software framework for runtime-Adaptive and secure deep Learning On Heterogeneous Architectures

Project Number 780788

Project Acronym ALOHA

D4.3	First release of the hardware abstraction layer utilities		
Work Package:	WP4	**Lead Beneficiary:**	ST-I
Type:	Demonstrator	**Dissemination level:**	Public
Due Date:	31/12/2018	**Delivery:**	31/12/2018
Version:	v1.0		

Brief description:

The purpose of this deliverable is to release a first version of the utilities and libraries implementing hardware abstraction layer for ALOHA reference computing platforms. It contains simple and static implementation of DL algorithms on each considered platform, making use of the techniques and libraries under development.

Deliverable Author(s):

Name	Beneficiary
Giulio Urlini	ST-I
Paolo Meloni	UNICA

Deliverable Revision History:

Reviewer Beneficiary	Issue Date	Version	Comments
UNICA	20/12/2018	v0.1	Released of the Table of Content and introduction
ST-I	28/12/2018	v0.2	Draft version
ST-I	31/12/2018	v1.0	Final Version reviewed by the coordinator

Disclaimer

This document may contain material that is copyright of certain ALOHA beneficiaries, and may not be reproduced, copied, or modified in whole or in part for any purpose without written permission from the ALOHA Consortium. The commercial use of any information contained in this document may require a license from the proprietor of that information. The information in this document is provided "as is" and no guarantee or warranty is given that the information is fit for any particular purpose. The user thereof uses the information at its sole risk and liability.

The ALOHA Consortium is the following:

#	Participant Legal Name	Acronym	Country
1	STMICROELECTRONICS SRL	ST-I	Italy
2	UNIVERSITA' DEGLI STUDI DI CAGLIARI	UNICA	Italy
3	UNIVERSITEIT VAN AMSTERDAM	UVA	Netherlands
4	UNIVERSITEIT LEIDEN	UL	Netherlands
5	EIDGENOESSISCHE TECHNISCHE HOCHSCHULE ZUERICH	ETHZ	Switzerland
6	UNIVERSITA' DEGLI STUDI DI SASSARI	UNISS	Italy
7	PKE ELECTRONICS AG	PKE	Austria
8	CA TECHNOLOGIES DEVELOPMENT SPAIN SA	CA	Spain
9	SOFTWARE COMPETENCE CENTER HAGENBERG GMBH	SCCH	Austria
10	SANTER REPLY SPA	REPLY	Italy
11	IBM ISRAEL - SCIENCE AND TECHNOLOGY LTD	IBM	Israel
12	SYSTHMATA YPOLOGISTIKIS ORASHS IRIDA LABS AE	IL	Greece
13	PLURIBUS ONE SRL	P-ONE	Italy
14	MAXQ ARTIFICIAL INTELLIGENCE, LTD (formerly MEDYMATCH TECHNOLOGY, LTD)	MaxQ-AI (formerly MM)	Israel

Table of Contents

1 Executive Summary 6
1.1 Acronyms and abbreviations 6
2 Demonstrator overview 7

Figures

Figure 1: Overview of the ALOHA toolflow. The components related with the porting of the inference application on the target architecture are represented as blue boxes. 7

Tables

Table 1: WP4 activity products at M12 7

1 Executive Summary

This document is aimed to be related to Deliverable D4.3 due at M12. This demonstrator consists in a set of examples and utilities, in a first usable standalone version, implementing the output of the ALOHA toolflow on the target reference platforms. During the development activities and for release purpose the code implementing the examples has been stored and shared using the https://gitlab.com/aloha.eu repository. The repository has been made accessible for project partners and for reviewers. To obtain access to the repository, please fill in the form available at https://www.aloha-h2020.eu/project/get-involved

1.1 Acronyms and abbreviations

Acronym	Meaning
CNN	Convolution Neural Network
DL	Deep Learning
DNN	Deep Neural Network
WP	Work Package

2 Demonstrator overview

Figure 1 shows the positioning of the utilities and libraries collected in WP4 inside the ALOHA toolflow. The aim is to adapt the DNN configuration selected by upper levels of the design flow to the target underlying platform. At M12 we have studied the reference programming models and available computing resources, looking for the definition of a common abstraction methodology. We have implemented and tested commonly known CNN benchmarks on the available hardware and started automation of translation between CNN specification and programming interfaces.

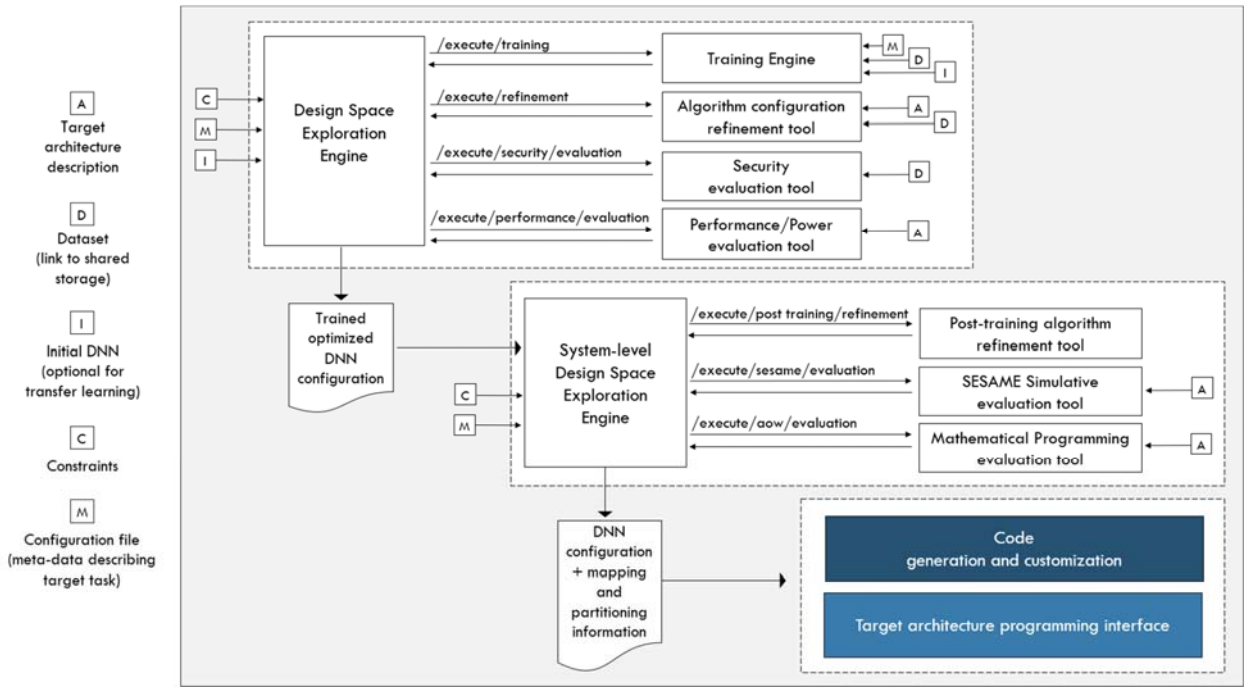

Figure 1: Overview of the ALOHA toolflow. The components related with the porting of the inference application on the target architecture are represented as blue boxes.

At this stage the repository contains implementation examples for the NEURAghe platform and for simpler ARM-based processing systems (ST Micro's Sensortile). Release plan for code-level examples targeting the Orlando platform must be refined in the coming months. A video showing a demonstration based on Resnet-16 on NEURAghe is available at https://ibm.ent.box.com/s/3jd16abflzio85zcgrfodaypuxed1lvr. Several demonstrators have been created by ST-I using Orlando (see a publicly available example at https://youtu.be/FHfyq91tTzk). Finally, we release the first version of the utility needed to generated C code from ONNX-based network specification.

Here follows a list of the related projects inside the repository. A detailed description of the main features of each block has been provided on Deliverable D4.1 at M9.

Table 1: WP4 activity products at M12

Product description	Related task in DoA	Link to Gitlab project of tool components	Main contributor
Resnet on NEURAghe	T4.1	https://gitlab.com/aloha.eu/neuraghedemo (DEMO code) https://gitlab.com/aloha.eu/neuragheconvnet (Software/Firmware environment)	UniCA/ETHZ

Keyword Spotting on Sensortile	T4.3	https://gitlab.com/aloha.eu/kws-on-sensortile	UniCA/UniSS/ST-I
ONNX to C code translation utility	T4.1/T4.2/T4.3	https://gitlab.com/aloha.eu/onnx2c	UniCA