VLSI-COMPATIBLE IMPLEMENTATIONS FOR ARTIFICIAL NEURAL NETWORKS

by

Sied Mehdi Fakhraie
Kenneth C. Smith

KLUWER ACADEMIC PUBLISHERS

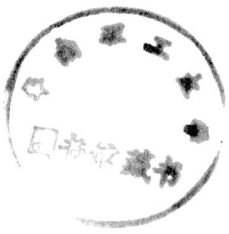
VLSI - COMPATIBLE IMPLEMENTATIONS FOR ARTIFICIAL NEURAL NETWORKS

by

Sied Mehdi Fakhraie
University of Tehran

Kenneth Carless Smith
University of Toronto
Hong Kong University of Science & Technology

KLUWER ACADEMIC PUBLISHERS
Boston / Dordrecht / London



Distributors for North America:
Kluwer Academic Publishers
101 Philip Drive
Assinippi Park
Norwell, Massachusetts 02061 USA

Distributors for all other countries:
Kluwer Academic Publishers Group
Distribution Centre
Post Office Box 322
3300 AH Dordrecht, THE NETHERLANDS

Library of Congress Cataloging-in-Publication Data

A C.I.P. Catalogue record for this book is available
from the Library of Congress.

Printed on acid-free paper.

Printed in the United States of America

VLSI - COMPATIBLE IMPLEMENTATIONS FOR ARTIFICIAL NEURAL NETWORKS

THE KLUWER INTERNATIONAL SERIES IN ENGINEERING AND COMPUTER SCIENCE

ANALOG CIRCUITS AND SIGNAL PROCESSING

Consulting Editor

Mohammed Ismail

Ohio State University

Related Titles:

CHARACTERIZATION METHODS FOR SUBMICRON MOSFETs, edited by Hisham Haddara
ISBN: 0-7923-9695-2

LOW-VOLTAGE LOW-POWER ANALOG INTEGRATED CIRCUITS, edited by Wouter Serdijn
ISBN: 0-7923-9608-1

INTEGRATED VIDEO-FREQUENCY CONTINUOUS-TIME FILTERS: ***High-Performance Realizations in BiCMOS***, Scott D. Willingham, Ken Martin
ISBN: 0-7923-9595-6

FEED-FORWARD NEURAL NETWORKS: ***Vector Decomposition Analysis, Modelling and Analog Implementation,*** Anne-Johan Annema
ISBN: 0-7923-9567-0

FREQUENCY COMPENSATION TECHNIQUES LOW-POWER OPERATIONAL AMPLIFIERS, Ruud Easchauzier, Johan Huijsing
ISBN: 0-7923-9565-4

ANALOG SIGNAL GENERATION FOR BIST OF MIXED-SIGNAL INTEGRATED CIRCUITS, Gordon W. Roberts, Albert K. Lu
ISBN: 0-7923-9564-6

INTEGRATED FIBER-OPTIC RECEIVERS, Aaron Buchwald, Kenneth W. Martin
ISBN: 0-7923-9549-2

MODELING WITH AN ANALOG HARDWARE DESCRIPTION LANGUAGE, H. Alan Mantooth,Mike Fiegenbaum
ISBN: 0-7923-9516-6

LOW-VOLTAGE CMOS OPERATIONAL AMPLIFIERS: ***Theory, Design and Implementation***, Satoshi Sakurai, Mohammed Ismail
ISBN: 0-7923-9507-7

ANALYSIS AND SYNTHESIS OF MOS TRANSLINEAR CIRCUITS, Remco J. Wiegerink
ISBN: 0-7923-9390-2

COMPUTER-AIDED DESIGN OF ANALOG CIRCUITS AND SYSTEMS, L. Richard Carley, Ronald S. Gyurcsik
ISBN: 0-7923-9351-1

HIGH-PERFORMANCE CMOS CONTINUOUS-TIME FILTERS, José Silva-Martínez, Michiel Steyaert, Willy Sansen
ISBN: 0-7923-9339-2

SYMBOLIC ANALYSIS OF ANALOG CIRCUITS: Techniques and Applications, Lawrence P. Huelsman, Georges G. E. Gielen
ISBN: 0-7923-9324-4

DESIGN OF LOW-VOLTAGE BIPOLAR OPERATIONAL AMPLIFIERS, M. Jeroen Fonderie, Johan H. Huijsing
ISBN: 0-7923-9317-1

STATISTICAL MODELING FOR COMPUTER-AIDED DESIGN OF MOS VLSI CIRCUITS, Christopher Michael, Mohammed Ismail
ISBN: 0-7923-9299-X

SELECTIVE LINEAR-PHASE SWITCHED-CAPACITOR AND DIGITAL FILTERS, Hussein Baher
ISBN: 0-7923-9298-1

ANALOG CMOS FILTERS FOR VERY HIGH FREQUENCIES, Bram Nauta
ISBN: 0-7923-9272-8

ANALOG VLSI NEURAL NETWORKS, Yoshiyasu Takefuji
ISBN: 0-7923-9273-6

ANALOG VLSI IMPLEMENTATION OF NEURAL NETWORKS, Carver A. Mead, Mohammed Ismail
ISBN: 0-7923-9049-7

AN INTRODUCTION TO ANALOG VLSI DESIGN AUTOMATION, Mohammed Ismail, José Franca
ISBN: 0-7923-9071-7

To our families, our teachers, and our students.

S. Mehdi Fakhraie

K. C. Smith

Foreword

This book introduces several state-of-the-art VLSI implementations of artificial neural networks (ANNs). It reviews various hardware approaches to ANN implementations: analog, digital and pulse-coded. The analog approach is emphasized as the main one taken in the later chapters of the book.

The area of VLSI implementation of ANNs has been progressing for the last 15 years, but not at the fast pace originally predicted. Several reasons have contributed to the slow progress, with the main one being that VLSI implementation of ANNs is an interdisciplinary area where only a few researchers, academics and graduate students are willing to venture. The work of Professors Fakhraie and Smith, presented in this book, is a welcome addition to the state-of-the-art and will greatly benefit researchers and students working in this area. Of particular value is the use of experimental results to backup extensive simulations and in-depth modeling. The introduction of a synapse-MOS device is novel. The book applies the concept to a number of applications and guides the reader through more possible applications for future work.

I am confident that the book will benefit a potentially wide readership.

M. I. Elmasry

University of Waterloo

Waterloo, Ontario

Canada

Preface

Neural Networks (NNs), generally defined as parallel networks that employ a large number of simple processing elements to perform computation in a distributed fashion, have attracted a lot of attention in the past fifty years. As the result, many new discoveries have been made. For example, while conventional serial computational techniques are reaching intrinsic physical limits of speed and performance, parallel neural-computation techniques introduce a new horizon directing humans towards the era of tera-computation: The inherent parallelism of a neural-network solution is one of the features that has attracted most attention. In addition, the learning capability embedded in a neural solution provides techniques by which to adapt potentially low-cost low-precision hardware in ways which are of great interest on the implementation side.

Although the majority of advancements in neuro-computation have resulted from theoretical analysis, or by simulation of parallel networks on serial computers, many of the potential advantages of neural networks await effective hardware implementations. Fortuitously, rapid advancement of VLSI technology has made many of the previously-impossible ideas now quite feasible.

The first transistor was invented nearly fifty years ago; yet it took more than a decade until early integrated circuits began to appear. Since then, the dimensions of a minimum-size device on an integrated circuit chip have shrunk dramatically. Nowadays, to have a billion transistors on a chip seems quite possible. However, such transistors have their own limitations which impose special conditions on their effective use.

The inherent parallelism of neural networks and their trainable-hardware implementation provide a natural means for employment of VLSI technologies. Analog and digital storage techniques are conveniently available in VLSI circuits. Also implementation of addition, multiplication, division, exponentiation and threshold operations have proved to be possible. As well, through advances in VLSI technology, multilayer metal and polysilicon connecting lines have eased the communication problems inherent in any implementation of artificial neural networks (ANNs). Thus the VLSI environment naturally suits neural-network implementation. Moreover, fortuitously, tolerances, mismatches, noise, and other hardware imperfections in VLSI can be best accommodated in the adaptive-training process which ANNs naturally incorporate.

Thus, it appears that these two landmark technologies, ANNs and VLSI, rapidly emerging in the final years of the second millennium, must be united for mutual benefit: one provides a seemingly never-satisfiable demand for more and more processing elements organizable to deal with real-world information-processing problems; the other delivers an apparently ever-increasing number of resources on a

chip. One provides a reasonable performance with a great degree of tolerance to the operation of any single device; the other can best take advantage of this property to increase the overall acceptable yield of working systems, despite the increasing level of imperfections which accompanies the ever-shrinking dimensions of a single device on a chip of potentially ever-increasing area. Locality of individual operations with global communication of information, and possible modularity of system-level designs, in combination with unsupervised training algorithms, are among the many interesting features that encourage the expectation of a better future available through the combination of these fields.

However, one annoying problem, which shows itself every now and then, is that although the marriage of VLSI implementation and neural-information-processing techniques seems natural and inevitable, one which is publicly encouraged and even announced as attractive by most researchers in these fields, their union is not without difficulty. In fact, at some level of detail, it may be quite unnatural! The problem is that the basic premises on which each works are quite different.

VLSI technology has its own way of expressing relations and connections. In reality, a VLSI layout is composed of different layers of metals, other conductors, various insulators, semiconductors, and so on. When several of these basic elements are combined, one of several basic electrical elements is formed (in particular, resistors, capacitors, inductors, diodes, and transistors), each represented by a small number of characteristic equations whose role and validity are generally understood and not simply proprietary knowledge of the layout engineer. Designers use these abstract representations to realize their circuits as larger abstractions, which, after several levels of translation and compilation, are finally mapped onto a piece of silicon or other electronic medium.

Correspondingly, but differently, neural networks have their own block structure for structuring, expressing, and processing the facts they embody. Their structure is normally quite simple, repeatable, and usually complete with some appropriate analytical interpretation. Each processing element employs the basic characteristics of its building blocks, whose origin, naturally enough, is in the modelling of real biological systems. Logically enough, the simplicity of these building blocks has a dramatic impact on the feasibility of the construction of a computational medium composed of potentially billions of basic elements. Ironically, while the simplicity of neural characteristics is well-adapted to biological implementations, it is not so suited to silicon ones!

The disturbing fact is that the normal characteristic equation of a processing element (a neuron), or of an interconnection (a synapse) in an artificial neural network, is not at all similar to the characteristic equation or physical behavior of any basic building element easily available in any current VLSI technology. This is in sharp contrast to the structure of biological neural networks where a startling natural harmony exists between the resources available in biological implementation media

(including liquids, ions, membranes,...), and the abstract system-level behavioral models describing their operation.

Following the lead represented by this observation, it appears logical that instead of continuing to work on pure biologically-inspired models, one should concentrate on another, more global, aspect of biological neural networks: the fact that they are parallel interconnected networks of simple processing and interconnecting elements that naturally employ existing resources available in their implementation media. Therefore, one can conclude that the simplicity of the processing elements, and the way their natural and intrinsic characteristic equations are employed, are important elements in attempting to make the most out of what is readily available.

These biologically-inspired observations encouraged us to view the VLSI implementation of artificial neural networks in another light: It was first to understand the optimality principles governing the development of biological neural networks; Then, second, with these principles as a guideline, it was to consider the resources available in our VLSI implementation medium, with a view to designing distributed parallel networks in a way that makes the most of them.

Correspondingly, in the presentation to follow, the major inspiration to be derived from biological neural networks will be the simple idea of employing parallel networks composed of distributed simple processing elements. However, rather than on biologically-inspired models and equations, the reader will find the emphasis to be placed on the intrinsic characteristics of the electronics building blocks available in the CMOS-VLSI technology being used.

Through this approach, we believe that an effective union of the VLSI and neural network fields can be achieved in which natural simplicity is emphasized. Correspondingly, rather than a direct synapse equivalent being implemented as a system composed of dozens of transistors, as has been done conventionally, in our work, the intrinsic operating equation of a simple MOS transistor will be employed as the basic synaptic operation.

In this book, we introduce the basic premise of our approach to biologically-inspired and VLSI-compatible definition, simulation, and implementation of artificial neural networks. As well, we develop a set of guidelines for general hardware implementation of ANNs. These guidelines are then used to find solutions for the usual difficulties encountered in any potential work, and as guidelines by which to reach the best compromise when several options exist. As well, system-level consequences of using the proposed techniques in future submicron technologies with almost-linear MOS devices are discussed.

While the major emphasis in this book is on our desire to develop neural networks optimized for compatibility with their implementation media, we have also extended this work to the design and implementation of a fully-quadratic ANN based

on the desire to have network definitions optimized for both efficient discrimination of closed-boundary circular areas and ease of implementation in a CMOS technology.

Overall, this book implements a comprehensive approach which starts with an analytical evaluation of specific artificial networks. This provides a clear geometrical interpretation of the behavior of different variants of these networks. In combination with the guidelines developed towards a better final implementation, these concepts have allowed us to conquer various problems encountered and to make effective compromises. Then, to facilitate the investigation of the models needed when more difficult problems must be faced, a custom simulating program for various cases is developed. Finally, in order to demonstrate our findings and expectations, several VLSI integrated circuits have been designed, fabricated, and tested. While these results demonstrate the feasibility of our approach, we emphasize that they merely show a direction in which to go, rather than the realization of the ultimate destination!

Finally, it is our hope that while providing our readership with the results of advanced ongoing research, they will also find here the outlines of a comprehensive framework within which they can develop, examine, and firmly analyze their own innovative neural-network models and ideas. Further, it is our hope that, in this respect, theoretical neural-network researchers and software developers might also find the proposed methodology quite useful for their own purposes.

Organization of the Book

After discussing the underlying motivation and objectives in Chapter 1, we will review various existing hardware-implementation techniques in Chapter 2. In Chapter 3, we present the general model we have used in developing our neural networks together with the idea of employing MOS-transistor-like processing elements directly in a neural network. Our simulation program and the test problems used in developing and verifying the ideas presented in the book are discussed as well. In Chapter 4, the foundation elements for the architectural design are described. To begin, conventional linear- and quadratic-synapse networks are introduced together with simple geometrical interpretations of their operation. A detailed analysis of the operation of our proposed single-transistor-based network follows, and related problems and potentials are examined. Then, architectures by which to take full advantage of the proposed synaptic elements, and to solve their associated problems, are described. Next, the performances of different architectures are compared based on extended simulations, and the effectiveness of our new direction is shown. In Chapter 5, we highlight our expectations for a low-level silicon device, one which we call a Synapse-MOS or SyMOS device, which can be employed in a range of networks. Then, our present approach to implementing such a SyMOS device in a standard double-polysilicon CMOS process, is described. Experimental results based on fabricated chips completes Chapter 5. In Chapter 6, further details of the VLSI implementation of proposed Synapse-MOS Artificial Neural Networks (SANNs) are

discussed, and experimental results are reported. In Chapter 7, we describe a second approach, the parallel development of a novel fully-quadratic analog neural network, for which we show circuit-design detail and the results of VLSI implementation. As well, the applied advantages of this type of network in unsupervised-competitive-learning and function-approximation problems are discussed. Finally, Chapter 8 concludes this book by providing an overall view and summary, together with the introduction of directions for possible future work.

In addition, several appendices have been added to further facilitate and extend future application of the techniques introduced in this book. Appendix A provides some information about different approaches to implementation of nonvolatile semiconductor devices. Appendix B describes our view of possible utilization of the techniques developed in this book in coming submicron CMOS technologies. Finally, in Appendix C, based on power, speed, and chip-area performance measures, some of the practical advantages of our approach are discussed.

Acknowledgments

The research reflected in this manuscript has been undertaken partly as the Ph.D. work of S. Mehdi Fakhraie at the University of Toronto. The authors wish to extend their sincere gratitude to Professor J. M. Xu for his many useful suggestions and his support during its progress. Also, they wish to express their appreciation to professors B. Benhabib, M. I. Elmasry, G. E. Hinton, D. A. Johns, W. T. Ng, B. E. Shi, and A. N. Venetsanopoulos for their careful review and fruitful discussion during its completion.

We also thank the staff, students, and our colleagues in the VLSI Research Group, and in the Optoelectronics, and Computer Integrated Manufacturing Laboratories at the University of Toronto for their cooperation and assistance during the course of this work. As well, the authors acknowledge fabrication support of the Canadian Microelectronics Corporation (CMC).

We would also like to extend our appreciation to our colleagues at the Hong Kong University of Science and Technology and the University of Tehran for their support during final preparation of this work.

Especial thanks go to Mrs. F. Fatemi-Dezfoli for typing this manuscript. Finally, we thank our respective family members for the patience, support, and encouragement they have provided.

S. Mehdi Fakhraie

K. C. Smith

Contents