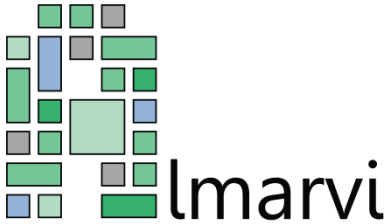


ALMARVI

"Algorithms, Design Methods, and Many-Core Execution Platform for Low-Power Massive Data-Rate Video and Image Processing"

Project co-funded by the ARTEMIS Joint Undertaking under the

ASP 5: Computing Platforms for Embedded Systems

ARTEMIS JU Grant Agreement n. 621439

D7.8 - ALMARVI Project Booklet

Due date of deliverable: M36

Start date of project: 1 April, 2014	**Duration:** 39 months
Organisation name of lead contractor for this deliverable:	VTT
Author(s):	Janne Keränen
Validated by:	
Version number:	0.2
Submission Date:	24-04-2017
Doc reference:	ALMARVI - D7.8 ALMARVI Project Booklet v2.docx
Work Pack./ Task:	WP7 task 7.2
Description: ***(max 5 lines)***	The ALMARVI project booklet summarises the ALMARVI's technical contribution, developments, research and technological achievements.

Nature:	**R**		
Dissemination Level:	**PU**	Public	**X**
	PP	Restricted to other programme participants (including the JU)	
	RE	Restricted to a group specified by the consortium (including the JU)	
	CO	Confidential, only for members of the consortium (including the JU)	

DOCUMENT HISTORY

Release	Date	Reason of change	Status	Distribution
V0.1	13/04/2017	First draft	Draft	Frank van der Linden
v0.2	24/04/2017	Second draft according to comments	Draft	PMT

Table of Contents

1. Introduction

ALMARVI aims at providing cross-domain many-core platform solution, system software stack, tool chain, and adaptive algorithms that will enable massive data-rate image/video processing with high energy efficiency. ALMARVI provides mechanisms and support for high degree of adaptivity at various system layers that abstract the variations in the underlying platforms (e.g., due to imperfections in the fabrication process), communication channels (e.g., available bandwidth), application behavior (dynamic workloads, changing requirements) from the application developer. This is crucial for providing consistent performance efficiency in an interoperable manner when considering heterogeneous platform options and dynamic operating conditions. The key is to leverage image/video content-specific properties, application-specific features, and inherent resilience properties of image/video processing applications.

This document is an information source for ALMARVI project results in a booklet form. The ALMARVI project booklet summarises the ALMARVI's technical contribution, developments, research and technological achievements, etc. The booklet is based on demonstrators presented in ALMARVI project and it explicitly illustrates the results and developments of the ALMARVI project using multiple demonstrators developed in Work Package 5.

This document starts with presenting the contact lists of the project and the ALMARVI project structure in Section 2 and 3, respectively. Then, Section 4 presents the ALMARVI technological achievements as they relate to the objectives of the project and organized along the lines of the various work packages in the project. The demonstrators developed by the project partners in the three application domains (healthcare, surveillance and mobile) are discussed in Section 5. Public deliverables prepared within the project are presented in Section 6. Finally, dissemination activities and ALMARVI partner logos are shown in Section 7 and 8, respectively.

2. Contact Lists

This chapter provides contact lists for the ALMARVI project.

2.1 Project and country coordinators

Project coordinators lead the administrative and scientific developments of the project and ensure timely progress towards the project objectives. They are also the link between project partners and the Public Authorities.

Task	Organization	Name
Overall project coordinator	Philips	Frank van der Linden
Scientific and technical coordinator	TUDelft	Zaid Al-Ars

Country coordinators contact to the national Public Authorities and co-ordinate the national progress reporting. Country coordinators also ensure consistency between National and Artemis reporting.

Country	Organization	Name
The Netherlands	Philips	Frank van der Linden
Finland	UEF	Pekka Toivanen
Turkey	Aselsan	Toygar Akgun
Czech Republic	UTIA	Jiri Kadlec

2.2 ALMARVI web access

ALMARVI websites	
Websites - external	http://www.artemis-ia.eu/project/index/view?project=55 http://www.almarvi.eu

2.3 Contact persons for each partner

Partner	**Name**
Philips Healthcare	Frank van der Linden
TU Delft	Zaid Al-Ars
TU Eindhoven	Dip Goswami
Nokia	Heikki Berg
Hurja Solutions	Vili Harkonen
Visidon	Markus Turtinen
UEF	Pekka Toivanen
VTT	Janne Keranen
Tampere UT	Pekka Jääskeläinen

Partner	Name
Turun yliopisto	Lauri Koskinen
UTIA	Jiri Kadlec
Brno UT	Pavel Svoboda
CAMEA	Lukas Marsik
ASELSAN	Toygar Akgun
Ozyegin University	Fatih Ugurdag

3. ALMARVI Project Structure

3.1 Project members

The ALMARVI project member organizations are the following:

Participant No.	Participant Organisation Name	Short Name	Country
1 (Co-ordinator)	Philips Healthcare	PHILIPS	NL
2	Technische Universiteit Delft	TUDelft	NL
3	Technische Universiteit Eindhoven	TUE	NL
4	Nokia Oyj	NOK	FI
5	Hurja Solutions Oy	HURJA	FI
6	Visidon Oy	VIS	FI
7	ITA-SUOMEN YLIOPISTO	UEF	FI
8	VTT Technical Research Center of Finland	VTT	FI
9	Tampere University of Technology	TUT	FI
10	Turun yliopisto	UTURKU	FI
11	Ustav teorie informace a automatizace AV CR, v.v.i.	UTIA	CZ
12	Brno University of Technology	BUT	CZ
13	CAMEA, spol. s r.o. (ltd.)	CAMEA	CZ
14	ASELSAN	ASEL	TR
15	Ozyegin University	OZYEGIN	TR

3.2 Work package leaders

In ALMARVI, the work package (WP) leaders take care of day to day management of the WPs. WP leaders track WP progress and report to the project manager and technical manager. WP leaders also organize WP meetings when needed.

Work Package	Work Package Title	Lead Partic. Short Name & Person	Start Month	End Month
WP-1	ALMARVI System Architecture	ASEL Toygar Akgun	M01	M30
WP-2	Development of Scalable Low-Power Image/Video Processing Algorithms	UEF Pekka Toivanen	M07	M30
WP-3	ALMARVI Execution Platform and Design Tools	TUDelft Zaid Al-Ars	M07	M30
WP-4	ALMARVI System Software Stack	TU Tampere Pekka Jääskeläinen	M07	M32
WP-5	ALMARVI Demonstrators	PHILIPS Mathijs Visser	M02	M36
WP-6	Project Management	PHILIPS Frank van der Linden	M01	M36
WP-7	Exploitation and Dissemination	UTIA Jiri Kadlec	M01	M36

4. ALMARVI Technological Achievements

4.1 ALMARVI objectives

The overall objective of ALMARVI project is to provide foundation and platform solution to enable massive date rate image/video processing at low power budgets under variability conditions. The key is to leverage joint hardware-software adaptations and properties of image/video content and algorithms. Various applications scenarios from healthcare, security/surveillance/monitoring, and mobile industrial domains will be considered. Specific objectives of the ALMARVI project are:

1) ***Objective-1 – Enabling Massive Data Rate Processing:*** *ALMARVI enables massive data rate image/video processing on embedded devices executing advanced high-efficiency image/video processing algorithms. The goal is to develop:*
    - *(i) adaptive many-core execution platforms that are scalable with heterogeneous acceleration fabrics (like FPGAs, DSPs, GPUs, etc.);*
    - *(ii) application-specific image/video processing cores and coprocessors;*
    - *(iii) design tools to expedite the development flow and to enable the IP reusability;*
    - *(iv) application-specific parallelisation; and*
    - *(v) methods for joint hardware-software adaptations based on algorithm resilience properties.*
2) ***Objective-2 – Achieving Low-Power Consumption:*** *ALMARVI aims at enabling cross-domain power-efficient techniques and methods for efficient and lightweight resource and power management at various system layers, while jointly accounting for the architectural features, algorithmic properties, and image/video content characteristics. The goal is to develop:*
    - *(i) low-power architectures of image/video processing cores;*
    - *(ii) low-power adaptive algorithms with run-time quality vs. energy trade-off;*
    - *(iii) novel hardware-software-collaborative power-management techniques; and*
    - *(iv) methods to exploit algorithmic resilience for increased power-efficiency.*
3) ***Objective-3 – Composability and Cross-Domain Applicability:*** *To enable independent development of various system components and smooth integration for demonstrators from different domains. One of the key objectives of ALMARVI is to provide a cross-domain scalable platform solution with efficient design tool chain, IP reuse, composability, and system software stack for seamless interoperability, and scalability on commercially available heterogeneous acceleration fabrics.*
4) ***Objective-4 – Robustness to Variability:*** *ALMARVI targets consistent and predictable system performance and power consumption over different product categories and application domains that are subjected to variability in underlying processing hardware, communication channels, application workload behaviour, system state (available resources and energy budgets), environmental factors, etc. To achieve this, the goal is to devise power-aware scalability and adaptivity at the algorithm and system levels while exploiting inherent resilience properties of image and video processing applications for adaptive resource and power management.*

General objectives are achieving low-cost solutions, high product quality, lifetime, high yield, and high engineering efficiency which will be accomplished through design tools, efficient algorithm designs, scalable and adaptive execution platforms, support for interoperability, etc.

4.2 Cross-layer models for estimating system properties/parameters

Cross-Layer Models for estimating System Properties/Parameters describes the hierarchical performance, power and error resiliency models to be developed when taking into account properties of algorithms and the nature of the underlying hardware. Prospective uses of these performance and power models involve:

(i) tuning of architectural parameters of the adaptive cores (WP-3);

(ii) identification of the power-wise or performance-wise optimal mapping of an application kernel onto a set of cores (WP-4);

(iii) efficient resource and power management by the run-time system in the system software stack (WP-4);

(iv) developing power-aware adaptive image/video processing algorithms (WP-2).

Specifically, the work includes three models: performance, power, and error resilience which have been cross-modeled on three layers: component-layer, application-layer and multiple-applications layer. Additionally, the work includes validating the models.

Cross-layer modeling: Heterogeneous many-core ALMARVI system architecture is modeled in three layers: component-layer, application-layer and multiple-applications layer. The developed models will be utilized for characterization, optimization and trade-off analysis of streaming applications (e.g., image/video processing algorithms) and feedback control applications that will share the ALMARVI platform. The models are either at a high-level abstraction hiding implementation details or at the implementation layer dealing with the source-code level details.

Performance models: Hierarchical performance models will be developed across the layers. At the component-layer, models of computational components (computation, communication, storage) will be developed towards analyzing timing properties such as worst-case execution time (WCET) and response time. At the application-layer, higher-level performance metrics of the applications, i.e., throughput and quality of control (QoC) will be analyzed utilizing the component-layer models. Further, our models will answer questions related to resource requirements and various tradeoff between performance and resource requirements. At the multiple-applications layer, utilizing the models of components and applications, questions related to resource sharing between streaming applications and feedback control applications will be addressed. Design space exploration will be performed towards obtaining the optimal resource allocation and platform configuration.

Power models: Parallelization tooling, for C and C++, targeting OpenMP and Open-CL will be developed for homogeneous shared-memory multi-cores. Using the source code of a given image/video processing algorithm, the developed tools will parallelize code, analyze and profile the application behavior.

Error resilience models: In this context, low-voltage operations and channel fluctuations typically induce errors in the algorithm output. Normally these are handled by margining; the system is operated at a voltage / frequency point which guarantees correct operation in worst-case situations. In all other situations power is lost. However, many applications can handle small error (e.g. LSB-level noise in a DCT). The error resilience properties will be modeled on the component layer as a function of correct output under ultra-low-power operational modes and lossy channels. Further, we will investigate the effect of such errors on the feedback control applications and how the feedback control algorithms can be designed/adapted in an error-resilient fashion in order to tolerate such hardware errors.

In the table below, we summarize our overall modeling approaches adopted in this project. The models use mainly two abstractions: First, the models derived based on the source codes running on certain computation platform, e.g., the experimental execution times of a code on a given platform. Second, the higher level of abstraction based on a set of given source-code parameters, e.g., throughput analysis for a given task graph with execution times. Obviously, the parameters from the first category of modeling might be used in the second category. Further, the models are validated either experimentally or analytically.

Performance models		Abstraction		Validation		Tooling
		Source-code/ Model	Higher level	Analytical	Experimental	
Cross-layer	Multiple-applications -layer		TUE	TUE		SDF³
	Application-layer	VF	TUE	TUE	VF	SDF³+TRACE + Pareon tool
	Component-layer		TUE	TUE		SDF³
Power models		**Abstraction**		**Validation**		**Tooling**
		Source-code/ Model	Higher level	Analytical	Experimental	
Cross-layer	Multiple-applications layer					
	Application-layer	VF			VF	Pareon tool
	Component layer				UTURKU	Spice
Error-resilience models		**Abstraction**		**Validation**		**Tooling**
		Source-code/ Model	Higher level	Analytical	Experimental	
Cross-layer	Multiple-applications layer					
	Application-layer		TUE	TUE		Matlab
	Component-layer	UTURKU			UTURKU	Spice

4.3 Algorithms architecture and design

This chapter describes research and development of the methods and advanced algorithms for image and video. The key focus is on developing algorithms for massive data-rate image/video processing and analysis, tailored towards the specific ALMARVI requirements and system specifications to support for performance and power scalability. The high-performance and low-power algorithms are amenable to parallelization using ALMARVI multi-/many-core execution platform instances (including CPU + GPU/FPGA + TTA + rVEX). Based on this, parallelized low-power algorithms and implementations for different image/video processing functions will be consolidated in this deliverable. Furthermore, the work performed in this document is aimed at presenting challenges in the low quality images/videos in surveillance and healthcare applications. Therefore, solutions to the enhancement and evaluation of the image/video quality are proposed. Different advanced image segmentation, feature extraction, and clustering methods will be utilized.

One of the ALMARVI project's main goals is to develop adaptive, scalable, and parallelized algorithms for massive data-rate image/video processing and analysis to enable high-end services in key industrial domains, such as healthcare, security/surveillance, and mobile devices. This development enables to combine the power and performance challenges at all system layers, i.e. hardware, system software, and algorithms, while adapting each system layer to react to the run-time changing scenarios of available energy budgets, resources, user-defined constraints, etc.

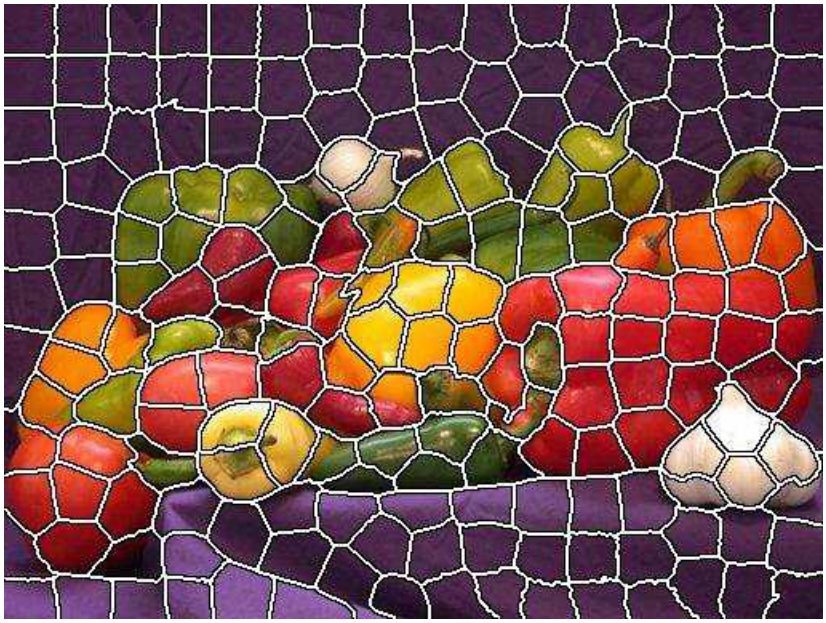
The key focus is on developing algorithms for massive data-rate image/video processing and analysis, tailored towards the specific ALMARVI requirements and system specifications to support for performance and power scalability. To increase the potential of parallelization, the architectural features of the heterogeneous acceleration fabrics will be exploited in designing these algorithms using ALMARVI multi-/many-core execution platform instances (including CPU + GPU/FPGA + TTA + rVEX). Based on this, parallelized low-power algorithms and implementations for different image/video processing functions will be consolidated in this deliverable. Furthermore, scalability will be considered as an important design consideration in these image/video segmentation, feature extraction, and clustering algorithms.

Currently, several methods have been developed for image/video segmentation. Therefore, it is necessary to be able to evaluate the performance of image segmentation algorithms objectively. Image segmentation is the process of partitioning a digital image/video into multiple segments (set of pixels, superpixels). The goal of

segmentation is to simplify and/or change the representation of an image/video into something that is more meaningful and easier to analyze. Correspondingly, the goal of feature extraction is to extract useful information from an input image/video. The image/video is transformed into a reduced set of features (also named a "features vector"). The extracted features are expected to contain the relevant information from the input data, so that the desired task can be performed by using this reduced representation instead of the complete initial data.

The goal of clustering is to group a set of objects from an image/video in such a way that objects in the same group (called a “cluster”) are more similar (in some sense or another) to each other than to those in other groups (clusters). It is a main task of exploratory data mining and a common technique for statistical data analysis, used in many fields, such as image/video analysis, machine learning, pattern recognition, information retrieval, and bioinformatics. In this respect, the research and development of algorithms for image/video segmentation, feature extraction, and clustering will bring significant additional value to the goals of ALMARVI project.

E.g. Simple Linear Iterative Clustering (SLIC) method is a superpixel algorithm that can be leveraged for effective and efficient image segmentation. Superpixel algorithms group pixels into perceptually meaningful atomic regions that can be used to replace the rigid structure of the pixel grid is presented in figure below.



Parallel Object Recognition and Tracking, Motion Analysis Algorithms provide architecture and design details including architecture and design figures for various algorithms for object recognition and tracking, motion analysis, behavior analysis, etc. As for the algorithm architecture is concerned, all of the ALMARVI participants took the approach of integrated processing of the object detection and/or recognition fused with the tracking and motion analysis. This trend corresponds to modern trends in tracking where “tracking by learning” that uses aspects of object detection as well as the more traditional tracking techniques. Acceleration and parallelization aspects were evaluated as well wherever applicable.

Also architecture and design details for algorithms for image enhancement and image restoration were researched. The development of adaptive, scalable and parallelizable next-generation algorithms for image/video restoration and enhancement is one of the key objectives for image processing. The research describes architecture and design details of algorithms for image enhancement, image restoration, and image fusion, having in mind their scalability, parallelised design, and power-aware scalability. These three categories represent main approaches for improvement of the image/video content.

The quality increase can be driven by aesthetic or application oriented criteria, then we speak about image/video enhancement or fusion. The other possibility, image restoration, aims to go beyond camera resolution, recover missing or corrupted information (typically high frequency components suppressed by low pass filtering or corrupted by aliasing) that may not even be visually present in any of the input images. Both categories are included.

Input images can be captured by different types of sensors, from different angles, under varying illumination conditions or exposure settings. The output image or video enhancement/restoration should have better perceivable image content to the end user. The key issues to be addressed in the ALMARVI project are characteristics of the algorithms with respect to their scalability, to their ability to be parallelized, and to their power consumption. The overall goal is to develop novel parallelizable algorithms (techniques such as combined multi-channel multi-frame video enhancement, denoising, and de-convolution; image fusion) for selective image enhancement and restoration, which will lead into higher quality. The multimodal information (near infrared) is incorporated too, in order to increase the output quality even more. There are three categories of methods described:

Image enhancement

- Multigrid non-linear anisotropic diffusion filtering optimized denoising method for microscopic cancer images

- Non-local means algorithm (NLMS) noise reduction algorithm implementation and testing of the algorithm for mobile devices

Image restoration

- Space-variant deconvolution computationally effective space variant version of the deconvolution

- Multi-frame super-resolution space variant version of the super-resolution and demosaicing

Image fusion

- Thermal and day light camera fusion fusion method based on combination of multiple single channel images with RGB colour characteristics of a similar looking scene

Proposed algorithms are addressing full range of image enhancement/restoration/fusion tasks. The goal to improve the quality of the input image/video data in order to create better perceivable image content to the end user was fulfilled in all cases. The adaptability, scalability, and parallelizability were discussed and tested.

4.4 Abstracting heterogeneous hardware architectures

The Almarvi project aims to develop an approach that allows for portable application software, across a range of modern high performance and energy efficient heterogeneous computing architectures. An important aspect of portable applications are the APIs and libraries used to express the application functionality. This chapter describes the selected APIs for the Almarvi project, which allow to interface and integrate heterogeneous acceleration fabrics: A software layer abstracts from details in the heterogeneous accelerators, enabling re-use through design methods and tools. The Almarvi software stack is mainly based on OpenCL. Choosing OpenCL as baseline set of APIs greatly helps in creating portable applications, both towards our novel Almarvi target platforms, as well as other (classic and novel) off-the-shelf 3rd party platforms. That is because:

- OpenCL is growing in popularity for programming heterogeneous devices, in particular in the embedded world.
- The OpenCL programming model is -in general- a good match with the application domains that Almarvi targets, on high-performance imaging and video analysis and mobile multi-media.

OpenCL will be the key mechanism to allow portability of applications over a range of potential target architectures in the Almarvi project. Beyond the classic application partitioning between a host CPU and a GP-GPU extension card, Almarvi studies the use of recent OpenCL-2.0 extensions: shared-memory architectures, pipelining, and abstract accelerators, and utilization in embedded computer architectures. The project use-cases allow to explore these options and report on their merits for different application domains. This work is partially supported

by developments in OpenCL itself, in the context of the pocl open-source implementation. Next to (and in combination with) OpenCL, application concurrency is deployed through OpenMP, which is a convenient and efficient parallelization method to exploit multi-core homogeneous host processors. As we focus on OpenCL, the project will not adopt recent OpenMP extensions from its version 4.0 that also target heterogeneous architectures. Finally, vectorization is a well-known mechanism to obtain throughput improvements in combination with power efficiency. It is deployed in a few applications, but most prominently visible in the wide-data-path transport triggered architecture (TTA), where the code portability issues are addressed by challenging compiler auto-vectorization.

4.5 Scalability, quality and usability of the execution platform

The ALMARVI executions platforms cover a wide spectrum of flexibility and customizability for the ALMARVI applications. This deliverable described the various quality aspects of the three ALMARVI specific hardware platform configurations developed by Nokia, TUT and UTURKU; by the TUDelft; and by UTIA. Four quality aspects for each platform have been discussed:

- performance improvements,
- power/energy efficiency,
- scalability, and
- usability.

In the following, a number of these aspects are discussed for a couple of applications on the platforms to show the ALMARVI targets that have been achieved so far.

Nokia/TUT/UTURKU platform

With regards to performance, the processor core designed within the ALMARVI project (called LordCore) is based on the Transport Triggered Architecture (TTA) paradigm. To enable high performance computation, it contains a 512bit wide SIMD datapath, which is able to calculate 32 lanes of 16-bit wide half-precision floating-point values in parallel. The architecture is also scalable to other SIMD widths: 8- and 16-lane versions were also developed for lower performance usage such as for LTE-M applications.

With regard to efficiency, the two core version of the processor for the LTE receiver was synthesized with Synopsys Design Compiler version I-2013.12, using 28nm Fully Depleted Silicon-On-Insulator (FDSOI) process technology. The estimated power consumption for the two-core version is 137 mW (MMSE) and 163 mW (LORD), which are well below the targeted 1W performance boundary. Extrapolated power consumption for a four-core version is about 270 mW, which would deliver the targeted LTE device category 11 performance of 600 Mbit/s. For the audio signal processor, the estimated power consumption for two-core version is 300 µW, which is well below our targeted 1mW upper boundary.

With regard to scalability, the architecture of the processor was designed to be scalable in various ways. Most notably, the memory architecture allows both efficient scaling to multiple cores, as well as to multiple SIMD widths of the processor, and the number of functional units.

With regard to usability, the platform can programmed using the OpenCL language for the LTE receiver. OpenCL allows using vector data types to execute the same code on many SIMD lanes of the processor, and also allows easy parallelization of the workload over multiple cores. The audio processor, on the other hand, can be programmed in the C language. C was selected due to the size of the application and possibility for lower level control than OpenCL.

TUDelft platform

By implementing a new VLIW instruction encoding, the performance of the rVEX processor has been increased by up to a factor of three while maintaining compatibility with the processor's dynamic parametrizability.

With regards to scalability, the rVEX processor is both run-time parametrizable and design-time configurable (using VHDL generics). Therefore, scalability is a metric in the design-space when choosing the right parameters for the application.

On the usability level, the rVEX platform has matured immensely since the start of the ALMARVI project. In 2015, a new core has been redesigned with various improvements: 1) advanced debugging hardware/software, 2)

hardware tracing functionality and performance counters, 3) core structure and pipeline organization are now (easily) design-time configurable, 4) core is now (easily) design-time extensible, 5) number of execution lanes are run-time parametrizable, and 6) dynamic cache that supports the varying number of execution lanes of the core.

UTIA platform

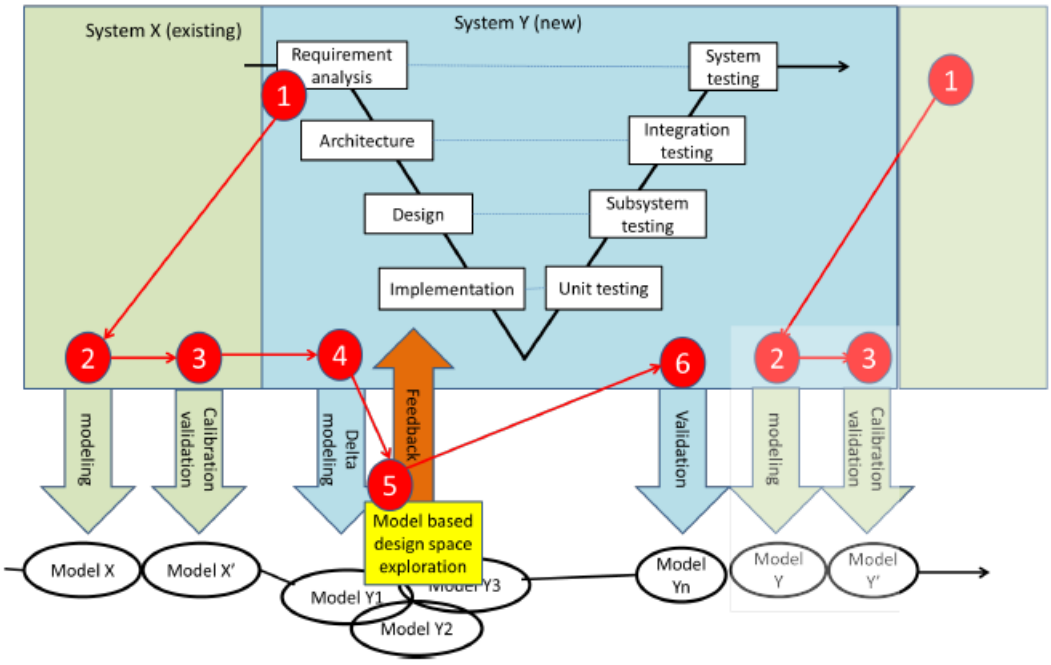
With regard to performance and to allow for real-time processing, the Full-HD motion detection algorithm had to be accelerated approximately 60 times using the 666MHz ARM processor on the programmable logic part (PL) of the Xilinx ZYNQ device. The PL logic can implement IP cores clocked at 150 MHz.

UTIA performed design exploration in the in the Xilinx SDSoC 2015.4 environment. Two parallel chains of SDSoC managed to generate accelerators (at 150MHz) delivering 57.09 FPS. This represents an acceleration of 48 times. The energy needed by the complete board to compute the Full---HD motion detection frame was reduced 45 times.

4.6 Design space exploration

The ALMARVI project aims to develop an approach that allows for portable application software, across a range of modern high performance and energy efficient heterogeneous computing architectures. The aim is to develop analysis techniques for systematic design space exploration (DSE) methods dealing with task mapping, scheduling and resource arbitration. This task is built upon the models developed in Task 1.3 to provide the right abstractions of the underlying heterogeneous hardware, applicable at the development level. The DSE targets multiple objectives, performance being the prime objective (often a constraint) in view of various trade-off between resource usage (cores, memory, cost) and embedded performance.

Design-space-exploration presented modelling, analysis, evaluation, and implementation of single design point targeting multi-processors both at the model level and at the source code level. There are a number of tools extended and used in this context by ALMARVI partners. Various results are presented showing the improvement in terms of resource utilizing the models at different levels component, application and multi applications. The application development process of ALMARVI follows the V model for performance engineering as illustrated in figure below. The following elaborates various steps in the development process.

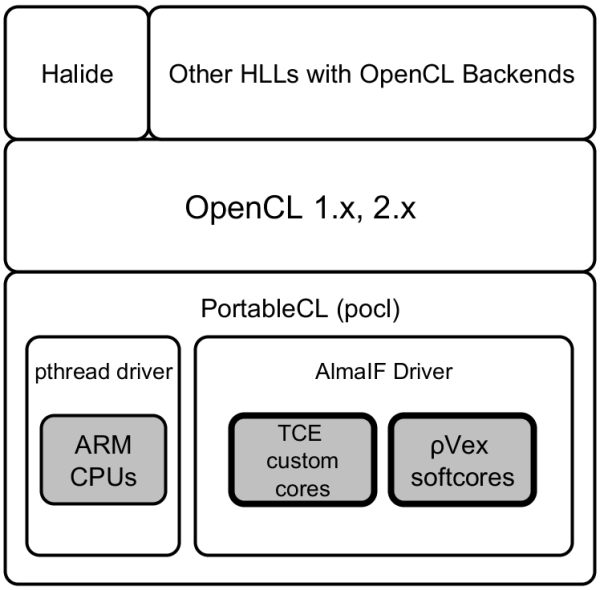


4.7 Integrated system software stack

In the ALMARVI project, a customizable image/video processing hardware platform with easy integration of components such as hardware accelerators and application-specific processors which work together in executing applications with multiple tasks in a coordinated manner was developed. The platform was designed to be programmable by utilizing OpenCL as a middleware API. On top of OpenCL, it is possible to implement higher-level programming models as long as the underlying devices support OpenCL programming. In order to make hardware integration cleaner, a common ALMARVI IP block interface called AlmaIF was developed. This interface was designed with OpenCL support in mind, specifically with the intention to make OpenCL driver development and their use from application programs more unified across the ALMARVI custom devices. The work presents integrated system software stack of ALMARVI, of which key components include pocl, an open source OpenCL implementation and the AlmaIF that provides a common control interface utilized by the AlmaIF pocl driver. We also presented the software integration of the Dyplo eco-system and programming model from Topic Embedded Products, used by Philips.

Furthermore, in this work we proposed a concept for sequencing FSM-SADF scenarios that can be analysed and executed of TDMA-scheduled platforms. The concept for programming and analysing FSM-SADF on multi-processor platforms that we presented allows to capture input-dependent application behaviour. Each scenario describes one such behaviour and can have a different topology and different rates, WCETs and persistent tokens to do so.

We also presented an implementation with three novel aspects. Firstly we glue all scenarios together by implementing the switch and select actors as (de-)multiplexers by disconnecting and reconnecting FIFO channels. The timing behaviour of these actors is modelled in the scenario graphs. Secondly we propose a rolling static-order scheduler that automatically extends the schedule each time a scenario is detected. Thirdly we map persistent tokens that are shared in multiple scenarios to the same physical FIFO. The platform-aware analysis model annotates the scenarios with the exact timing behaviour of the implementation. We extended an existing HSDF model of the platform to suit FSM-SADF analysis. A more sophisticated method for encoding SO schedules in dataflow graphs was integrated in the tool-flow to account for multi-rate actors and their DMAs. We mapped the SUSAN edge detection algorithm to a platform with two processors and found that the real throughput is slightly higher than the constraint given to the analysis tool. This proves that the model is both conservative and precise. The implementation cost in terms of timing and memory footprint is marginal. Figure below illustrates ALMARVI OpenCL-based System Software Stack.



5. Partner Demonstrators

To demonstrate and validate the project developments and results, the ALMARVI concepts are evaluated using demonstrators from three different domains:

1) Medical Imaging and Healthcare Demonstrators: Medical imaging assisted diagnosis with a mixture of real-time and non-real-time image-processing tasks for different healthcare applications like minimal invasive treatment for cardiovascular diseases.

2) Security/Surveillance and Monitoring Demonstrators: Distributed Monitoring using Mobile and Fixed Video Sensor Nodes; Continuous Monitoring of Industrial Processes.

3) Mobile Handset Demonstrators: Novel ultra-energy-efficient high performance heterogeneous multicore platform for smart phones and nomadic embedded devices of the future.

For each domain, the common researched concepts for algorithms, design tools, system software stack, and many-core execution platform will be instantiated.

5.1 Medical imaging and healthcare demonstrators

In the healthcare domain, ALMARVI has developed two main demonstrators: Interventional X-Ray and Breast Cancer Diagnostics.

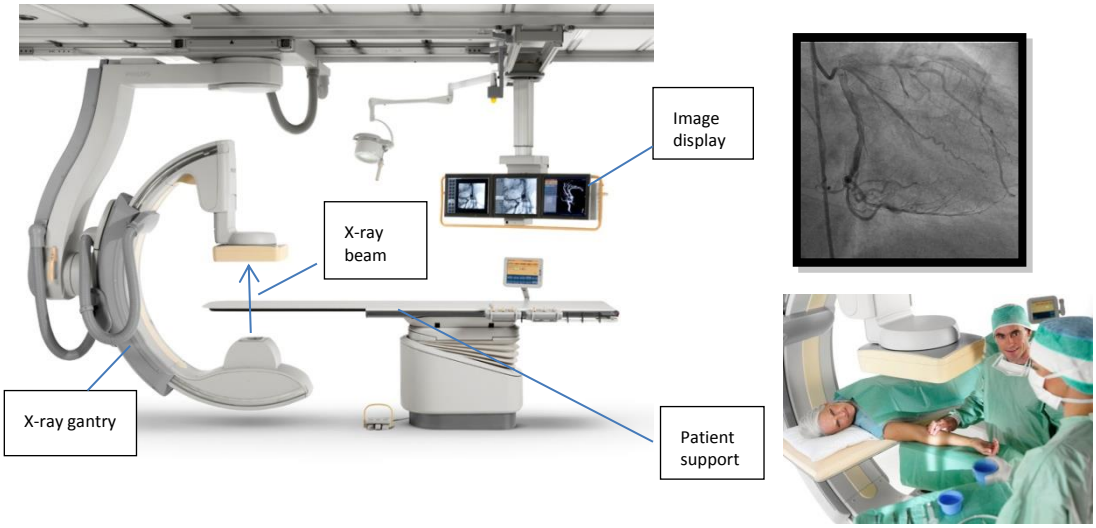
5.1.1 Interventional X-Ray

The goal of an interventional X-ray system is to provide the doctor with real-time images from the anatomy of the patient while performing a medical intervention. Typical interventions on the system include repairing blood vessel deformations such as aneurisms by positioning stents or replacing heart valves. During these procedures, blood vessels are filled with a contrast medium, which is visualized by X-rays and shown in real time as high resolution video images to the doctor.

The interventional X-ray system contains functions like the acquisition and display of an X-ray image series, the movement control of patient support and X-Ray gantry, patient administration, communication to hospital servers and many others. Though some tasks may be performed on a “best effort” base (for instance: patient administration, export of an image series to the hospital servers), others must satisfy strict real time requirements. Handling images during acquisition and movement control belong to the latter category.

Failure to meet movement control timing requirements may lead to injury to or even death of a patient, for example when the detector hits and hurts the patient. Another critical parameter is the image pipeline latency, which is the time between the irradiation of the body with X-rays and the display of the corresponding image on the display. In general no delay is perceived if the latency is 150ms or less. If the latency is significantly larger, the hand-eye coordination of the doctor is disrupted and treatment of the patient is impaired.

In general we see good opportunities to reduce classical FPGA development time by using High Level Synthesis tooling and network on chip like solutions thus increasing the engineering efficiency and at the same time making designs better testable, increasing the composability and making designs easier to maintain. In the future we wish to abstract even further from FPGA designs and enable portability from OpenCL and other heterogeneous languages, such tools are currently entering the high performance computing market and we see good opportunities for these tools for embedded devices (SoC) as well. Figure below presents an interventional X-ray system and a typical still from a X-ray video. In the inset the patient is on the patient support while an interventions is performed.



5.1.2 Breast Cancer Diagnostics

IBCAS (Intelligent Breast Cancer Analysis System) is a Matlab-based demo system for conducting image analysis experiments on histopathological breast cancer images. With this system, we have experimented image enhancement, feature extraction, segmentation, and classification methods for medical imagery. In addition, we have implemented an easy-to-use stand-alone C-based application of the system for demonstrating and concretizing how to utilize the research results in real-world medical analyzes without the need to perform any complex choices in selecting feature extraction, segmentation, or classification methods, i.e., “optimal” settings are automatically selected by the application. Moreover, we have also implemented an OpenCL-based version of the system that utilizes GPUs and its development is currently in its finalization phase. Furthermore, we have converted selected algorithms of our Matlab-based version of the IBCAS system into C/C++ for cooperation purposes aiming at running them on TUDelft’s rVEX platform, which is a reconfigurable and extensible Very-Long Instruction Word (VLIW) processor that is part of the overall “Liquid Architectures” research theme within the Quantum Engineering Lab at TUDelft, The Netherlands. The rVEX processor architecture is based on the VEX ISA.

Histopathological image analysis continues to be a very challenging task. The small variations in images (tissue, cells, and coloring technique) poses problems for accurate segmentation, feature extraction, and classification. These small variations makes it difficult for classifying malignant cancerous tissues from the image. The methods we were researching in ALMARVI project were not trivially parallelizable, because these methods require multiple iterations and also each iteration is dependent on the results from the previous iterations. However, despite of these challenging problems and obstacles, we managed to get promising results with lot of potential for future research work.

5.2 Security/surveillance and monitoring demonstrators

In the security domain, ALMARVI has developed five main demonstrators: large area video surveillance, road traffic surveillance, smart surveillance, logical analysis of multimodal camera data and the protection of walnut tree harvest against birds.

5.2.1 Large Area Video Surveillance

In large area video surveillance applications with many cameras, we apply an optical flow algorithm to all camera streams. If optical flow flags no motion, the respective stream is not dispatched to a server for further processing nor stored on disk. If enough motion is detected, then the respective stream is dispatched to servers for further processing in order to detect certain objects and is stored on disk. Also, the respective stream undergoes more precise optical flow analysis until there is again not enough motion. The vectors produced from optical flow are not only used to switch between these two modes but also are used by the specific detection and recognition algorithms run on the servers. Hence, optimized implementation (GPU and FPGA based) of optical flow in terms of usage of hardware resources and power, while being able to support high throughput and resolution (for the second mode after motion is detected), was critical in this demonstrator.

Because of the significant size of our FPGA designs, almost all the problems that can be encountered in an FPGA design and implementation process have been encountered. Seeking solutions to these problems has given the team considerable experience. Serious problems were encountered in the places shown as new methodologies above. Therefore, we have discovered that new design methods should be developed in the relevant stages. Especially multi-cycle floating point units may not be utilized all the time; turning these units off during periods when they are not active reduces power consumption. Another important factor in power consumption is the use of DRAM. During the early phases of our development work, DRAM bandwidth was not efficiently utilized in most cases. We had to optimize our DRAM traffic to lower power consumption.

5.2.2 Road Traffic Surveillance

This demonstrator includes real-time object detection in HD video using a low cost and low power compact system. Such embedded system will be used for road traffic surveillance applications as counting of vehicles, license plate detection (for further recognition) and many others. The core of the system – the object detector – is based on the WaldBoost algorithm and is trained for cars' license plates. We met our objectives by designing and manufacturing an embedded camera based on Xilinx Zynq SoC. This demonstrator is configured to detect licence plates and is using classifier trained in cooperation with BUT. The object detection is fully HW accelerated in FPGA and is performed by the IP core of Waldboost based detector (developed by BUT). Results of the detection are processed by ARM cores of Xilinx Zynq (drawing rectangles as results of the detection) which are also setting up and driving the CMOS sensor (Python 2k CMOS from On Semiconductor). CMOS data output is attached directly to the FPGA. The camera is also equipped with a H.264 hardware encoder from Fujistu. Resultant camera output enhanced by detection results is compressed and streamed in MPEG-TS over the Wi-fi and Ethernet network and can be displayed by most of the media players with H.264 decoding support.

The most efficient way of reducing power consumption of an image processing system is avoiding regular industrial/personal computers with connected cameras and going for an embedded solution. Using a FPGA we are able to build the all-in-one solution including an image sensor and all the controlling logic. Such a low power system has ability of autonomous operation including accelerated/parallelized image processing e.g. object detection. This is the key for a further application of cameras for traffic monitoring e.g. battery/solar powered sites on roads or vehicles equipped with multicamera systems for mobile surveillance.

Using HDR image capturing (simple multi-exposition or more advanced composition) has many advantages when lightning or weather conditions get worse. This is not rare case and sometimes strong sunlight or rain makes the system blind unpredictably. Not with HDR on board that shows the processing algorithms more details. This definitely allows better results in real scenarios. Figure below presents camera based on Xilinx Zynq.

5.2.3 Smart Surveillance (HSS)

We managed to build a pedestrian detection and counting system with off-the-shelf components. The algorithm running on the Raspberry Pi runs OpenCV based algorithm to determine if there's anything interesting (i.e. large moving object) in the image. We came to conclusion that by using a pre-trained classifier running on an external server to combine all the detection results, it's possible to count multiple unique pedestrians across multiple camera systems by running the same basic detection algorithm.

The objective was to identify and count persons who are entering or leaving an area or gateway. The server collects data from the detections including but not limited to the timestamp, location and the running count of detections from multiple units with cameras. Some cropped images will also be saved for external usage (such as showing the detection on GUI). No video or full size frames are stored from the camera streams.

When working on HSS system development we learned that it is possible to utilize off-the-shelf components and open source software libraries to build a system for tracking and counting pedestrians with multiple camera modules. The components were selected focusing to the portability and low power requirements, which was a learning experience on its own. We also learned that some of the used open source algorithms were combined in a way that had been used before in different purposes such as background subtraction with hog detection

5.2.4 Logical Analysis of Multimodal Camera Data

The ALMARVI contribution of VTT addresses the security/surveillance and monitoring demonstrator with an application use case that focuses on logical analysis in a multi-camera surveillance system. The logical analysis entails the usage of camera data to extract machine level data, and analyses this data to support the surveillance system users in the scope of the ALMARVI platform.

The demonstrator focuses on multi-camera object recognition and tracking in cameras and between cameras to get 3D context awareness of the person's motions. The demonstrator utilizes many camera nodes that are connected to a central processing unit, and therefore presents distributed and parallel processing. The demonstrator presents algorithms such as automatic calibration to get sense of 3D space, and person recognition and tracking. The challenge is to get the 3D sense and object recognition tracking accurate enough and tackling performance challenge of integration of many video streams.

The demonstrator utilizes many camera nodes that are connected to a central processing unit, and presents algorithms such as automatic calibration to get sense of 3D space, and person recognition and tracking. The demonstrator is up and running, a 3D sense is created, and persons are identified and tracked from camera to another.

- The system is able to extract machine level data, and analyse this data to support the surveillance system users in the scope of the ALMARVI platform.
- The system draws 3D/depth information from normal video feed.
- The system is configurable to existing camera systems, and requires no extra devices to be acquired.
- The system utilizes different types of camera information from many cameras
- The 3D sense and object recognition tracking accuracy is satisfactory
- Performance challenge of integration of many video streams with heavy algorithm calculations remains to a degree, but the developed algorithms are made to be parallelizable which provides performance boost.

5.2.5 Protection of Walnut Tree Harvest Against Birds

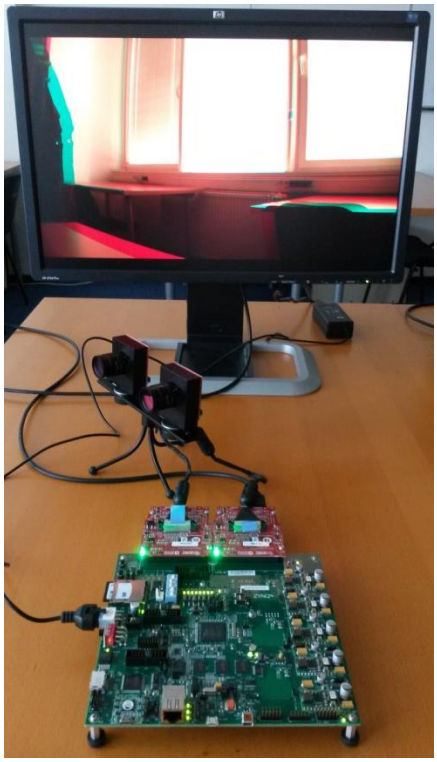
The UTIA demonstrator focuses on implementation of the execution platform with hardware acceleration of image/video processing algorithms as well as on design methodology used for implementation of algorithms for the platform to be used in the ALMARVI demonstrator Walnut Harvest Protection against birds. Special attention is focused on portability of results between FPGA boards, low power and high performance requirements posed on video processing solutions. We are demonstrating capabilities of our platform by implementing three hardware accelerated algorithms. They can be demonstrated individually or implemented together if ported to bigger FPGA device. The first algorithm is a simple Sobel filter based motion detection, the second is background subtraction derived from OpenCV and the third is object detection derived from the OpenCV implementation as well.

The SDSoC tool provides unprecedented possibility to create hardware accelerators written in “C/C++” language and to automatically generate complete dataflow network around them. The implementation of complex image or video processing for FPGA, however, can be still complex and demanding task. Thus the availability of image/video processing library with HLS synthesizable functions is a key to rapid development. The support for Open Source OpenCV library was tested by our team. We have found that the integration of system composed of Xilinx Video library cores (HLS synthesizable OpenCV cores), hand coded cores and SDSoC tool provides efficient design flow with possibility to debug in software, test in hardware and to incrementally extend the hardware accelerated part of the design.

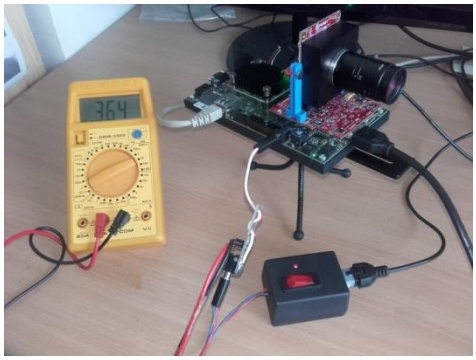
On the other hand the SDSoC tool introduces limitations on accelerator interfaces and global memory accesses. In our development we have also bumped into some errors in Vivado HLS itself (for example: if the child class of original template class is derived and in source code there is one instance of that child, all HLS pragmas, which are located in constructor of parent class are followed properly. But when the second instance of that class is created, all pragmas seem to be ignored by HLS for that instance).

We have also learned a possibility to take stream processing cores from SDSoC and to make from them pure Vivado HLS based with input/output stream. This allows using them also in FPGAs outside Zynq family. However, we cannot recommend to compose larger systems of such cores since it is almost impossible to foresee all buffering and handshake back pressure effects to integrate them all reliably to one system.

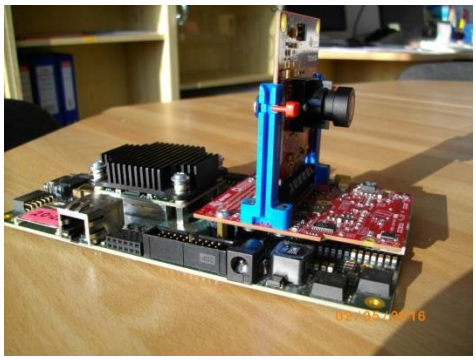
In the development of the object detection algorithm we have found that the different cascade classifiers lead to different optimal number of fully pipelined stages in terms of resources used per FPS speed improvement achieved. We have implemented different combinations of stages for bird and face detector cases. In figures below: hardware used for camera platform alternatives in ALMARVI.



a) Stereo Vita 2000 Camera Sensors (synchronous capture), ZC702 Development board



b) ALMARVI Python Camera Platform (APCP), Power measurement setup



c) Toshiba Full HD Camera, TE0701-5 Carrier with TE0715-3-30 SoM

5.3 Mobile Demonstrators

We have applied the ALMARVI design methodologies for selected use cases of mobile handset domain and demonstrated customized architectures for the domain using tools designed in ALMARVI. Further, the demonstrator has shown that the developed toolset and methodologies can be leveraged across multiple domains.

Objective was to demonstrate application of ALMARVI tools to cross-domain and scalable execution platform implementation for multimedia and radio processing that can serve different product categories in performance, power, development time, and cost constraints ranging from Radio and Imaging use cases to ultra-low power wearable computers. ALMARVI approach supports a broad selection of heterogeneous acceleration fabrics such as CPUs, DSPs, GPUs, FPGA, but in addition application specific programmable co-processors etc. to provide a good trade-off between the performance/throughput, energy efficiency and reuse via programmability. All targets should be supported from the same C/C++/OpenCL application source code. The selected algorithm benchmarks were used to set the minimum performance requirements for the implemented solutions

Customized processors provide a middle ground between fixed function accelerators and generic programmable cores. They bring benefits of hardware tailoring to programmable designs, while adding new advantages such as reduced implementation verification effort. The hardware of customized processor is optimized for executing a predefined set of applications, while allowing the very same design being used to run other, close enough routines by switching the executed software in the instruction memory. The degree of processor hardware tailoring is dictated by the use case and the targeted product.

In any case, the processor customization process is highly demanding and error-prone, with high non-recurring engineering costs. Moreover, as the design process of customized processors is usually iterative in nature, porting the required software program codes to new processor variations needs either assembly language rewrites or retargeting the compiler. One approach to simplifying the processor customization process is to compose the processor from a set of component libraries and other verified building blocks, thereby reducing the required

verification effort. The software porting problem can be alleviated with automatically retargeted software development kits.

TTA-Based Co-Design Environment (TCE) is a processor design and programming toolset which is based on a processor template that supports different styles of parallelism efficiently. TCE enables rapid design of cores ranging from tiny scalar microcontrollers to multicore vector machines with a resource oriented design methodology that emphasizes reuse of components.

5.3.1 Image Segmentation and LTE receiver

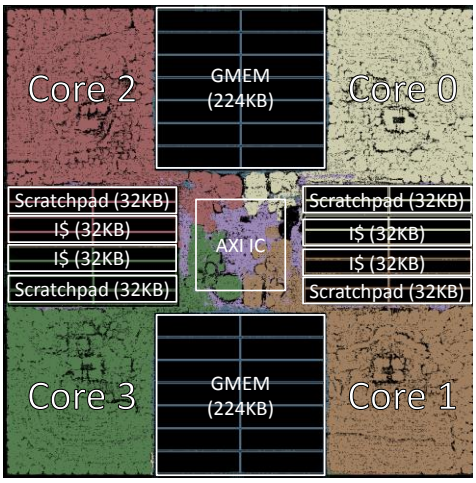
4G LTE is a standard of high-speed, low latency, data for wide-area cellular communications, and builds upon the technologies developed by 3GPP project. The most demanding and compute intensive algorithms in modern radio receiver relate to signal detection and demodulation. MIMO technique employs multiple transmitter and receiver antennas for transmitting multiple parallel data streams. As the system employs M x N different paths for the signal higher diversity, more reliable communications and higher throughput is achieved. The expense to be paid is the higher receiver complexity with exponential complexity increase to the MIMO and modulation order. LTE provides 10 device categories up-to 452.2 Mbit/s downlink data rate with 4 MIMO layers, and up to 102 Mbit/s uplink data rate with 2 MIMO layers in a single 20 MHz LTE carrier. For 5G systems researchers have suggested massive MIMO approaches, meaning tens of parallel antennas and parallel data streams. State-of-the-art implementations even for 2x2 MIMO receivers employ either custom HW or vector DSPs for signal processing. For implementer, it means RTL level coding or assembly level coding with compiler intrinsic extensions. High-level programming languages or standard OpenCL based programming models are not supported, meaning that there is no portability of implementation, yet alone scalability.

In the second co-processor design, we targeted audio signal processing in a wearable, always-on type of a device. The processor is implementing audio signal processing algorithms such as IIR bi-quads, linear filters, spectral analysis and adaptive filters. The main use case for the audio processing is isolation of the headphone user from environment with active noise cancellation. For realistic virtual reality immersion, the audio plays very critical role as sound-scene will inform the user to which direction to look. The input sample rate of such systems is quite modest compared to wideband radio transceivers, but the use-case requires extremely low energy consumption as well as very short processing latency of below (1/48000) seconds. The main target was to implement the low latency algorithms, such as active noise cancellation, in programmable TTA co-processor with significantly lower latency and with lesser power consumption than with traditional CPU.

The FFT as well as image segmentation use cases were implemented as OpenCL source codes. However, Nokia internal customer finally dropped these requirements and clear focuses were to focus on MIMO Detection from the high-throughput scenario perspective and to audio signal processing from ultra-low power wearable perspective.

As always, the devil is in the details. It is very rapid to do the initial designs for programmable co-processors; however, it takes quite a long time to mature the designs and go through the design iterations. The “final” iteration thus the layout and synthesis of IC in our case still revealed architectural bottlenecks and even bugs, which were unobserved in the FPGA design. One must reserve at least a same amount of time to IC design as for architecture and functional verification on FPGA.

LTE MIMO DETECTION: The four-core ASIC configuration of the LTE processor was synthesized and placed and routed with Synopsys tools using a 28nm Fully Depleted Silicon On Insulator (FDSOI) process technology. Clock gating and multi-threshold voltage optimizations were enabled in synthesis. Operating conditions were set to 1V supply voltage and 25◦C temperature. The routed design achieves a clock rate of 968 MHz and has a cell area of 2.47mm2 at utilization of 71%. A layout image is shown in figure below.



5.3.2 Image and video Enhancement

This demonstrator focuses on implementation of image enhancement algorithms for standard of the-shelves mobile platform. The main purpose of the demonstrator is to find performance and power consumption differences between various heterogeneous processing units available on the platform, namely ARM CPU, ARM NEON, and GPU. The non-local means image de-noising software implemented and optimized for different processing units is used for experimenting processing speed and power consumption to asses ALMARVI objectives 1 and 2. In addition design and software portability related to ALMARVI objective 3 is considered and robustness to variability (ALMARVI objective 4) is discussed.

The most important findings that were made with this demonstrator were:

- Mobile GPU is more efficient in terms of power consumption for certain kind of image processing task compared to multi-core mobile CPU.
- OpenCL API is useful for mobile imaging and can be supported by the modern mobile platforms.
- ARM NEON SIMD is relatively efficient in terms of processing speed, but when used in multi-core scenario it can consume more power compared to mobile GPU.
- Power consumption of the COTS SoC platform (Snapdragon 820 in this demonstrator) is higher than expected when memory intensive algorithms are run (access to image pixels).
- Standard APIs (ANSI C/C++, OpenCL, ARM NEON) are very useful when implementing software solutions for different mobile platforms. The same implementation can be run on different generation of the devices.

6. Public Deliverables

Deliverables will be produced at the end of each work package and milestone. Deliverables at the end of milestones illustrate the project achievements in a periodic way. Deliverables will also be produced during the WP development in order to ensure incremental updates, integration, and validation. The following deliverables are publicly accessible.

Del. No.	**Deliverable Name**	**WP No.**	**Lead Partic.**	**Nature**
D7.1	*Project Website and Initial Project Presentation*	7	UEF, TUE	O, R
D7.3	*Dissemination Plan and Strategies*	7	TUE	R
D1.3	*Cross-Layer Models for estimating System Properties/Parameters*	1	UTURKU	R
D2.4	*Parallel and Power-Aware Image Segmentation Algorithms (Architecture and Design)*	2	UTIA	R
D2.5	*Parallel Object Recognition and Tracking, Motion Analysis Algorithms (Architecture and Design)*	2	CAMEA	R
D2.7	*Parallel Image Enhancement, Restoration, and Fusion Algorithms (Architecture and Design)*	2	UEF	R
D3.3	*Abstracting heterogeneous hardware architectures*	3	TUT	R
D6.4	*Progress Efficiency Report-1*	6	PHILIPS	R
D7.6	*Dissemination Report (Intermediate)*	7	TUE	R
D3.5	*Scalability, quality and usability of the execution platform*	3	TUDelft	R
D4.3	*Design Space Exploration*	4	TUE	P, R
D4.6	*Integrated System Software Stack*	4	PHILIPS	P, R
D5.7	*Evaluation of the ALMARVI Demonstrators*	5	PHILIPS	R
D6.5	*Progress Efficiency Report-2*	6	PHILIPS	R
D6.9	*Final Project Report*	6	PHILIPS	R
D7.7	*Dissemination Report (Final)*	7	TUE	R
D7.8	*ALMARVI Project Booklet*	7	VTT	R
D7.9	*Standardisation Efforts*	7	PHILIPS	R

7. Dissemination Activities

The ALMARVI project performed at a large-scale world-wide dissemination of the project results and development. The dissemination activity was driven by the entire ALMARVI team in order to ensure a high-degree of visibility and awareness of the project concept, development, and results in order to reach the largest possible number of stakeholders like industrial partners, partner chains like suppliers, manufacturers, technology providers, etc. for the systems based image/video processing; suppliers of hardware/software components like RTOS, algorithm libraries, etc. Additional stakeholders are: academic partners, organisations, research organisations, educational policy makers, students, etc.

The ALMARVI project team has devised two main dissemination groups to achieve a large-scale world-wide dissemination of the project results and development:

1) Industrial Partners (LEs, SMEs): these partners in the industrial chain (suppliers, technology providers, manufacturers, solution provider, service providers, HW/SW component suppliers, tool suppliers, etc.), will focus on dissemination through application in their product lines, demonstrations to interested parties, exhibitions, tradeshows, market contacts, demonstrator workshops, exhibitions, etc. Since ALMARVI is an industry-driven R&D project, this group will be an important target in the ALMARVI dissemination plans.

2) Research Partners (universities, research organisations): a high-grade contribution to the research, development, and technological communities through top-class publications in premier conferences and journals, technical workshops, keynotes, invited talks, special sessions at premier conferences and forums, panels, exhibitions, ARTEMIS events, etc.

8. ALMARVI Partner Logos