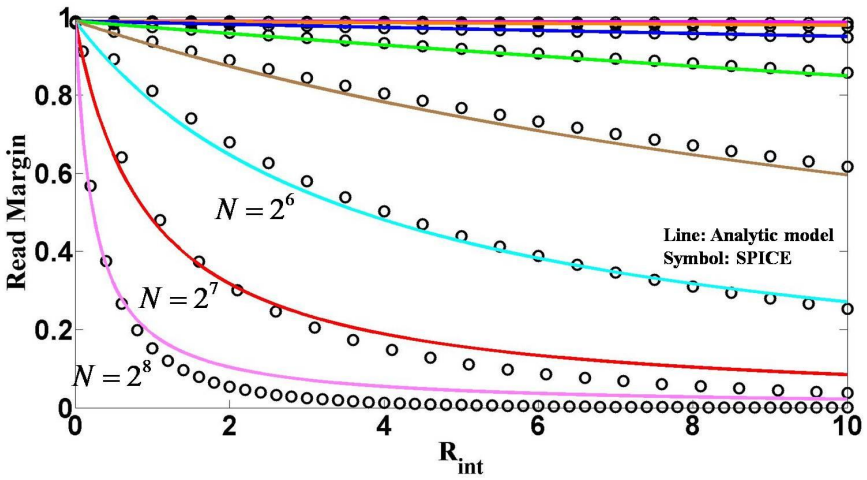

Figure 5.7: Comparison of the read margin between the analytic model and simulation.

The results shown in Figs. 5.4, 5.6, and 5.7 illustrate the sensitivity of the write and read operations to increasing interconnect resistance and array size. This sensitivity is more acute for the read case since cell nonlinearity at the selected row is not exploited due to selecting an entire row. For small array sizes, the interconnect resistance reduces the read margin due to IR losses across the interconnects, and the ineffectiveness of the cell selectors in this particular biasing scheme. To reduce the effect of the interconnect resistance to mitigate both the read margin and voltage degradation, higher nonlinearity factors are required. A different biasing scheme for read and write therefore needs to be considered. In the following section, the biasing scheme proposed in [112], based on floating unselected rows and columns of an array, is applied to the read operation while the biasing scheme proposed in [118], based

on applying one third of a write voltage across the unselected cells to enhance the nonlinearity factor, is applied to the write operation.

5.2 Enhancement of Nonlinearity Factor

Models for the driver resistance, worst case voltage drop, and read margin are provided in this section for the aforementioned floating scheme during a read operation [112], and $V/3$ during a write operation [118]. The biasing schemes are illustrated in Fig. 6.1. During a write operation, one third of the write voltage is applied to

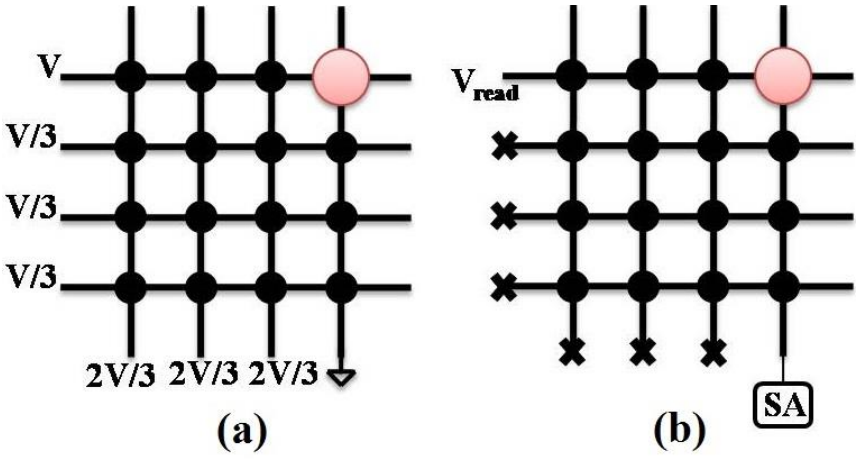

Figure 5.8: Enhancing cell nonlinearity for (a) write operation with $V/3$ biasing scheme, and (b) read operation with floating biasing scheme.

the unselected columns when grounding the selected column, and two thirds of the write voltage are applied to the unselected rows when applying a full write voltage to the selected row. The benefit of this biasing scheme is that only one third of the write voltage is across the half selected cells during a write operation rather than half

of the write voltage. The nonlinearity factor is therefore much higher and typically on the order of 10^3 to 10^4 [121–123]. Nonlinearity factor as high as 10^7 have been demonstrated [124]. Moreover, due to the decreased voltage across the half selected cells (from $\frac{V}{2}$ to $\frac{V}{3}$), the write disturbance improves [125]. One third of the write voltage is across the remaining unselected cells, as compared to the previous case (ideally, zero voltage drop), possibly increasing the leakage current. The current flowing through the unselected rows and columns is however greatly reduced due to the higher nonlinearity factor. The advantage of the V/3 biasing scheme is therefore only beneficial with high nonlinearity factors. A high nonlinearity factor needs to be sufficiently high to suppress the current flowing through the unselected rows and columns, ensuring that the effect of the interconnect resistance and therefore the IR losses is negligible.

During a read operation, a read voltage is applied to the selected row while connecting the selected column to the sense amplifier and floating the unselected rows and columns. The cell selectors effectively suppress the current flowing through the selected row, thereby reducing IR losses since half of the read voltage is dropped across the unselected cells at the selected row and column. However, as compared to the grounded biasing scheme [see Fig. 5.1(b)], only a single cell can be read at a time.

In Sections 5.2.1 to 5.2.3, models for the driver resistance, worst case voltage drop, and read margin are provided for the biasing schemes illustrated in Fig. 6.1. These models provide intuition while characterizing the limitations of the crossbar array and estimating requirements for the device parameters (K_r, R_{ON}, R_{OFF}).

5.2.1 Driver size

For the same worst case conditions assumed for $R_{driver(write)}$ and $R_{driver(read)}$, as described in Section 5.1, the driver resistance during a write operation under a $V/3$ biasing scheme and a read operation with a floating biasing scheme is, respectively,

$$R_{driver(write\_V/3)} = \frac{R_{ON}(\frac{V_{driver}}{V_{cell}} - 1)}{\frac{N-1}{K_{r(write)}} + 1}, \tag{5.11}$$

$$R_{driver(read\_float)} = \frac{R_{ON}(\frac{V_{driver}}{V_{cell}} - 1)}{\frac{N-1}{K_{r(read)}} + 1}, \tag{5.12}$$

where $K_{r(write)}$ and $K_{r(read)}$ are, respectively,

$$K_{r(write)} = \frac{I_{cell}(V_{write})}{I_{cell}(V_{write}/3)} = 3 \times \frac{R_{ON|V_{write}/3}}{R_{ON}}, \tag{5.13}$$

$$K_{r(read)} = \frac{I_{cell}(V_{read})}{I_{cell}(V_{read}/2)} = 2 \times \frac{R_{ON|V_{read}/2}}{R_{ON}}, \tag{5.14}$$

where $R_{ON|V_{write}/3}$ and $R_{ON|V_{read}/2}$ are the on-state cell resistances when the voltage across the cell equals to, respectively, one third of the write voltage and half of the read voltage. Unlike the driver resistance for reading with the grounded biasing scheme, cell selectors are used. Moreover, since the read operation uses lower voltages as compared to the write operation, the nonlinearity factor is higher than K_r, as described by (5.2). Similarly, the driver resistance during a write operation under the $V/3$ biasing scheme is also greatly enhanced due to the increased nonlinearity factor. The degradation of the driver resistance with increasing array size is therefore not as severe as the biasing schemes described in Section 5.1.

5.2.2 Voltage degradation across selected cell

To determine the worst case voltage drop across the selected cell, the cell farthest from the write (read) voltage source at the selected row, and farthest from the ground (sense amplifier) at the selected column is evaluated, as illustrated in Fig. 6.1. The voltage drop across the worst case selected cell during the write and read operation is, respectively,

$$\frac{V_{cell}}{V_{write}} = \frac{1}{\frac{NR_{int}}{R_{ON}}(\frac{N-1}{K_{r(write)}}+2)+1}, \tag{5.15}$$

$$\frac{V_{cell}}{V_{read\_float}} = \frac{1}{1+(N\frac{R_{int}}{R_{ON}}(2+\frac{N-1}{K_{r(read)}}))+(\frac{R_{sense}}{R_{ON}}(1+\frac{N-1}{K_{r(read)}}))}. \tag{5.16}$$

The models provided in (5.15) and (5.16) are in good agreement with SPICE, exhibiting a maximum error of, respectively, 10% for voltage ratios above 0.5, and 6.6% for voltage ratios above 0.35, as shown, respectively, in Figs. 5.9 and 5.10.

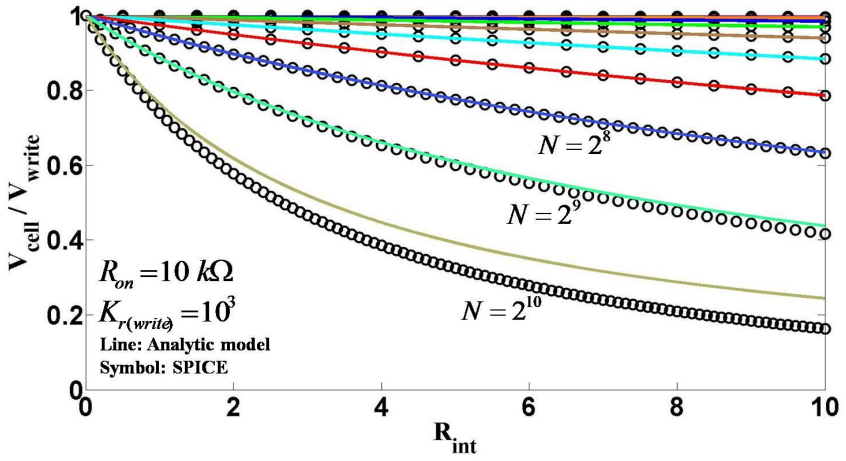

Figure 5.9: Ratio of the voltage drop across the worst case selected cell to the driver voltage during a write operation under the $V/3$ biasing scheme.

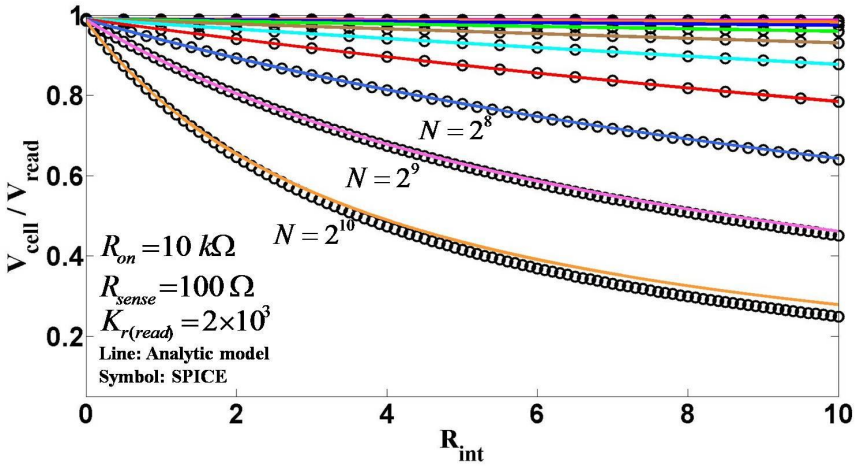

Figure 5.10: Ratio of the voltage drop across the worst case selected cell to the driver voltage during a read operation under the floating biasing scheme.

As the nonlinearity factor decreases, the accuracy of these models also decreases. Hence, (5.4) is relatively less accurate as compared to (5.15) and (5.16). This inaccuracy is due to ignoring the parasitic interconnect resistance along the unselected rows and columns. As the nonlinearity factor increases, the current flow through those lines decreases, making the parasitic interconnect resistance and hence the IR losses negligible.

5.2.3 Read margin

Expression (5.7) is used to evaluate the read margin, where $I_{sense(L)}$ and $I_{sense(H)}$ are, respectively,

$$I_{sense(L)\_float} = \frac{V_{read}}{NR_{int} + R_{sense} + \frac{1}{\frac{1}{R_{ON}+NR_{int}} + \frac{N-1}{K_{r(read)}R_{ON}}}}, \tag{5.17}$$

$$I_{sense(H)\_float} = \frac{V_{read}}{NR_{int} + R_{sense} + \frac{1}{\frac{1}{R_{OFF}+NR_{int}} + \frac{N-1}{K_{r(read)}R_{ON}}}}. \tag{5.18}$$

Assuming all of the cell selectors are off, the resistance is dominated by the selector resistance. The worst case condition becomes data pattern independent since a single cell is selected while the other cells are at a high resistance. Based on this condition, (5.17) and (5.18) exhibit good agreement with SPICE, exhibiting a maximum error of 0.12% (based on the parameters listed in Table 5.1), as illustrated in Fig. 5.11.

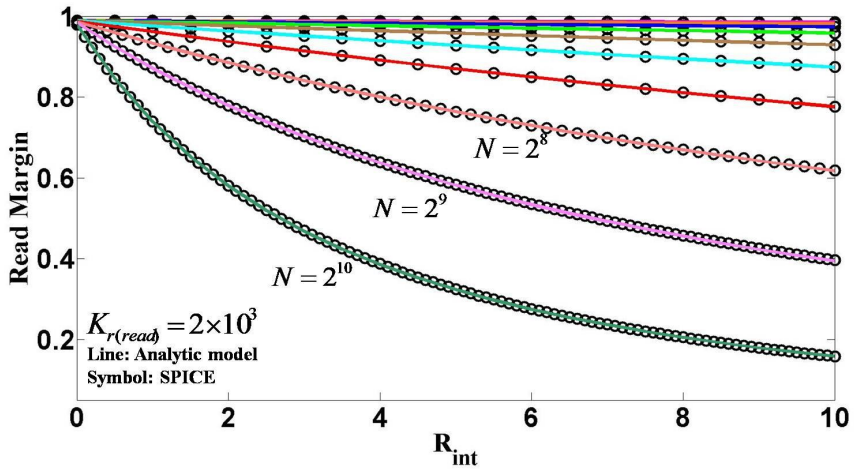

Figure 5.11: Comparison of the read margin between the model and simulation for the floating biasing scheme.

5.3 Design Requirements for Varying Array Size

An important aspect of these models is computational efficiency while providing physical intuition into crucial parameters such as K_r, R_{ON}, R_{driver}, R_{int}, R_{sense}, and N during the design process of a crossbar array. The area of the drivers (R_{driver} dependent), process technology (R_{int} dependent), and device requirements (K_r and R_{ON} dependent) can be extracted for a target crossbar array size N. Moreover, these models describe the device and circuit requirements for scaled array sizes and interconnect resistance. In this section, design requirements for large arrays are projected.

5.3.1 Driver Resistance

The driver resistance during both read and write operations based on the biasing schemes described in Sections 5.1 and 5.2 for different array sizes is shown in Fig. 5.12.

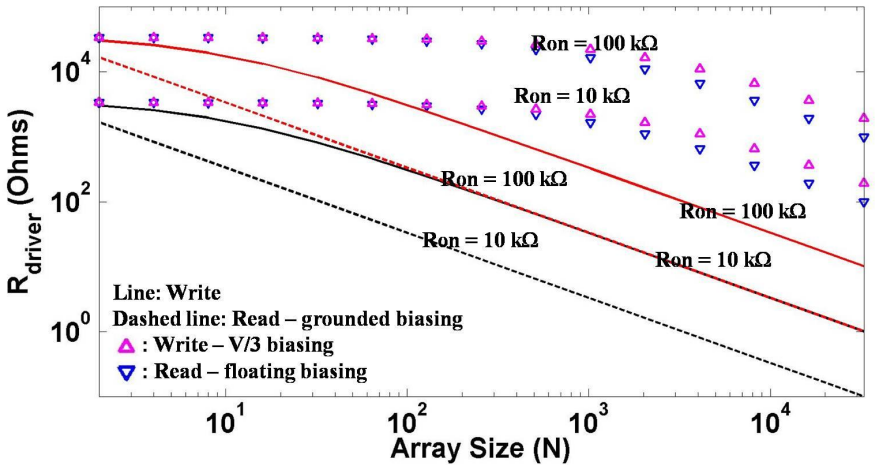

Figure 5.12: Analytic model of driver resistance with respect to varying array sizes for $K_r = 10$, $K_{r(write)} = 2 \times 10^3$, and $K_{r(read)} = 10^3$ that satisfies $\frac{V_{driver}}{V_{read}} = \frac{4}{3}$.

From Fig. 5.12, the driver resistance during a read operation based on the grounded biasing scheme should be below 10 Ω for a large scale crossbar array (> 1 Mbits) with an R_{ON} of 10,000 Ω for a V_{driver} to V_{read} ratio of 4/3. This severe degradation in driver resistance is due to the connection of N resistive devices in parallel with a full read voltage across them. This stringent constraint requires a large area dedicated to the peripheral circuitry, degrading the $4F^2$ density advantage of RRAM crossbar arrays. Due to the grounded biasing scheme during a read operation, the read voltage across a single cell selects all of the other cells on the same row, causing

the input resistance of the selected row to be inversely proportional to the array size. By choosing the floating biasing scheme, illustrated in Fig. 6.1(b), the required driver resistance is greatly increased. Reading a single cell in a specific row does not require the other cells on that row to be read since the untargeted cells are half selected and undisturbed due to the cell selectors.

During a write operation under the $V/2$ biasing scheme, due to the nonlinearity of the selector devices, the half selected cell remains at a higher resistance. The input resistance is therefore much higher as compared to reading with the grounded biasing scheme. The input resistance however is much lower as compared to reading with the floating biasing scheme. This behavior occurs since the nonlinearity factor decreases when the operating voltage increases when switching from the read voltage to the write voltage. For the same reason, the driver resistance during a write operation under the $V/3$ biasing scheme becomes higher since the nonlinearity factor increases due to the greater voltage difference between the unselected cells and the selected cells ($2V/3$) despite the higher write voltages.

5.3.2 Voltage Degradation and Device Nonlinearity

An implication of (5.15) and (5.16) is that a high nonlinearity factor is insufficient in large crossbar arrays. Nonlinearity factors are typically on the order of 10^3 to 10^4. Hence, a significantly high R_{ON} is essential for large crossbar arrays to maintain a

reasonable ratio between the cell voltage and the read/write voltage. These qualities are noted in Fig. 5.13. A nonlinearity factor greater than 10^4 only slightly improves

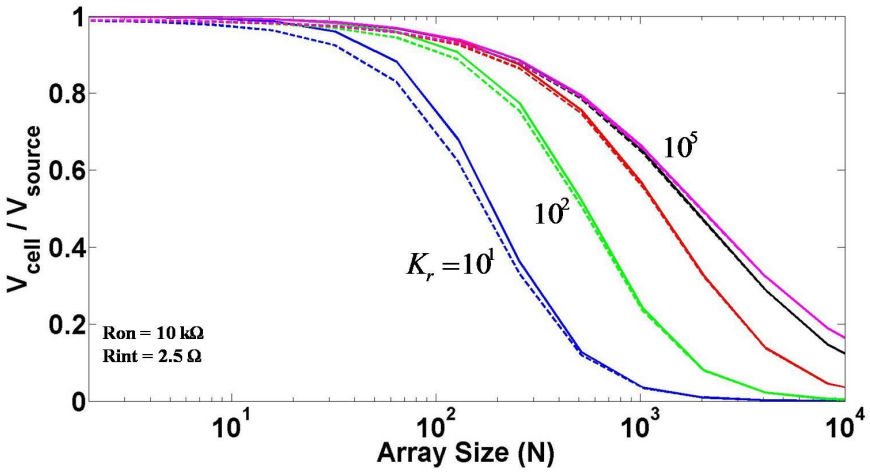

Figure 5.13: Voltage degradation vs. array size where $V_{source} = V_{write}$ (solid lines) and $V_{source} = V_{read}$ (dashed lines). $R_{sense} = 100\ \Omega$.

the voltage across the worst case selected cell. Beyond 10^4, a higher R_{ON} is required to produce a larger voltage across the selected cell.

5.3.3 Read operation

Considering the read margin when using the grounded biasing scheme, the denominator of (5), (8), and (9) consists of two different parts. One part considers the loss due to the interconnect resistance while the other part considers the loss due to sneak path currents. The resistance between the selected column and unselected rows R_{sneak} creates a sneak path. Since a voltage exists at the node that connects the column to the sense amplifier, the current flowing through the selected cell is

partially lost due to the current flow through R_{sneak}. This loss caused by the sneak path however has a negligible effect on the read margin since R_{sneak} remains at a high resistance due to the small voltage across the sense amplifier. Degradation in the read margin is therefore primarily due to IR losses across the interconnect rather than sneak current paths. When using the open circuit biasing scheme, however, any degradation in the read margin is primarily due to sneak path leakage current rather than due to IR losses, as illustrated in Fig. 5.14. Since the sneak current path is the dominant factor for read margin degradation under the open circuit biasing scheme, for negligible interconnect resistances, the grounded biasing scheme performs better [see Fig. 5.14(a)]. For significant interconnect resistance, however, since IR losses is the dominant factor for read margin degradation under the grounded biasing scheme, the open circuit biasing scheme performs better. With the open circuit biasing scheme, the read margin increases by 3.4 times for small array size ($N = 128$), and as much as 85 times for larger array sizes ($N = 1024$) for $R_{ON} = 10,000$, as compared to the grounded biasing scheme [see Fig. 5.14(b)].

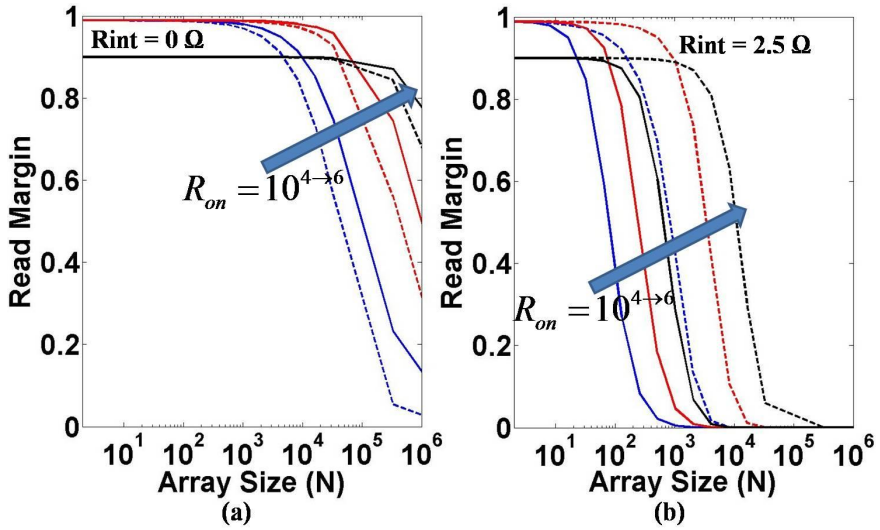

Figure 5.14: Read margin with respect to array size based on the parameters listed in Table 5.1 for (a) $R_{int} = 0\ \Omega$, and (b) $R_{int} = 2.5\ \Omega$. The solid lines describes the grounded biasing scheme whereas the dashed lines describes the floating biasing scheme.

5.4 Design Of A Crossbar Array Based On These Models

The design of an example crossbar array using these models is demonstrated in this section. A resistive cell based on the RRAM described in [121] with a 14 nm metal1 half pitch is considered. Moreover, the V/3 biasing scheme and floating biasing scheme are used, respectively, for the write and read operations. Based on these assumptions and decisions, the extracted device and interconnect parameters together with the assumed circuit parameters are listed in Table 5.2. This analysis focuses on Mbit capacity array sizes.

Table 5.2: Design parameters

Values from [121]		**Circuit level choices**	
Parameters	**Value**	**Parameters**	**Value**
R_{ON}	24 kΩ	R_{sense}	100 Ω
R_{OFF}	1.5 MΩ	R_{tran}	R_{ON}
$K_{r(read)}$	1000	Tolerable Read Margin	0.5
$K_{r(write)}$	1100	Tolerable $\frac{V_{cell}}{V_{write}}$	0.75
R_{int}[82]	8 Ω	Tolerable $\frac{V_{driver}}{V_{cell}}$	$\frac{4}{3}$

For the parameters listed in Table 5.2, the maximum array size (N) is 420 (176.4 kbit) and is limited due to the voltage degradation across the worst case selected cell during a write operation. Increasing the nonlinearity factor has a negligible effect. Two options therefore remain to mitigate this voltage degradation; enhance the device to provide a higher R_{ON} or place the crossbar array within the higher metal levels to decrease R_{int}. In Table 5.3, the effect of different R_{int} and R_{ON} on the array size N is listed. From a driver area perspective, it is beneficial to increase R_{ON} rather

Table 5.3: Varying array sizes to satisfy $V_{cell}/V_{write} = 0.75$

	R_{ON}		
Array Size (N)	24 **kΩ**	36 **kΩ**	72 **kΩ**
$R_{int} = 2.5$ Ω (22 nm)	1075	1448	2331
$R_{int} = 4$ Ω	746	1024	1695
$R_{int} = 8$ Ω (14 nm)	420	591	1024

than decrease R_{int}. While the output resistance of the driver should be 4 kΩ for R_{ON} = 24 kΩ, this resistance increases to 12 kΩ for R_{ON} = 72 kΩ. If R_{ON} and

R_{int} cannot be changed, increasing the write voltage is preferable. This method can however consume significant power and limits the usage of more advanced technology nodes due to low breakdown voltages of sub 45 nm MOS transistors (below 1.1 volts) [120]. To overcome low breakdown limitations of thin oxide MOS transistors, cascoded topologies as well as breakdown voltage multiplying circuits have been demonstrated [126]. These circuits however require increased driver area, exacerbating the area efficiency of a crossbar array.

5.5 Summary

Design models for three important metrics in crossbar arrays are provided, the driver resistance, voltage across the worst case cell (during both writes and reads), and read margin. These metrics provide intuition into the design of resistive crossbar arrays with unipolar or bipolar memory elements. The models exhibit good accuracy as compared to simulations and can be used to project the performance characteristics of large crossbar arrays. For nonlinearity factors greater than 10^4, the voltage degradation during a write and read operation can no longer be mitigated for, respectively, the V/3 biasing and floating biasing schemes. Thus, R_{ON} needs to be increased to prevent voltage degradation due to IR losses. For the read margin, under the grounded biasing scheme, sneak path leakage current is not the primary

source of signal degradation but rather the interconnect resistance. For a read operation under the floating biasing scheme, the primary source of signal degradation is sneak path leakage current. Moreover, a write operation under the V/3 biasing scheme can be advantageous as compared to the V/2 biasing scheme if the cell selectors provide a significantly higher nonlinearity factor for a smaller voltage drop across the unselected cells. These models demonstrate that a higher R_{ON} can greatly benefit all three critical metrics that limit the size of crossbar arrays.

Chapter 6

Energy Efficient Write Scheme for Non-Volatile Resistive Crossbar Arrays with Selectors

Resistive memories are expected to replace charge based conventional memories due to scalability limitations and energy benefits due to non-volatility characteristics. Resistive memory devices such as resistive RAM (RRAM), phase change memory (PCM), and magnetoresistive RAM (MRAM) have been explored for non-volatile memories [68, 69, 72, 111]. To achieve high density, these resistive devices are placed within a crossbar array structure. The area of a memory cell in a RRAM based crossbar array utilizing a two terminal one-selector-one-resistor (1S1R) configuration can be as low as $4F^2$, where F is the minimum feature size of a technology node [68]. These arrays can be placed within the metal layers, supporting cell placement above the CMOS logic, further reducing area. Moreover, a crossbar array can be configured as a logic gate, providing a path to non-von Neumann in-memory computing [127].

To enable this capability, however, the energy consumption of a crossbar array should be within practical limits due to thermal design power (TDP) envelope constraints in high performance integrated circuits (ICs) and the limited battery size of mobile devices. The energy consumption of 1S1R memories increases significantly as the size of the array grows. In particular, the write energy is a large portion of the total energy and is significantly greater than the read energy [79]. This difference is due to the long switching times of the selected devices whereas the read latency primarily depends upon the sense amplifier which improves with technology scaling. The write latency is typically on the order of a few hundred nanoseconds whereas the read latency can be as low as 5 ns [128].

The energy consumption of a crossbar array during a write operation depends upon the bias scheme, typically a $V/2$ or $V/3$ bias scheme [118, 129] (see Section 6.1). While most of the work described in the literature considers the $V/2$ bias scheme [79, 81], the advantages of one bias scheme over the other bias scheme are unclear in terms of energy efficiency. Furthermore, the $V/2$ bias scheme is often claimed to be more energy efficient than the $V/3$ bias scheme [72, 130]. The most energy efficient bias scheme can however vary depending upon the circuit and device characteristics. In particular, the selector device has a profound effect on the energy consumption since the leakage currents due to partially biased cells increases with array size. The selector device within a crossbar array suppresses the currents under a

low voltage bias while supporting higher currents under a high voltage bias. Different three terminal devices such as an MOS transistor and bipolar junction transistor as well as two terminal devices such as a silicon-based diode and metal-insulator-metal tunneling barrier have been considered [71, 83]. The three terminal transistor based selector devices provide greater isolation between the selected cells and unselected cells within a crossbar array. This solution however significantly increases cell area and inhibits scalability. Two terminal selectors can however be vertically integrated within a non-volatile resistive cell, preserving the area. A wide range of two terminal selectors exist which can be classified into two categories, unipolar and bipolar. In addition, depending upon the material and the non-volatile resistive cell, the selector can be a silicon based diode, self-rectifying device, or metal-insulator-metal (MIM) with different kinds of tunneling mechanisms depending upon the thickness of the insulator material [84]. In this chapter, a 1S1R element is used to refer to a non-volatile resistive cell integrated with a two terminal selector device. To incorporate the effects of the selector within an array, the nonlinearity factor is used as the primary metric to quantify the isolation capability (see Section 6.1).

In this chapter, the write bias schemes are compared from an energy efficiency point of view for 1S1R crossbar arrays with two terminal selectors. It is shown here that the bias scheme that provides the highest energy efficiency depends upon several parameters such as the nonlinearity factor of the selectors, size of the array, and

number of selected cells during a write operation. Simple closed-form expressions that model the write energy of a crossbar array in terms of these parameters, excluding the peripheral circuitry, are provided for the case where the interconnect resistance is negligible. The models are applicable to both unipolar and bipolar devices. Most of the existing work described in the literature does not consider the effects of writing multiple bits (i.e., multiple selected cells) on the energy consumption of an array. In [79] and [81], the power consumption when selecting multiple bits is considered; however, only for the $V/2$ bias scheme. In this work, the effects of writing multiple bits on the energy efficiency of different bias schemes are explored for the first time. Moreover, the effects of leakage current on energy consumption are discussed. In addition, an energy efficient write scheme is proposed that adaptively utilizes both the $V/2$ and the $V/3$ bias schemes to lower the write energy, extending the preliminary work described in [131]. Based on the proposed write operation, the bias scheme is altered for maximum energy efficiency depending upon the number of selected bits, which can vary for different write operations. In Section 6.1, the bias schemes during a write operation are reviewed. In Section 6.2, models of the energy consumption are described. In Section 6.3, the proposed energy efficient write scheme is described. The potential challenges and overhead are also discussed. In Section 6.4, some conclusions are offered.

6.1 Write Operations

The two types of write bias schemes, $V/2$ and $V/3$, are illustrated in Fig. 6.1. For

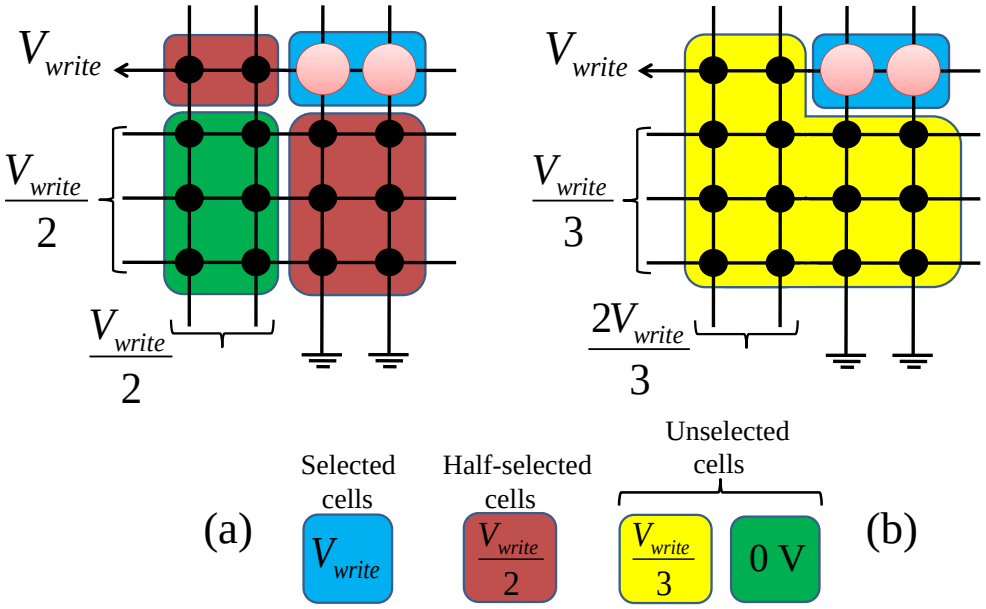

Figure 6.1: Bias schemes for a two bit write operation, (a) $V/2$ bias scheme, and (b) $V/3$ bias scheme.

the $V/2$ bias scheme, the selected wordline is connected to the write voltage while the selected bitlines are grounded. The unselected wordlines as well as bitlines are biased to half of the write voltage. Similarly, for the $V/3$ bias scheme, the selected wordline is connected to the write voltage while the selected bitlines are grounded. The unselected wordlines are biased at one third of the write voltage whereas the unselected bitlines are biased at two thirds of the write voltage. The voltage drop across the unselected cells along the selected wordline and selected bitlines, also called the half-selected cells, are therefore biased at one half of the write voltage for the $V/2$

bias scheme. For the $V/3$ bias scheme, this voltage decreases to one third of the write voltage. More importantly, the cells on the unselected wordlines and bitlines are at zero voltage for the $V/2$ bias scheme and at one third of the write voltage for the $V/3$ bias scheme, resulting in a large number of cells leaking current when the $V/3$ bias scheme is applied.

The leakage current of the unselected cells depends upon the nonlinearity factor of the selector. The two terminal selector is placed above a resistive cell to form a nonlinear I–V characteristic. A selector with a higher nonlinearity factor further decreases the current of the cell when biased below the threshold voltage of the selector [123]. The leakage current due to the partially biased unselected cells is therefore suppressed, decreasing IR voltage drops and supporting larger array sizes [132]. The nonlinearity factor of a selector is the ratio of the current passing through a selected cell to the current passing through a half-selected cell. The nonlinearity factor of the $V/2$ and $V/3$ bias schemes are, respectively,

$$K_{V/2} = \frac{I_{cell}(V_{write})}{I_{cell}(V_{write}/2)} = 2 \times \frac{R_{on@V_{write}/2}}{R_{on}}, \tag{6.1}$$

$$K_{V/3} = \frac{I_{cell}(V_{write})}{I_{cell}(V_{write}/3)} = 3 \times \frac{R_{on@V_{write}/3}}{R_{on}}, \tag{6.2}$$

where $I_{cell}(V_{write})$, $I_{cell}(V_{write}/2)$, and $I_{cell}(V_{write}/3)$ are, respectively, the current passing through the cell when the cell voltage is equal to the write voltage, one half of the write voltage, and one third of the write voltage. R_{on}, $R_{on@V_{write}/2}$, and $R_{on@V_{write}/3}$ are, respectively, the cell resistance during an on-state when the cell voltage is equal to the write voltage, one half of the write voltage, and one third of the write voltage. The leakage current therefore depends upon the bias scheme which is related to the nonlinearity factor.

The nonlinearity factor $K_{V/2}$ of a one-selector-one-resistor (1S1R) device is typically on the order of 10^1 to 10^2, whereas $K_{V/3}$ is on the order of 10^3 to 10^4 [121–123, 133–137]. A selector device with an on/off ratio as high as 10^8 has recently been demonstrated [138]. The choice of bias scheme can therefore greatly affect the energy consumption.

6.2 Energy Models

In this section, a model of the energy consumption of the $V/2$ and $V/3$ bias schemes is provided. A design guideline for choosing the proper bias scheme is explained in Section 6.2.1. Moreover, the effect of the nonlinearity factor on the choice of bias scheme is explored in Section 6.2.2. The effect of leakage current on the total energy consumption is discussed in Section 6.2.3.

To provide an intuitive closed-form expression that models the energy consumption of a crossbar array, the interconnect resistance is assumed to be negligibly small. Although this assumption is not always practical in large arrays, it permits the effects of the critical parameters on the energy consumption, such as the nonlinearity factor, size of the array, number of selected cells, and bias scheme, to be captured while retaining simplicity and providing intuitive expressions. An array with an equal number of rows and columns biased according to the $V/2$ and $V/3$ bias schemes, as illustrated in Fig. 6.1, is considered. The selected devices are modeled based on the VTEAM model [139] considering linear switching, and the remaining devices are modeled as resistors. The switching devices are considered to be symmetric with equal on/off threshold voltages and equal set/reset times. The switching energy during set and reset operations is therefore the same (see the Appendix). Based on these considerations and assumptions, the energy consumption of a crossbar array for the $V/2$ and $V/3$ bias schemes are, respectively,

$$E_{V/2} = V_{write}\frac{I_{on}}{K_{V/2}}\frac{(Nn+N-2n)}{2}t_{sw} + nE_{sw}, \tag{6.3}$$

$$E_{V/3} = V_{write}\frac{I_{on}}{K_{V/3}}\frac{(N^2-n)}{3}t_{sw} + nE_{sw}, \tag{6.4}$$

where V_{write} is the write voltage, I_{on} is the cell current when biased at the write voltage during the on state, N is the number of rows and columns, n is the number of selected cells, t_{sw} is the switching time, and E_{sw} is the switching energy consumption of the selected device,

$$E_{sw} = \frac{V_{write}^2}{R_{off} - R_{on}} ln(\frac{R_{off}}{R_{on}}) t_{sw}. \tag{6.5}$$

R_{on} and R_{off} are, respectively, the 1S1R cell resistance during the on and off states (for more details see the Appendix). The resistance of the selector device due to the limited current density is assumed to be considered in R_{on} and R_{off}. Note that the second term in (6.3) and (6.4) is the dynamic portion of the total energy due to switching the selected cells while the first term is due to the leakage current of the half-selected and unselected cells.

The closed-form expressions are in good agreement with SPICE, exhibiting an average error of 0.04% and a maximum error of 0.74%, as shown in Fig. 6.2. The energy consumption scales differently with respect to array size for different bias schemes. The $V/2$ bias scheme follows a linear trend whereas the $V/3$ bias scheme scales superlinearly with array size ($\sim N^2$). Moreover, while the energy consumption for the $V/2$ bias scheme is strongly dependent on the number of selected cells, the $V/3$ bias scheme is constant for large arrays ($N >> n$). Note that $E_{V/3}$ quadratically scales with N, exhibiting a near constant profile with respect to n if $N >> n$. Under

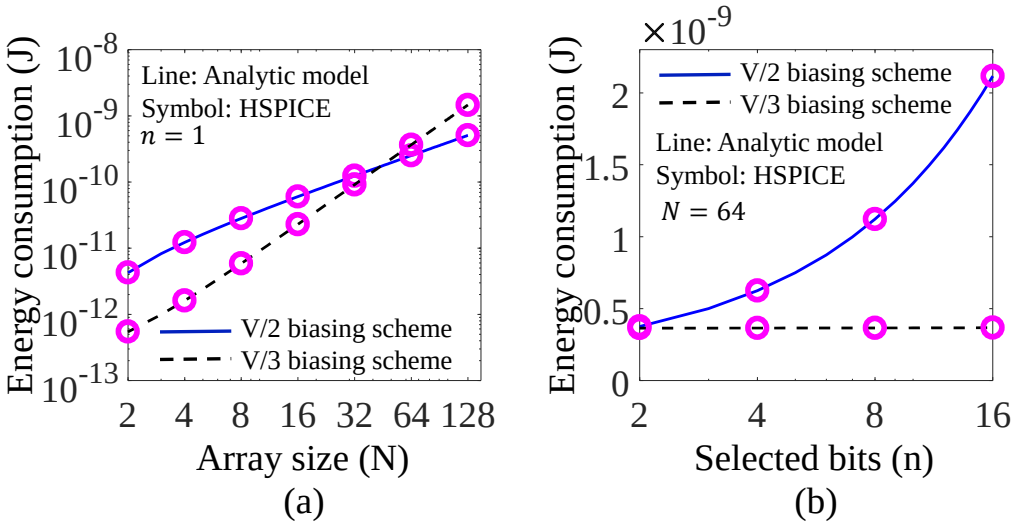

Figure 6.2: Energy consumption of a crossbar array with respect to (a) array size, and (b) number of selected cells, assuming $R_{on} = 10^4\ \Omega$, $R_{off} = 10^7\ \Omega$, $K_{V/2} = 20$, $K_{V/3} = 1,000$, and $V_{write} = 2\ V$.

this condition, while the switching energy continues to grow with large n since the leakage energy dominates for large array sizes, E_{sw} remains insignificant. The effect of n on the energy consumption for different array sizes is illustrated in Fig. 6.3. A summary of the parameters used in the following simulations are provided in Table 6.1 (unless otherwise noted).

Table 6.1: Summary of parameters for write operation

Parameters	Values
R_{on}	$10^4\ \Omega$
R_{off}	$10^7\ \Omega$
t_{sw}	$100\ ns$
V_{write}	$4\ V$

The increasing number of selected bits per write operation significantly adds to the energy consumption of the $V/2$ bias scheme. The $V/3$ bias scheme remains relatively

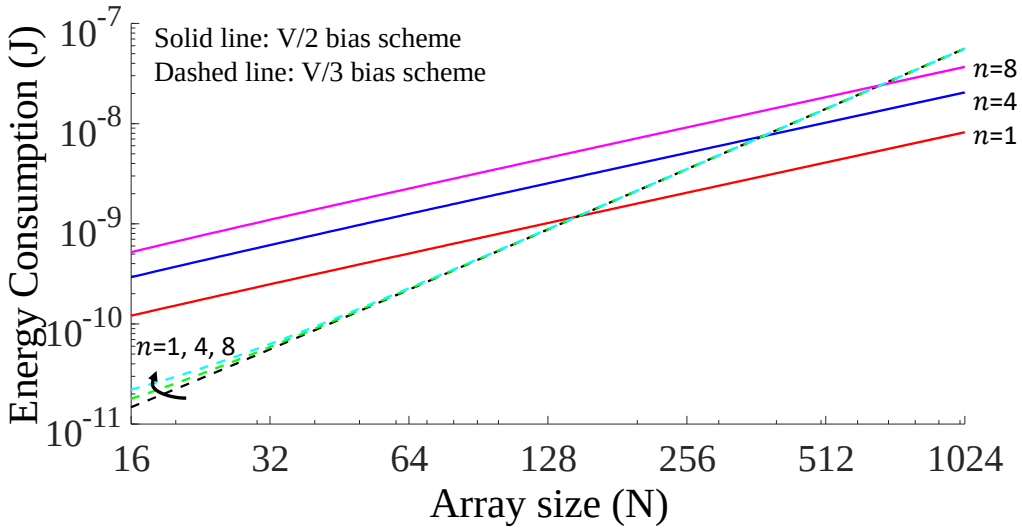

Figure 6.3: Effect of the number of selected cells on the energy consumption of a crossbar array for the $V/2$ and $V/3$ bias schemes, assuming $K_{V/2} = 20$ and $K_{V/3} = 1,000$.

constant for large array sizes. This behavior is due to the increasing number of half-selected cells for the $V/2$ bias scheme with increasing n. In contrast, for the $V/3$ bias scheme, the variation in the number of unselected cells become negligible as n increases if the size of the array N is much larger than n.

One method to decrease the energy consumption is by using selectors with a higher nonlinearity factor. A higher nonlinearity factor decreases the leakage current of the unselected cells, improving the ability of the selector to isolate the switching cell from the rest of the unselected array. The effect of the nonlinearity factor on the energy consumption is shown in Fig. 6.4. Note that with increasing nonlinearity factor, the energy consumed during both bias schemes decreases since (6.3) and (6.4) are, respectively, inversely proportional to $K_{V/2}$ and $K_{V/3}$.

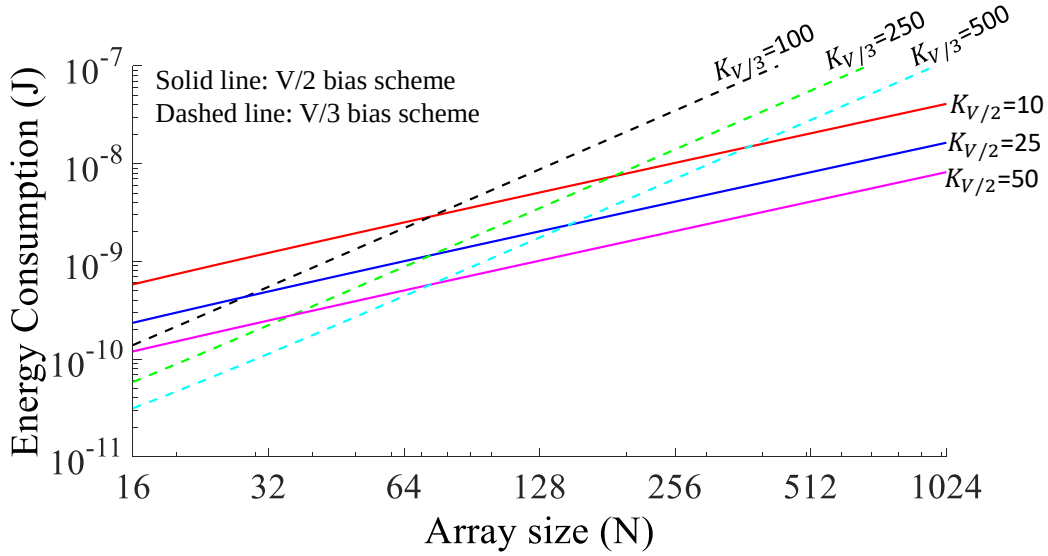

Figure 6.4: Effect of the nonlinearity factor on the energy consumption of a crossbar array for the $V/2$ and $V/3$ bias schemes, assuming $n = 4$.

6.2.1 Energy Efficient Bias Scheme

Depending upon the array size, one bias scheme is more efficient than the other bias scheme. The number of selected cells n during a write operation may however alter the most energy efficient bias scheme, as shown in Fig. 6.5. Note that the line of intersection between the two bias schemes (where $E_{V/2} = E_{V/3}$) spans a range of array sizes ($N = 128, 256$, and 512) depending upon the number of selected bits. Since the $V/2$ bias scheme scales with the number of selected cells as opposed to the $V/3$ bias scheme which remains relatively constant, the line of intersection bends for different values of n.

Extra energy is expended due to an incorrect choice of bias scheme, wasting significant power during a write operation. The ratio of the energy consumption between

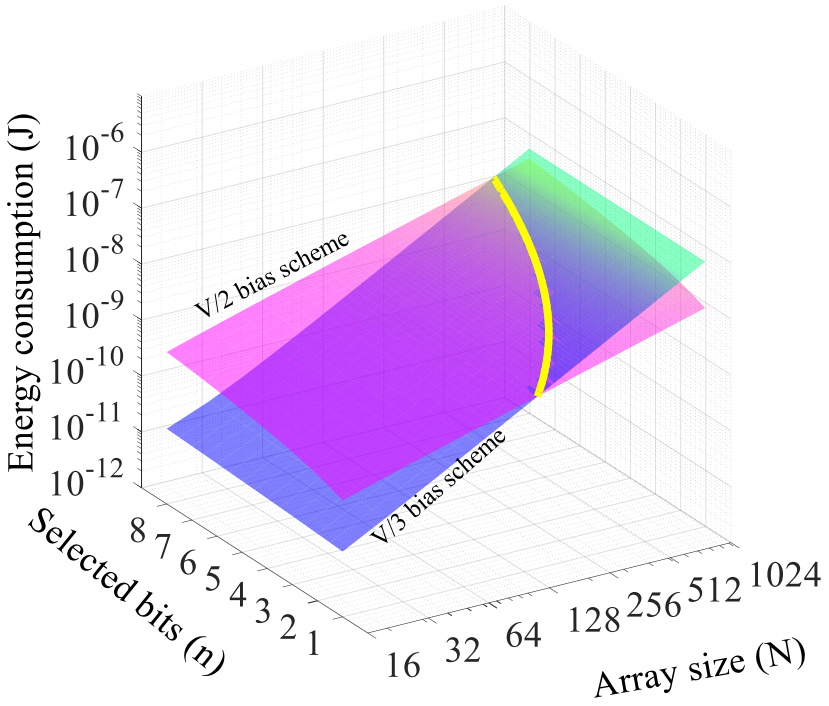

Figure 6.5: Comparison of the energy consumption in terms of the array size and number of selected cells for the $V/2$ and $V/3$ bias schemes, assuming $K_{V/2} = 20$ and $K_{V/3} = 1,000$.

the two bias schemes is shown in Fig. 6.6. The right side of the contour is the region where the $V/2$ bias scheme is more efficient than the $V/3$ bias scheme, and the left side is where the $V/3$ bias scheme is more efficient than the $V/2$ bias scheme. Since increasing the number of selected cells consumes more energy for the $V/2$ bias scheme for low n, the $V/2$ bias scheme remains more energy efficient over a wider range of array sizes. In contrast, for high n, the $V/3$ bias scheme is more energy efficient over a wider range of array sizes. The write energy can be as much as 5x lower for a 128 x 128 array and 10x lower for a 64 x 64 array using the $V/3$ bias scheme with eight

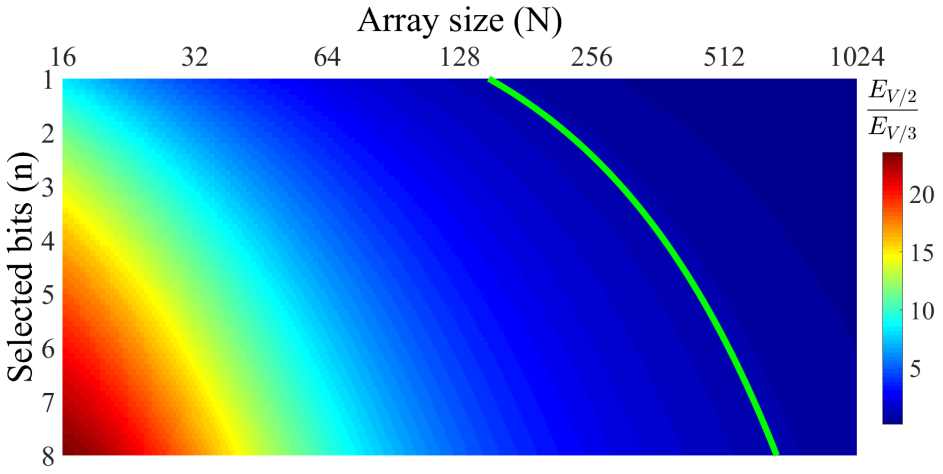

Figure 6.6: Energy savings of the $V/3$ bias scheme as compared to the $V/2$ bias scheme assuming the same parameters listed in Fig. 6.5. The solid line is the contour where the energy consumption between the two bias schemes is equal.

selected bits. For large arrays, however, since the number of cells leaking current during the $V/3$ bias scheme scales with N^2, the $V/2$ bias scheme can consume as much as 7x lower energy for an array size of 1024 x 1024 with single bit operation.

The interconnect resistance changes the location of the contour (see Fig. 6.6) where the energy for both bias schemes is equal. Since the leakage current due to the half-selected cells for the $V/2$ bias scheme is significantly greater than the leakage current of the cells biased at one third of the write voltage, the IR voltage drops are greater for the $V/2$ bias scheme [132]. Thus, the voltage drop across the selected cells for the $V/2$ bias scheme is smaller than for the $V/3$ bias scheme. The switching time of the selected cells for the $V/2$ bias scheme is therefore longer, increasing the energy consumption [139] and resulting in the $V/3$ bias scheme being more energy

efficient. This effect is more pronounced with larger IR voltage drops, resulting in slower switching times.

6.2.2 Impact of Nonlinearity Factor

The bias scheme affects the total leakage current due to the difference between the nonlinearity factors and the number of leaking cells. While the size of the array as well as the number of selected bits affect the choice of energy efficient bias scheme, the difference between the nonlinearity factors ($K_{V/2}$ and $K_{V/3}$) determines the range of N and n at which the two energy consumptions, $E_{V/2}$ and $E_{V/3}$, are equal. For

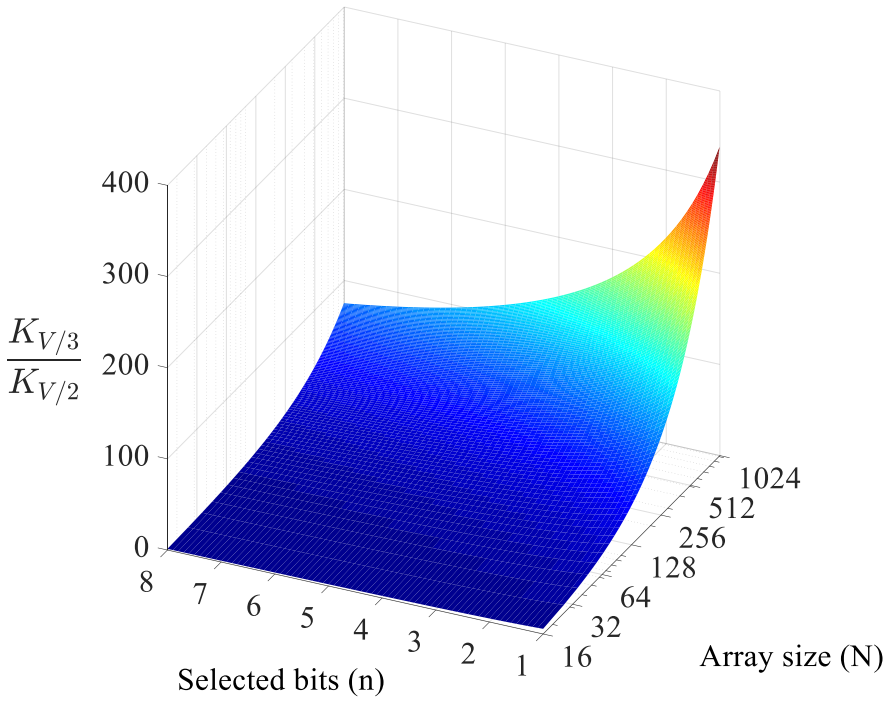

Figure 6.7: Ratio of the nonlinearity factors $K_{V/3}$ to $K_{V/2}$ to maintain equal energy consumption for the $V/2$ and $V/3$ bias schemes in terms of the array size and number of selected cells.

instance, if one nonlinearity factor is much greater than the other nonlinearity factor, the bias scheme that provides the higher nonlinearity factor will be the most energy efficient bias scheme for a wide range of N and n. The ratio of the two nonlinearity factors, $K_{V/2}$ and $K_{V/3}$, is a function of the array size and number of selected cells. Based on this ratio, for the $V/3$ bias scheme to be more energy efficient than the $V/2$ bias scheme, the following condition must be satisfied,

$$\frac{K_{V/3}}{K_{V/2}} \geqslant \frac{2}{3}\frac{N^2 - n}{Nn + N - 2n}. \tag{6.6}$$

Note that for negligible parasitic interconnect resistance, (6.6) is a function of the size of the array and number of selected cells. The variation of $K_{V/3}$ to satisfy (6.6) is shown in Fig. 6.7. The $V/3$ bias scheme is more energy efficient if $K_{V/3}$ is at least two orders of magnitude greater than $K_{V/2}$ for array sizes up to 1024 x 1024 with six selected bits or an array size up to 256 x 256 with a single selected bit.

6.2.3 Write Pulse Width

The pulse width to successfully program the selected cells depends upon the switching time of the cells. While shorter pulses may produce write failures, extended pulse widths may consume excessive power, degrading the energy efficiency. Due to the significance of the leakage current of the unselected cells, it is crucial to accurately set the pulse width with high precision. For large arrays, the leakage current

portion of the total energy dominates, making the switching energy E_{sw} negligible, as shown in Fig. 6.8.

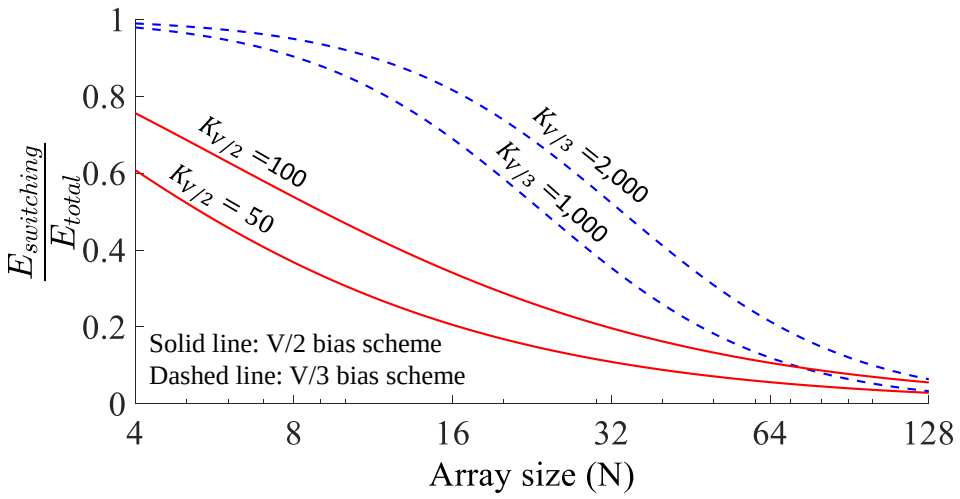

Figure 6.8: Ratio of the switching energy to the total energy in terms of the array size, $R_{on} = 10^4\ \Omega$, $R_{off} = 10^6\ \Omega$, and $n = 4$.

Note that the switching energy for the $V/3$ bias scheme is a larger portion of the total energy as compared to the $V/2$ bias scheme. This difference is due to the smaller leakage current for the $V/3$ bias scheme due to the larger nonlinearity factor, $K_{V/3}$. Similarly, a higher nonlinearity factor reduces the leakage energy, resulting in the switching energy being more pronounced and exhibiting greater energy efficiency. The switching energy is less than 10% of the total energy for array sizes exceeding $N = 128$.

To lower the energy due to leakage currents, the pulse width is set as precisely as possible, sufficient to switch the selected cells. This excess energy due to leakage

currents requires write termination circuitry to isolate the write voltage from the array once successful switching is achieved. While write termination techniques have been adopted for resistive cells based on STT-MRAM due to the stochastic nature of the switching process [140], a similar approach in RRAM based 1S1R crossbar arrays can be useful to save energy since an over extended write pulse can significantly reduce the energy efficiency due to the large leakage currents. The write termination circuitry exhibits a negligible energy overhead of, on average, less than 100 fJ [140].

6.3 Energy Efficienct Hybrid Write Scheme

In this section, a write scheme is proposed to improve the energy efficiency of a crossbar array during write operations. The optimal choice of the energy efficient bias scheme is explained in Section 6.3.1. The overhead and challenges of the proposed system are discussed in Section 6.3.2.

The number of selected cells affects the energy of an array and can be used to determine the most energy efficient bias scheme. The proposed write scheme improves the energy efficiency by adaptively switching between the $V/2$ and $V/3$ bias schemes depending upon the number of selected cells during a write operation. The number of selected bits during a write operation depends upon the difference between the patterns of the old data and the new data, as shown in Fig. 6.9. Consider a word size of eight bits. If the new data are the same as the old data, the number of selected

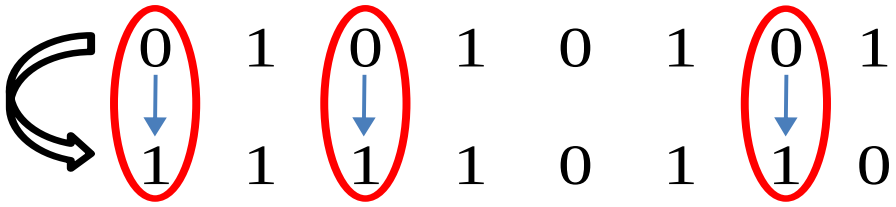

Figure 6.9: Writing an eight bit word. Four bits of the new string are the same as the old string; however, only three bits are selected since one bit requires a reset whereas the other three bits require a set operation.

cells is equal to zero. If however the new data are different than the previous data, the number of selected cells depends separately upon the number of sets and resets, since in resistive memories, writing a 1 or a 0 requires two different write operations. To determine the number of bits, a read-before-write technique is typically used [141]. This approach detects those cells that require switching, reducing excessive energy consumption during a write operation. By adopting a similar approach to monitor the number of selected cells during each write operation, the optimal energy efficient bias scheme can be determined.

The steps summarizing the write process using the energy efficient write scheme is shown in Fig. 6.10. The initial step is a read-before-write operation followed by counting the number of cells that will switch for the new string of data. Once the number of selected cells n is known, n is compared to n_{th} (see Section 6.3.1) for a specific array. Following this step, the power delivery system is configured to support either the $V/2$ or $V/3$ bias scheme to lower the energy. During this step the crossbar array remains idle, no energy is therefore consumed. Finally, once the regulator

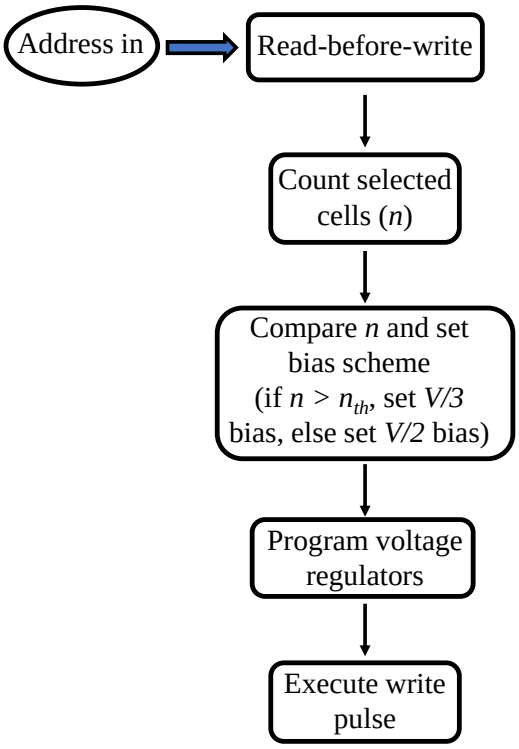

Figure 6.10: Steps during the proposed energy efficient write scheme.

voltage converges to the appropriate bias scheme, the write pulse is executed to write the new data and complete the write process.

6.3.1 Optimal Choice of Bias Scheme

The bias scheme of a crossbar array is altered when the number of selected cells n crosses a threshold, n_{th}. At this threshold, the write energy of the $V/2$ and $V/3$ bias schemes is equal. Since the energy for the $V/2$ bias scheme grows with increasing n, if $n < n_{th}$, the power delivery system switches to the $V/2$ bias scheme. If $n > n_{th}$, the power delivery system switches to the $V/3$ bias scheme. The energy savings in

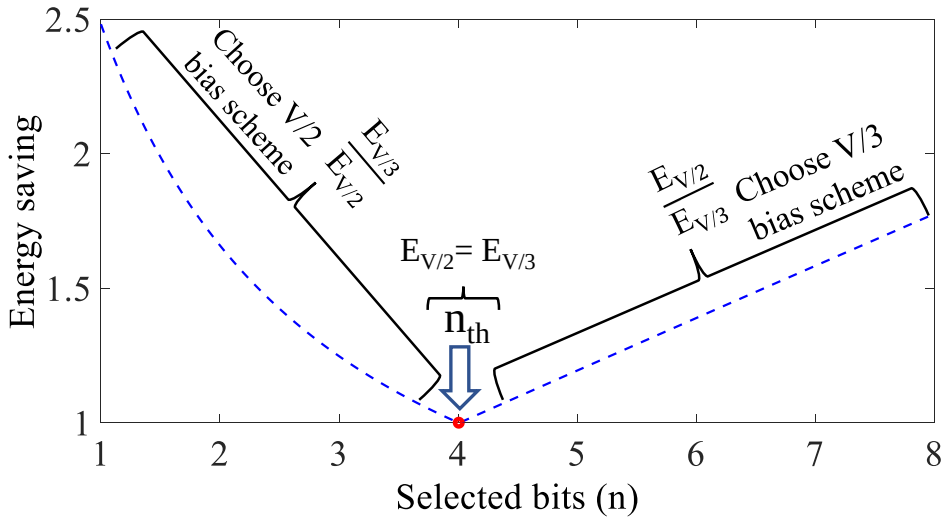

Figure 6.11: Energy improvement in terms of the number of selected cells, assuming $N = 128$, $K_{V/2} = 20$, and $K_{V/3} = 345$. The proposed write operation chooses the most energy efficient bias scheme based on the number of selected cells n with respect to n_{th}.

terms of the number of selected cells n is shown in Fig. 6.11. Note that if n_{th} is four, the $V/2$ bias scheme provides as much as a 2.5x energy improvement for a 128 x 128 array when a single bit is selected. The $V/3$ bias scheme provides up to a 1.8x savings in energy when eight bits are selected. Note, however, the size of the array N as well as the ratio of the nonlinearity factor K_r can affect the most energy efficient bias scheme. Depending upon N and K_r, n_{th} may reside outside the range of allowed values of n.

The effect of N and K_r on the energy savings is shown in Fig. 12. Note that the hybrid bias scheme only benefits specific array sizes for a fixed value of K_r. For instance, according to Figs. 12a, 12b, and 12c, the hybrid bias scheme can be used

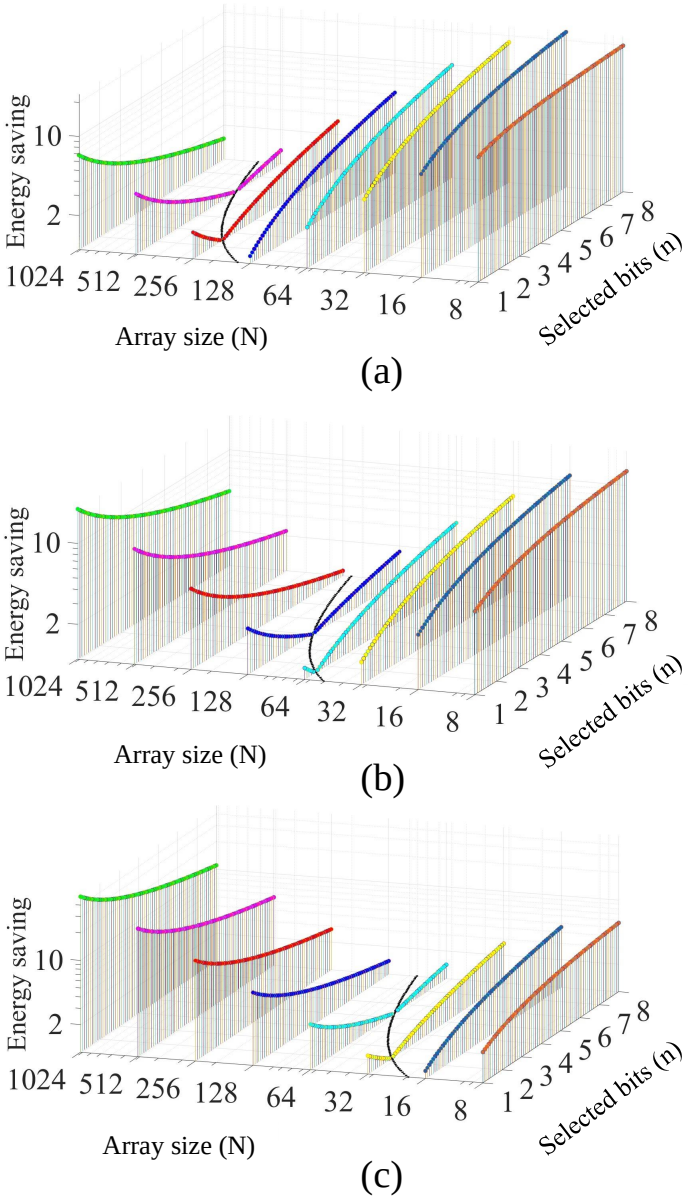

Figure 6.12: Energy savings for different array sizes and number of selected cells considering, (a) $K_r = 1000/20$, (b) $K_r = 345/20$, and (c) $K_r = 345/50$.

for an array size of, respectively, 512 x 512 or 256 x 256, 128 x 128 (same as shown in Fig. 11), and 64 x 64. The curve along the N and n axes spans the regions where no energy savings exist (i.e., unity). If the array size is above this curve, the bias scheme

is set to $V/2$. If below this curve, the bias scheme is set to $V/3$. If the array size is neither above nor below this curve, the hybrid bias scheme can be used to improve the energy efficiency.

By setting the energy for both bias schemes, (6.3) and (6.4), equal, the number of bits in which both bias schemes consume the same energy n_{th} can be determined. Based on this equality, n_{th} is

$$n_{th} = \frac{2N^2 - 3K_r N}{3K_r N - 6K_r + 2}, \tag{6.7}$$

where K_r is the ratio of the nonlinearity factors,

$$K_r = \frac{K_{V/3}}{K_{V/2}}. \tag{6.8}$$

Note that n_{th} is a function of K_r and the array size N when the interconnect resistance is negligible. The change of n_{th} as a function of K_r and the array size N are shown in Fig. 6.13. For large arrays with low K_r, n_{th} increases significantly, reaching 16. This effect is due to the diminishing savings in energy of the $V/3$ bias scheme with increasing array size, resulting in a large number of unselected cells leaking current, which scales with N^2. Furthermore, a lower K_r means the difference in leakage current between the half-selected cells for both bias scheme decreases. Thus, the $V/2$ bias scheme is more energy efficient for a wider number of selected cells. Since the leakage

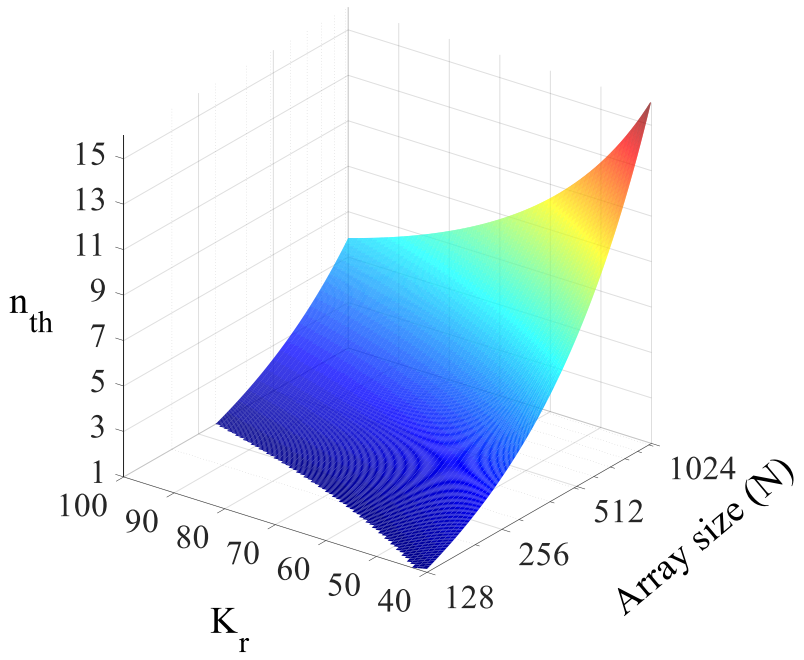

Figure 6.13: Number of selected cell in which the energy for both bias schemes are equal with respect to K_r and the array size N.

current of the unselected cells for the $V/3$ bias scheme decreases relative to the leakage current of the $V/2$ bias scheme, as K_r increases, the $V/3$ bias scheme becomes more energy efficient for a wider number of selected cells, hence decreasing n_{th}. In addition, if the interconnect resistance incurs significant IR voltage loss, n_{th} decreases since the switching time for the $V/2$ bias scheme is larger than the $V/3$ bias scheme due to the voltage degradation across the selected cells [132]. Increasing K_r from a few tens to 100 can reduce n_{th} from 16 to six. If n_{th} is larger or equal to the maximum number of selected cells (i.e., word size), the array is biased with only the $V/2$ bias scheme rather than the hybrid bias scheme (see Fig. 6.11). The nonlinearity factor of a 1S1R cell for the $V/2$ bias scheme is typically less than 100 whereas the nonlinearity factor

for the $V/3$ bias scheme reaches a few thousands. K_r is typically in the range of a few tens to several hundreds.

6.3.2 Overhead and Challenges

While the $V/2$ bias scheme requires two voltages, V_{write} and $V_{write}/2$, the $V_{write}/3$ bias scheme requires three voltages, namely, V_{write}, $V_{write}/3$, and $2V_{write}/3$. A hybrid solution using both bias schemes requires four voltage levels. Providing a large number of heterogeneous on-chip voltages is challenging due to the limited board area for the off-chip power supplies and the limited number of power I/Os. In [142], a boost converter with a charge pump is used to bias the array. This approach is however not feasible for a hybrid bias scheme with multiple voltage levels since the switching converter requires large off-chip inductors as well as large capacitors, greatly increasing the area and therefore the cost [143]. Linear regulators, alternatively, are less power efficient as opposed to switching converters; however, linear converters are much smaller since bulky capacitors or inductors are not required [98]. Heterogeneous power delivery systems with a large number of voltages using on-chip linear regulators have been proposed [95, 144]. These on-chip voltage regulators can be placed close to the load, further reducing the response time while providing fast local power management to control the bias scheme (as opposed to an off-chip power management solution which exhibits higher latency) [90]. By programming the reference

voltage of the on-chip regulators, the bias scheme can be altered between $V/2$ and $V/3$ [90, 145, 146].

The proposed energy efficient write scheme provides energy savings as high as 2.5x as compared to a conventional system with a single bias scheme. The write process however incurs additional steps as compared to a conventional write operation with a constant bias scheme (see Fig. 6.10), increasing the write latency. The write latency is typically the switching time of the 1S1R cell. In the proposed hybrid write scheme, however, the read-before-write operation necessitates a read operation for every write operation. The time required to compute and compare n with respect to n_{th} has to be considered in addition to the switching time of the 1S1R cell.

In memory systems, the read operation is typically a primary performance bottleneck. If however the write latency increases significantly, it can inhibit memory performance. Thus, a fast power delivery system is required for time constrained memory applications such as DRAM and cache memory. For slower memory systems, such as flash, the stringent timing requirements can be relaxed. While the read latency is significantly smaller than the write latency [79] and can be as low as five nanoseconds [128], the time required to program the voltage regulators has to be within a few nanoseconds to prevent write dependent performance limitations. Hence, the need for an on-chip voltage regulator (as opposed to an off-chip regulator)

becomes necessary since, unlike on-chip local regulation, off-chip power management and regulation cannot provide sub-μs bandwidth [90].

The energy overhead of the energy efficient write scheme is insignificant. The write operation for a 1S1R crossbar array is typically on the orders of hundreds of nanojoules [79]. The read operation during the read-before-write requires negligible energy, typically less than one nanojoule since the read latency is significantly less than the write latency. The programmable CMOS reference voltage consumes a few picojoules [146], assuming a switching time on the order of hundreds of nanoseconds. The primary challenge for the proposed write scheme is lowering the overhead of the write latency in time constrained memory applications.

6.4 Summary

The energy consumption of a 1S1R crossbar array for two bias schemes, $V/2$ and $V/3$, for optimal energy efficiency is discussed. Closed–form expressions that intuitively model the energy consumption in terms of the nonlinearity factor, size of the array, and number of selected cells are presented. The most energy efficient bias scheme depends upon the size of the array as well as the number of selected cells during a write operation. The energy consumed during both bias schemes scales differently. The $V/2$ bias scheme is more energy efficient for large arrays. As the number of selected cells increases, however, the $V/3$ bias scheme achieves greater

energy efficiency. The $V/3$ bias scheme provides higher efficiency, decreasing the energy consumption by an order of magnitude for a 64 x 64 array with eight selected cells. As the array size increases and the number of selected cells decreases, the energy benefits of the $V/3$ bias scheme diminish.

For the $V/3$ bias scheme to be as energy efficient as the $V/2$ bias scheme for large arrays ($N > 128$), $K_{V/3}$ should be two orders of magnitude greater than $K_{V/2}$. The appropriate choice of bias scheme can save an order of magnitude of energy. A higher nonlinearity factor significantly decreases the energy consumption by suppressing leakage currents within the half-selected and unselected cells. The switching energy is a negligible portion of the total energy for large arrays ($N > 128$). To prevent excess energy consumption due to leakage currents, write termination circuitry can be used to prevent over extended write pulses.

To improve the energy efficiency during write operations, an energy efficient write scheme is proposed. The write operation uses a hybrid bias scheme to exploit both the $V/2$ and $V/3$ bias schemes to enhance the energy efficiency based on the number of selected cells. The critical number of selected cells in which the bias scheme switches (n_{th}) is characterized. Energy improvements provided by the hybrid write scheme can be as high as 2.5x. To effectively exploit the energy efficient write scheme in time constrained memory systems, the program time of the voltage regulators and the time to compute n_{th} need to be on the order of a few nanoseconds. The proposed write

scheme incurs negligible energy overhead. Future work will focus on integrating the interconnect resistance into the energy models to capture the effects of IR voltage drops on the switching time of the selected cells and the energy consumption of the crossbar array.

Chapter 7

Stability of On-Chip Power Delivery Systems with Multiple Low Dropout Regulators

Voltage regulators are fully integrated on-chip to enable granular power management without communicating off-chip and reducing power consumption with fast dynamic voltage scaling [147]. In addition, a wide range of heterogeneous voltages can be generated on-chip without increasing the off-chip board area [148]. These benefits have led to many industrial products that incorporate different types of on-chip voltage regulators [58, 60, 87, 149], containing as many as 64 regulators in a single power domain [58].

Fully integrated linear regulators as well as switching converters deliver on-chip power [59, 60, 88, 150]. While switching and switched capacitor regulators occupy significant space due to the bulky inductors and capacitors, capacitor-less LDOs occupy small area at the expense of lower power efficiency, thereby supporting the deployment

of a larger number of on-chip regulators [44]. Due to the low integration overhead of linear regulators, capacitor-less low dropout regulators (LDOs) are widely preferred over fully integrated switching converters [59, 96, 150].

A power delivery system that contains multiple on-chip LDO regulators can exhibit instability [151–155]. This phenomenon is a result of the interaction between the resonance of the off-chip parasitic network and the actively regulated on-chip power grid. One of the primary requirements of on-chip capacitorless LDOs is ensuring stability in the absence of a large output capacitor. Conventionally, satisfying stability requirements under light load conditions (< 1 mA) is sufficient to ensure stability for a wide range of load variations [98, 108]. The general assumption is once the regulator is stable under light load conditions, the stability is guaranteed as long as the regulator operates above the minimum load condition for a range of output capacitance. Ensuring stability under light loads however does not guarantee stability when multiple LDO regulators share the same power delivery network (as discussed in Section 7.1). In this chapter, this second source of instability is demystified. The effect of resonance due to the parasitic impedances of the power delivery network is shown to depend on the number of voltage regulators, thereby affecting the stability of the overall system.

In section 7.1, instability due to the increasing number of LDOs is described. In section 7.2, a summary of the relevant works from the literature is provided. In section

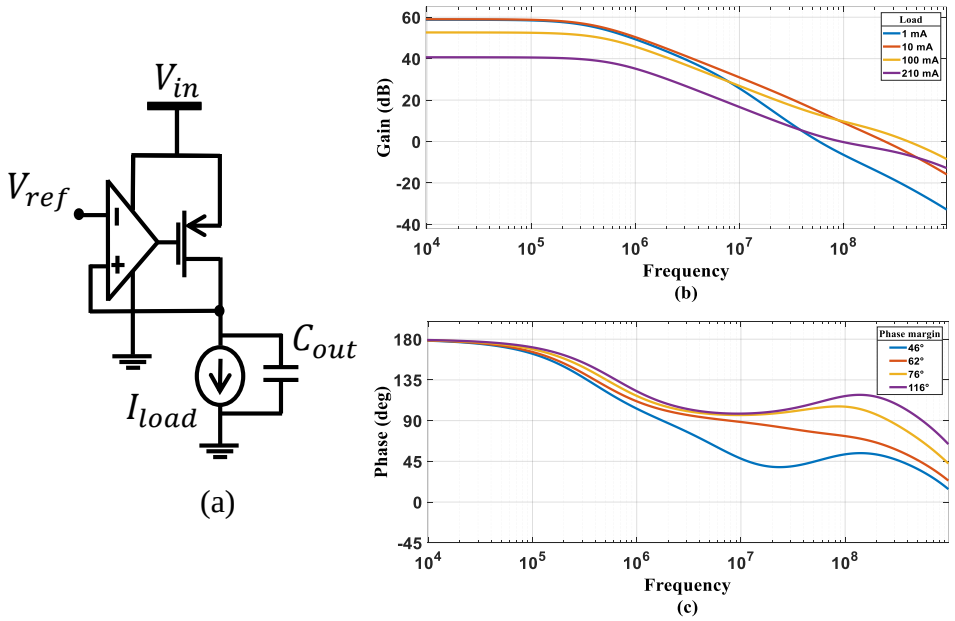

Figure 7.1: Linear regulator used to analyze the stability of multiple connected LDOs, (a) conventional low dropout regulator, and (b) Bode plot of a regulator under different load conditions.

7.3, the relationship between grid stability and the number of LDOs is explored. In section 7.4, the effect of circuit design parameters on the stability of the power grid is discussed. In section 8.3, some conclusions are offered.

7.1 Stability of Parallel Connected LDOs

A conventional LDO architecture consisting of a single closed loop with an error amplifier and a pass transistor is considered here to simplify the stability analysis and produce an analytic relationship between the system stability and number of LDOs, as shown in Fig. 7.1. The error amplifier is a two stage operational transconductance

amplifier with Miller compensation incorporating a nulling resistor [156]. These conventional LDOs are typically modified to support higher phase margin under light load conditions (1 mA) since a fully integrated regulator does not incorporate a large output capacitor, producing a non-dominant output pole [99]. To separate the stability concerns under light load conditions from the stability concerns when considering multiple LDOs, a heavy load condition is assumed (on the order of a few hundred mA). The standalone regulator exhibits a highly stable system, as shown in Fig. 7.1b. The phase margin is as low as 46^o at 1 mA and as high as 116^o at 210 mA considering a 50 pF output capacitance.

The LDO regulator produces a stable transient response to a step load variation from 175 mA to 210 mA, when considered alone. The same regulator co-existing with 14 other LDOs sharing the same input grid however yields an unstable system, as shown in Fig. 7.2. Note that a single LDO, exhibiting a stable transient response, produces an oscillatory response when operating under the same conditions with a larger number of LDOs. This phenomenon is due to the additional LDOs which exacerbate the interactions with the off-chip parasitic impedances. Specifically, the parasitic impedance at the input of each LDO changes with an increasing number of LDOs, shifting the resonant frequency, thereby increasing the phase shift of the regulator and producing an unstable power grid (see Section 7.3.1). Moreover, the power delivery system is stable under light load conditions and unstable under heavy

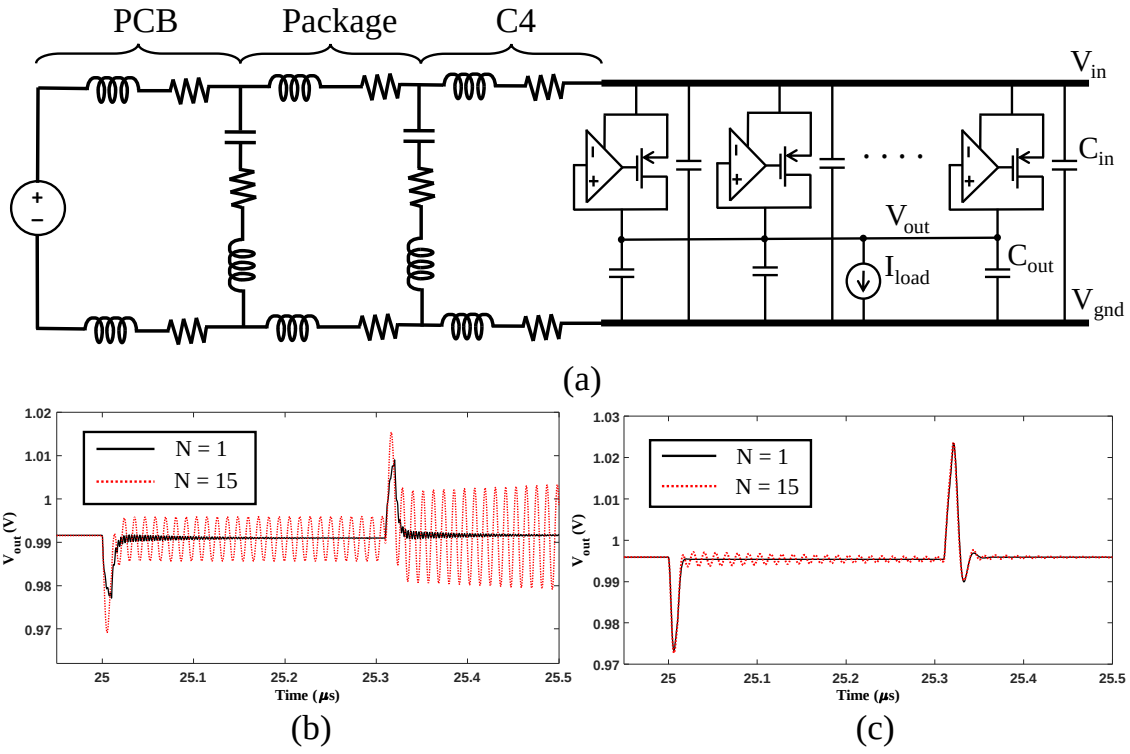

Figure 7.2: Comparison of single LDO to multiple connected LDOs sharing a common power grid, (a) power delivery network with an input voltage of 1.2 volts and an output voltage of 1 volt, (b) transient response to a load varying from 175 mA to 210 mA in 10 ns considering one and 15 LDOs, and (c) transient response to a load varying from 1 mA to 3 mA in 10 ns considering one and 15 LDOs. The parasitic impedances are listed in Table B.1 in the Appendix. The input and output capacitance are, respectively, 1 nF and 50 pF per LDO. The load current as well as the input and output capacitors proportionally increase with the number of regulators. Each regulator therefore operates under the same load conditions and AC characteristics (see Fig. 7.1b).

load conditions despite lower phase margins under light load conditions (see Fig. 7.1b). The phase margin of a standalone LDO without considering the power delivery network is therefore not a proper stability metric.

7.2 Existing Work

Most existing work on the stability of capacitor-less on-chip LDOs considers a single LDO system [98]. One exception is [154], where the effect of multiple on-chip LDOs on grid stability is reported. The inductive off-chip network used in the analog power delivery system produces an unstable grid when shared with 16 LDO regulators, each operating under a load of 20 mA. This work however does not explain the reason for the degradation in stability with an increasing number of on-chip LDO regulators.

In [151], the stability of six distributed LDOs sharing a common grid is considered. A passivity based stability criterion based upon the output impedance of the LDOs is proposed. This work however assumes an ideal voltage supply at the input of the on-chip regulators, disregarding the parasitic effects of the off-chip power delivery network. The source of instability is described as due to the particular load conditions that the regulators are exposed to under unbalanced current sharing, causing the LDOs to exhibit low or negative phase margin. The source of grid instability with multiple on-chip regulators is however not due to unbalanced current sharing if

the regulators exhibit positive phase margin across any load condition. Stable capacitorless on-chip LDOs have been reported under no load conditions (0 A) [108] where most of the capacitorless LDOs exhibit increasing phase margin with increasing load. Since maintaining sufficient phase margin under any load condition is a primary design requirement, unbalanced current sharing in the presence of on-chip regulators cannot by itself cause grid instability as long as the regulators retain a minimum phase margin under stringent load conditions.

The analyses proposed in [152] consider the parasitic network of the off-chip as well as the on-chip interconnects, and provide a comprehensive discussion on the evaluation and optimization of grid stability with multiple digital LDO regulators. A complex power grid consisting of multiple digital LDOs is evaluated based on the signal flow graph of an entire system. The stability is evaluated based on a transfer function constructed from a signal flow graph of the grid system using Mason's gain formula. In [153], a stability checking methodology based on a hybrid stability constraint is provided, which considers a comprehensive power delivery network including parasitic impedances. A hybrid stability margin based on the hybrid passivity and finite gain stability theorem is established to evaluate the stability of a power grid. This stability constraint can be separately considered for each LDO, enabling the LDOs to be individually tuned to enhance the stability of the grid. In [157], the hybrid stability constraint is considered to explore the relationship between different

circuit level parameters and power grid stability. The effect of different LDO topologies, LDO parameters such as the unity gain frequency, and decoupling capacitors on grid stability are evaluated.

In this work, the phenomenon of a decreasing resonant frequency due to an increasing number of LDOs is described using a different approach. By separating the individual LDOs sharing a common input grid while considering the impedance of the power delivery network, the open loop characteristics are evaluated in terms of the number of LDOs. Furthermore, a passivity based stability criterion similar to [151] is used to evaluate the relationship between the phase margin and stability of the power grid. The degradation in the resonant frequency caused by the off-chip parasitic impedances is shown to be the primary factor affecting the stability of a grid with multiple LDOs.

7.3 Evaluating the Stability of Multiple LDOs

To evaluate the stability of a power grid, the parasitic impedance of the power delivery network needs to be considered. To use the classical phase margin of an open loop, single-input-single-output (SISO) system, the power delivery network is separated, as shown in Fig. 7.3. To detach an LDO from the grid while including the impedance of the power grid, the impedance of the power delivery network is split per each regulator. For example, considering the three LDOs shown in Fig. 7.3 with

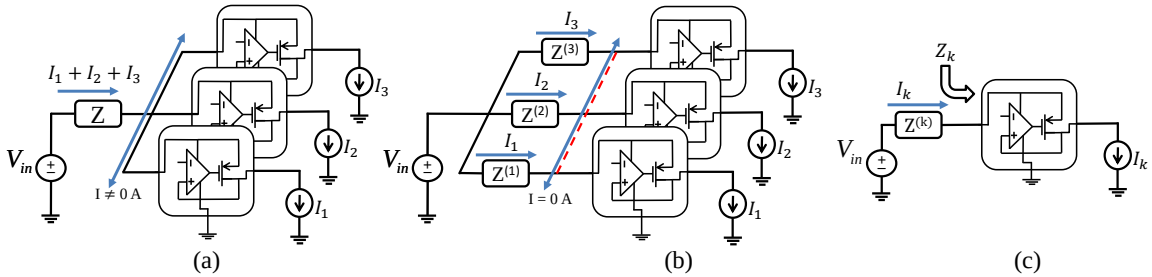

Figure 7.3: Model of a power delivery system with parallel connected LDOs. The impedance of the power delivery network observed from the input of the LDOs is represented as a lumped impedance Z, (a) multiple LDOs attached to the same power grid, and (b) the grid impedance split per LDO, and (c) each LDO is separated based on the corresponding grid impedance at the input of the LDO.

an input grid impedance Z, the impedances at the input of the LDOs are

$$Z^{(1)} = \frac{Z_1 Z}{Z_{123}},\ Z^{(2)} = \frac{Z_2 Z}{Z_{123}},\ Z^{(3)} = \frac{Z_3 Z}{Z_{123}}, \tag{7.1}$$

$$Z_{123} = Z_1 || Z_2 || Z_3, \tag{7.2}$$

where $Z^{(k)}$ is the impedance of the power delivery network observed from the input of an LDO (k), and Z_k is the input impedance of an LDO (k). For N LDO regulators sharing a common input grid, the impedance at the input of an LDO is

$$Z^{(k)} = \frac{Z_k Z}{Z_{123\ldots N}}, \tag{7.3}$$

$$Z_{123...N} = Z_1||Z_2||...||Z_N. \tag{7.4}$$

Note that the impedance at the input of each LDO depends upon the parasitic impedance of the power delivery network and the input impedance of each LDO. The open loop response and phase margin of the SISO LDO system considering the input impedance $Z_{(k)}$ produce useful insight into the relationship between the number of LDOs and the stability of the power grid. To intuitively relate the number of LDOs to the grid stability, multiple LDOs consisting of the same topology and bias conditions is considered. Note that under these conditions, the impedance at the input of each LDO simplifies to $N \times Z$ where N is the number of LDO regulators.

7.3.1 Effect of Number of LDOs on Grid Stability

To evaluate the stability of a grid composed of multiple LDOs, the circuit shown in Fig. 7.4a is considered. The input to a number of parallel connected LDOs is loaded with an RLC network characterizing the parasitic impedances. A circuit composed of multiple LDOs sharing a common grid can be simplified under the condition that each regulator is symmetric and operates under balanced current sharing, as illustrated in Fig. 7.4b. Under this condition, the parasitic impedance can be divided into N sections where N is the number of LDO regulators, each conducting the same current. As a result, each section of impedance is directly connected to an LDO and detached

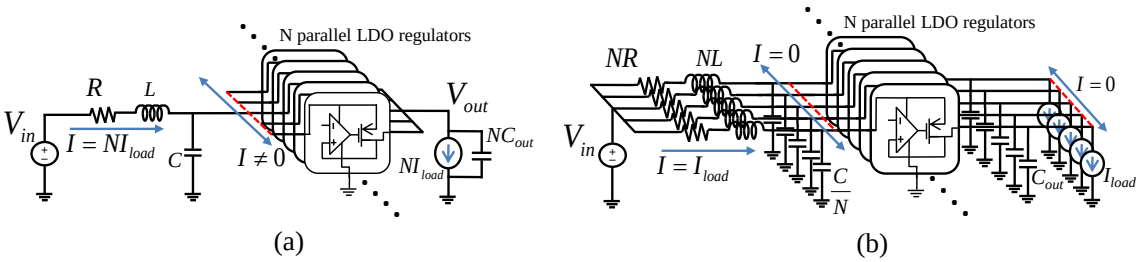

Figure 7.4: Model of a power delivery system with parallel connected LDOs, (a) with an off-chip parasitic impedance, and (b) distribution of the parasitic impedance when the LDOs operate under the same load conditions [158]. The quiescent current of the LDOs is assumed to be negligibly small.

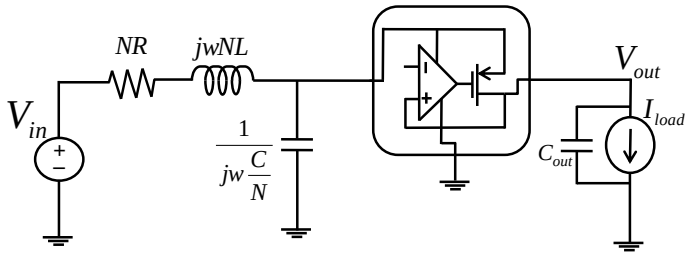

Figure 7.5: Reduction of parallel connected LDOs operating under the same load conditions [158].

from the rest of the circuit, leading to the circuit shown in Fig. 7.5. Note that this circuit produces the identical response to load variations if the system with multiple LDOs remains balanced, regardless of whether the output load is shared or located across separated grids, as shown in Fig. 7.6. The instability due to the increasing number of LDOs can therefore be evaluated by this simplified single loop system consisting of a single LDO using the phase margin as a metric. Note that although this condition reflects a constrained case for regulators in different blocks, the condition of balanced loads is more common in shared output grids. In the remainder of this

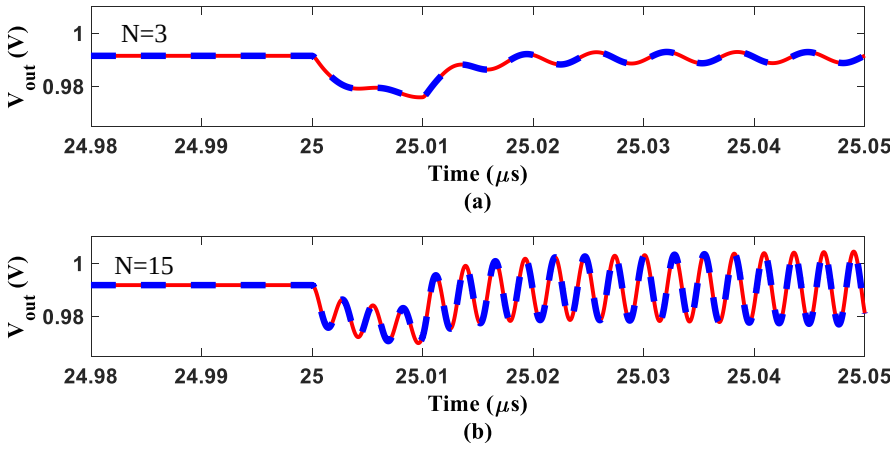

Figure 7.6: Transient simulation of multiple connected LDO regulators where the solid line describes the circuit shown in Fig. 7.4a and the dashed line describes the circuit shown in Fig. 7.5, (a) three LDOs, and (b) 15 LDOs.

section, this system is used to provide intuition behind the degradation in stability of a multi-LDO system, providing insight into the relationship between grid stability and the number of LDO regulators.

7.3.2 Source of Instability

The stability of a system depends upon the open loop gain and phase of that system. The system becomes unstable if the open loop gain is greater than unity when the phase shift is greater than 180^o. To evaluate the stability, a small-signal model of the simplified circuit shown in Fig. 7.7 is used. The LDO is modeled as a two pole system [97]. The dominant pole is generated by the error amplifier whereas the second pole is due to the output capacitor C_{out} of the LDO regulator. The open

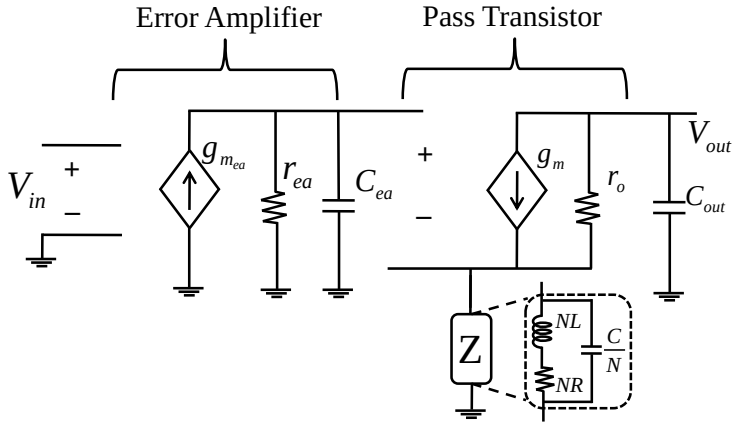

Figure 7.7: Small-signal model of the simplified circuit shown in Fig. 7.5 [158].

loop gain of the LDO regulator is

$$H(s) = \frac{V_{out}}{V_{in}}(s) = \frac{-A_{ol}}{(1 + sC_{ea}r_{ea})(1 + sC_{out}(r_o + g_m r_o Z + Z))}, \tag{7.5}$$

$$A_{ol} = g_m r_o A_{ea}, \tag{7.6}$$

$$A_{ea} = g_{m_{ea}} r_{ea}, \tag{7.7}$$

$$Z = \frac{NR + sNL}{1 + sRC + s^2 LC}, \tag{7.8}$$

where R, L, and C are, respectively, the parasitic resistance, inductance, and capacitance at the input, and A_{ol} and A_{ea} are, respectively, the open loop gain of the LDO regulator and error amplifier over the midband frequency range. The parasitic impedance at the input of the LDO regulator adds two additional poles and zeros, producing a biquad characteristic. The open loop transfer function can be re-written

as

$$H(s) = \frac{V_{out}}{V_{in}}(s) \approx \frac{-A_{ol}(1 + sRC + s^2LC)}{(1 + sC_{ea}r_{ea})(1 + sC_{out}r_o)(1 + \frac{s}{w_oQ} + \frac{s^2}{w_o^2})}, \tag{7.9}$$

$$w_o \approx \frac{1}{\sqrt{L(C + NC_{out})}}, \tag{7.10}$$

$$Q \approx \frac{\sqrt{L(C + NC_{out})}}{RC + C_{out}(NR + r_o)}, \tag{7.11}$$

where w_o and Q are, respectively, the resonant angular frequency and quality factor of the complex poles, produced by the interaction between the LC impedances and the LDO.

The complex poles, due to the RLC impedances, produces a resonant spike in the open loop characteristic of the regulator, as shown in Fig. 7.8. Note that the resonant peak changes with the number of LDOs. As the number of regulators that share a common input grows from one to ten, the gain at the resonant frequency increases beyond 0 dB above the initial unity gain frequency (UGF), resulting in multiple zero crossings. The significant reduction in phase at the resonant frequency therefore produces an unstable system when the number of LDOs is sufficiently high to shift the resonant frequency close to the UGF of the LDO. If the resonant frequency is not sufficiently apart from the UGF, the stability of the system degrades even under heavy load conditions ($>$ 1 mA). The degradation of the resonant frequency with an

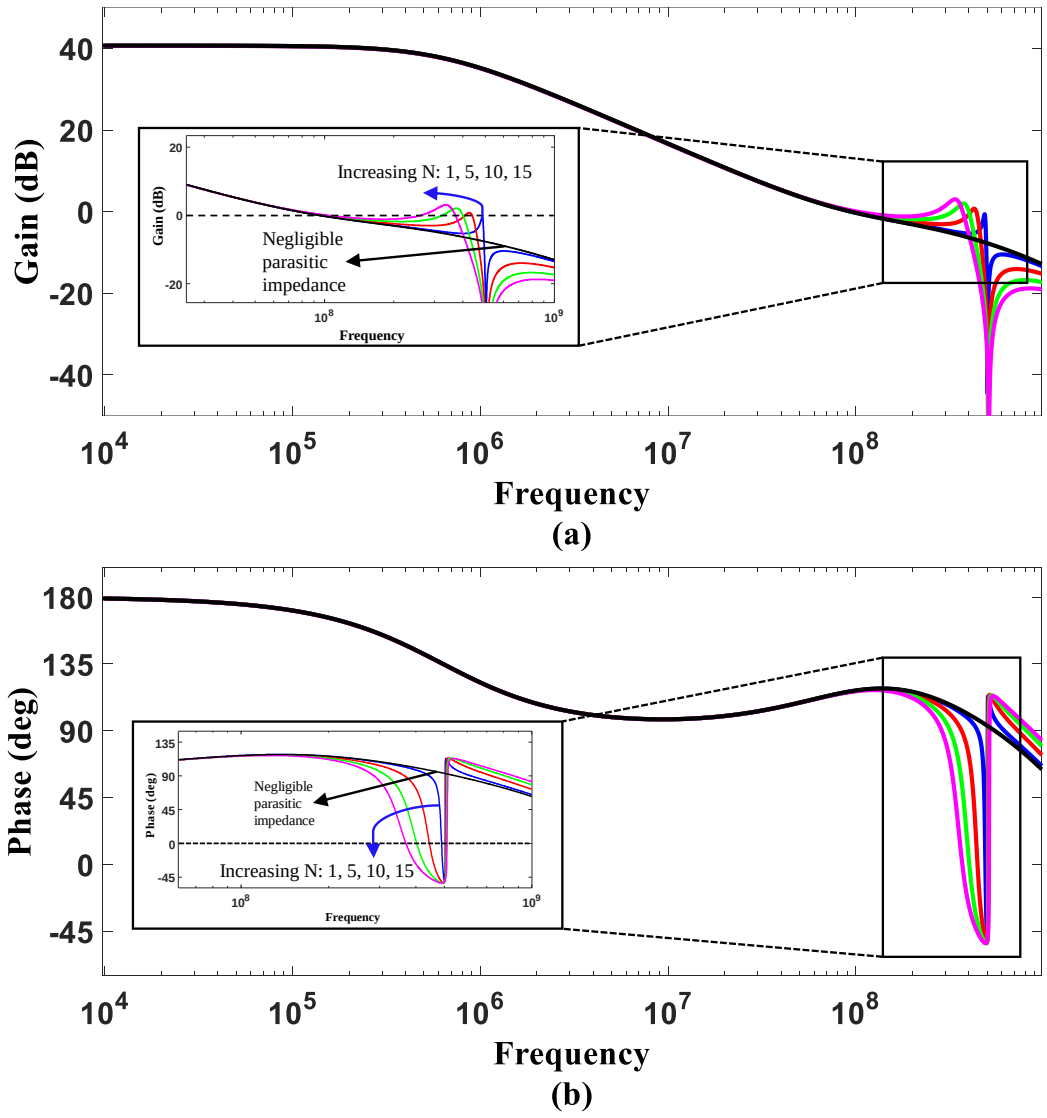

Figure 7.8: Bode plot of circuit model shown in Fig. 7.5. The effect of an increasing number of parallel LDOs, (a) open loop gain, and (b) phase. With five parallel LDOs, the open loop gain rises above 0 dB beyond the initial unity gain frequency, producing an unstable system. $C = 1$ nF, $L = 100$ pH, $R = 100$ $\mu\Omega$, $C_{out} = 50$ pF, and $I_{load} = 210$ mA per LDO.

increasing number of LDOs is the source of instability when the input power grid is shared among multiple LDOs [1].

Note that this effect resembles the resonant behavior steming from the high quality factor under light load conditions [98]. Depending upon the circuit architecture, the LDO can exhibit complex poles under light load conditions (< 1 mA), potentially producing a high quality factor (> 0.707), degrading the stability. Since a heavy load is assumed, the biquad characteristic is due to the parasitic impedance at the input of the LDO (210 mA). Under this condition, the LDO exhibits a single pole behavior below the UGF when the input parasitic impedance is neglected (see Fig. 7.8). The resonant behavior is therefore due to the RLC impedance of the power grid.

7.3.3 Degradation of Resonant Frequency

The resonant frequency and quality factor of the parasitic impedance are, respectively,

$$f_{res} = \frac{1}{2\pi\sqrt{LC}}, \tag{7.12}$$

[1] The on-chip power delivery network consists of input and output power grids. The input power grid connecting the off-chip power network to the inputs of the LDOs is shared among all the regulators. The output power grid delivers the output voltage from the individual regulators to the distributed loads across the integrated circuit.

$$Q = \frac{1}{R}\sqrt{\frac{L}{C}}. \tag{7.13}$$

These standard parameters characterizing a power grid remain constant irrespective of the number of LDOs. When the parasitic impedance of the power grid is combined with the LDO regulators, however, the resonant frequency as well as the quality factor changes with respect to the number of LDOs. The spike in the resonant frequency shifts to a lower frequency when the number of LDOs increases, degrading the grid stability.

The variation in the resonant frequency is due to the output capacitance of the LDOs that interacts with the input power grid. Note that if the output capacitance is negligible, the transfer function does not exhibit any complex poles (thus, no resonant peak), as described by (7.9) and (7.11). To understand the interaction of the output capacitance with the input power grid, the input impedance of the small-signal model (see Fig. 7.7), described in (7.14)[2], is considered. The magnitude of the input impedance $|Z_{in}|$ is shown in Fig. 7.9c. Note that the input impedance simplifies to $\frac{1}{sC_{out}} + (r_o||\frac{1}{g_m})$ under high frequencies (above the corner frequency of $|Z_{in}|$). The LDO shown in Fig. 7.9a can be represented as an RC circuit from the input port, as shown in Fig. 7.9b. Under heavy load conditions, since the pass gate enters the linear region, the input resistance of the pass transistor (from the source terminal,

[2]A parallel resistor R_{ea} is considered to account for the additional path to ground through the error amplifier.

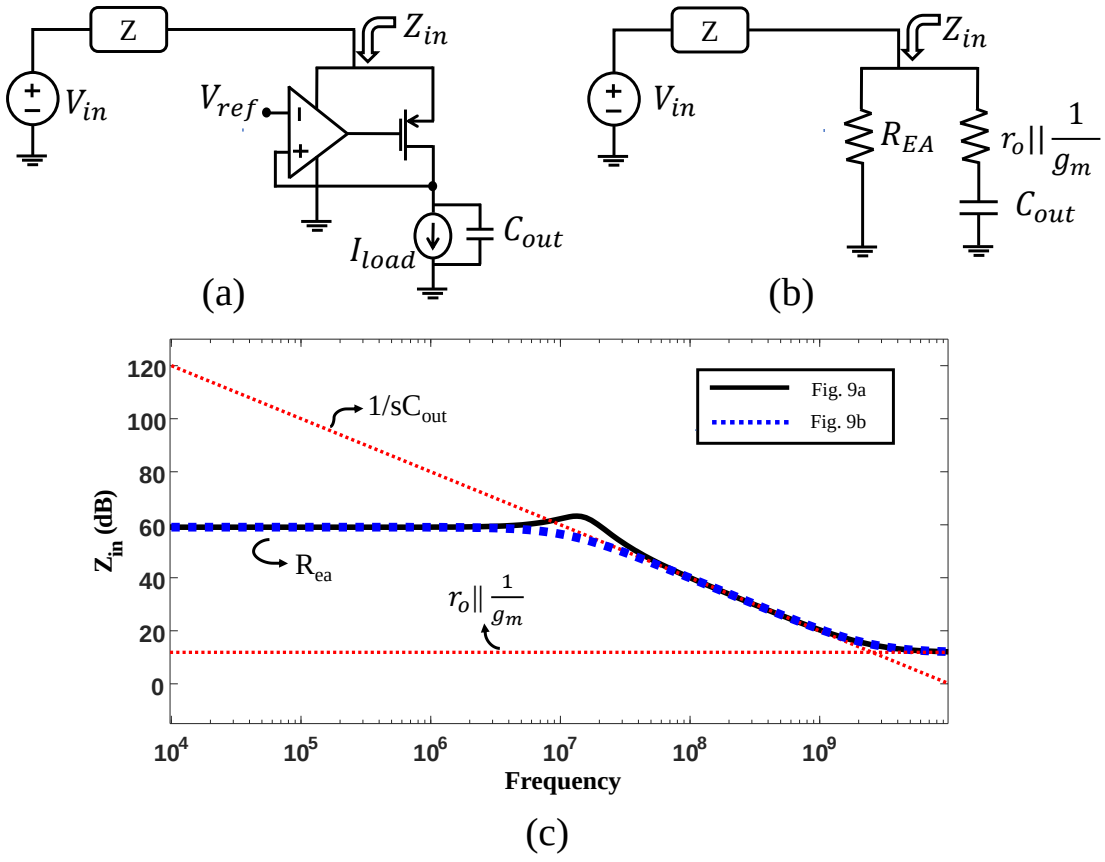

Figure 7.9: Input impedance, (a) considering the LDO regulator, (b) model of the input impedance as an RC circuit, and (c) comparison of the magnitude of the input impedances.

$r_o||\frac{1}{g_m}$) significantly decreases. As a result, the input of the LDO is exposed to the output capacitance. The capacitance at the input of the grid therefore increases with additional LDOs (due to the increasing output capacitance), shifting the resonant frequency when the corner frequency of $|Z_{in}|$ is below the resonant frequency (7.12).

The effective resonant frequency based on the small-signal model in (7.9) is approximately

$$f_{res\_eff} \approx \frac{1}{2\pi\sqrt{L(C+NC_{out})}}. \tag{7.15}$$

This expression is consistent with the remarks on the interactions between the input and output capacitance of the LDO. When the input power grid is exposed to the output capacitors, the input capacitance with N LDOs increases from C to approximately $C + NC_{out}$, leading to the new resonant frequency described in (7.15). The relationship between the number of on-chip LDOs and the resonant frequency and the quality factor (based on (7.9) and (7.15)) is illustrated in Fig. 7.10.

$$Z_{in}(s) = \frac{(1+g_m r_o g_{m_{ea}} r_{ea}) + s(r_{ea}C_{ea} + r_o C_{out}) + s^2(r_{ea}C_{ea}r_o C_{out})}{(\frac{1}{R_{ea}} + \frac{1}{R_{ea}} g_m r_o g_{m_{ea}} r_{ea}) + s(\frac{1}{R_{ea}} r_{ea}C_{ea} + \frac{1}{R_{ea}} r_o C_{out} + C_{out} + C_{out} g_m r_o) + s^2(\frac{1}{R_{ea}} r_{ea}C_{ea}r_o C_{out} + r_{ea}C_{ea}C_{out} + r_{ea}C_{ea}C_{out} g_m r_o)} \tag{7.14}$$

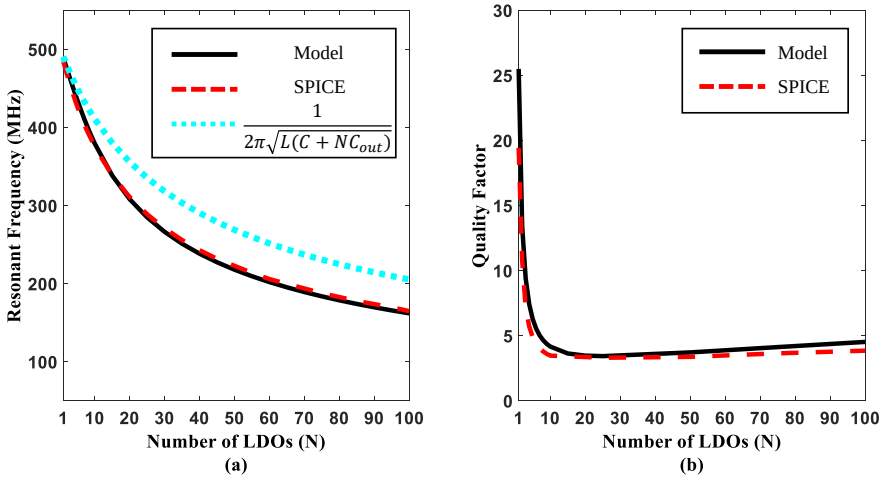

Figure 7.10: Effect of different number of LDO regulators, (a) resonant frequency, and (b) quality factor, based on the circuit characteristics considered in Fig. 7.8. The model is based on the small-signal circuit shown in Fig. 7.7.

Note that increasing the number of LDOs from one to ten lowers the resonant frequency by roughly 100 MHz when considering an input capacitance of 1 nF and an output capacitance of 50 pF (per LDO).

The degradation in resonant frequency causes instability, particularly in wide bandwidth LDOs [89, 108] where the closed loop UGF of the circuit is on the order of tens to hundreds of MHz (e.g., [104, 106, 108, 159–161]). Low bandwidth LDOs are therefore resilient to instability caused by an increasing number of LDOs. LDOs with high UGF are used in applications such as microprocessors where fast response times are necessary [108], or analog circuits requiring high power supply rejection

across a wide range of frequencies [160, 161]. In these applications, careful design of the power network is necessary to ensure sufficient separation between the LDO UGF and the resonant frequency of the network.

7.3.4 Condition for Stability

The traditionally assumed worst case stability condition of an LDO under light load conditions changes when the input of an LDO is coupled to an LC network. The complex poles due to the input impedances increases the phase shift by 180^o, significantly lowering the phase margin if the resonant frequency is close to the UGF.

If the resonant frequency is close to the UGF, the open loop characteristics of the LDO produces a non-monotonic response (see Fig. 7.8). At the resonant frequency, the gain increases while lowering the phase below 0^o. As a result, multiple zero crossings are observed. With non-monotonic behavior of the loop gain, a positive phase margin at the initial unity gain frequency does not guarantee system stability. To ensure stability, sufficient phase margin is required at the last 0 dB crossing of the Bode diagram. Equivalently, the following condition must be satisfied,

$$\angle H(max\{f_{_{0dB}}\}) > 0, \tag{7.16}$$

where $H(f)$ is the open loop transfer function of the LDO and $max\{f_{_{0dB}}\}$ is the highest unity gain frequency (i.e., last 0 dB crossing).

If the phase margin is nonpositive, the power grid is unstable. This conclusion is based on the passivity based stability criterion in [151], considering two observations proved in [162] and [163]:

1. An LTI system, when coupled to a passive system, is stable if and only if the driving point impedance is passive [162].

2. An impedance Z(s) cannot be passive if Z(s) exhibits imaginary or right half plane (RHP) poles [163].

Note that the second condition also implies bounded-input-bounded-output (BIBO) stability, i.e., a passive impedance $Z(s)$ is bounded for all bounded inputs [163]. Hence, the output impedance Z_{out} of N parallel LDOs, shown in Fig. 7.11, needs to satisfy the BIBO stability, exhibiting no RHP poles.

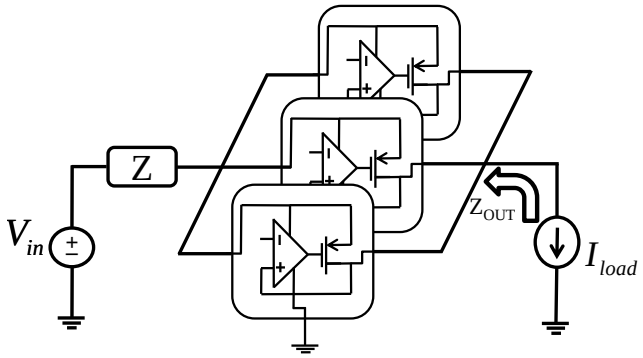

Figure 7.11: Output impedance of multiple LDOs sharing a common load.

The output impedance $Z_{out}(s)$ of N LDOs sharing a common output under balanced loads is

$$Z_{out}(s) = \frac{(\frac{1}{N})(r_o + Zg_m r_o + Z)(1 + sr_{ea}C_{ea})}{A_{ol} + (1 + sr_{ea}C_{ea})(1 + sC_{out}(r_o + Zg_m r_o + Z))}. \tag{7.17}$$

To relate the open loop phase margin to the passivity constraints, Z_{out} is re-written as

$$Z_{out}(s) = Z'_{out}(s)\frac{1}{|H(s)|e^{j(\pi+\theta(s))} + 1}, \tag{7.18}$$

where $\theta(s)$ is the phase of the open loop transfer function ($\angle H(s)$), and $Z'_{out}(s)$ is

$$Z'_{out}(s) = \frac{(\frac{1}{N})(r_o + Zg_m r_o + Z)}{(1 + sC_{out}(r_o + Zg_m r_o + Z))}. \tag{7.19}$$

Evaluating (7.18) at the UGF yields

$$Z_{out}(jw_{UGF}) = Z'_{out}(jw_{UGF})\frac{1}{1 \cdot e^{j(\pi+PM)} + 1}, \tag{7.20}$$

where phase margin PM is

$$PM = 180^o - \Delta\theta(jw_{UGF}). \tag{7.21}$$

Note that when the phase margin decreases to 0, the output impedance diverges to infinity, violating the passivity criterion and producing an unstable power grid. The

effect of a decreasing phase margin on the complex poles of the output impedance is shown in Fig. 7.12.

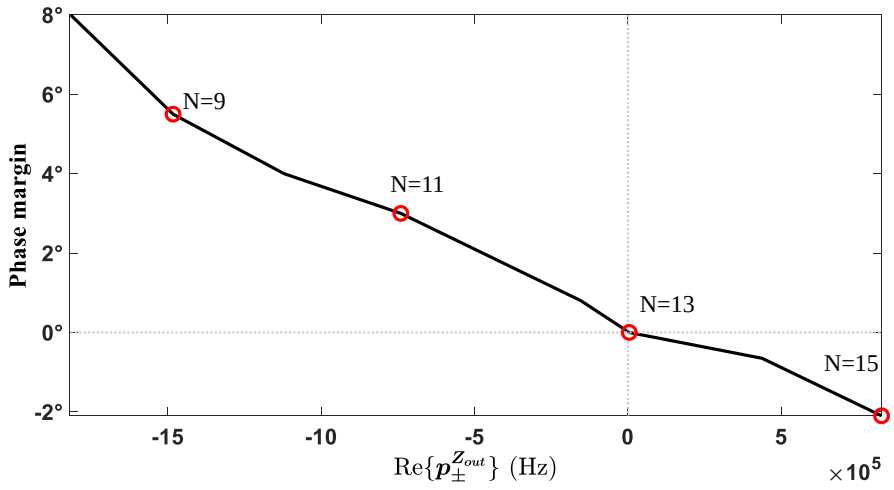

Figure 7.12: Decreasing phase margin shifts the complex poles of the output impedance $p_{\pm}^{Z_{out}}$ to the RHP.

Note that when the phase margin is nonpositive, the complex poles exhibit non-negative real parts, producing an unstable power grid. A positive phase margin of the open loop LDO when considering the input impedance is therefore a necessary condition to ensure a stable power delivery network.

7.4 Effect of Design Parameters on Grid Stability

In this section, the relationship between the circuit level parameters and the stability of the power delivery network is explored. In Section 7.4.1, LDO design parameters

such as the UGF, output capacitance, and input impedance, and in Section 7.4.2, the power grid design parameters such as the input parasitic impedances are considered.

7.4.1 LDO Design Parameters

The effect of several LDO design parameters on the stability of a power grid is explored. The output capacitance, UGF, and input impedance of the LDO are considered, respectively, in Sections 7.4.1.1, 7.4.1.2, and 7.4.1.3.

7.4.1.1 Output Capacitance

The effective resonant frequency is influenced by the output capacitance C_{out}, as described in (7.15). A small output capacitance reduces the variations in the resonant frequency while a large output capacitance increases the variations with respect to the number of LDOs, as shown in Fig. 7.13. Variations in the resonant frequency become negligible when $C >> NC_{out}$. While a smaller output capacitance may be desired for this reason, decreasing the output capacitance C_{out} can increase the quality factor, exacerbating the peak resonant frequency. This effect is due to the complex poles of the open loop circuit (see (7.9)) merging with the complex zeros with decreasing C_{out}, as shown in Fig. 7.14. Note that if the output capacitance is zero, the complex poles and zeros are equal, thereby canceling (see (7.11)). For zero output capacitance, the input impedance has therefore no effect on the open loop transfer characteristics,

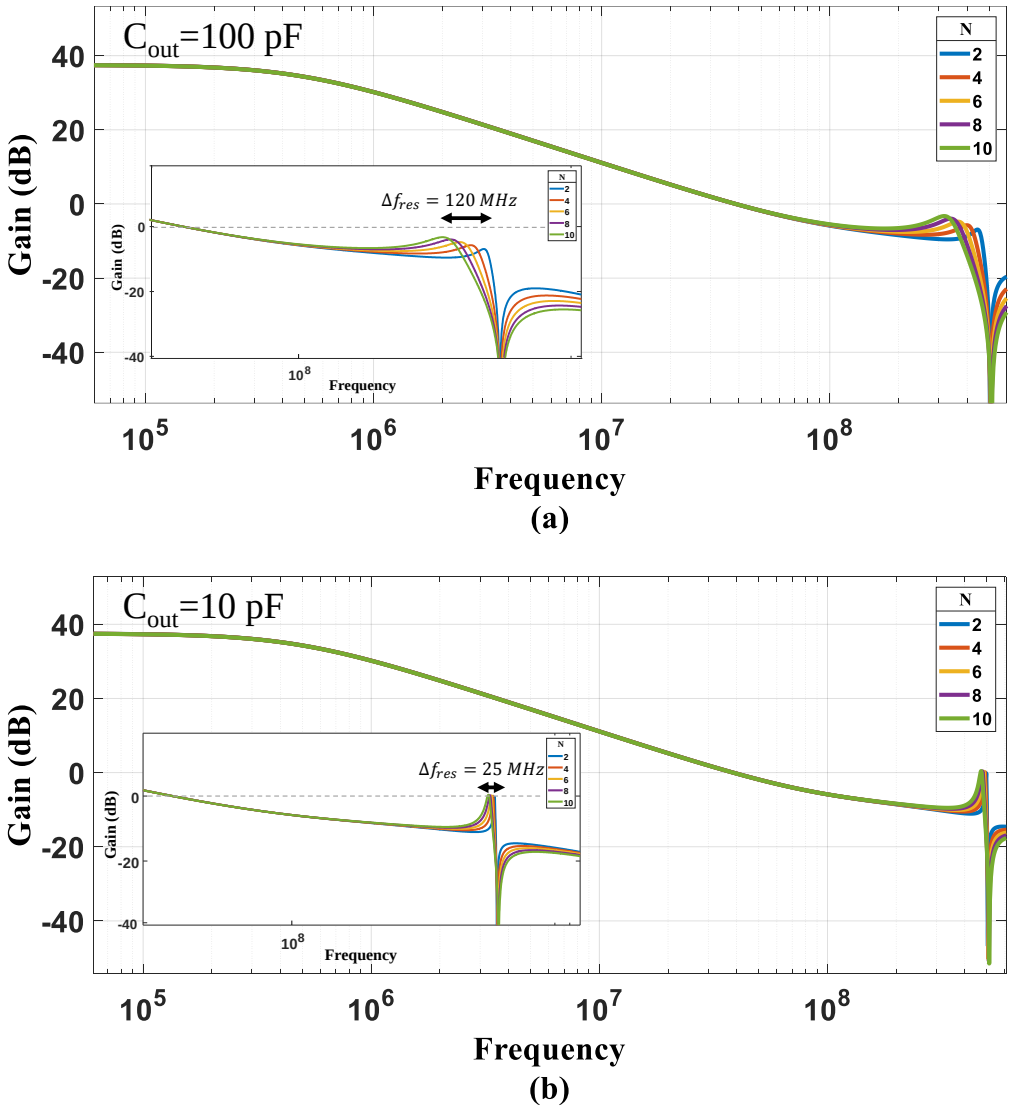

Figure 7.13: Open loop transfer characteristics of an LDO assuming the simulation setup shown in Fig. 7.8, (a) a large output capacitance, and (b) a small output capacitance.

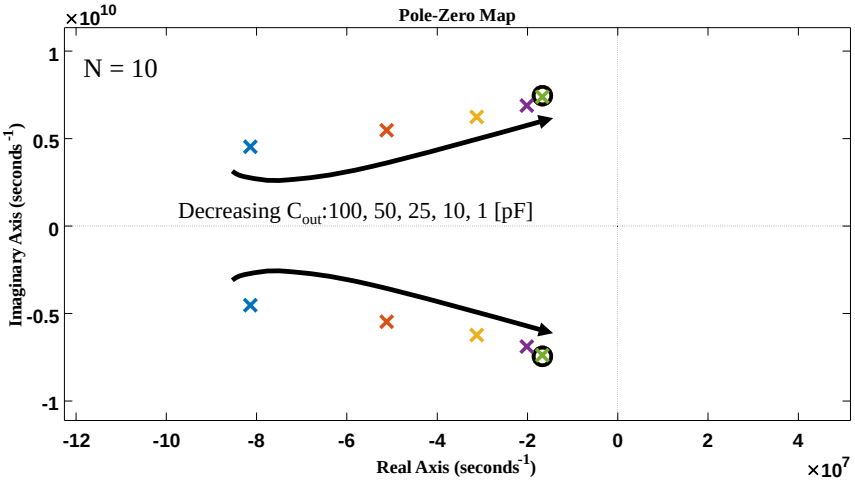

Figure 7.14: Complex poles of the open loop transfer function $H(s)$ become equal to the complex zeros with decreasing output capacitance.

as noted in (7.9). Under this ideal condition, the stability of the power grid does not decrease with an increasing number of LDOs. Since some output capacitance is always present however to either reduce the voltage droop or due to the parasitic capacitance of the load, the resonance effect needs to be considered.

7.4.1.2 Unity Gain Frequency

Sufficient separation between the UGF and the resonant frequency is necessary to ensure a stable power delivery network. Under a fixed resonant frequency, the LDOs can be designed for a lower UGF to increase this separation. For instance, the gain or 3 dB frequency can be decreased to reduce the UGF, as shown in Fig. 7.15. Note that a sufficient reduction in gain lowers the resonant spike below 0 dB at a cost of

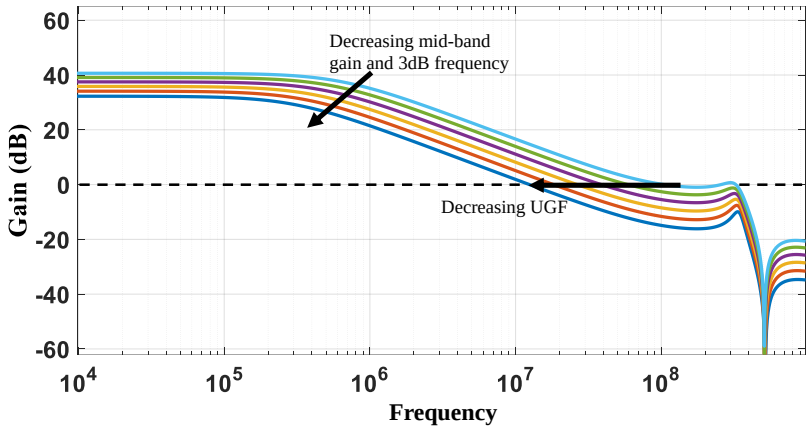

Figure 7.15: Reduction in the open loop gain of the LDO significantly lowers the UGF. Decreasing the mid-band gain from 40 dB to 32 dB reduces the UGF (considering the first 0 dB crossing) from 100 MHz to 12 MHz.

lower load regulation. Furthermore, the lower the UGF, the slower the response time of the LDO, increasing the voltage droop. A tradeoff therefore exists between the power grid stability and power noise due to voltage droops.

The effect of a decreasing UGF on the voltage of the power delivery network is shown in Fig. 7.16. Note that as the UGF decreases, the output response of the regulators become more stable. This effect is due to improved phase margins as the peak resonant frequency falls below the unity gain of the regulator, farther from the UGF.

7.4.1.3 Input Impedance

The resonant frequency of the input power grid changes with respect to the number of LDOs due to the interactions between the input power grid and the output

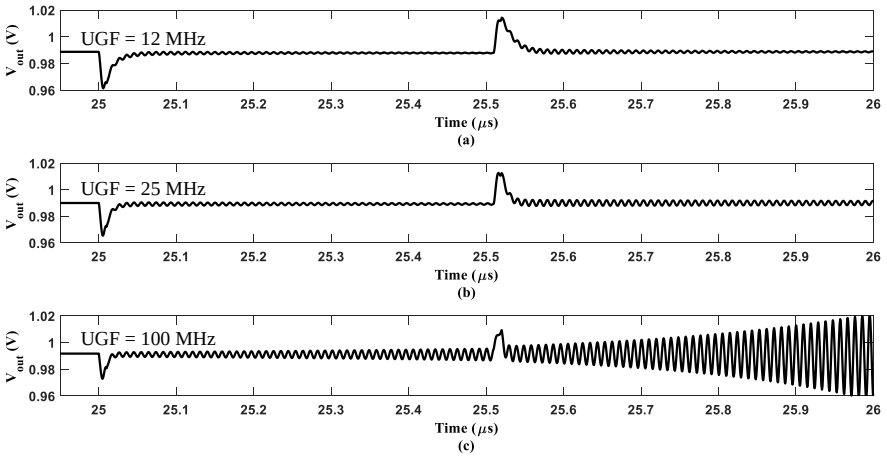

Figure 7.16: Effect of the UGF on the output power grid shared by ten LDOs, (a) UGF at 12 MHz, (b) UGF at 25 MHz, and (c) UGF at 100 MHz. An output capacitance of 100 pF and an input capacitance of 1 nF per LDO is considered with the off-chip power grid described in the Appendix. A total load variation from 1.75 A to 2.1 A is assumed (equally divided among the LDOs).

capacitors of the LDOs, as described in Section 7.3.3. The input impedance of an LDO Z_{in} is equal to the output capacitance above the corner frequency of $|Z_{in}|$ under heavy load conditions, increasing the total capacitance on the input power grid (from C to $C+NC_{out}$) and decreasing the resonant frequency with an increasing number of LDOs. To prevent degradation of the resonant frequency with additional LDOs, the corner frequency of $|Z_{in}|$ is increased, as shown in Fig. 7.17. The corner frequency is

$$f_{corner} \approx \frac{1}{2\pi R_{ea} C_{out}}, \tag{7.22}$$

where R_{ea} is the resistance of the path to ground through the error amplifier (see Fig. 7.9). Note that R_{ea} (and therefore the corner frequency of Z_{in}) depend upon the

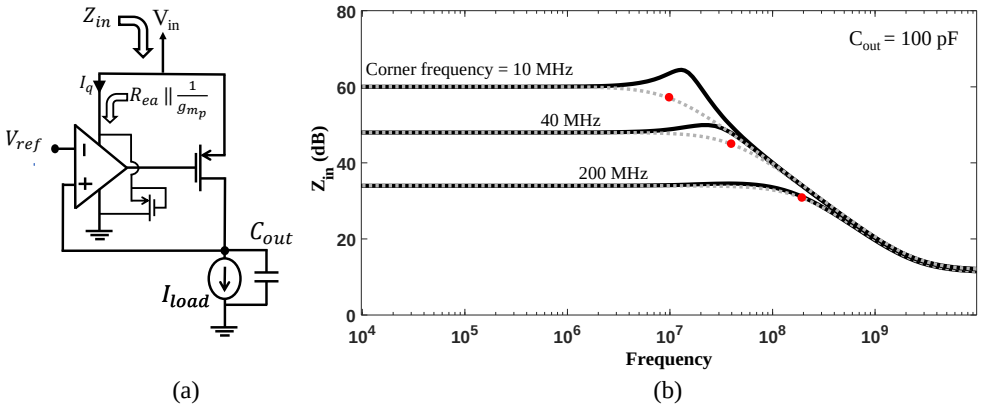

Figure 7.17: Increasing the corner frequency of $|Z_{in}|$ reduces the interaction between the input and output power grids over a wider range of frequencies, (a) LDO with an additional pull-up transistor to reduce the resistance of the error amplifier R_{ea} observed from the input power network, and (b) different corner frequencies of $|Z_{in}|$ under a fixed output capacitance by increasing the quiescent current I_q of the LDO.

topology and quiescent current of the error amplifier. To evaluate the relationship between the corner frequency of $|Z_{in}|$ and the stability of the power grid, a pull-up transistor is placed in parallel to the error amplifier, as shown in Fig. 7.17a. For a fixed output capacitance, R_{ea} is decreased to increase the corner frequency. The size of the pull-up transistor is increased to reduce R_{ea} at a cost of higher quiescent current I_q.

The additional path to ground supplied by the error amplifier limits the frequency range where the input impedance of the LDO is equivalent to the impedance of the output capacitance, preventing the input power grid to be loaded by the output capacitor of the LDOs. Increasing the corner frequency f_{corner} therefore mitigates the stability of the power grid, as shown in Fig. 7.18. The stability of the power grid

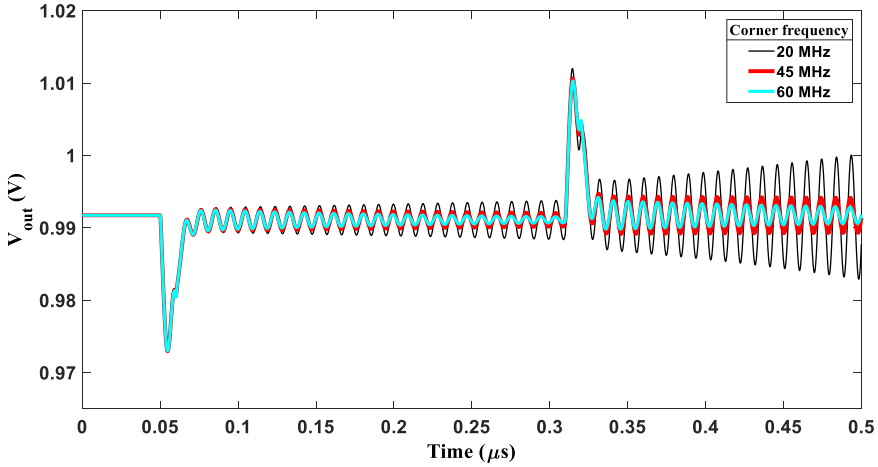

Figure 7.18: Increasing the corner frequency of Z_{in} reduces the interaction of the output capacitance with the input power grid, improving the stability of the power delivery network. The same power delivery network described in Fig. 7.15 is assumed with a UGF of 100 MHz.

improves as the corner frequency increases from 20 to 60 MHz at a cost of higher I_q. A tradeofff therefore exists between the current efficiency and the stability of the power delivery network.

7.4.2 Power Grid Parameters

A power delivery network can support a more stable multi-LDO system by exploiting the strong dependence between the grid stability and the parasitic impedance of the power delivery network. A complex power delivery network typically consists of parasitic board and package impedances as well as the impedance of the controlled collapse chip connections (C4s), as shown in Fig. B.1. While the package and board impedances have little effect on system stability due to the large off-chip capacitors

(typically tens to hundreds of μF), the resonance produced by the C4 inductance and on-chip capacitance is often significant [164–166]. This resonance, typically a few hundred MHz, exhibits a higher quality factor due to the small on-chip capacitance (from tens to hundreds of nF). In this section, the power delivery network described in the Appendix is considered. The number of LDOs N is ten and the UGF is 100 MHz (see Fig. 7.1b, assuming a load condition of 210 mA). The input capacitance is 1 nF per LDO. The resonant frequency due to the C4s is 112 MHz.

To improve the stability of the power grid, either the quality factor is reduced to increase the system damping or the resonant frequency is increased to further separate the UGF of the LDOs. One approach to increase the damping is using a more resistive power delivery network. Resistive power grids are more stable at a cost of greater power noise, as shown in Fig. 7.19. Note that increasing the C4 parasitic resistance by an order of magnitude stabilizes the power grid. This effect is due to the quality factor Q that is inversely proportional to the resistance, as described in (7.11). This approach is suitable in those systems with low current demand (e.g., 1 to 2 A) where the IR drop is negligible.

Alternatively, reducing the C4 parasitic inductance L_{C4} improves the grid stability. The benefits of decreasing L_{C4} are twofold. A lower parasitic inductance increases the resonant frequency, thereby increasing the separation with the UGF. Moreover, the lower the inductance, the smaller the quality factor. Power delivery networks that

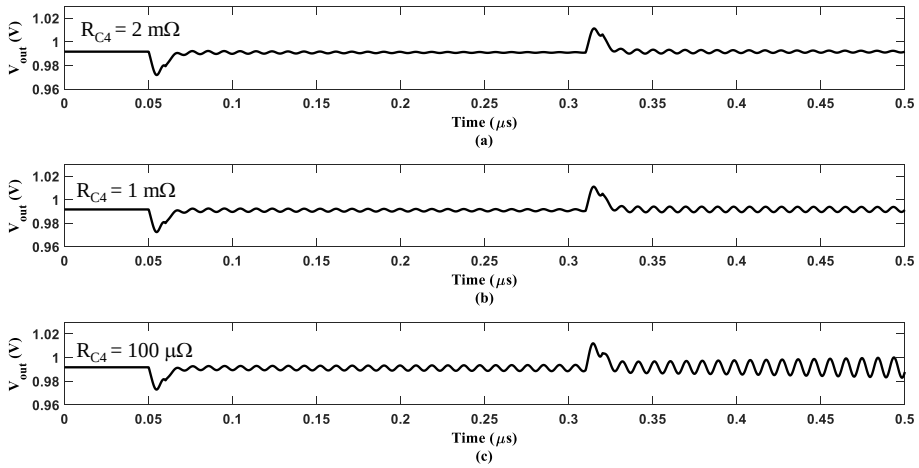

Figure 7.19: C4 parasitic resistance R_{C4} is increased to reduce the quality factor and improve the stability of the power grid. Transient response of the output grid considering, (a) $R_{C4} = 2$ mΩ, (b) $R_{C4} = 1$ mΩ, and (c) $R_{C4} = 0.1$ mΩ. The power delivery network and parasitic impedances are listed in the Appendix with a load variation of 1.75 A to 2.1 A evenly distributed among the ten LDOs. The resonant frequency is 112 MHz.

are less inductive can therefore support a greater number of LDOs, as shown in Fig. 7.20. For a parasitic inductance of 100 pH, five LDOs sharing the same power grid is sufficient to produce RHP poles and destabilize the power grid. Moreover, to support ten LDOs, this parasitic inductance needs to be reduced to a few tens of pH or less. The LDOs operating with a UGF of 100 MHz are exposed to a resonant frequency of approximately 110 MHz when $L_{C4} = 100$ pH and $N = 10$. Reducing the inductance to 10 pH increases the resonant frequency to 350 MHz, improving the stability of the power delivery network.

Lastly, adding on-chip capacitance to the input power grid reduces the quality factor. Due to the weak relationship between the quality factor and the input capacitance

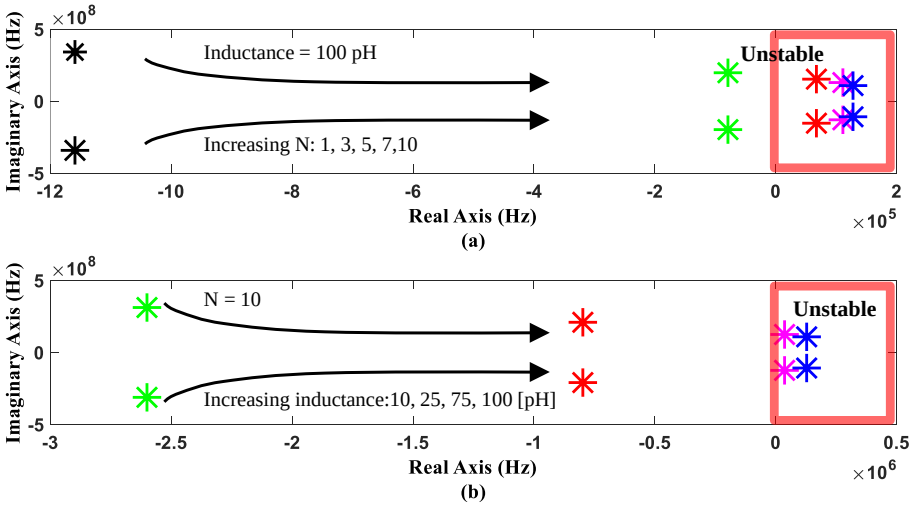

Figure 7.20: Pole movement of the output impedance at the driving point, (a) considering single and multiple LDOs, and (b) several C4 parasitic inductances.

($Q \propto \frac{1}{\sqrt{C}}$), a large input capacitance is needed to increase the damping characteristics. Note that increasing the capacitance decreases the resonant frequency. The input capacitance therefore needs to be sufficiently high to not exacerbate the grid stability. Specifically, if $C >> NC_{out}$, the complex poles and zeros are approximately similar, reducing the phase roll off at the resonant frequency. The required input capacitance in terms of the number of LDOs and input parasitic inductance is shown in Fig. 7.21. Note that the required capacitance at the input of the regulators reaches a few tens of nanofarads per LDO when the number of regulators is more than ten. Reducing the parasitic inductance significantly lowers the required input capacitance since the difference between the complex poles and zeros decreases with a smaller

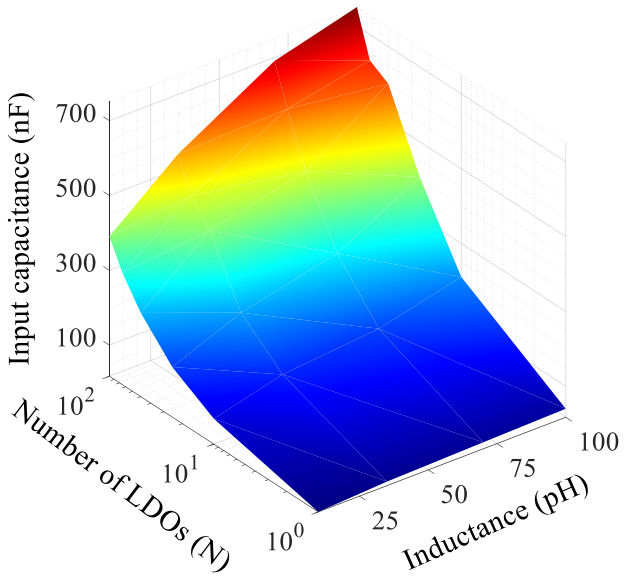

Figure 7.21: Effect of inductance and number of LDOs on the input capacitance required to prevent a phase shift of more than 45^o at the resonant frequency (considering the circuit characteristics used in Fig. 7.5).

inductance. The overhead of large input capacitance can therefore be mitigated by reducing the parasitic inductance.

7.5 Summary

The stability of a power delivery system composed of multiple on-chip LDO regulators is explored. Ensuring stability under light load conditions is insufficient to guarantee the stability of a power delivery system shared by multiple LDO regulators. The phase margin of a standalone LDO, without considering the impedance of the input power grid, yields a misleading stability metric. The stability of a power network is shown to degrade as the number of regulators increases due to the resonance

generated by the RLC parasitic impedances. With a greater number of LDOs, the resonant frequency decreases, degrading the stability of the power grid when the unity gain of the regulator and resonant frequency of the input grid are insufficiently apart. Ensuring sufficient separation between the UGF and resonant frequency is critical to maintain a stable power grid while supporting a larger number of on-chip LDOs.

Chapter 8

Distributed Pass Gates in Power Delivery Systems with Digital Low Dropout Regulators

In high performance, high complexity integrated systems, power noise is a primary challenge due to increasing load currents and parasitic interconnect resistances [44]. Furthermore, lower supply voltages have made digital circuits highly sensitive to power noise. In a deeply scaled integrated system (for example, below 28 nm), a noise voltage of a few tens of millivolts can be sufficient to produce a timing failure. The on-chip power delivery network has therefore become a primary concern [167,168].

To reduce power consumption in high performance, high complexity systems, digital low dropout (LDO) regulators are integrated on-chip to enable fast dynamic voltage scaling [89,169]. A digital LDO consists of a single controller driving multiple pass gates. A digital LDO controller (typically a bidirectional shift register) receives a control signal from a comparator which compares a reference voltage to the output

voltage of the regulator, sampled from the power grid, as shown in Fig. 8.1 [170]. Depending upon the difference between the reference and grid voltages, the controller

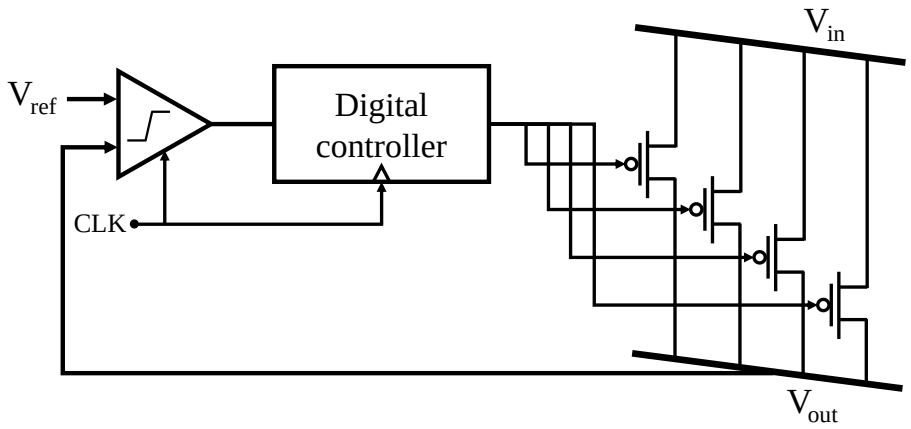

Figure 8.1: A digital low dropout regulator.

either turns on or off a set of pass transistors. The digital LDO includes a digital controller with a dedicated clock signal. A set of pass gates is switched by the active edge of the clock signal.

To reduce power noise, a low parasitic resistance between the pass gates and the load is desirable. The parasitic resistance can be reduced by either decreasing the distance between the pass gates and the load or by utilizing the low resistance, upper metal interconnects. In practice, the upper metal layers are limited and shared among the clock distribution network, input power grid, and global signals [169]. Due to these limited metal resources, the higher resistance, lower metal layers are often used to route the pass gates to the loads. As a result, the pass gates need to be placed close to the loads [169]. In this chapter, a methodology to distribute the pass gates of

a digital LDO to reduce IR drops across a power grid is proposed. Based on the load distribution, a centroid that represents the region of a grid with the largest current demand is introduced. These grid centroids are used to determine the location of the pass gates.

In Section 8.1, the proposed pass gate distribution methodology is described. In Section 8.2, an analysis of power noise is provided. The proposed methodology is also compared to other conventional distribution techniques. In Section 8.3, some conclusions are offered.

8.1 Distributed Pass Gates

A digital LDO compares a reference voltage to a voltage sampled from a single node within a power grid. A small portion of the power grid is therefore actively regulated. One approach to enlarge the regulated portion of a power grid is distributing multiple LDOs across a grid. This approach however requires additional area and reduces the energy efficiency due to higher quiescent currents. Furthermore, multiple LDOs can decrease the stability of a power grid [151, 158]. To maintain a regulated voltage across the power grid using a single digital LDO, the IR drops need to be low. To reduce the IR drops, the proposed digital LDO methodology switches on the pass gates at the centroid of the power delivery network, as shown in Fig. 8.2. In the case of a single centroid, the current is injected into the grid from the node where

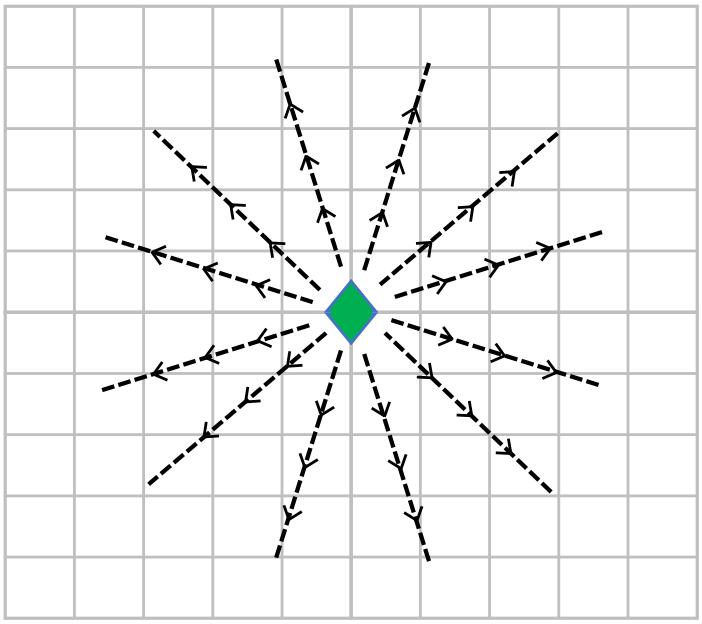

Figure 8.2: Pass gates located at the centroid of a grid to source the distributed load currents.

the centroid is located. In subsection 8.1.1, the centroid of a grid is described, and a heuristic to estimate the centroid is provided in subsection 8.1.2.

8.1.1 Grid Centroid

The centroid of a grid is the location within a power grid that minimizes the maximum IR drop when connected to either a voltage or current source. Consider a network consisting of two current loads, as shown in Fig. 8.3. The centroid between these two current sources, I_1 and I_2, is located to ensure that, if connected to a current source, the maximum voltage drop is minimized. To minimize the maximum IR drop, the centroid is placed between two loads to ensure that $I_1R_1 = I_2R_2$, where

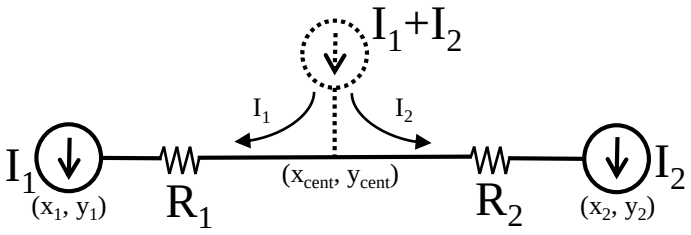

Figure 8.3: Centroid of two load currents.

R_1 is the resistance between I_1 and the centroid, and R_2 is the resistance between I_2 and the centroid.

Suppose I_1 and I_2 are located within a grid structured power network. The location of the centroid between two current sources is

$$\left(x_1 + \frac{(x_2 - x_1)I_2}{I_1 + I_2}, y_1 + \frac{(y_2 - y_1)I_2}{I_1 + I_2}\right), \tag{8.1}$$

where (x_1, y_1) and (x_2, y_2) are, respectively, the location of current sources I_1 and I_2 on the grid. The resistance between two nodes, (x_1, y_1) and (x_2, y_2), is assumed to be linearly proportional to the Manhattan distance between these nodes ($\propto |x_1 - x_2| + |y_1 - y_2|$). Note that the grid centroid gravitates towards the load requiring higher current.

8.1.2 Proposed Heuristic

To estimate the centroid of a grid consisting of multiple current sources, an iterative approach is considered here. Pseudocode for estimating the centroid is shown in

Heuristic 1. The runtime for determining the centroid of n current loads is $O(n-1)$.

Heuristic 1 Calculate (x_{cent}, y_{cent})

INPUT: Set of n current loads and coordinates
OUTPUT: Centroid coordinates
$I_{cent} \leftarrow I_1$ ▷ Initialize centroid
$(x_{cent}, y_{cent}) \leftarrow (x_1, y_1)$
for all i = 2 to n **do**
    $I_1 \leftarrow I_{cent}$
    $I_2 \leftarrow I_i$
    $(x_1, y_1) \leftarrow (x_{cent}, y_{cent})$
    $(x_2, y_2) \leftarrow (x_i, y_i)$
    $I_{cent} \leftarrow I_1 + I_2$ ▷ Update centroid
    $(x_{cent}, y_{cent}) \leftarrow \left(x_1 + \frac{(x_2 - x_1) I_2}{I_1 + I_2}, y_1 + \frac{(y_2 - y_1) I_2}{I_1 + I_2} \right)$
end for
return (x_{cent}, y_{cent})

During each step of the iteration, a new centroid between the old centroid and a load current is determined based on (8.1). This iterative process is illustrated in Fig. 8.4. The heuristic places the centroid closer to those regions loaded with a higher current to minimize the maximum IR drop. Alternatively, the centroid of n load currents can be determined using

$$(x_{cent}, y_{cent}) = \left(\frac{\sum_{i=1}^{n} x_i I_i}{\sum_{j=1}^{n} I_j}, \frac{\sum_{i=1}^{n} y_i I_i}{\sum_{j=1}^{n} I_j} \right). \tag{8.2}$$

Note that (8.2) is based on Heuristic 1, assuming the centroid between two loads is (8.1).

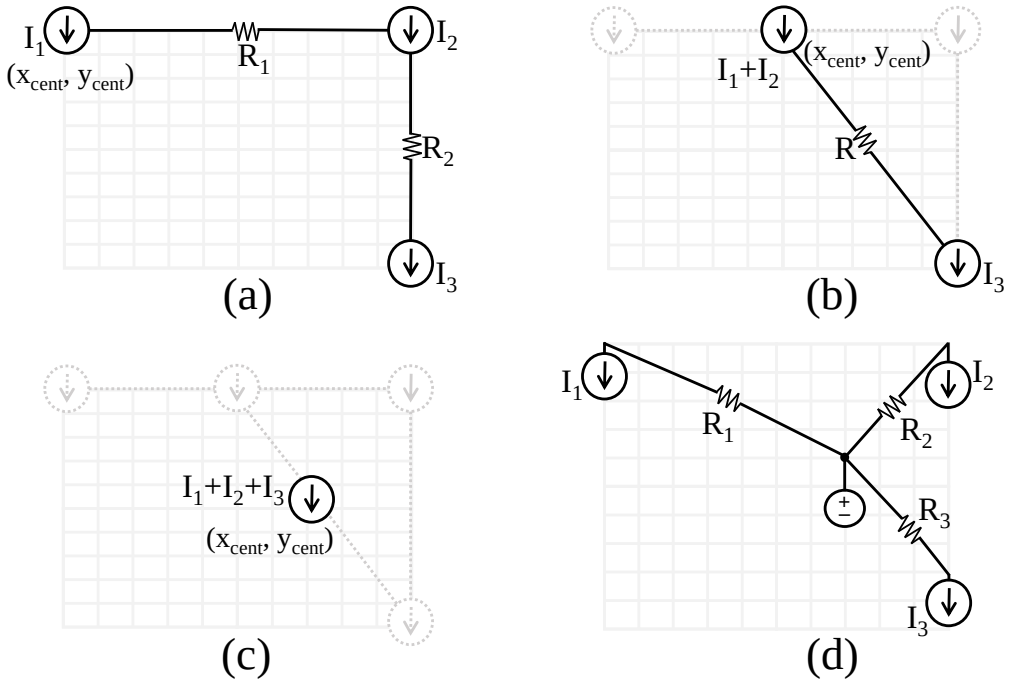

Figure 8.4: Iterative process for determining the centroid of three load currents, (a) the centroid is initially assigned to load current I_1, (b) a new centroid between I_1 and I_2 is determined and replaces the old centroid, (c) a new centroid is determined between the current centroid and I_3, replacing the old centroid, and (d) the final centroid is replaced by a source connected to the load currents.

To reduce IR drops across the grid, the pass gates are placed at the centroid of the grid (see Fig. 8.2). If the maximum IR drop is greater than a critical threshold V_{drop}^{th} (which depends upon the technology node), the grid is partitioned into quadrants to support distributed centroids. The steps describing the partitioning process is shown in Fig. 8.5. For example, a power grid divided into four different quadrants, each with a unique centroid is illustrated in Fig. 8.6a. A centroid is determined within each quadrant using the procedure outlined in Heuristic 1, as shown in Fig. 8.6b. The grid can be further divided into finer granularity by breaking each quadrant into an

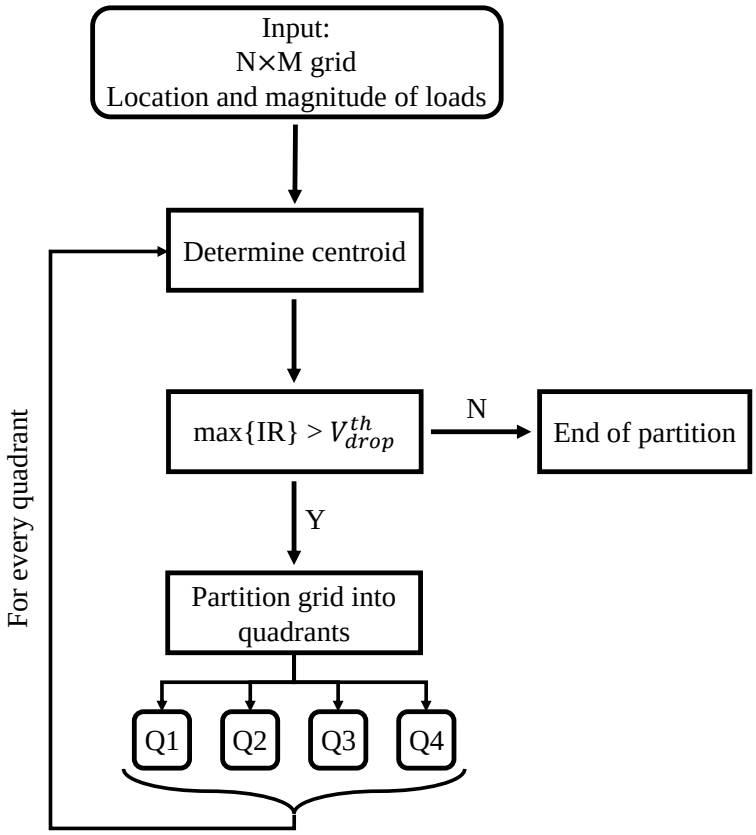

Figure 8.5: Recursive process to determine multiple centroids.

additional four quadrants, leading to 16 regions. The pass gates are placed at these quadrant centroids to distribute the current across the grid, as shown in Fig. 8.7. A power grid consisting of Q centroids and a digital LDO with an N-bit bidirectional shift register embodies $Q \times N$ pass gates. A set of Q pass gates of $Q \times N$ pass gates are turned on at the active edge of the clock signal of the digital LDO, each located at a different centroid within the Q quadrants. The pass gates are sized in proportion to the current demand within the specific centroid to reduce the amplitude of any limit

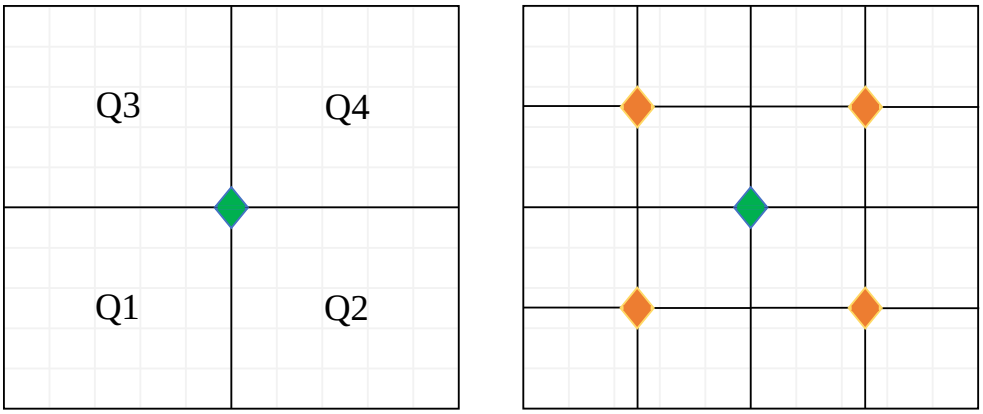

Figure 8.6: Iterative process for determining the location of the quadrant centroids, (a) power grid divided into four quadrants, and (b) a centroid is placed within each quadrant. A diamond represents an individual centroid.

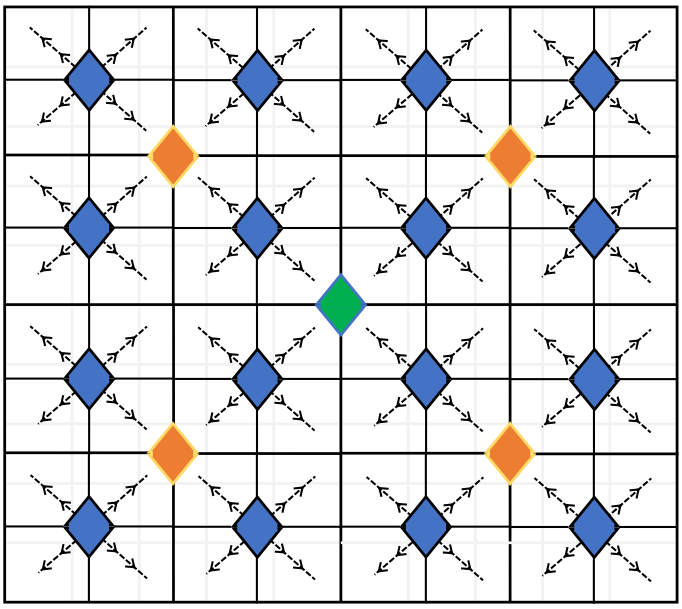

Figure 8.7: A power grid composed of 16 regions with separate centroids.

cycle oscillations [171]. In the next section, the proposed digital LDO is compared to conventional placement methodologies.

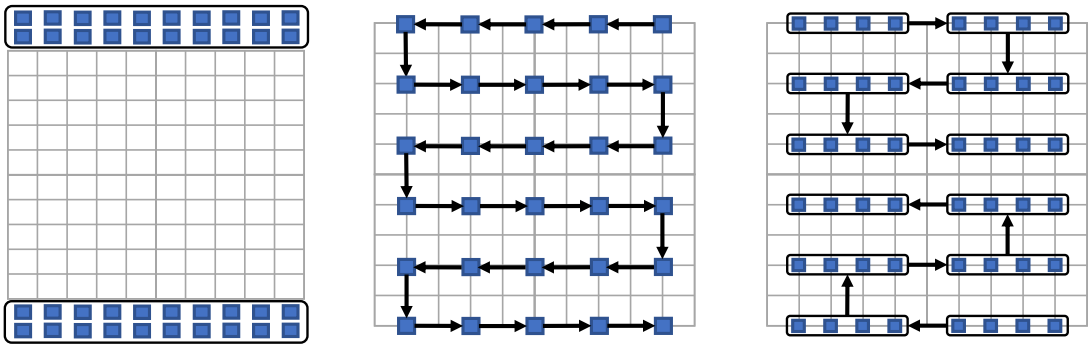

Figure 8.8: Different pass gate distribution topologies, (a) top–bottom [96, 150], (b) daisy chain [172], and (c) distribution from [169].

8.2 Power Grid Analysis

The proposed pass gate distribution topology is compared to three different distribution topologies. These distribution topologies are introduced in subsection 8.2.1. The power grid analysis and a comparison of different distribution topologies are explored in subsection 8.2.2.

8.2.1 Distribution Topologies

Three different pass gate distribution topologies have been considered; top–bottom, daisy chain, and the distribution topology described in [169], as shown in Fig. 8.8. In the top–bottom topology, the pass gates are clustered in the upper and lower sections of a power grid. The low resistance, upper metal layers are used to distribute the current from the top and bottom portions of the grid to the loads, ensuring low IR

drops. This technique is prohibitive in those cases where the upper metal layers are largely used by the clock distribution network, global signals, and the input portion of the power distribution network [169]. Alternatively, the daisy chain topology is typically considered in power gating schemes. Since the pass gates used for power gating are also used for digital LDOs [105], the daisy chain distribution is considered here. In the daisy chain configuration, the pass gates are serially turned on to reduce the transient current ($C\frac{dV}{dt}$ current where C is the power grid capacitance) (see Fig. 8.8b). The propagation delay between the two ends of the chain prevents the pass gates to simultaneously turn on, limiting the $C\frac{dV}{dt}$ current. The topology illustrated in Fig. 8.8c has recently been proposed to distribute the pass gates of a digital LDO [169]. Similar to the daisy chain topology, a clustered group of pass gates is switched in a serial fashion at the active edge of the clock signal. Two clusters switch from the top and bottom sections of the grid toward the center of the grid.

8.2.2 Comparison of Pass Gate Distribution Topologies

In this subsection, the proposed distribution topology is compared to the topologies described in subsection 8.2.1. A power analysis is conducted considering a 32×32 output grid with a 1 Ω resistance between adjacent grid nodes. The resistance of the input and ground grids are assumed to be negligible (due to the low resistance, upper metal layers and the large number of distributed C4 power connections). The input

voltage is 1.2 volts and the output voltage is 1 volt. A total current load of 150 mA is assumed. The total size of the pass gates is set according to the total current load and is maintained equal across all topologies. A digital LDO with two controllers of 32-bit bidirectional shift registers is considered to enable both coarse and fine modes of operation for mitigating limit cycle oscillations [171]. In the proposed topology, 16 centroids are considered (see Fig. 8.7), and the grid voltage is sampled from the primary centroid of the grid (the initial centroid before partitioning). The grid voltage is sampled from the center of the grid in the other topologies.

A voltage map of the power grid assuming a uniform load distribution is shown in Fig. 8.9. The output grid voltage is averaged across time to determine the DC voltage (filtering out the limit cycle oscillations). The difference between the maximum and minimum voltages are 5 mV for the proposed distribution topology and the topology described in [169], whereas this difference is 18 and 22 mV for, respectively, the top–bottom and daisy chain topologies.

The daisy chain topology suffers from significant IR drops, on the order of tens of millivolts. Since the grid voltage is sampled at the center of the grid, the digital LDO stops turning on additional pass gates once the voltage at the center node is equal to the reference voltage. The pass gates toward the end of the chain therefore remain closed, resulting in an unbalanced distribution of pass gates. The current therefore

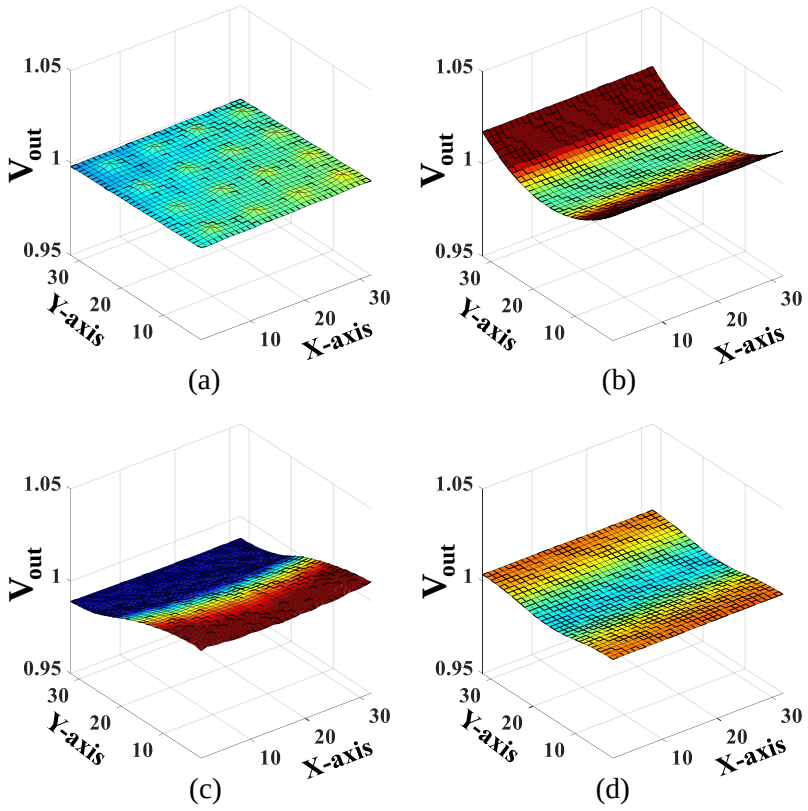

Figure 8.9: Power grid analysis assuming a uniform load distribution, (a) proposed centroid-based distribution topology, (b) top–bottom topology [96, 150], (c) daisy chain topology [172], and (d) distribution topology from [169].

flows from the upper portion of the chain towards the lower portion, leading to the voltage degradation shown in Fig. 8.9c.

The top–bottom topology also suffers from significant IR drops since the pass gates are not distributed across the power grid. A set of pass gates are simultaneously switched from both clusters on the top and bottom regions of the grid, producing a symmetric voltage map (see Fig. 8.9b). The current flowing from the upper and

lower sections towards the grid center produces a higher voltage at the edge of the grid, forming a valley pattern.

The high IR drop of the top-bottom topology is mitigated using the topology described in [169]. Since the clustered pass gates are distributed and simultaneously controlled from both the upper and lower parts of the grid, the distance the currents travel within the grid is reduced. Under a uniform load condition, both the proposed centroid-based topology and [169] significantly eliminate any IR drops.

Under nonuniform load conditions where the portion of the grid is loaded with large currents, the daisy chain, top-bottom, and the topology described in [169] exhibit significantly higher IR drops, as shown in Fig. 8.10. The difference between the maximum and minimum voltages after re-evaluating the centroids are 9.5 mV for the proposed distribution topology, 25 mV for the topology described in [169], and 38 and 33 mV for, respectively, the top–bottom and daisy chain topologies. The distribution topology described in [169], daisy chain, and top–bottom are sensitive to nonuniform load distributions since the location of the pass gates do not consider the location of the load currents. The proposed centroid-based distribution topology however determines the regions loaded with the highest currents, and places and sizes the pass gates according to the location and current demand of the centroids. A low IR drop across the power grid is therefore maintained despite highly nonuniform load distributions, as shown in Fig. 8.10a.

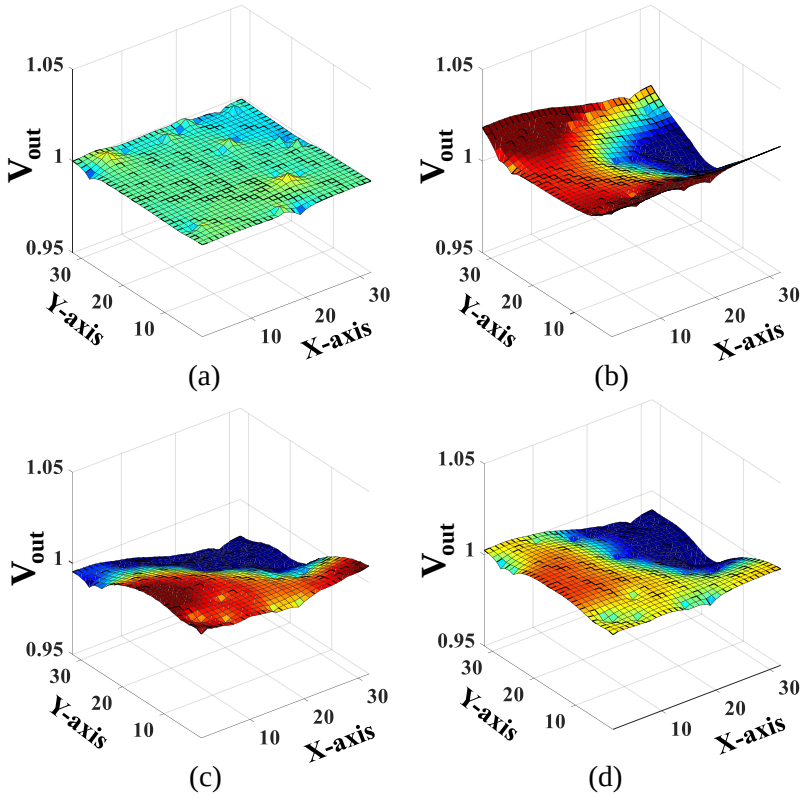

Figure 8.10: Power grid analysis considering nonuniform load distribution, (a) proposed centroid-based distribution topology, (b) top–bottom topology [96, 150], (c) daisy chain topology [172], and (d) distribution topology from [169].

To explore the effect of a nonuniform load distribution on the steady state voltage variations, a set of Monte Carlo simulations is described, as shown in Fig. 8.11. The magnitude of the load between each node of the output and ground grids is treated as an independent random variable. These random variables are assigned a number sampled from a Gaussian distribution with a mean equal to the current load when the load is uniformly distributed ($\frac{0.15}{32\times32}$ amperes per node). For each sample, the centroid is re-evaluated using the procedure outlined in Heuristic 1. A Monte Carlo

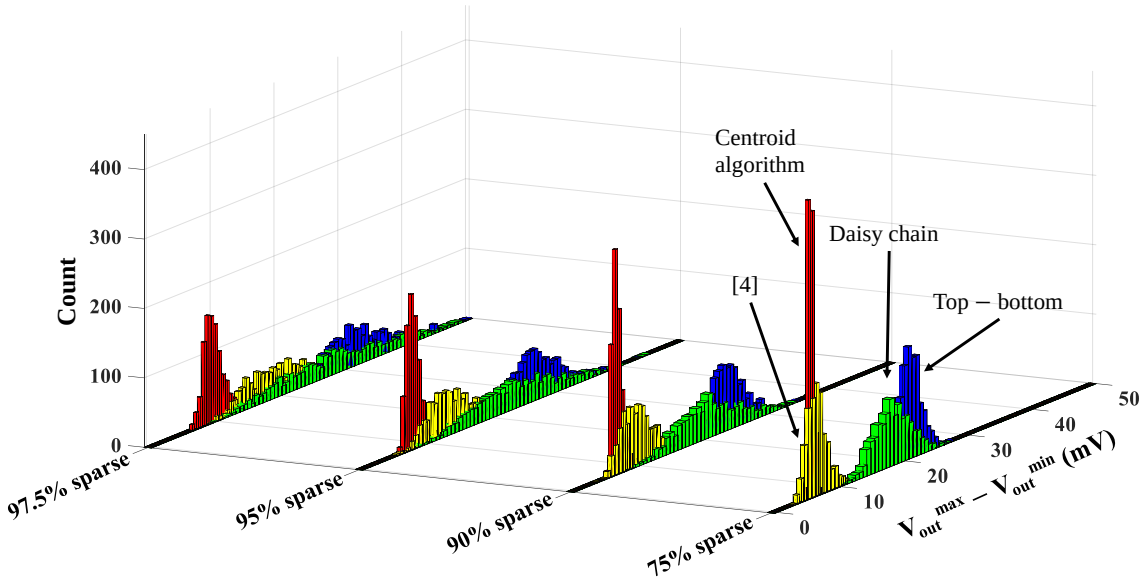

Figure 8.11: Monte Carlo simulations evaluating the difference between the maximum and minimum voltage across a power grid considering four different pass gate distribution topologies.

simulation is conducted for four cases where, in each case, the sparsity of the current distribution is altered to produce different degrees of nonuniform load distribution. For example, considering 75% sparsity, 256 ($0.25 \times 32 \times 32$) different load currents are distributed across the grid (as opposed to $32 \times 32 = 1,024$ nonzero current loads). Note that the total load current is maintained at 150 mA. A higher sparsity therefore exacerbates the load nonuniformity across the power grid. The mean and variance of the histograms are summarized, respectively, in Tables 8.1 and 8.2.

The voltage variations across the power grid significantly increases for the topology described in [169], top–bottom, and daisy chain since these distribution topologies do

Table 8.1: Mean of distributions

Sparsity (%)	Centroid method	[169]	Top-Bottom	Daisy Chain
75	6.07	7.04	21.85	18.68
90	7.25	10.61	24.76	21.19
95	8.61	14.82	27.94	23.85
97.5	10.55	21.19	33.02	28.88

Table 8.2: Variance of distributions

Sparsity (%)	Centroid method	[169]	Top-Bottom	Daisy Chain
75	0.15	2.40	2.22	7.26
90	0.65	7.71	6.55	22.42
95	1.95	16.37	12.78	38.55
97.5	5.43	35.08	26.57	70.79

not consider the location of the loads. The mean of the differences between the maximum and minimum voltages on the grid increases by 200%, from sub-ten millivolts up to more than 20 millivolts for the distribution topology described in [169]. For the proposed centroid-based distribution topology, top–bottom, and daisy chain, the mean increases by, respectively, 72%, 55%, and 56%. Under a highly nonuniform load distribution, the proposed centroid topology retains an average voltage difference two to three times lower than the average voltage difference of the other three topologies. Furthermore, the variance significantly increases for the topology described in [169], top–bottom, and daisy chain. As a result, a large number of cases exist in which these topologies lead to higher voltage variations across the grid under steady state conditions. In contrast, the proposed centroid-based approach maintains a relatively low variance of less than 10 millivolts under highly nonuniform load conditions. The

proposed pass gate distribution topology is therefore more suitable to ensure minimal voltage variations across a power grid. Note however that the grid centroids are based on a specific load condition. The loads considered here represent average load conditions (magnitude and location) within a power grid.

8.3 Summary

In this work, the effect of different pass gate distribution topologies on a power distribution grid utilizing a digital LDO is evaluated. The concept of a power grid centroid is introduced to determine those regions within a power grid loaded with the highest currents. Based on this centroid approach, a pass gate distribution topology is presented. The proposed topology is compared to three different pass gate distribution topologies. Since the existing topologies do not consider the location of the loads, the voltage variations across the power grid are greater as compared to the proposed topology based on centroids. Statistical simulations conclude that the voltage variations across the grid significantly increase when using existing distribution topologies under highly nonuniform load conditions. In contrast, the proposed centroid topology demonstrates significantly lower voltage variations across the grid despite a highly nonuniform load distribution.

Chapter 9

Conclusions

The grid is an indispensable apparatus in the design toolbox of engineers. The grid is a structure commonly considered across a broad range of fields to aid the analysis and development of complex systems, from architecting buildings and cities to integrated circuits. Grid structures have been prevalent in integrated circuits since the early days of the semiconductor industry in the 1960's through today in state-of-the-art, highly complex VLSI systems. From memory arrays to power delivery systems, from floorplanning to automated routing, grids are applied across many facets of integrated circuits. The regularity, density, and path diversity of grids, highly desirable qualities in complex integrated systems, are exploited to improve numerous design and analysis challenges. In this dissertation, the underlying reasons for using a grid structure in VLSI systems is investigated along with a set of challenges in designing grid-based circuits and systems such as in resistive memory arrays and on-chip power grids.

Two issues in the design of nonvolatile resistive memories with selectors have been addressed; the computational complexity of large arrays and the energy consumption during write operations. To reduce the computational complexity of designing and analyzing large nonvolatile resistive arrays with cell selectors, a set of closed-form expressions characterizing the size limitations of these arrays has been developed for both $V/2$ and $V/3$ bias schemes. The expressions model the size of the array in terms of circuit and device parameters such as the interconnect resistance, nonlinearity factor, and on–off resistance. The models, exhibiting good agreement with SPICE, provide intuitive insight into the relationships between the array performance (e.g., voltage degradation and read margin) and certain circuit parameters without increasing computational complexity. These models consider the steady state conditions and are therefore not suitable for capturing switching dynamics that are necessary to characterize the energy consumed by a resistive array. To address this deficiency, the energy consumption of resistive memories with cell selectors during write operation has also been explored.

The write energy is typically orders of magnitude greater than the read energy and can significantly affect the total energy consumption of a resistive memory array. The write energy of a resistive array is modeled considering two different bias schemes, the $V/2$ and $V/3$ bias schemes. A critical insight gained from these models is the $V/2$ bias scheme, which is typically assumed to be the most energy efficient bias

scheme, can be less energy efficient than the $V/3$ bias scheme depending upon the array size, nonlinearity factor, and number of selected cells. By exploiting the effect of the number of selected cells on the most energy efficient bias scheme, a hybrid bias scheme, which uses both $V/2$ and $V/3$ bias schemes, is proposed. Energy savings of as much as 2.5x have been demonstrated using this hybrid bias scheme.

Beyond resistive memory, design challenges of on-chip power grids with integrated LDOs have also been addressed. Specifically, the stability of power grids when shared among multiple analog LDOs has been explored. Conventionally, the stability of an on-chip, capacitorless LDO is guaranteed by ensuring sufficient phase margin under light load conditions since light loads typically produce the worst case stability conditions. Instability in a power grid under heavy load conditions has however also been demonstrated when the grid is shared among multiple LDOs. The fundamental cause of instability in a grid with multiple on-chip LDOs has been revealed. The resonant frequency of the power grid is shown to decrease with a larger number of LDOs. The decreasing separation between the unity gain and resonant frequency is the source of instability within a power grid.

Lastly, the challenge of integrating digital LDOs with on-chip power distribution networks is explored. The pass gates of a digital LDO, when clustered within the upper and lower regions of a power grid, produces significant voltage variations due to IR drops across the power grid. The IR drops increase when the load currents

are distributed in a nonuniform fashion. The low resistance, upper metal layers can be leveraged to reduce these IR drops. However, since these metal layers are shared among different networks and are therefore limited, the high resistance, lower metal layers are often used, exacerbating the voltage variations across the power grid. To reduce these IR drops, a topology for distributing the pass gates that considers the location of the loads is proposed. The concept of a grid centroid is introduced to describe the region within a grid loaded with the highest currents. Based on these centroids, the pass gates are distributed across the grid to reduce IR drops. Monte Carlo simulations show a significant reduction in IR drops for a large number of different load distributions as compared to existing pass gate distribution topologies.

In conclusion, grids are amazing. The simple and regular structure that enhances density has tamed the design complexity of VLSI systems for decades, enabling large robust computing platforms available to a huge number of people. The grid however is by no means the silver bullet to design complex VLSI systems. For example, the density of the grid within a resistive memory produces other challenges such as sneak current paths and half-select cells, limiting array performance. It is therefore critical to understand the application-specific nature of grid structures to satisfy design constraints.

As J. Muller-Brockmann once stated, "... one must learn how to use the grid; it is an art that requires practice [23]."

Chapter 10

Future Work

While grids can be used to improve a variety of different characteristics of VLSI systems (see Chapter 2), a grid structure can also inhibit system performance. For example, sneak paths, half-selected cells, and leakage currents in resistive memory arrays (see Chapters 4 and 5), which hinder array performance (latency and energy consumption), are caused by the grid structure. A tradeoff therefore exists in resistive memory systems between density and performance. In addition to these issues, a frequently selected row in a memory system can interact with unselected adjacent lines, unintentionally changing the state of the unselected cells. This undesirable effect is known as the RowHammer effect [173–175]. In this chapter, future research directions are offered to address these issues in nonvolatile resistive memories. In Section 10.1, the RowHammer effect is described. In Section 10.2, the RowHammer phenomenon in resistive memory arrays is explained. In Section 10.3, a research

direction to address this issue in resistive memories is discussed. In Section 10.4, some conclusions are offered.

10.1 RowHammer Effect

The RowHammer effect, first reported in commodity DRAM systems [174], is the frequent access (at the DRAM refresh rate) of the same row (hence, row hammering) within a memory array, resulting in altered data on adjacent, unselected rows. This issue produces both erroneous data within the memory as well as a security vulnerability. In fact, a malicious attack leveraging the RowHammer effect is demonstrated in [174]. The RowHammer effect is therefore a critical reliability and security concern in memory systems.

The number of affected cells due to the RowHammer effect significantly increases in advanced technology nodes [173]. The underlying reason for the RowHammer effect is the small physical distance between the word lines, leading to unintended interactions. Similarly, the RowHammer effect can occur in nonvolatile resistive memories where the cell area can be as small as $4F^2$. Under this situation, the distance between adjacent rows is equal to the minimum feature size, F. Noise coupling between adjacent lines therefore needs to be evaluated, in particular when an aggressor row

(selected wordline) is consecutively activated, injecting noise into a victim row (an unselected adjacent wordline). In the next section, the RowHammer effect is discussed in the context of nonvolatile resistive arrays.

10.2 RowHammer Effect in Nonvolatile Resistive Arrays

With device and interconnect scaling, the space between interconnects decreases while the aspect ratio of the interconnect increases [18]. The coupling capacitance therefore significantly increases in advanced technology nodes. In resistive memory arrays where the cell area is as small as $4F^2$ (e.g., considering 1R or 1S1R memory cells), the coupling capacitance between adjacent interconnects can be large, increasing power consumption and, more importantly, coupling noise from the selected rows or columns, as shown in Fig. 10.1. When writing to or reading from a resistive memory array, a select voltage is applied to the selected row. During the transition from an unselected state to a selected state, the selected row can inject noise into an adjacent line, producing a voltage spike on the unselected rows (see Fig 10.1). The larger the coupling capacitance, the higher the noise spike because the impedance between the two adjacent lines is lower. More importantly, the amplitude of the noise spike increases on the portion of the interconnect farther from the voltage source (the

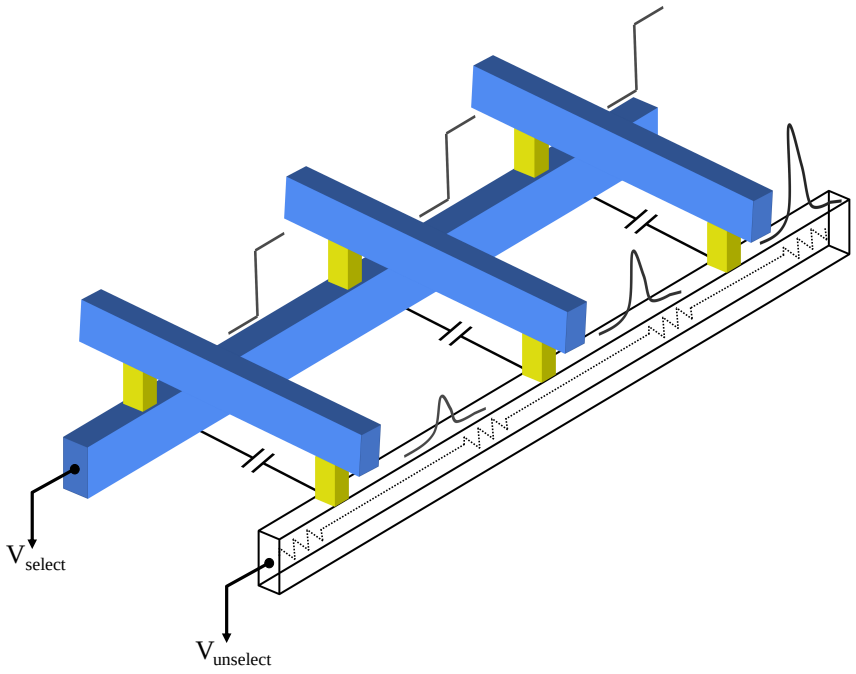

Figure 10.1: Noise coupling from a selected row to an unselected adjacent row.

driver on the unselected row). This effect is due to the increasing resistance as the physical distance from the driver increases [18].

If the amplitude of the noise spike is greater than the threshold voltage of a resistive cell (assuming 1R or 1S1R resistive cells), the resistance states begin to drift. To switch the cell from a high resistance state to a low resistance state, the amplitude of the noise spike has to be at least as wide as the switching latency of the memory cell. While the noise amplitude is typically smaller than the switching latency, consecutive access of a selected row, producing frequent spikes, can significantly drift the resistance state of the unselected cells on adjacent, unselected rows. Degradation of the read margin or even unintended bit flipping (i.e., the RowHammer effect) can

be produced. This issue is further exacerbated in multi-bit resistive memory (e.g., RRAM and PCM) where multiple resistance states are leveraged to enable multi-bit memory cells. Since the noise margin is lower in a multi-bit resistive cell, capacitive coupling becomes a more critical reliability concern.

10.3 Proposed Research Direction

To address the RowHammer effect in resistive arrays to improve memory reliability, two tasks requiring further research are described in this section. First, a set of guidelines is necessary to evaluate the coupling noise between selected and unselected rows. The number of consecutive read and write operations to flip the worst case cells on an adjacent row or column (e.g., the cells farther from the source) needs to be modeled in terms of the coupling capacitance and parasitic resistance, device threshold (either voltage or current), and array size. These models are needed to characterize the tradeoff between array density and noise immunity. Furthermore, multiple bias schemes (e.g., $V/2$, $V/3$, and floating bias schemes) need to be separately considered to determine the most and least reliable bias schemes in terms of noise immunity.

In addition, the magnitude of the access voltage needs to be evaluated to ensure low noise coupling. The magnitude of the read voltage is typically set high enough to maintain sufficient read margin but low enough to prevent cell disturbances. Similarly, the write voltage is set sufficiently high to ensure a low write latency but sufficiently

low to prevent any cell disturbances. In the context of the RowHammer effect, the write and read voltages need to be re-evaluated to ensure the cells on the adjacent unselected lines are also not disturbed.

The second research task aims to devise techniques to reduce the coupling noise or improve the noise immunity. The most commonly used solution in DRAM systems to mitigate the RowHammer effect is to increase the refresh rate at a cost of higher power consumption [173]. In resistive memories however the memory cells are not refreshed since most resistive memory technologies are nonvolatile. Imposing a refresh routine on a nonvolatile resistive memory is infeasible since the write energy is typically high (particularly for RRAM and PCM) and the refresh latency would impose significant performance overhead. Alternatively, increasing the interconnect pitch to reduce the coupling capacitance increases the cell area beyond $4F^2$, greatly increasing the cost per bit. A potential solution to reduce this overhead is to recognize those cells most sensitive to resistance drift. Those cells closest to the source on the unselected lines are more noise immune since the parasitic resistance between a cell and driver is smaller. Those cells farther from the source are therefore exposed to greater noise spikes and resistance drift. By only targeting those cells beyond a critical distance from the driver, the overhead of the refresh can be mitigated. These critical cells can be more frequently read to detect the level of resistance drift to decide whether or not to impose a refresh routine. Those cells exposed to large noise spikes on

unselected lines is however unclear and need to be studied to determine the energy and performance overhead to refresh the nonvolatile resistive cells. Coupling these two tasks of noise immunity characterization and mitigation is necessary to provide an energy efficient, low latency solution to the RowHammer effect.

10.4 Summary

Grid based structures in resistive memory arrays improve density but produce several issues such as the half-select problem, leakage currents, and sneak paths. In addition to these issues, the RowHammer effect in resistive memories is addressed here. The highly dense nature of a resistive array requires a short distance between the rows and columns, thereby degrading the reliability of the memory due to coupling noise between the selected and unselected lines. While this issue is well studied in DRAM systems [173–175], the effect on nonvolatile resistive memories remains unclear. To address this gap of knowledge, a possible research path is proposed. Specifically, two critical tasks need to be considered. The relationship between array density and size with respect to the number of unintentional bit flips needs to be characterized. The region within the array most vulnerable to large noise spikes also needs to be determined. Second, based on guidelines developed from the first task, design solutions that reduce the overhead of conventional solutions (e.g., refreshing all of the rows or increasing the metal pitch) need to be investigated. The insights

produced from these tasks will greatly improve the reliability of highly dense, grid structured nonvolatile resistive memory arrays.

Bibliography

[1] J. D. Cressler, *Silicon Earth: Introduction to the Microelectronics and Nanotechnology Revolution*, Cambridge University Press, 2009.

[2] D. Silver *et al.*, "Mastering the Game of Go with Deep Neural Networks and Tree Search," *Nature*, Vol. 529, No. 7587, pp. 484–489, January 2016.

[3] F. H. Hsu, "IBM's Deep Blue Chess Grandmaster Chips," *IEEE Micro*, Vol. 19, No. 2, pp. 70 – 81, March 1999.

[4] International Technology Roadmap for Semiconductors, 2015.

[5] Statista, "Semiconductor Industry - Statistics and Facts," 2017. [Online]. Available: https://www.statista.com/topics/1182/semiconductors/.

[6] F. Faggin, "The Making of the First Microprocessor," *IEEE Solid-State Circuits Magazine*, Vol. 1, No. 1, pp. 8–21, February 2009.

[7] F. Faggin, "4004 Q&A," [Online]. Available: www.intel4004.com/qa4004.htm.

[8] F. Faggin, "The New Methodology for Random Logic Design Used in the 4004 and in All the Early Intel Microprocessors: The Silicon Gate Design Methodology," n.d. [Online]. Available: http://www.intel4004.com/mrld.htm.

[9] S. Borkar, "Design Challenges of Technology Scaling," *IEEE Micro*, Vol. 19, No. 4, pp. 23 – 29, July 1999.

[10] S. Borkar, "Designing Reliable Systems from Unreliable Components the Challenges of Transistor Variability and Degradation," *IEEE Micro*, Vol. 25, No. 6, pp. 10 – 16, November 2005.

[11] D. MacMillen *et al.*, "An Industrial View of Electronic Design Automation," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, Vol. 19, No. 12, pp. 1428–1448, December 2000.

[12] P. Gelsinger, D. Kirkpatrick, A. Kolodny, and G. Singer, "Coping with the Complexity of Microprocessor Design at Intel – A CAD History," *IEEE Solid-State Circuits Magazine*, Vol. 2, No. 3, pp. 1–10, Summer 2010.

[13] A. B. Kahng, "Machine Learning Applications in Physical Design: Recent Results and Directions," *Proceedings of the ACM International Symposium on Physical Design*, pp. 68–73, March 2018.

[14] A. B. Kahng, "Design Challenges at 65nm and Beyond," *Proceedings of the ACM/IEEE Design, Automation and Test in Europe*, pp. 1466–1467, April 2007.

[15] R. E. Bryant, "Limitations and Challenges of Computer-Aided Design Technology for CMOS VLSI," *Proceedings of the IEEE*, Vol. 89, No. 3, pp. 341–365, March 2001.

[16] J. T. Kong, "CAD for Nanometer Silicon Design Challenges and Success," *IEEE Transactions on Very Large Scale Integration Systems*, Vol. 12, No. 11, pp. 1132–1147, November 2004.

[17] A. S. Vincentelli, "The Tides of EDA," *IEEE Design and Test of Computers*, Vol. 20, No. 6, pp. 59–75, November 2003.

[18] E. Salman and E. G. Friedman, *High Performance Integrated Circuit Design*, McGraw Hill, 2012.

[19] M. Rostami, F. Koushanfar, and R. Karri, "A Primer on Hardware Security: Models, Methods, and Metrics," *Proceedings of the IEEE*, Vol. 102, No. 8, pp. 1283–1295, July 2014.

[20] F. Pompei, "Architecture Grids," 2018. [Online]. Available: http://www.francescapompei.it/.

[21] "Manhattan Grid," [Online]. Available: www.architectural-review.com/.

[22] "The New York Times Front Page," 2018. [Online]. Available: https://static01.nyt.com/images/2018/04/15/nytfrontpage/scan.pdf.

[23] J. M. Brockmann, *Grid Systems in Graphic Design: A Visual Communication Manual for Graphic Designers, Typographers and Three Dimensional Designers*, Niggli, 1996.

[24] D. Stanislawski, "The Origin and Spread of the Grid-Pattern Town," *Geographical Review*, Vol. 36, No. 1, pp. 105–120, January 1946.

[25] J. Chilton, *Space Grid Structures*, Architectural Press, 2000.

[26] Intel, "Intel 4004 – 45th Anniversary Project," 2011. [Online]. Available: http://www.4004.com/.

[27] C. Gonzalez *et al.*, "POWER9: A Processor Family Optimized for Cognitive Computing with 25Gb/s Accelerator Links and 16Gb/s PCIe Gen4," *Proceedings of the IEEE International Solid-State Circuits Conference*, pp. 50–51, February 2017.

[28] J. M. Rabaey, A. P. Chandrakasan, and B. Nikolic, *Digital Integrated Circuits*, Prentice Hall, 2002.

[29] H. Fleisher and L. I. Maissel, "An Introduction to Array Logic," *IBM Journal of Research and Development*, Vol. 19, No. 2, pp. 98–109, March 1975.

[30] R. A. Henle, I. T. Ho, G. A. Maley, and R. Waxman, "Structured Logic," *Proceedings of the Fall Joint Computer Conference*, pp. 61–67, November 1969.

[31] S. M. Kang and Y. Leblebici, *CMOS Digital Integrated Circuits Analysis and Design*, McGraw Hill, 2003.

[32] Y. Kambayashi, "Logic Design of Programmable Logic Arrays," *IEEE Transactions on Computers*, Vol. 28, No. 9, pp. 609–617, September 1979.

[33] N. A. Sherwani, *Algorithms for VLSI Physical Design Automation*, Kluwer Academic Publishers, 2002.

[34] D. Wentzlaff *et al.*, "On-Chip Interconnection Architecture of the Tile Processor," *IEEE Micro*, Vol. 27, No. 5, pp. 15 – 31, November 2007.

[35] S. L. Teig, "The X Architecture: Not your Father's Diagonal Wiring," *Proceedings of the ACM/IEEE International Workshop on System Level Interconnect Prediction*, pp. 33–37, April 2002.

[36] S. R. Vangal *et al.*, "An 80-Tile Sub-100-W TeraFLOPS Processor in 65-nm CMOS," *IEEE Journal of Solid-State Circuits*, Vol. 43, No. 1, pp. 29 – 41, January 2008.

[37] J. Balkind *et al.*, "OpenPiton: An Open Source Manycore Research Framework," *Proceedings of the ACM International Conference on Architectural Support for Programming Languages and Operating Systems*, pp. 217–232, March 2016.

[38] B. Bohnenstiehl *et al.*, "KiloCore: A 32-nm 1000-Processor Computational Array," *IEEE Journal of Solid-State Circuits*, Vol. 52, No. 4, pp. 891 – 902, February 2017.

[39] S. M. Tam *et al.*, "SkyLake-SP: A 14nm 28-Core Xeon Processor," *Proceedings of the IEEE International Solid-State Circuits Conference*, pp. 34–35, February 2018.

[40] D. Geer, "Chip Makers Turn to Multicore Processors," *Computer*, Vol. 38, No. 5, pp. 11 – 13, May 2005.

[41] D. W. Bailey and B. J. Benschneider, "Clocking Design and Analysis for a 600-MHz Alpha Microprocessor," *IEEE Journal of Solid-State Circuits*, Vol. 33, No. 11, pp. 1627–1633, November 1998.

[42] H. Parvez and H. Mehrez, *Application-Specific Mesh-Based Heterogeneous FPGA Architectures*, Springer, 2011.

[43] B. Jacob, S. W. Ng, and D. T. Wang, *Memory Systems: Cache, DRAM, Disk*, Elsevier, 2008.

[44] I. Vaisband, R. Jakushokas, M. Popovich, A. V. Mezhiba, S. Kose, and E. G. Friedman, *On-Chip Power Delivery and Management*, Springer, 2016.

[45] T. Bjerregaard and S. Mahadevan, "A Survey of Research and Practices of Network-on-Chip," *ACM Computing Surveys*, Vol. 38, No. 1, pp. 1–51, June 2006.

[46] L. T. Wang, Y. W. Chang, and K. T. Cheng, *Electronic Design Automation: Synthesis, Verification, and Test*, Elsevier, 2009.

[47] I. Koren and A. D. Singh, "Fault Tolerance in VLSI Circuits," *IEEE Computer*, Vol. 23, No. 7, pp. 73–83, July 1990.

[48] P. Gupta and A. B. Kahng, "Manufacturing-Aware Physical Design," *Proceedings of the IEEE/ACM International Conference on Computer-Aided Design*, pp. 681–687, November 2003.

[49] V. Rantala, T. Lehtonen, and J. Plosila, "Network on Chip Routing Algorithms," University of Turku, Turku, Finland, August 2006.

[50] G. Strang, *Introduction to Linear Algebra*, Wellesley Cambridge Press, 2016.

[51] D. B. West, *Introduction to Graph Theory*, Pearson Education, 2001.

[52] C. A. Mead and M. Rem, "Cost and Performance of VLSI Computing Structures," *IEEE Transactions on Electron Devices*, Vol. 26, No. 4, pp. 533–540, April 1979.

[53] S. Powell, E. Iodice, and E. G. Friedman, "An Automated, Low Power, High Speed Complementary PLA Design System for VLSI Applications," *Microelectronics Journal*, Vol. 15, No. 4, pp. 47–54, July 1984.

[54] R. Amann, B. Eschermann, and U. G. Baitinger, "PLA Based Finite State Machines using Johnson Counters as State Memories," *Proceedings of the IEEE International Conference on Computer Design: VLSI in Computers and Processors*, pp. 267–270, October 1988.

[55] G. D. Micheli, A. Sangiovanni-Vincentelli, and T. Villa, "Computer-Aided Synthesis of PLA-Based Finite State Machines," *Proceedings of the IEEE*, pp. 154–156, September 1983.

[56] S. Baranov, I. Levin, O. Keren, and M. Karpovsky, "Designing Fault Tolerant FSM by Nano-PLA," *Proceedings of the IEEE International On-Line Testing Symposium*, pp. 229–234, June 2009.

[57] A. Kaveh, *Optimal Analysis of Structures by Concepts of Symmetry and Regularity*, Springer, 2013.

[58] E. J. Fluhr *et al.*, "The 12-Core POWER8 Processor with 7.6 Tbs IO Bandwidth, Integrated Voltage Regulation, and Resonant Clocking," *IEEE Journal of Solid-State Circuits*, Vol. 50, No. 1, pp. 10–23, January 2015.

[59] C. Gonzalez *et al.*, "The 24-Core POWER9 Processor with Adaptive Clocking, 25-Gb/s Accelerator Links, and 16-Gb/s PCIe Gen4," *IEEE Journal of Solid-State Circuits*, Vol. 53, No. 1, pp. 91–101, January 2018.

[60] A. Varma *et al.*, "Power Management in the Intel Xeon E5 v3," *Proceedings of the ACM/IEEE International Symposium on Low Power Electronics and Design*, pp. 371–376, July 2015.

[61] C. Berry *et al.*, "IBM z14TM: 14nm Microprocessor for the Next-Generation Mainframe," *Proceedings of the IEEE International Solid-State Circuits Conference*, pp. 36–38, February 2018.

[62] C. L. Wey, "On Yield Consideration for the Design of Redundant Programmable Logic Arrays," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, Vol. 7, No. 4, pp. 528–535, April 1988.

[63] S. Y. K. Kuo and W. K. Fuchs, "Fault Diagnosis and Spare Allocation for Yield Enhancement in Large Reconfigurable PLAs," *IEEE Transactions on Computers*, Vol. 41, No. 2, pp. 221–226, February 1992.

[64] A. B. Kahng, J. Lienig, I. L. Markov, and J. Hu, *VLSI Physical Design: From Graph Partitioning to Timing Closure*, Springer, 2011.

[65] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms*, The MIT Press, 2009.

[66] F. O. Hadlock, "A Shortest Path Algorithm for Grid Graphs," *Networks*, Vol. 7, No. 4, pp. 323–334, December 1977.

[67] M. Hanan, "On Steiners Problem with Rectilinear Distance," *SIAM Journal on Applied Mathematics*, Vol. 14, No. 2, pp. 255–265, March 1966.

[68] H. S. P. Wong *et al.*, "Metal–Oxide RRAM," *Proceedings of the IEEE*, Vol. 100, No. 6, pp. 1951–1970, May 2012.

[69] H. S. P. Wong *et al.*, "Phase Change Memory," *Proceedings of the IEEE*, Vol. 98, No. 12, pp. 2201–2227, October 2010.

[70] J. G. Zhu, "Magnetoresistive Random Access Memory: The Path to Competitiveness and Scalability," *Proceedings of the IEEE*, Vol. 96, No. 11, pp. 1786–1798, December 2008.

[71] A. Chen, "A Review of Emerging Non-Volatile Memory (NVM) Technologies and Applications," *Solid-State Electronics*, Vol. 125, pp. 25–38, November 2016.

[72] S. Yu and P. Y. Chen, "Emerging Memory Technologies: Recent Trends and Prospects," *IEEE Solid-State Circuits Magazine*, Vol. 8, No. 2, pp. 43–56, June 2016.

[73] T. M. Maffitt *et al.*, "Design Considerations for MRAM," *IBM Journal of Research and Development*, Vol. 50, No. 1, pp. 25–39, January 2006.

[74] Y. Huai, "Spin-Transfer Torque MRAM (STT-MRAM): Challenges and Prospects," *AAPPS Bulletin*, Vol. 18, No. 6, pp. 33–40, December 2008.

[75] H. S. P. Wong *et al.*, "Stanford Memory Trends," 2017. [Online]. Available: https://nano.stanford.edu/stanford-memory-trends.

[76] Everspin Technologies, "DDR3 DRAM Compatible MRAM - Spin Torque Technology," 2016. [Online]. Available: https://www.everspin.com/ddr3-dram-compatible-mram-spin-torque-technology-0.

[77] Tech Insight, "Intel 3D XPoint Memory Die Removed from Intel OptaneTM PCM (Phase Change Memory)," 2016. [Online]. Available: http://www.techinsights.com/about-techinsights/overview/blog/intel-3D-xpoint-memory-die-removed-from-intel-optane-pcm/.

[78] T. Liu *et al.*, "A 130.7-mm2 2-Layer 32-Gb ReRAM Memory Device in 24-nm Technology," *IEEE Journal of Solid-State Circuits*, Vol. 49, No. 1, pp. 140–153, January 2014.

[79] D. Xiangyu, C. Xu, Y. Xie, and N. P. Jouppi, "NVSim: A Circuit-Level Performance, Energy, and Area Model for Emerging Nonvolatile Memory," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, Vol. 31, No. 7, pp. 994–1007, July 2012.

[80] C. Xu, X. Dong, N. P. Jouppi, and Y. Xie, "Design Implications of Memristor-Based RRAM Cross-Point Structures," *Proceedings of the IEEE Design, Automation and Test in Europe Conference and Exhibition*, pp. 1–6, March 2011.

[81] X. Cong *et al.*, "Overcoming the Challenges of Crossbar Resistive Memory Architectures," *Proceedings of the IEEE International Symposium on High Performance Computer Architecture*, pp. 476–488, February 2015.

[82] C. W. Stanley and S. S. Wong, "Compact One-Transistor-N-RRAM Array Architecture for Advanced CMOS Technology," *IEEE Journal of Solid-State Circuits*, Vol. 50, No. 5, pp. 1299 – 1309, May 2015.

[83] R. Aluguri and T. Y. Tseng, "Overview of Selector Devices for 3-D Stackable Cross Point RRAM Arrays," *IEEE Journal of the Electron Devices Society*, Vol. 4, No. 5, pp. 294–306, September 2016.

[84] B. Song, Q. Li, H. Liu, and H. Liu, "Exploration of Selector Characteristic Based on Electron Tunneling for RRAM Array Application," *IEICE Electronics Express*, Vol. 14, No. 17, pp. 1–8, August 2017.

[85] S. H. Jo, T. Kumar, S. Narayanan, and H. Nazarian, "Cross-Point Resistive RAM Based on Field-Assisted Superlinear Threshold Selector," *IEEE Transactions on Electron Devices*, Vol. 62, No. 11, pp. 3477–3481, November 2015.

[86] Vicor Corporation, "Power-on-Package: Enabling Higher Performance in Artificial Intelligence Applications," 2017. [Online]. Available: http://powerblog.vicorpower.com/2017/08/power-package-higher-performance-artificial-intelligence/.

[87] A. J. D'Souza *et al.*, "A Fully Integrated Power-Management Solution for a 65nm CMOS Cellular Handset Chip," *Proceedings of the IEEE International Solid-State Circuits Conference*, pp. 382–384, February 2011.

[88] E. A. Burton *et al.*, "FIVR–Fully Integrated Voltage Regulators on 4th Generation Intel Core SoCs," *Proceedings of the IEEE Applied Power Electronics Conference and Exposition*, pp. 432–439, March 2014.

[89] J. F. Bulzacchelli *et al.*, "Dual-Loop System of Distributed Microregulators with High DC Accuracy, Load Response Time Below 500 ps, and 85-mV Dropout Voltage," *IEEE Journal of Solid-State Circuits*, Vol. 47, No. 4, pp. 863 – 874, April 2012.

[90] Z. Toprak-Deniz *et al.*, "Distributed System of Digitally Controlled Microregulators Enabling Per-Core DVFS for the POWER8TM Microprocessor," *Proceedings of the IEEE International Solid-State Circuits Conference*, pp. 98–99, February 2014.

[91] H. P. Le, S. R. Sanders, and E. Alon, "Design Techniques for Fully Integrated Switched-Capacitor DC-DC Converters," *IEEE Journal of Solid-State Circuits*, Vol. 46, No. 9, pp. 2120 – 2131, September 2011.

[92] H. K. Krishnamurthy *et al.*, "A Digitally Controlled Fully Integrated Voltage Regulator with On-Die Solenoid Inductor with Planar Magnetic Core in 14-nm Tri-Gate CMOS," *IEEE Journal of Solid-State Circuits*, Vol. 53, No. 1, pp. 8 – 19, January 2018.

[93] D. El-Damak, S. Bandyopadhyay, and A. P. Chandrakasan, "A 93% Efficiency Reconfigurable Switched-Capacitor DC-DC Converter using On-Chip Ferroelectric Capacitors," *Proceedings of the IEEE International Solid-State Circuits Conference*, pp. 374–375, February 2013.

[94] J. Wibben and R. Harjani, "A High-Efficiency DC–DC Converter using 2 nH Integrated Inductors," *IEEE Journal of Solid-State Circuits*, Vol. 43, No. 4, pp. 844 – 854, April 2008.

[95] I. Vaisband and E. G. Friedman, "Heterogeneous Methodology for Energy Efficient Distribution of On-Chip Power Supplies," *IEEE Transactions on Power Electronics*, Vol. 28, No. 9, pp. 4267–4280, September 2013.

[96] R. Muthukaruppan *et al.*, "A Digitally Controlled Linear Regulator for Per-Core Wide-Range DVFS of Atom Cores in 14nm Tri-Gate CMOS Featuring Non-Linear Control, Adaptive Gain and Code Roaming," *Proceedings of the IEEE European Solid State Circuits Conference*, pp. 275–278, September 2017.

[97] G. A. Rincon-Mora, *Current Efficient, Low Voltage, Low Dropout Regulators*, Ph.D. Thesis, Georgia Institute of Technology, 1996.

[98] J. Torres *et al.*, "Low Drop-Out Voltage Regulators: Capacitor-less Architecture Comparison," *IEEE Circuits and Systems Magazine*, Vol. 14, No. 2, pp. 6–26, May 2014.

[99] R. J. Milliken, S. Martinez, and E. Sanchez-Sinencio, "Full On-Chip CMOS Low-Dropout Voltage Regulator," *IEEE Transactions on Circuits and Systems I: Regular Papers*, Vol. 54, No. 9, pp. 1879–1890, September 2007.

[100] P. R. Gray, P. J. Hurst, S. H. Lewis, and R. G. Meyer, *Analysis and Design of Analog Integrated Circuits*, Wiley, 2009.

[101] T. C. Carusone, D. A. Johns, and K. W. Martin, *Analog Integrated Circuit Design*, Wiley, 2012.

[102] "Technical Review of Low Dropout Voltage Regulator Operation and Performance: Application Report," Texas Instruments, Dallas, Texas, August 1999.

[103] E. N. Y. Ho and P. K. T. Mok, "A Capacitor-Less CMOS Active Feedback Low-Dropout Regulator with Slew-Rate Enhancement for Portable On-Chip Application," *IEEE Transactions on Circuits and Systems II: Express Briefs*, Vol. 57, No. 2, pp. 80–84, February 2010.

[104] I. Vaisband, B. Price, S. Kose, Y. Kolla, E. G. Friedman, and J. Fischer, "Distributed LDO Regulators in a 28 nm Power Delivery System," *Analog Integrated Circuits and Signal Processing*, Vol. 83, No. 3, pp. 295–309, June 2015.

[105] K. Luria, J. Shor, M. Zelikson, and A. Lyakhov, "Dual-Mode Low-Drop-Out Regulator/Power Gate with Linear and OnOff Conduction for Microprocessor Core On-Die Supply Voltages in 14 nm," *IEEE Journal of Solid-State Circuits*, Vol. 51, No. 3, pp. 752–762, March 2016.

[106] S. Kose, S. Tam, S. Pinzon, B. McDermott, and E. G. Friedman, "Active Filter-Based Hybrid On-Chip DC-DC Converter for Point-of-Load Voltage Regulation," *IEEE Transactions on Very Large Scale Integration Systems*, Vol. 21, No. 4, pp. 680–691, April 2013.

[107] J. Guo and K. N. Leung, "A 6-uW Chip-Area-Efficient Output-Capacitorless LDO in 90-nm CMOS Technology," *IEEE Journal of Solid-State Circuits*, Vol. 45, No. 9, pp. 1896–1905, September 2010.

[108] P. Hazucha *et al.*, "Area-Efficient Linear Regulator with Ultra-Fast Load Regulation," *IEEE Journal of Solid-State Circuits*, Vol. 40, No. 4, pp. 933–940, April 2005.

[109] C. A. David and B. Feldman, "High-Speed Fixed Memories using Large-Scale Integrated Resistor Matrices," *IEEE Transactions on Computers*, Vol. C-17, No. 8, pp. 721–728, August 1968.

[110] W. T. Lynch, "Worst-Case Analysis of a Resistor Memory Matrix," *IEEE Transactions on Computers*, Vol. C-18, No. 10, pp. 940–942, October 1969.

[111] D. B. Strukov, G. S. Snider, D. R. Stewart, and R. S. Williams, "The Missing Memristor Found," *Nature*, Vol. 453, No. 7191, pp. 80–83, May 2008.

[112] A. Flocke and T. G. Noll, "Fundamental Analysis of Resistive Nano-Crossbars for the Use in Hybrid Nano/CMOS-Memory," *Proceedings of the IEEE Solid State Circuits Conference*, pp. 328–331, September 2007.

[113] P. O. Vontobel *et al.*, "Writing to and Reading From a Nano-Scale Crossbar Memory Based on Memristors," *Nanotechnology*, Vol. 20, No. 42, pp. 425204, September 2009.

[114] J. Liang and H. S. P. Wong, "Cross-Point Memory Array Without Cell Selectors Device Characteristics and Data Storage Pattern Dependencies," *IEEE Transactions on Electron Devices*, Vol. 57, No. 10, pp. 2531–2538, October 2010.

[115] P. Y. Chen and S. Yu, "Impact of Vertical RRAM Device Characteristics on 3D Cross-Point Array Design," *Proceedings of the IEEE International Memory Workshop*, pp. 1 – 4, May 2014.

[116] A. Chen, "A Comprehensive Crossbar Array Model with Solutions for Line Resistance and Nonlinear Device Characteristics," *IEEE Transactions on Electron Devices*, Vol. 60, No. 4, pp. 1318–1326, April 2013.

[117] A. Ciprut and E. G. Friedman, "Design Models of Resistive Crossbar Arrays with Selector Devices," *Proceedings of the IEEE International Symposium on Circuits and Systems*, pp. 1250 – 1253, May 2016.

[118] Y. C. Chen *et al.*, "An Access-Transistor-Free (0T/1R) Non-Volatile Resistance Random Access Memory (RRAM) using a Novel Threshold Switching, Self-Rectifying Chalcogenide Device," *Proceedings of the IEEE International Electron Devices Meeting*, pp. 4–37, December 2003.

[119] J. Mustafa, *Design and Analysis of Future Memories Based on Switchable Resistive Elements*, Ph.D. Thesis, RWTH Aachen University, Aachen, Germany, 2006.

[120] International Technology Roadmap for Semiconductors, 2007.

[121] J. J. Huang *et al.*, "One Selector-One Resistor (1S1R) Crossbar Array for High-Density Flexible Memory Applications," *Proceedings of the IEEE International Electron Devices Meeting*, pp. 7–31, December 2011.

[122] W. Lee *et al.*, "High Current Density and Nonlinearity Combination of Selection Device Based on TaOx/TiO2/TaOx Structure for One Selector One Resistor Arrays," *ACS Nano*, Vol. 6, No. 9, pp. 8166–8172, August 2012.

[123] J. J. Huang, Y. M. Tseng, C. W. Hsu, and T. H. Hou, "Bipolar Nonlinear Ni/TiO2/Ni Selector for 1S1R Crossbar Array Applications," *IEEE Electron Device Letters*, Vol. 32, No. 10, pp. 1427–1429, October 2011.

[124] Q. Luo *et al.*, "Cu BEOL Compatible Selector with High Selectivity (10^7), Extremely Low Off-Current (pA) and High Endurance (10^{10})," *Proceedings of the IEEE International Electron Devices Meeting*, pp. 10.4.1 – 10.4.4, December 2015.

[125] L. Zhang *et al.*, "One-Selector One-Resistor Cross-Point Array with Threshold Switching Selector," *IEEE Transactions on Electron Devices*, Vol. 62, No. 10, pp. 3250 – 3257, October 2015.

[126] S. Mandegaran and A. Hajimiri, "A Breakdown Voltage Multiplier for High Voltage Swing Drivers," *IEEE Journal of Solid-State Circuits*, Vol. 42, No. 2, pp. 302 – 312, February 2007.

[127] S. Kvatinsky, G. Satat, N. Wald, E. G. Friedman, A. Kolodny, and U. C. Weiser, "Memristor-Based Material Implication (IMPLY) Logic: Design Principles and Methodologies," *IEEE Transactions on Very Large Scale Integration Systems*, Vol. 22, No. 10, pp. 2054–2066, October 2014.

[128] T. Na, B. Song, J. P. Kim, S. H. Kang, and S. O. Jung, "Offset-Canceling Current-Sampling Sense Amplifier for Resistive Nonvolatile Memory in 65 nm CMOS," *IEEE Journal of Solid-State Circuits*, Vol. 52, No. 2, pp. 496–504, February 2017.

[129] C. Yong *et al.*, "Nanoscale Molecular-Switch Crossbar Circuits," *Nanotechnology*, Vol. 14, No. 4, pp. 462–468, March 2003.

[130] S. Yu, *Resistive Random Access Memory (RRAM) From Devices to Array Architectures*, Morgan & Claypool, 2016.

[131] A. Ciprut and E. G. Friedman, "On the Write Energy of Non-Volatile Resistive Crossbar Arrays with Selectors," *Proceedings of the IEEE International Symposium on Quality Electronic Design*, pp. 184–188, March 2018.

[132] A. Ciprut and E. G. Friedman, "Modeling Size Limitations of Resistive Crossbar Array with Cell Selectors," *IEEE Transactions on Very Large Scale Integration Systems*, Vol. 25, No. 1, pp. 286–293, January 2017.

[133] Q. Luo *et al.*, "Demonstration of 3D Vertical RRAM with Ultra Low-Leakage, High-Selectivity and Self-Compliance Memory Cells," *Proceedings of the IEEE International Electron Device Meeting*, pp. 10.2.1–10.2.4, December 2015.

[134] M. Wang, J. Zhou, Y. Yang, S. Gaba, M. Liu, and W. D. Lu, "Conduction Mechanism of a TaOx-Based Selector and its Application in Crossbar Memory Arrays," *Nanoscale*, Vol. 7, No. 11, pp. 4964–4970, February 2015.

[135] B. J. Choi *et al.*, "Trilayer Tunnel Selectors for Memristor Memory Cells," *Advanced Materials*, Vol. 28, No. 2, pp. 356–362, January 2016.

[136] C. Y. Lin *et al.*, "Attaining Resistive Switching Characteristics and Selector Properties by Varying Forming Polarities in a Single HfO2-Based RRAM Device with a Vanadium Electrode," *Nanoscale*, Vol. 9, pp. 8586–8590, May 2017.

[137] Q. Luo *et al.*, "Super Non-Linear RRAM with Ultra-Low Power for 3D Vertical Nano-Crossbar Arrays," *Nanoscale*, Vol. 8, pp. 15629–15636, July 2016.

[138] R. Midya *et al.*, "Anatomy of Ag/Hafnia-Based Selectors with 10^{10} Nonlinearity," *Advanced Materials*, Vol. 29, No. 12, pp. 1–8, January 2017.

[139] S. Kvatinsky, M. Ramadan, E. G. Friedman, and A. Kolodny, "VTEAM: A General Model for Voltage-Controlled Memristors," *IEEE Transactions on Circuits and Systems II: Express Briefs*, Vol. 62, No. 8, pp. 786–790, August 2015.

[140] H. Farkhani, M. Tohidi, A. Peiravi, J. K. Madsen, and F. Moradi, "STT-RAM Energy Reduction using Self-Referenced Differential Write Termination Technique," *IEEE Transactions on Very Large Scale Integration Systems*, Vol. 25, No. 2, pp. 476–487, February 2017.

[141] H. L. Lung, "Method, Apparatus and Computer Program Product for Read Before Programming Process on Multiple Programmable Resistive Memory Cell," U.S. Patent No. 7,433,226, October 7, 2008.

[142] T. Ishii, S. Ning, M. Tanaka, K. Tsurumi, and K. Takeuchi, "Adaptive Comparator Bias-Current Control of 0.6 V Input Boost Converter for ReRAM Program Voltages in Low Power Embedded Applications," *IEEE Journal of Solid-State Circuits*, Vol. 51, No. 10, pp. 2389–2397, October 2016.

[143] G. Villar-Pique, H. J. Bergveld, and E. Alarcon, "Survey and Benchmark of Fully Integrated Switching Power Converters: Switched-Capacitor Versus Inductive Approach," *IEEE Transactions on Very Large Scale Integration Systems*, Vol. 28, No. 9, pp. 4156–4167, September 2013.

[144] S. Kose and E. G. Friedman, "Distributed On-Chip Power Delivery," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, Vol. 2, No. 4, pp. 704–713, December 2012.

[145] Y. C. Wu, C. Y. Huang, and B. D. Liu, "A Low Dropout Voltage Regulator with Programmable Output," *Proceedings of the IEEE Conference on Industrial Electronics and Applications*, pp. 3357–3361, June 2009.

[146] V. Srinivasan, G. Serrano, C. M. Twigg, and P. Hasler, "A Floating-Gate-Based Programmable CMOS Reference," *IEEE Transactions on Circuits and Systems I: Regular Papers*, Vol. 55, No. 11, pp. 3448–3456, December 2008.

[147] W. Kim, M. S. Gupta, and D. Brooks, "System Level Analysis of Fast, Per-Core DVFS using On-Chip Switching Regulators," *Proceedings of the IEEE International Symposium on High Performance Computer Architecture*, pp. 123–134, February 2008.

[148] H. Per *et al.*, "Haswell: The Fourth-Generation Intel Core Processor," *IEEE Micro*, Vol. 34, No. 2, pp. 6–20, March 2014.

[149] Y. H. Lee *et al.*, "A DVS Embedded Power Management for High Efficiency Integrated SoC in UWB System," *Proceedings of the IEEE Asian Solid-State Circuits Conference*, pp. 321–324, December 2009.

[150] T. Singh *et al.*, "Zen: An Energy-Efficient High-Performance 86 Core," *IEEE Journal of Solid-State Circuits*, Vol. 53, No. 1, pp. 102–114, January 2018.

[151] I. Vaisband and E. G. Friedman, "Stability of Distributed Power Delivery Systems with Multiple Parallel On-Chip LDO Regulators," *IEEE Transactions on Power Electronics*, Vol. 31, No. 8, pp. 5625–5634, August 2016.

[152] S. B. Nasir, Y. Lee, and A. Raychowdhury, "Modeling and Analysis of System Stability in a Distributed Power Delivery Network with Embedded Digital Linear Regulators," *Proceedings of the IEEE International Symposium on Quality Electronic Design*, pp. 68–75, March 2014.

[153] S. Lai, B. Yan, and P. Li, "Localized Stability Checking and Design of IC Power Delivery with Distributed Voltage Regulators," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, Vol. 32, No. 9, pp. 1321–1334, September 2013.

[154] J. Shor, "Low Noise Linear Voltage Regulator for Use as an On-Chip PLL Supply in Microprocessors," *Proceedings of the IEEE International Symposium on Circuits and Systems*, pp. 841–844, May 2010.

[155] L. Wang *et al.*, "Efficiency, Stability, and Reliability Implications of Unbalanced Current Sharing Among Distributed On-Chip Voltage Regulators," *IEEE Transactions on Very Large Scale Integration Systems*, Vol. 25, No. 11, pp. 3019–3032, November 2017.

[156] A. D. Grasso, G. Palumbo, and S. Pennisi, "Comparison of the Frequency Compensation Techniques for CMOS Two-Stage Miller OTAs," *IEEE Transactions on Circuits and Systems II: Express Briefs*, Vol. 55, No. 11, pp. 1099–1103, November 2008.

[157] X. Zhan, J. Riad, P. Li, and E. Sanchez, "Design Space Exploration of Distributed On-Chip Voltage Regulation Under Stability Constraint," *IEEE Transactions on Very Large Scale Integration Systems*, Vol. 26, No. 8, pp. 1580–1584, August 2018.

[158] A. Ciprut and E. G. Friedman, "On the Stability of Distributed On-Chip Low Dropout Regulators," *Proceedings of the IEEE Midwest Symposium on Circuits and Systems*, pp. 217–220, August 2017.

[159] S. Lai and P. Li, "A Fully On-Chip Area-Efficient CMOS Low-Dropout Regulator with Fast Load Regulation," *Analog Integrated Circuits and Signal Processing*, Vol. 72, No. 2, pp. 433–450, August 2012.

[160] Y. Lu *et al.*, "A Fully-Integrated Low-Dropout Regulator with Full-Spectrum Power Supply Rejection," *IEEE Transactions on Circuits and Systems I: Regular Papers*, Vol. 62, No. 3, pp. 707–716, March 2015.

[161] S. Bu, J. Guo, and K. N. Leung, "A 200-ps-Response-Time Output-Capacitorless Low-Dropout Regulator with Unity-Gain Bandwidth >100 MHz in 130-nm CMOS," *IEEE Journal of Solid-State Circuits*, Vol. 33, No. 4, pp. 3232–3246, April 2017.

[162] J. E. Colgate, *The Control of Dynamically Interacting Systems,*, Ph.D. Thesis, Massachusetts Institute of Technology, 1988.

[163] P. Triverio, S. Grivet-Talocia, M. S. Nakhla, F. G. Canavero, and R. Achar, "Stability, Causality, and Passivity in Electrical Interconnect Models," *IEEE Transactions on Advanced Packaging*, Vol. 30, No. 4, pp. 795–808, November 2007.

[164] M. S. Gupta *et al.*, "Understanding Voltage Variations in Chip Multiprocessors using a Distributed Power-Delivery Network," *Proceedings of the IEEE Design, Automation and Test in Europe Conference and Exhibition*, pp. 1–6, April 2007.

[165] "Intel Pentium 4 Processor in the 423 pin package / Intel 850 Chipset Platform Design Guide," Intel, Santa Clara, CA, February 2002.

[166] "Voltage Regulator Module (VRM) and Enterprise Voltage Regulator-Down (EVRD) 11.1," Intel, Santa Clara, CA, February 2009.

[167] M. Swinnen, "Designers Face Growing Problems with On-Chip Power Distribution," 2018. [Online]. Available: https://semiengineering.com/designers-face-growing-problems-with-on-chip-power-distribution/.

[168] A. S. Mutschler, "Variation in Low-Power FinFET Designs," 2018. [Online]. Available: https://semiengineering.com/variation-in-low-power-finfet-designs/.

[169] P. A. Meinerzhagen *et al.*, "An Energy-Efficient Graphics Processor in 14-nm Tri-Gate CMOS Featuring Integrated Voltage Regulators for Fine-Grain DVFS, Retentive Sleep, and VMIN Optimization," *IEEE Journal of Solid-State Circuits*, Vol. 54, No. 1, pp. 144–157, January 2019.

[170] Y. Okuma *et al.*, "0.5-V Input Digital LDO with 98.7% Current Efficiency and 2.7-A Quiescent Current in 65nm CMOS," *Proceedings of the IEEE Custom Integrated Circuits Conference*, pp. 1–4, November 2010.

[171] M. Huang, Y. Lu, S. W. Sin, S. P. U, and R. P. Martins, "A Fully Integrated Digital LDO with Coarse–Fine-Tuning and Burst-Mode Operation," *IEEE Transactions on Circuits and Systems II: Express Briefs*, Vol. 63, No. 7, pp. 683–687, February 2016.

[172] L. M. G. Gomes, "Power Reduction of a CMOS High-Speed Interface using Power Gating," M.S. Thesis, University of Porto, 2013.

[173] O. Mutlu, "The RowHammer Problem and Other Issues we may Face as Memory Becomes Denser," *Proceedings of the ACM Conference on Design, Automation and Test in Europe*, pp. 1116–1121, March 2017.

[174] Y. Kim *et al.*, "Flipping Bits in Memory Without Accessing Them: An Experimental Study of DRAM Disturbance Errors," *Proceedings of the ACM International International Symposium on Computer Architecuture*, pp. 361–372, June 2014.

[175] S. K. Gautam, A. Kumar, and S. K. Manhas, "Improvement of Row Hammering using Metal Nanoparticles in DRAM–a Simulation Study," *IEEE Electron Device Letters*, Vol. 39, No. 9, pp. 1286–1289, September 2018.

Appendix A

Derivation of Switching Energy Consumption

To estimate the switching energy of a resistive cell, the resistance is modeled as a linear function during the switching interval. The resistance during a set operation is

$$R(t) = R_{off} + \frac{t}{t_{set}}(R_{on} - R_{off}), \tag{A.1}$$

assuming the set operation is initiated between $t = 0$ and $t = t_{set}$. If the interconnect resistance is negligible, the voltage across the cell is equal to the write voltage V_{write}. The power consumption is

$$P_{set}(t) = \frac{V_{write}^2}{R(t)}. \tag{A.2}$$

Integrating (A.2) over the set period, the energy consumption is

$$E_{set} = \int_{t=0}^{t_{set}} P_{set}(t)dt = \frac{V_{write}^2}{R_{on} - R_{off}} ln(\frac{R_{on}}{R_{off}})t_{set}. \tag{A.3}$$

For a symmetric resistive cell where the set and reset voltages as well as switching times are equal, the set and reset energy consumption is also the same.

Appendix B

Off-chip Power Delivery Network

The off-chip power delivery network considered in Sections 7.1 and 7.4 is shown in Fig. B.1. The value of the individual impedances is listed in Table B.1. The value of the board and package impedances are from [165]. The lumped parasitic inductance and resistance connecting the package to the integrated circuit (e.g., using C4s) are assumed to be, respectively, 100 pH and 100 $\mu\Omega$.

Table B.1: Off-chip parasitic impedances

Parameters	**Value**	**Parameters**	**Value**
${L_{pcb}}^S$	21 pH	${C_{pcb}}^P$	240 μF
${R_{pcb}}^S$	94 $\mu\Omega$	${R_{pcb}}^P$	165.4 $\mu\Omega$
${L_{pckg}}^S$	120 pH	${L_{pcb}}^P$	33.92 nH
${R_{pckg}}^S$	1.1 mΩ	${C_{pckg}}^P$	26.4 μF
L_{C4}	100 pH	${R_{pckg}}^P$	541.5 $\mu\Omega$
R_{C4}	100 $\mu\Omega$	${L_{pckg}}^P$	4.61 μH

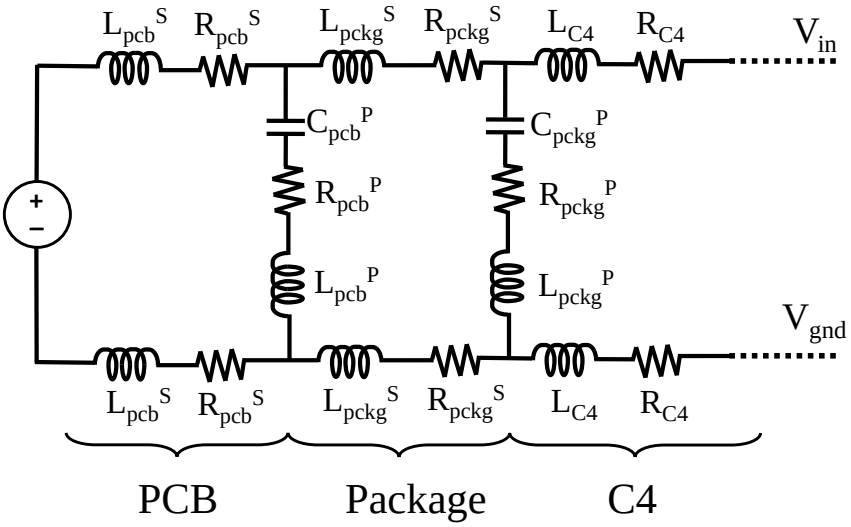

Figure B.1: Off-chip power delivery network model considered in Sections 7.1 and 7.4 [165].