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

WP2 New accelerator designs exploiting mixed precision

D2.6 IP with data compression, 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: D2.6 IP with data compression, 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	D2.6
WP No:	WP2
WP Leader:	CINI-UNIPI
Due date:	M18 (September 30, 2022)
Delivery date:	13/10/2022

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:	D2.6					
Deliverable title:	IP with data compression, 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:	16					
Authorized (date):	13/10/2022					
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	2022-10-12	F. Rossi	First version completed
0.3	2022-10-12	D. Gregori	First Review
1.0	2022-10-13	F. Rossi	Final version completed

Quality Control:

Checking process	Who	Date
Checked by internal reviewer	Daniele Gregori	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 13th, 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 IP with data compression.

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

The IP with data compression is designed according to the specifications defined in D2.1 [1].

1. Introduction

This document D2.6 reports the activities done by Textarossa partner CINI (UNIPISA) in WP2.

D2.6 deals with the preliminary HDL design, using SystemVerilog, verification and synthesis of an IP for data compression.

The main innovation is in the hardware support of a new arithmetic format called Posit that particularly for ML and DNN applications has been proved to have a high compression effect: 4x compression for the same quality.

The IP with data compression 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 data compression IP can be integrated with RISC-V cores like Ariane RISC-V 64 bits.

Conclusions are drawn in Section 3.

2. IP with data compression

The goal of this work the design of an IP core for lightweight PPU (Posit Processing Unit) to be connected to a 64b RISC-V processor in the form of a co-processor with an extension of the Instruction Set Architecture.

Please note that the theory of Posit arithmetic, the structure of Posit numbers and their compression data benefit vs. classic integer and floating-point formats, particularly for DNN computation, have been already discussed by us in published works such as [2-4]. Therefore, the goal of this deliverable is on the design and verification of digital IPs supporting data compression for DNN thanks to the light support of Posits.

We focus on the compression abilities of posits by providing a co-processor with only conversions in mind, called light PPU, (as in Figure 2.1 below).

The difference between D2.6 and D2.2 is that D2.2. presents an AI accelerator IP giving full hardware support to posits number and hence using the IP in D2.2 the FPU of a RISC-V processor can be eliminated, while in D2.2 the idea is minimizing the circuit complexity overhead and hence the proposed IP supports only the data compression using Posits with Float to/from Posit translation but operations have to be done still in the FPU.

We can convert binary32 floating point numbers to posit numbers with 16 and 8 total bits (and a changing number for the exponent since Posit 16,0 and Posit16,1 are considered).

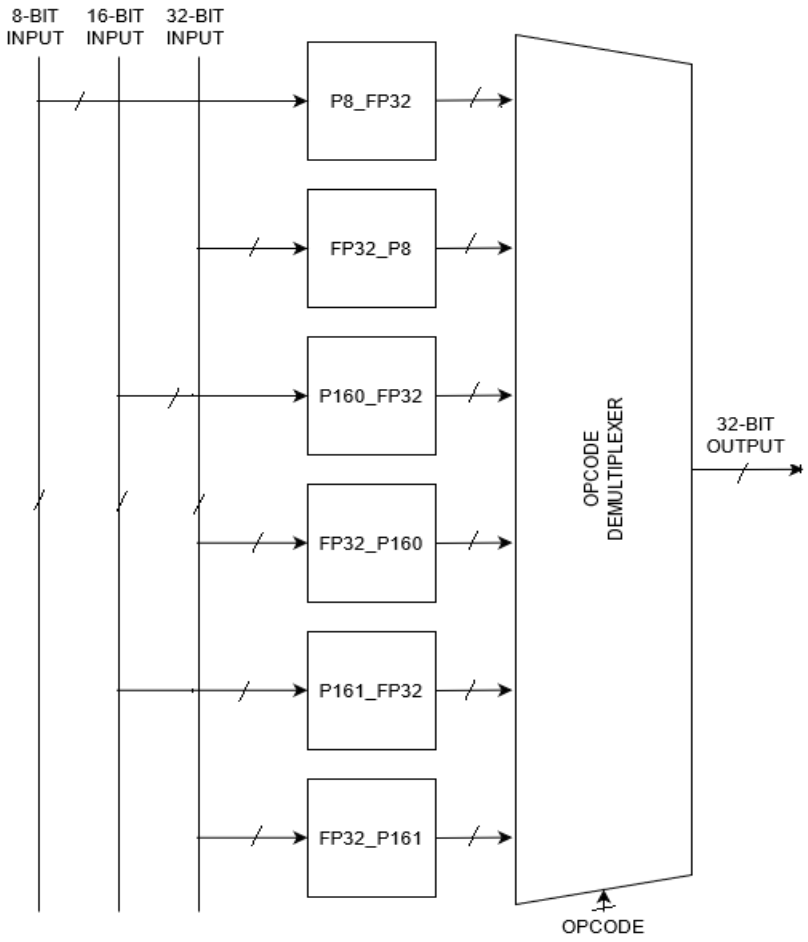

Figure 2.1: light PPU architecture

This co-processor, as shown in Figure 2.2, can be paired with a RISC-V core that already has a floating-point unit (e.g., the Ariane 64b RISC-V) without interrupting the existing pipeline.
On the other hand, we can use this unit to enable ALU computation of posit numbers with the posit-to-fixed conversion modules on a RISC-V core that does not support floating-point.
We investigated the first use-case by outfitting a CVA6 core with our PPU co-processor and synthesizing it for a Xilinx Genesys 2 FPGA, resulting in a working RISC-V core capable of running a general-purpose Linux distribution.

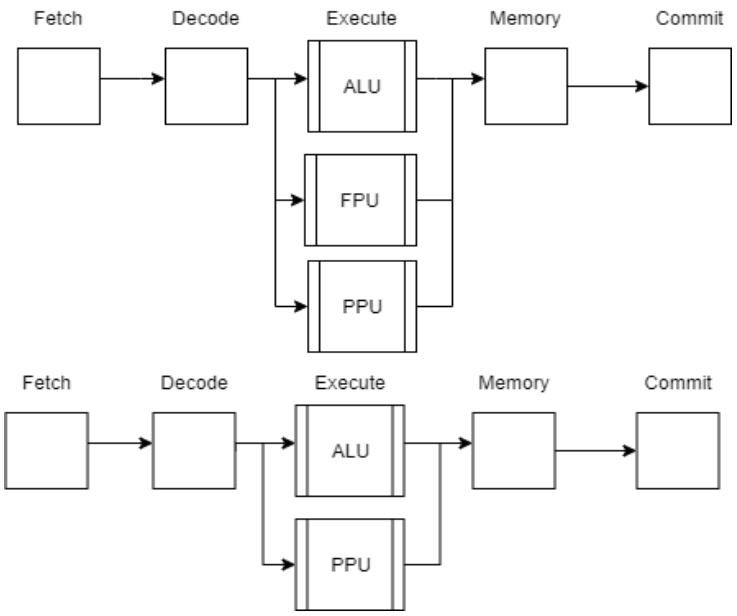

Figure 2.2: PPU possible integration modes within the RISC-V instruction set (with/without the FPU)

We chose the Xilinx Genesys 2 board for the hardware implementation (equipped with a Kintex 7 XC7K325T-2FFG900C FPGA component). We chose this board to reduce the work required to construct our PPU inside a RISC-V core. We did, in fact, use the ARIANE RISC-V core that was originally built for this board.
The resulting design was then implemented on the same board.
For the PPU component, we performed power, circuit complexity, and propagation delay (worst case combinatorial propagation delay of the PPU) reports:
1. Look-up table (LUT) utilization: 747/203800 (0.36%) LUTs used.
2. Component latency: 6.332ns (worst propagation delay).

Finally, the new instruction set architecture was merged into the Ariane RISC-V core and synthesized for the Xilinx Genesys 2.
The following quality parameters were obtained:
1. Clock frequency: 125MHz
2. Total power on FPGA component (Kintex 7): 2.056W
3. Look-up table (LUT) utilization: 63805/203800 (31.54%) LUTs used.

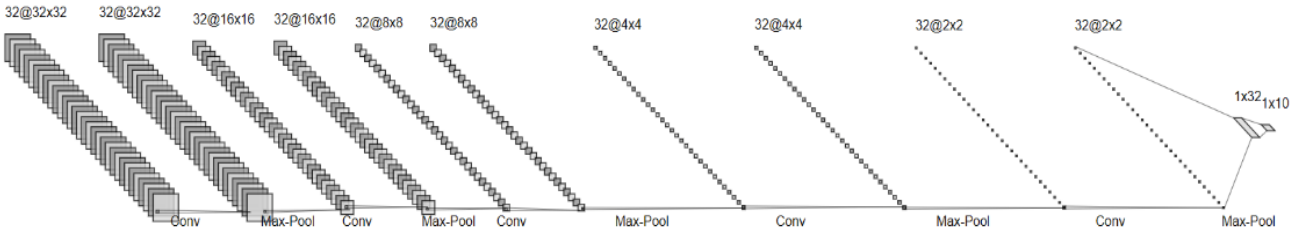

Figure 2.3: example neural network used for test of the light PPU data compression IP

We used the following neural network (LeNet-5) as benchmark base for weight compression:

We also give earlier accuracy results on this neural network in Figure 2.3 utilizing both posit and binary32 integers with varying workloads for completeness. This result was reached by using posit compression for the weights and computing using binary32 (fp32) format. We measured the relative speedup to a software-based posit compression.
As a result, we investigated the overall system compression times with the weights of a tiny LeNet-5 neural network, yielding the result displayed in Table 2.2.

	LeNet	
	MNIST	GTRSB
FP32	98.83%	91.8%
posit⟨16, 1⟩	98.83%	91.8%
posit⟨16, 0⟩	98.50%	90.5%
posit⟨8, 0⟩	98.34%	90.4%

Table 2.1: Accuracy performance of Posit vs FP32 for different benchmark data sets (MNIST [5] and the German Traffic Road Sign Benchmark [6])

	w/ PPU (s)	wo/ PPU (s)	Speedup
posit⟨8, 0⟩	5.4	58.87	**10.90**
posit⟨16, 0⟩	11.6	64.54	**5.56**

Table 2.2: Speed up performance of Posit vs FP32 for a DNN

	Time (s)	DNN size (bytes)	Compression
IEEE FP32	2,1	224894	-
posit⟨16, 0⟩	11.6	112874	**1.99**
posit⟨8, 0⟩	5.4	56864	**3.95**

Table 2.3: Compression performance of Posit vs FP32 for a DNN

Integration with RISC-V CPUs

The integration within the RISC-V core (Ariane CVA6 code, [7, 8]) can be done using the possibility to customize the instruction set.
The posit-based compression IP proposed in D2.6 can be integrated in addition to the Ariane integer ALU and in addition to the Ariane FPU.

3. Conclusions

In this work we dealt with the preliminary HDL design, using SystemVerilog, verification and synthesis of an IP for data compression exploiting posit arithmetic.

The key novelty is the hardware support for a new arithmetic format called Posit, which has been shown to have a strong compression effect, notably for ML and DNN applications: 4x compression for the same quality.

We designed the data compression IP using FPGA technology targeting the Xilinx Genesys 2 FPGA platform. Furthermore, we tested the IP design against a golden-model software library for posit arithmetic.

The data compression IP 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).

Finally, we integrated the data compression IP within the ARIANE 64-bits 6-stage RISC-V core.

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] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. "Gradient-based learning applied to document recognition." Proceedings of the IEEE, 86(11):2278-2324, November 1998.
[6] J. Stallkamp, M. Schlipsing, J. Salmen, C. Igel, Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition, Neural Networks, Available online 20 February 2012, ISSN 0893-6080, 10.1016/j.neunet.2012.02.016.
(http://www.sciencedirect.com/science/article/pii/S0893608012000457) Keywords: Traffic sign recognition; Machine learning; Convolutional neural networks; Benchmarking
[7] 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.
[8] https://github.com/openhwgroup/cva6.