Towards EXtreme scale Technologies and Accelerators for euROhpc hw/Sw Supercomputing Applications for exascale

WP2 New accelerator designs exploiting mixed precision

D2.2 AI Accelerator with mixed-precision including Posit, part 1

http://textarossa.eu

This project has received funding from the European Union's Horizon 2020 research and innovation programme, EuroHPC JU, grant agreement No 956831

TEXTAROSSA

Towards EXtreme scale Technologies and Accelerators for euROhpc hw/Sw Supercomputing Applications for exascale

Grant Agreement No.: 956831

Deliverable: D2.2 AI Accelerator with mixed-precision including Posit, part 1

Project Start Date: 01/04/2021 **Duration**: 36 months
Coordinator: *AGENZIA NAZIONALE PER LE NUOVE TECNOLOGIE, L'ENERGIA E LO SVILUPPO ECONOMICO SOSTENIBILE - ENEA , Italy.*

Deliverable No	D2.2
WP No:	WP2
WP Leader:	CINI-UNIPI
Due date:	M18 (September 30, 2022)
Delivery date:	11/10/2022

Dissemination Level:

PU	Public	X
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

This project has received funding from the European Union's Horizon 2020 research and innovation programme, EuroHPC JU, grant agreement No 956831

DOCUMENT SUMMARY INFORMATION

Project title:	Towards EXtreme scale Technologies and Accelerators for euROhpc hw/Sw Supercomputing Applications for exascale
Short project name:	TEXTAROSSA
Project No:	956831
Call Identifier:	H2020-JTI-EuroHPC-2019-1
Unit:	EuroHPC
Type of Action:	EuroHPC - Research and Innovation Action (RIA)
Start date of the project:	01/04/2021
Duration of the project:	36 months
Project website:	textarossa.eu

WP2 New accelerator designs exploiting mixed precision

Deliverable number:	D2.2					
Deliverable title:	AI Accelerator with mixed-precision including Posit, part 1					
Due date:	M18					
Actual submission date:	M18					
Editor:	Sergio Saponara					
Authors:	S. Saponara, F. Rossi					
Work package:	WP2					
Dissemination Level:	Public					
No. pages:	15					
Authorized (date):	11/10/2022					
Responsible person:	Sergio Saponara					
Status:	Plan	Draft	Working	Final	Submitted	Approved

Revision history:

Version	Date	Author	Comment
0.1	2022-09-30	S. Saponara	Draft structure
0.2	2022-10-11	F. Rossi	First version completed

Quality Control:

Checking process	Who	Date
Checked by internal reviewer	Paolo Palazzari	october 12th, 2022
Checked by Task Leader	Sergio Saponara	october 11th, 2022
Checked by WP Leader	Sergio Saponara	october 11th, 2022
Checked by Project Coordinator	Massimo Celino	october 12th, 2022

COPYRIGHT

ACKNOWLEDGEMENTS

This project has received funding from the European High-Performance Computing Joint Undertaking (JU) under grant agreement no 956831. The JU receives support from the European Union's Horizon 2020 research and innovation programme and Italy, Germany, France, Spain, Poland.
Please see http://textarossa.eu for more information on the TEXTAROSSA project.
The partners in the project are AGENZIA NAZIONALE PER LE NUOVE TECNOLOGIE, L'ENERGIA E LO SVILUPPO ECONOMICO SOSTENIBILE (ENEA), FRAUNHOFER GESELLSCHAFT ZUR FOERDERUNG DER ANGEWANDTEN FORSCHUNG E.V. (FHG), CONSORZIO INTERUNIVERSITARIO NAZIONALE PER L'INFORMATICA (CINI), INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET AUTOMATIQUE (INRIA), BULL SAS (BULL), E4 COMPUTER ENGINEERING SPA (E4), BARCELONA SUPERCOMPUTING CENTER-CENTRO NACIONAL DE SUPERCOMPUTACION (BSC), INSTYTUT CHEMII BIOORGANICZNEJ POLSKIEJ AKADEMII NAUK (PSNC), ISTITUTO NAZIONALE DI FISICA NUCLEARE (INFN), CONSIGLIO NAZIONALE DELLE RICERCHE (CNR), IN QUATTRO SRL (in4). Linked third parties of CINI are POLITECNICO DI MILANO (CINI-POLIMI), Università di Torino (CINI-UNITO) and Università di Pisa (CINI-UNIPI); linked third party of INRIA is Université de Bordeaux; in-kind third party of ENEA is Consorzio CINECA (CINECA); in-kind third party of BSC is Universitat Politècnica de Catalunya (UPC).
The content of this document is the result of extensive discussions within the TEXTAROSSA © Consortium as a whole.

DISCLAIMER

Table of contents

List of Acronyms

Acronym	Definition
ALU	Arithmetic Logic Unit
ASIC	Application Specific Integrated Circuit
CINI	Consorzio Interuniversitario Nazionale per l'Informatica
CPU	Central Processing Unit
DNN	Deep Neural Network
FP32	Floating Point 32 bit
FPGA	Field Programmable Gate Array
FTS	Fast Task Scheduler
HE	Homomorphic Encryption
HW	Hardware
HPC	High-Performance-Computing
INFN	Istituto Nazionale di Fisica Nucleare
IP	Intellectual Property
IPR	Intellectual Property Rights
XOF	eXtendable Output Function
PMB	Project Management Board
PPU	Posit Processing Unit
PQC	Post Quantum Cryptography
RISC	Reduced Instruction Set Computer
SEAL	Simple Encrypted Arithmetic Library
SW	Software
RLWE	Ring Learning With Errors
UART	Universal Asynchronous Receiver Transmitter interface
VHDL	VHSIC Hardware Description Language
VPU	Vector Processor Unit

Executive Summary

This document reports the activities done by Textarossa partner CINI (UNIPISA), with reference to preliminary HDL design, verification and synthesis of accelerator IPs in WP2 for AI accelerator with mixed-precision arithmetic including the new format called Posit.

The AI accelerator IP has been implemented in FPGA technology and can be integrated with RISC-V cores like Ariane RISC-V 64 bits.

The AI accelerator IP is designed according to the specifications defined in D2.1 [1].

1. Introduction

This document D2.2 reports the activities done by Textarossa partner CINI (UNIPISA) in WP2.

D2.2 deals with the preliminary HDL design, using SystemVerilog, verification and synthesis of accelerator IPs for AI acceleration.

The main innovation is in the hardware support of a new arithmetic format called Posit, suitable for acceleration of DNN computation.

The AI accelerator IP in Section 2 has been implemented in FPGA technology and it has been designed according to the specifications defined in D2.1.

Furthermore, it has been verified that the AI accelerator IP can be integrated with RISC-V cores like Ariane RISC-V 64 bits.

Please note that the theory of Posit arithmetic, the structure of Posit numbers and their benefits vs. classic integer and floating-point formats have been already discussed by us in published works such as [2, 3]. Therefore, the goal of this deliverable is on the design and verification of digital IPs supporting Posit.

Conclusions are drawn in Section 3.

2. AI accelerator IP with Posits

Goal of this work is to design and implement an AI accelerator exploiting Posit arithmetic as an extension of [4]. The Posit Processing Unit (hereafter called PPU) is capable of performing all arithmetic operations between posit numbers while being able, optionally (i.e. configurable at synthesis time) to perform conversion between binary32 (a.k.a IEEE754 32-bit floats) and posit numbers.

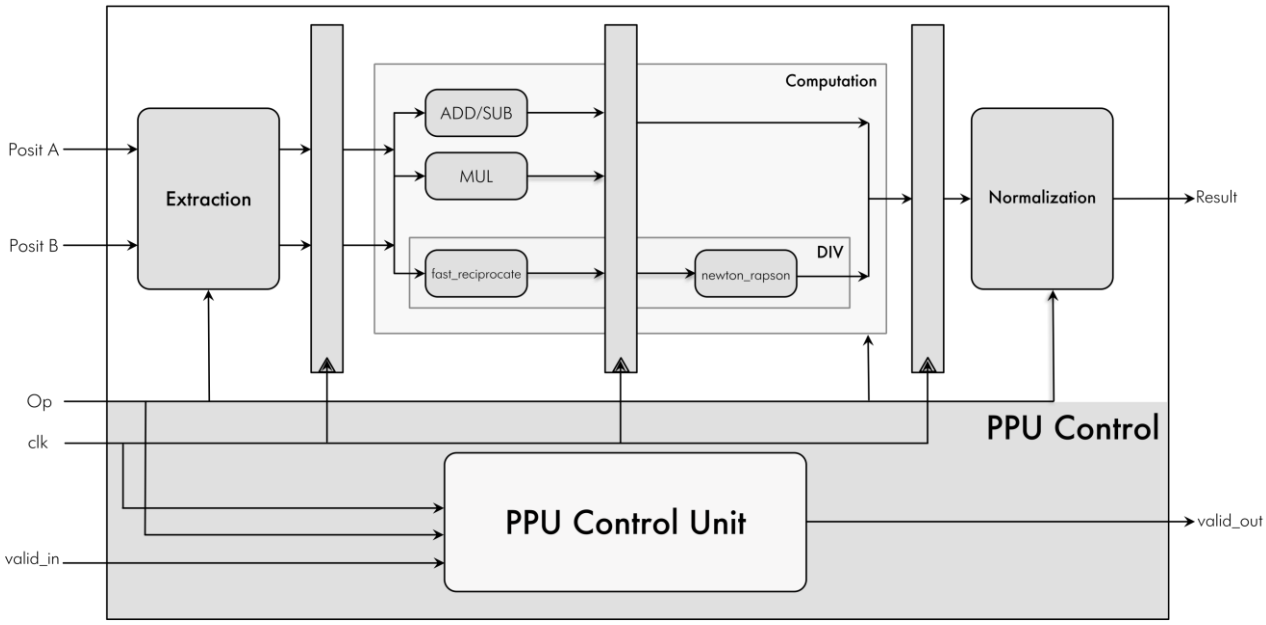
The target PPU can be configured at synthesis time to produce a module with different configurations. Hereafter we list the configuration parameters for the PPU design:

1. Number of posit bits (e.g. 16 bits)
2. Number of posit exponent bits (e.g. 2 bits)
3. Word size for the module interface (e.g. 32-bit or 64-bit)
4. Posit to binary32 conversions enabled or not.

The PPU was designed using a pipelined approach.
There are 4 stages in the pipeline, with two of them reserved for the computational part:

1. Extraction and decoding of the input operands (1 stage)
2. Computation of the arithmetic operation (2 stages)
3. Normalization and encoding of the output (1 stage)

The top-level module is shown in the Figure 1 below and includes:
- a PPU control unit which is a finite state machine that receives from the external IP of the processor core the clock, a data valid and the opcode of the type of operations to be implemented.
- internal units for hardware implementation of add, sub, mul, inverse (via fast reciprocate algorithm) and division operations



- extraction and normalization modules that manage the posit format with its constituents' parts: sign, mantissa, regime and exponent

Figure 2.1: The top-level module of the PPU

The PPU unit, with both functional RTL and synthetized gate-level netlists, was thoroughly tested against a golden model software posit library called CppPosit to check for implementation errors. We completely tested the PPU unit with the Posit16 and Posit8 configurations without errors both in Behavioral, Post-Synthesis and FPGA Post-Implementation evaluations.

We synthesized the unit targeting different FPGA platforms with different posit configurations. Some synthesis results are reported in Tables below.
Table 1 shows circuit complexity in terms of utilization (absolute value and %) of FPGA LUTs and registers for 3 different Xilinx FPGA devices and different configurations of Posits (with 16 and 8 bits).
Table 2 shows speed results in terms of clock frequency for 3 different Xilinx FPGA devices and different configurations of Posits (with 16 and 8 bits).
It is worth noting that Posit16 in literature [7] have been proved to be more effective, for DNN computation, than FP32 and that Posit8 in literature have been proved to be more effective than FP16.

Utilization

Part	POSIT	LUTS	TOTAL_LUTS	PERC_LUTS (%)	REGS	TOTAL_REGS	PERC_REGS (%)
xc7a12ticsg325-1L	P16E0	1249	8000	15.61	350	16000	2.19
	P16E1	1410	8000	17.63	363	16000	2.27
	P16E2	1412	8000	17.65	365	16000	2.28
	P8E0	453	8000	5.66	227	16000	1.42
	P8E1	444	8000	5.55	238	16000	1.49
	P8E2	449	8000	5.61	245	16000	1.53
xc7k70tfbv676-3	P16E0	1249	41000	3.05	350	82000	0.43
	P16E1	1410	41000	3.44	363	82000	0.44
	P16E2	1412	41000	3.44	365	82000	0.45
	P8E0	453	41000	1.10	227	82000	0.28
	P8E1	444	41000	1.08	238	82000	0.29
	P8E2	449	41000	1.10	245	82000	0.30
xc7s6ftgb196-2	P16E0	1319	3750	35.17	364	7500	4.85
	P16E1	1480	3750	39.47	377	7500	5.03
	P16E2	1475	3750	39.33	377	7500	5.03
	P8E0	453	3750	12.08	227	7500	3.03
	P8E1	444	3750	11.84	238	7500	3.17
	P8E2	449	3750	11.97	245	7500	3.27

Table 1: Circuit complexity of the PPU in different Xilinx FPGAs

Timing

Part	POSIT	clock (ns)	frequency (MHz)	slack (ns)	Max frequency (MHz)
xc7s6ftgb196-2	P16E1	20.000	50.000	2.273	56.411
xc7s6ftgb196-2	P16E2	20.000	50.000	2.309	56.526
xc7s6ftgb196-2	P16E0	20.000	50.000	2.309	56.526
xc7a12ticsg325-1L	P8E0	20.000	50.000	4.814	65.850
xc7a12ticsg325-1L	P8E2	20.000	50.000	5.285	67.958
xc7a12ticsg325-1L	P8E1	20.000	50.000	5.475	68.847
xc7k70tfbv676-3	P16E0	20.000	50.000	7.122	77.652
xc7k70tfbv676-3	P16E2	20.000	50.000	7.395	79.334
xc7k70tfbv676-3	P16E1	20.000	50.000	7.493	79.955
xc7s6ftgb196-2	P8E0	20.000	50.000	7.628	80.828
xc7s6ftgb196-2	P8E2	20.000	50.000	7.951	82.994
xc7s6ftgb196-2	P8E1	20.000	50.000	7.997	83.313
xc7k70tfbv676-3	P8E0	20.000	50.000	11.411	116.428
xc7k70tfbv676-3	P8E2	20.000	50.000	11.744	121.124
xc7k70tfbv676-3	P8E1	20.000	50.000	11.776	121.595

Table 2: Speed (frequency results) of the PPU synthesized in different Xilinx FPGAs

Integration with RISC-V CPUs

The integration within the RISC-V core (Ariane CVA6 code, [5, 6]) can be done using the possibility to customize the instruction set.
The PPU can be integrated in two ways:

- In addition to the Ariane integer ALU and Ariane FPU, in such case the optional FP32 to Posit conversion is not configured in the PPU;
- In substitution to the Ariane FPU (and in addition to the Ariane ALU). In this case to have compatibility with flat numbers the optional FP32 to/from Posits of the PPU has to be integrated

3. Conclusions

This document D2.2 has reported the activities done by Textarossa partner CINI (UNIPISA) in WP2.

D2.2 has presented the preliminary HDL design, using SystemVerilog, verification and synthesis of accelerator IPs for AI acceleration.

The main innovation is in the hardware support of a new arithmetic format called Posit, suitable for acceleration of DNN computation, via a unit called PPU.

The PPU has been implemented in different Xilinx FPGA devices and it has been designed according to the specifications defined in D2.1, aiming to demonstrate its platform independency. Future activities will expand the implementation to other platforms (e.g. the ALVEO U280 platform also selected by other partners of the project).

Furthermore, it has been verified that the AI accelerator IP can be integrated with RISC-V cores like Ariane RISC-V 64 bits.

4. References

[1] D21, “Consolidated specs of accelerators IPs”, Textarossa project, May 2022

[2] Marco Cococcioni, Federico Rossi, Emanuele Ruffaldi, & Sergio Saponara, Benoit De Dinechin, “Novel Arithmetics in Deep Neural Networks Signal Processing for Autonomous Driving: Challenges and Opportunities”, IEEE Signal Processing Magazine, Volume: 38, Issue: 1, 2021

[3] Marco Cococcioni, Federico Rossi, Emanuele Ruffaldi, Sergio Saponara, A Novel Posit-based Fast Approximation of ELU Activation Function for Deep Neural Networks, 2020 IEEE International Conference on Smart Computing (SMARTCOMP)

[4] Marco Cococcioni, Federico Rossi, Emanuele Ruffaldi, & Sergio Saponara. (2021). A Lightweight Posit Processing Unit for RISC-V Processors in Deep Neural Network Applications. IEEE Trans on Emerging topics in Computing, https://zenodo.org/record/7128760#.Y0ZfpC8QNbU

[5] F. Zaruba, and L. Benini, “The Cost of Application-Class Processing: Energy and Performance Analysis of a Linux-ready 1.7GHz 64bit RISC-V Core in 22nm FDSOI Technology”, arXiv e-prints, 2019.

[6] https://github.com/openhwgroup/cva6.

[7] M. Cococcioni, F. Rossi, E. Ruffaldi, S. Saponara and B. Dupont de Dinechin, "Novel Arithmetics in Deep Neural Networks Signal Processing for Autonomous Driving: Challenges and Opportunities," in IEEE Signal Processing Magazine, vol. 38, no. 1, pp. 97-110, Jan. 2021, doi: 10.1109/MSP.2020.2988436.