FINITE PRECISION NUMBER SYSTEMS AND ARITHMETIC

Fundamental arithmetic operations support virtually all of the engineering, scientific, and financial computations required for practical applications from cryptography, to financial planning, to rocket science. This comprehensive reference provides researchers with the thorough understanding of number representations that is a necessary foundation for designing efficient arithmetic algorithms.

Using the elementary foundations of radix number systems as a basis for arithmetic, the authors develop and compare alternative algorithms for the fundamental operations of addition, multiplication, division, and square root with precisely defined roundings. Various finite precision number systems are investigated, with the focus on comparative analysis of practically efficient algorithms for closed arithmetic operations over these systems.

Each chapter begins with an introduction to its contents and ends with bibliographic notes and an extensive bibliography. The book may also be used for graduate teaching: problems and exercises are scattered throughout the text and a solutions manual is available for instructors.

Encyclopedia of Mathematics and its Applications

All the titles listed below can be obtained from good booksellers or from Cambridge University Press. For a complete series listing visit

http://www.cambridge.org/uk/series/sSeries.asp?code=EOM

80 O. Stormark *Lie's Structural Approach to PDE Systems*
81 C. F. Dunkl and Y. Xu *Orthogonal Polynomials of Several Variables*
82 J. P. Mayberry *The Foundations of Mathematics in the Theory of Sets*
83 C. Foias *et al. Navier–Stokes Equations and Turbulence*
84 B. Polster and G. F. Steinke *Geometries on Surfaces*
85 R. B. Paris and D. Kaminski *Asymptotics and Mellin–Barnes Integrals*
86 R. McEliece *The Theory of Information and Coding, 2nd edn*
87 B. A. Magurn *An Algebraic Introduction to K-Theory*
88 T. Mora *Solving Polynomial Equation Systems I*
89 K. Bichteler *Stochastic Integration with Jumps*
90 M. Lothaire *Algebraic Combinatorics on Words*
91 A. A. Ivanov and S. V. Shpectorov *Geometry of Sporadic Groups II*
92 P. McMullen and E. Schulte *Abstract Regular Polytopes*
93 G. Gierz *et al. Continuous Lattices and Domains*
94 S. R. Finch *Mathematical Constants*
95 Y. Jabri *The Mountain Pass Theorem*
96 G. Gasper and M. Rahman *Basic Hypergeometric Series, 2nd edn*
97 M. C. Pedicchio and W. Tholen (eds.) *Categorical Foundations*
98 M. E. H. Ismail *Classical and Quantum Orthogonal Polynomials in One Variable*
99 T. Mora *Solving Polynomial Equation Systems II*
100 E. Olivieri and M. Eulália Vares *Large Deviations and Metastability*
101 A. Kushner, V. Lychagin and V. Rubtsov *Contact Geometry and Nonlinear Differential Equations*
102 L. W. Beineke and R. J. Wilson (eds.) with P. J. Cameron *Topics in Algebraic Graph Theory*
103 O. J. Staffans *Well-Posed Linear Systems*
104 J. M. Lewis, S. Lakshmivarahan and S. K. Dhall *Dynamic Data Assimilation*
105 M. Lothaire *Applied Combinatorics on Words*
106 A. Markoe *Analytic Tomography*
107 P. A. Martin *Multiple Scattering*
108 R. A. Brualdi *Combinatorial Matrix Classes*
109 J. M. Borwein and J. D. Vanderwerff *Convex Functions*
110 M.-J. Lai and L. L. Schumaker *Spline Functions on Triangulations*
111 R. T. Curtis *Symmetric Generation of Groups*
112 H. Salzmann *et al. The Classical Fields*
113 S. Peszat and J. Zabczyk *Stochastic Partial Differential Equations with Lévy Noise*
114 J. Beck *Combinatorial Games*
115 L. Barreira and Y. Pesin *Nonuniform Hyperbolicity*
116 D. Z. Arov and H. Dym *J-Contractive Matrix Valued Functions and Related Topics*
117 R. Glowinski, J.-L. Lions and J. He *Exact and Approximate Controllability for Distributed Parameter Systems*
118 A. A. Borovkov and K. A. Borovkov *Asymptotic Analysis of Random Walks*
119 M. Deza and M. Dutour Sikirić *Geometry of Chemical Graphs*
120 T. Nishiura *Absolute Measurable Spaces*
121 M. Prest *Purity, Spectra and Localisation*
122 S. Khrushchev *Orthogonal Polynomials and Continued Fractions*
123 H. Nagamochi and T. Ibaraki *Algorithmic Aspects of Graph Connectivity*
124 F. W. King *Hilbert Transforms I*
125 F. W. King *Hilbert Transforms II*
126 O. Calin and D.-C. Chang *Sub-Riemannian Geometry*
127 M. Grabisch *et al. Aggregation Functions*
128 L. W. Beineke and R. J. Wilson (eds.) with J. L. Gross and T. W. Tucker *Topics in Topological Graph Theory*
129 J. Berstel, D. Perrin and C. Reutenauer *Codes and Automata*
130 T. G. Faticoni *Modules over Endomorphism Rings*
131 H. Morimoto *Stochastic Control and Mathematical Modeling*
132 G. Schmidt *Relational Mathematics*
133 P. Kornerup and D. W. Matula *Finite Precision Number Systems and Arithmetic*
134 Y. Crama and P. L. Hammer (eds.) *Boolean Functions*
135 V. Berthé and M. Rigo (eds.) *Combinatorics, Automata and Number Theory*
136 A. Kristály, V. D. Rădulescu and C. Varga *Variational Principles in Mathematical Physics, Geometry, and Economics*

ENCYCLOPEDIA OF MATHEMATICS AND ITS APPLICATIONS

Finite Precision Number Systems and Arithmetic

PETER KORNERUP
University of Southern Denmark, Odense

DAVID W. MATULA
Southern Methodist University, Dallas

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore,
São Paulo, Delhi, Dubai, Tokyo, Mexico City

Cambridge University Press
The Edinburgh Building, Cambridge CB2 8RU, UK

Published in the United States of America by Cambridge University Press, New York

www.cambridge.org
Information on this title: www.cambridge.org/9780521761352

First published 2010

Printed in the United Kingdom at the University Press, Cambridge

A catalog record for this publication is available from the British Library

Library of Congress Cataloging in Publication data
Kornerup, Peter.
Finite precision number systems and arithmetic / Peter Kornerup, David W. Matula.
p. cm. – (Encyclopedia of mathematics and its applications ; 133)
Includes bibliographical references and indexes.
ISBN 978-0-521-76135-2
1. Arithmetic – Foundations. I. Matula, David W. II. Title.
QA248.K627 2010
513 – dc22 2010030521

ISBN 978-0-521-76135-2 Hardback

Additional resources for this publication at www.cambridge.org/9780521761352

Cambridge University Press
978-0-521-76135-2 - Finite Precision Number Systems and Arithmetic
Peter Kornerup and David W. Matula
Frontmatter
More information

CONTENTS

Cambridge University Press
978-0-521-76135-2 - Finite Precision Number Systems and Arithmetic
Peter Kornerup and David W. Matula
Frontmatter
More information

PREFACE

This book builds a solid foundation for finite precision number systems and arithmetic, as used in present day general purpose computers and special purpose processors for applications such as signal processing, cryptology, and graphics. It is based on the thesis that a thorough understanding of number representations is a necessary foundation for designing efficient arithmetic algorithms.

Although computational performance is enhanced by the ever increasing clock frequencies of VLSI technology, selection of the appropriate fundamental arithmetic algorithms remains a significant factor in the realization of fast arithmetic processors. This is true whether for general purpose CPUs or specialized processors for complex and time-critical calculations. With faster computers the solution of ever larger problems becomes feasible, implying need for greater precision in numerical problem solving, as well as for larger domains for the representation of numerical data. Where 32-bit floating-point representations used to be the standard precision employed for routine scientific calculations, with 64-bit double-precision only occasionally required, the standard today is 64-bit precision, supported by 128-bit accuracy in special situations. This is becoming mainstream for commodity microprocessors as the revised IEEE standard of 2008 extends the scalable hierarchy for floating-point values to include 128-bit formats. Regarding addressing large memories, the trend is that the standard width of registers and buses in modern CPUs is to be 64 bits, since 32 bits will not support the larger address spaces needed for growing memory sizes. It follows that basic address calculations may require larger precision operands.

The rising demand for mobile processors has made realizing "low-power" (less energy consumption) without sacrificing speed the preeminent focus of the next generation of many arithmetic unit designs. Multicore processors allow opportunities in heterogeneous arithmetic unit designs to be realized alongside legacy systems. All of these opportunities require a new level of understanding of number representation and arithmetic algorithm design as the core of new arithmetic architectures.

This book emphasizes achieving fluency in redundant representations to avoid the self imposed design straightjacket of prematurely forcing intermediate values into more familiar non-redundant forms. Exploiting parallelism is crucial for multicore processors and for realizing fast arithmetic on large word-size operands, and employing redundant radix representation of numbers allows addition to be performed in constant time, independent of the word size of the operands. Allowing intermediate results to remain in a redundant representation for use in subsequent calculations has to be exploited, avoiding slow (at best logarithmic time) conversion into non-redundant representations where possible. Fortunately, conversion between redundant representations in most cases (for "compatible" radix values) can be performed in constant time.

Radix representation remains the single most important and fundamental way of representing numbers, even serving as a foundation for most other number representations, some of which are presented in the later chapters of this book. We have chosen the foundations of radix arithmetic as the definitive topic for initiating the study of finite precision number systems and arithmetic. We provide a very thorough treatment of radix representations, looking into the implications of the choice of digit set for a given radix as well as the choice of radix. Properties of the resulting set of representable values in the system (its "completeness") and uniqueness of representations ("redundancy") are investigated in a detail not found elsewhere. Conversions between radix representations are analyzed as a separate topic of significant use and importance in the implementations of arithmetic algorithms. Our objective is to provide a substantive mathematical foundation for radix number systems and their properties rather than ad hoc developments tied to specific limited applications.

It is our belief that a detailed understanding of radix representations and conversions between these is of great importance when developing and/or implementing arithmetic algorithms. The results on these topics presented here form a "toolbox" no arithmetic "algorithm engineer" should be without. We have found these tools extremely useful over the more than 30 years of our own joint research on alternative number representation systems and their arithmetic, on algorithm engineering in general, and on developments for actual processor implementations. Being intimately involved with the organization of the bi-annual IEEE Symposia on Computer Arithmetic for a similar time frame, we have been able to follow the challenges and research in this area, often allowing us to improve on existing algorithms, or to explain fundamental issues. For example, writing this book has spawned ideas for several research papers, some of which appeared first in drafts of the book, but the book also includes results that we first presented at meetings, in journal papers, and in actual processor implementations. Participation in the arithmetic unit design and testing of several generations of commercially successful processors such as the Cyrix $\times 87$ coprocessor and the National Semiconductor/AMD Geode "one Watt" IEEE floating-point compliant processor chosen for the One Laptop per Child (OLPC) project has provided valuable feedback on the real world practicality of new arithmetic algorithms.

It has not been possible to include here all the developments presented on computer arithmetic and number systems over the past years; a selection has had to be made. But we claim in most cases to present both classical and up-to-date algorithms for the problems covered. Over the years of writing the book, we have constantly been monitoring the literature, modifying and updating the text with new results and algorithms as they became known.

The approach used in this book is quite mathematical when presenting and analyzing ideas and algorithms, but we go into very little detail on the logic design, and do not look at all at hardware implementations. Complexities of algorithms and designs are generally only specified in the mathematical O-notation, but occasionally we do count gates, just as hardware designs may be sketched as logic diagrams. However, it is definitely our intent that the presentations here should also be of great value for engineers selecting and designing actual VLSI or FPGA implementations of arithmetic algorithms.

The reader is not expected to have more depth of knowledge of logic design and electronics than would be gained from an undergraduate class on computer architecture. Nor is the reader expected to have a deep mathematical background, no more than is usually acquired from an undergraduate computer science curriculum. We do occasionally use terminology from abstract algebra, e.g., (denoting a mapping a homomorphism), or some number system to be a commutative ring, as a benefit to readers familiar with these concepts. We will not, however, use advanced properties beyond the few defined and described in the text. We extensively use the concept of sets and the standard notation for such, and, of course, assume knowledge of simple Boolean algebra. Our derivations and proofs will most often be based on elementary algebra and elementary number theory without recourse to calculus or analysis. The arguments should be quite accessible to those with a natural affinity for games and mathematical puzzles and the book should be invaluable to the serious student who may want to analyze or design such games and puzzles.

The contents of the book can be seen to consist of three parts. The first part comprising two chapters, covers the fundamentals of radix representation and conversion between these representations. For this part we introduce a formal notation for expressing radix polynomials, allowing us to distinguish between different radix representations of the same value, and of the value itself. This notation is used heavily in the first three chapters, but later our notation is more relaxed, when the interpretation should be implicitly obvious from the context. The second part covers in four chapters the basic arithmetic operations: addition, multiplication, division, and square root. These are the fundamental arithmetic operations that are standardized in the IEEE floating-point standard and are the subject of very competitive hardware implementations and much academic research regarding the practical performance of alternative algorithms. The third part presents examples of some special number systems, usually built on the fundamental radix systems. These are the floating-point systems, residue number systems, and finally rational number systems and arithmetic that were largely developed by the authors. The finite precision rational number systems are built on the number theoretic foundations of fractions and continued fractions. The chapter on residue number representations and modular arithmetic includes an extensive presentation of the basic modular operations, some of which are applicable to and important in cryptographic algorithms. The chapter on floating-point systems seeks to provide a foundation for these systems which have largely been ignored in the mathematical literature on number system foundations. We carefully distinguish floating-point number systems at three levels. First, we distinguish those subsets of real numbers which form a floating-point number system, then we specify individual floating-point numbers as real numbers characterized by their factorizations into component terms constituting a sign factor, a scale (or radix shift) factor, and a significand factor, and lastly we investigate the encodings of the various factors into component bit strings of a floating-point word in compliance with the IEEE floating-point standard. The first two levels treating floating-point numbers as reals characterized by a factorization allow the development of a number theoretic foundation for floating-point arithmetic similar to the foundation for rational arithmetic derived from reals characterized by being representable as fractions. The concept of precise roundings allows for a development of the best radix (and floating-point) approximation similar to the best rational approximation concept in the established number theoretic literature on continued fractions. Noteworthily absent among the special number systems are systems employing logarithmic representations, as well as error tolerant systems.

Each chapter begins with an introduction to its contents, and ends with bibliographic notes and a bibliography of publications and selected patents, pointing to the sources used for the ideas and presentations and to further reading. Most sections end with some problems and exercises, illustrating the material presented or further developing the topics. A solutions manual is available for instructors, describing possible solutions to (most of) the presented problems.

We are well aware that the contents of the book are beyond what can be covered in a normal (graduate) semester course. But it is feasible in that time to cover most of Chapters 1–4 and the early parts of Chapters 5 and 6. The remaining parts and chapters can be used for a follow-up course or for individual studies, possibly serving as an introduction to a master's student project, or as background for research towards a Ph D.

The content of this book is based on the last four decades of research in many of these topics, pursued in response to the explosive growth and omnipresence of digital computers. The content includes some of our own results over this period, although most material is based on what is found in the open literature and learned from active communication with colleagues in the international community of "arithmeticians," in particular through our active participation at the bi-annual IEEE Symposia on Computer Arithmetic attended by at least one of us since its initiation in 1969. We "stand on the shoulders" of many, including the very early pioneers of the field, but also many past and present colleagues and students have inspired our work in general, and in particular influenced the presentations here. We have tried to be very comprehensive in our coverage of the results on the topics presented, but it is, of course, not possible to cover everything.

It is our hope that the book may serve as a valuable resource for further academic research on these topics, and also as a useful bookshelf tool for practitioners in the industry who are building the processors of the future.

Despite our efforts, without doubt there are typos and possibly also more serious errors in this text. We apologize, and encourage the reader to contact us if such are found. We will establish a web page listing corrections to the book.

We would especially like to recognize the students who collaborated and contributed to the development of this book over the last two decades. They include D. DasSarma, M. Daumas, A. Fit-Florea, C. S. Iordache, C. N. Lyu, L. D. McFearin and S. N. Parikh in Dallas; and T. A. Jensen, S. Johansen, A. M. Nielsen, H. Orup in Aarhus and Odense. We also thank other graduate and postdoctoral students who have worked with us on this manuscript including S. Datla, G. Even, L. Li, J. Moore, A. Panhaleux, G. Wei and J. Zhang, as well as faculty collaborators W. E. Ferguson, M. A. Thornton, and P. -M. Seidel in Dallas, and R. T. Gregory, U. Kulisch, and J. -M. Muller at their institutions.

Finally we want to express our gratitude to our wives, Margot and Patricia, for their patience with our absence during the numerous hours spent on writing and discussions over many years, and during our mutual visits where they have generally followed us.

August MMX

Peter Kornerup
Dept. of Math. and Computer Science
University of Southern Denmark
Odense, Denmark
kornerup@imada.sdu.dk

David W. Matula
Dept. of Computer Science and Engineering
Southern Methodist University
Dallas, TX
matula@lyle.smu.edu