2019 IEEE 26th Symposium on Computer Arithmetic (ARITH 2019)

Kyoto, Japan
10 – 12 June 2019

IEEE

IEEE Catalog Number: CFP19121-POD
ISBN: 978-1-7281-3367-6

Copyright © 2019 by the Institute of Electrical and Electronics Engineers, Inc.
All Rights Reserved

Copyright and Reprint Permissions: Abstracting is permitted with credit to the source. Libraries are permitted to photocopy beyond the limit of U.S. copyright law for private use of patrons those articles in this volume that carry a code at the bottom of the first page, provided the per-copy fee indicated in the code is paid through Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923.

For other copying, reprint or republication permission, write to IEEE Copyrights Manager, IEEE Service Center, 445 Hoes Lane, Piscataway, NJ 08854. All rights reserved.

****** This is a print representation of what appears in the IEEE Digital Library. Some format issues inherent in the e-media version may also appear in this print version.***

IEEE Catalog Number:	CFP19121-POD
ISBN (Print-On-Demand):	978-1-7281-3367-6
ISBN (Online):	978-1-7281-3366-9
ISSN:	1063-6889

Additional Copies of This Publication Are Available From:

Curran Associates, Inc
57 Morehouse Lane
Red Hook, NY 12571 USA
Phone: (845) 758-0400
Fax: (845) 758-2633
E-mail: curran@proceedings.com
Web: www.proceedings.com

CURRAN ASSOCIATES INC.
proceedings
.com

26th IEEE Symposium on Computer Arithmetic

ARITH-26 (2019)

Table of Contents

Foreword x
Conference Organization xi
Program Committee xii
Steering Committee xiii
Sponsors xiv
Supporters xv

Keynote 1

Error Bounds for Computer Arithmetics 1
Siegfried M. Rump (Hamburg University of Technology)

Session 1: Numerical Computation and Floating-Point Arithmetic

Faster Arbitrary-Precision Dot Product and Matrix Multiplication 15
Fredrik Johansson (LFANT – Inria Bordeaux)

Accurate Complex Multiplication in Floating-Point Arithmetic 23
Vincent Lefèvre (Univ Lyon, EnsL, UCBL, CNRS, Inria, LIP) and Jean-Michel Muller (Univ Lyon, EnsL, UCBL, CNRS, Inria, LIP)

Exchange Algorithm for Evaluation and Approximation Error-Optimized Polynomials 30
Denis Arzelier (CNRS, LAAS and Univ de Toulouse), Florent Bréhard (ENS Lyon and LAAS-CNRS), and Mioara Joldes (CNRS, LAAS and Univ de Toulouse)

Reproducible Summation Under HUB Format 38
Julio Villalba-Moreno (University of Malaga), Javier Hormigo (University of Malaga), and Francisco Jaime (University of Malaga)

Session 2: Arithmetic for Cryptography 1

Hierarchical Approach in RNS Base Extension for Asymmetric Cryptography 46
Libey Djath (Université de Bretagne Occidentale UMR CNRS), Karim Bigou (Université de Bretagne Occidentale UMR CNRS), and Arnaud Tisserand (CNRS)

Efficient Implementation of Modular Division by Input Bit Splitting 54
Danila Gorodecky (National Academy of Science of Belarus) and Tiziano Villa (Verona University)

Session 3: Arithmetic for Machine Learning and Graphics

Scalar Arithmetic Multiple Data: Customizable Precision for Deep Neural Networks 61
Andrew Anderson (Trinity College Dublin), Michael Doyle (Trinity College Dublin), and David Gregg (Trinity College Dublin)

Leveraging the bfloat16 Artificial Intelligence Datatype For Higher-Precision Computations 69
Greg Henry (Intel Corporation), Ping Tak Peter Tang (Intel Corporation), and Alexander Heinecke (Intel Corporation)

New 3D Projection Transformation for Point Clouds 77
Alvaro Vazquez (University of Santiago de Compostela) and Elisardo Antelo (University of Santiago de Compostela)

Session 4: Special Session – Industrial Arithmetic

Optimized Fused Floating-Point Many-Term Dot-Product Hardware for Machine Learning Accelerators 84
Himanshu Kaul (Intel Corporation), Mark Anders (Intel Corporation), Sanu Mathew (Intel Corporation), Seongjong Kim (Intel Corporation), and Ram Krishnamurthy (Intel Corporation)

Bfloat16 Processing for Neural Networks 88
Neil Burgess (Arm), Jelena Milanovic (Arm), Nigel Stephens (Arm), Konstantinos Monachopoulos (Arm), and David Mansell (Arm)

DLFloat: A 16-b Floating Point Format Designed for Deep Learning Training and Inference 92
Ankur Agrawal (IBM), Bruce Fleischer (IBM), Silvia Mueller (IBM), Xiao Sun (IBM), Naigang Wang (IBM), Jungwook Choi (IBM), and Kailash Gopalakrishnan (IBM)

New Technologies for Improved Computing 96
Marius Cornea (Intel Corporation)

ARM Floating Point 2019: Latency, Area, Power 97
David Lutz (ARM Austin Design Center)

Keynote 2

Big Numbers for a Big Universe 99
Andrew Ensor (Auckland University of Technology)

Session 5: Short Paper and Student Session

Precise and Concise Graphical Representation of the Natural Numbers 100
David Matula (Southern Methodist University) and Zizhen Chen (Southern Methodist University)

Dynamic Precision Numerics Using a Variable-Precision UNUM Type I HW Coprocessor 104
Andrea Bocco (Univ. Grenoble Alpes, CEA, LETI), Yves Durand (Univ. Grenoble Alpes, CEA, LETI), and Florent de Dinechin (INSA-Lyon)

A Cost-Efficient Iterative Truncated Logarithmic Multiplication for Convolutional Neural Networks 108
HyunJin Kim (Dankook University), Min Soo Kim (University of California, Irvine), Alberto A. Del Barrio (Universidad Complutense de Madrid), and Nader Bagherzadeh (University of California, Irvine)

Under- and Overflow Detection in the Residue Logarithmic Number System 112
Mark Arnold (XLNS Research), Ioannis Kouretas (University of Patras), Vassilis Paliouras (University of Patras), and John Cowles (University of Wyoming)

Experimental Analysis of Matrix Multiplication Functional Units 116
Brian Hickmann (Intel Corporation) and Dennis Bradford (Intel Corporation)

Performance Evaluation of an Efficient Double-Double BLAS1 Function With Error-Free Transformation and its Application to Explicit Extrapolation Methods 120
Tomonori Kouya (Shizuoka Institute of Science and Technology)

Invited Student Presentation

A Perspective into Squarer Optimization 124
Katherine Parry (Central High School, Rapid City, South Dakota)

Session 6: Adders and Multipliers

An Ultra-Fast Parallel Prefix Adder 125
Kumar Sambhav Pandey (National Institute of Technology, Hamirpur and Indian Institute of Technology, Mandi), Dinesh Kumar B (Indian Institute of Technology, Mandi), Neeraj Goel (Indian Institute of Technology, Ropar), and Hitesh Shrimali (Indian Institute of Technology, Mandi)

Modulo-(2^n+3) Parallel Prefix Addition via Diminished-3 Representation of Residues 135
Ghassem Jaberipur (Shahid Beheshti University) and Sahar Moradi Cherati (Shahid Beheshti University)

High-Throughput Multiplier Architectures Enabled by Intra-Unit Fast Forwarding 143
Jihee Seo (Washington State University) and Dae Hyun Kim (Washington State University)

Table-Based versus Shift-And-Add Constant Multipliers for FPGAs 151
Florent de Dinechin (INSA Lyon), Silviu-Ioan Filip (INRIA, Rennes), Martin Kumm (University of Applied Science), and Luc Forget (INSA Lyon)

Session 7: Error Analysis and Verification

Optimal Bounds for Floating-Point Addition in Constant Time 159
Mak Andrlon (University of Melbourne), Peter Schachte (University of Melbourne), Harald Søndergaard (University of Melbourne), and Peter J. Stuckey (Monash University)

Semi-Automatic Implementation of the Complementary Error Function 167
Anastasia Volkova (Inria) and Jean-Michel Muller (CNRS)

Optimal Word-Length Allocation for the Fixed-Point Implementation of Linear Filters and Controllers 175
Hilaire Thibault (Sorbonne Université, LIP6 UMR), Ouzia Hacène (Sorbonne Université, LIP6 UMR), and Lopez Benoit (none)

Formal Verification of a State-of-the-Art Integer Square Root 183
Guillaume Melquiond (Inria) and Raphaël Rieu-Helft (TrustInSoft/Inria)

Session 8: Special Session – Automatic Datapath Generators

Reflections on 10 Years of FloPoCo 187
Florent de Dinechin (INSA Lyon)

DSL-Based Modular IP Core Generators: Example FFT and Related Structures 190
François Serre (ETH Zurich) and Markus Püschel (ETH Zurich)

Compile-Time Generation of Custom-Precision Floating-Point IP using HLS Tools 192
David B. Thomas (Imperial College London)

Hybrid Dot-Product Design for FP-Enabled FPGAs 194
Bogdan Pasca (Intel Corporation)

Keynote 3

Computer Arithmetic Research to Accelerate Privacy-Protecting Encrypted Computing Such as Homomorphic Encryption 197
Kurt Rohloff (New Jersey Institute of Technology)

Privacy-Preserving Deep Learning via Additively Homomorphic Encryption 198
Shiho Moriai (National Institute of Information and Communications Technology, Japan)

Session 9: Arithmetic for Cryptography 2

Randomization of Arithmetic Over Polynomial Modular Number System 199
Laurent-Stéphane Didier (Université de Toulon), Fangan-Yssouf Dosso (Université de Toulon), Nadia El Mrabet (Ecole des Mines de St Etienne), Jérémy Marrez (LIP6, Sorbonne Université), and Pascal Véron (Université de Toulon)

HyPoRes: An Hybrid Representation System for ECC 207
Paulo Martins (INESC-ID), Jérémy Marrez (LIP6), Jean-Claude Bajard (LIP6), and Leonel Sousa (INESC-ID)

Using the New VPMADD Instructions for the New Post Quantum Key Encapsulation Mechanism SIKE 215
Dusan Kostic (Ecole Polytechnique Federale de Lausanne) and Shay Gueron (University of Haifa and Amazon)

Author Index **219**