

2PARMA

PARallel PARadigms and Run-time MAnagement techniques for Many-core Architectures

KEYWORDS: Parallelisation and Programmability, Continuous Adaptation, Virtualization, Design Space Exploration, Run-time Resource Management, Multi-processor System-on-Chip Architectures



Project Coordinator

Name: Prof. Cristina Silvano
Institution: Politecnico di Milano
Email: silvano@elet.polimi.it

Project Technical Manager

Name: Prof. William Fornaciari
Institution: Politecnico di Milano
Email: fornacia@elet.polimi.it

Project website: <http://www.2parma.eu>

Partners:

Politecnico di Milano (IT)
STMicroelectronics (IT)
Fraunhofer – HHI (DE)
IMEC (BE),
ICCS (GR)
RWTH Aachen University (DE)
CoWare-Synopsys (BE)

Duration: 36 months

Start: 2010.01.01

Total Cost: €3 993 107

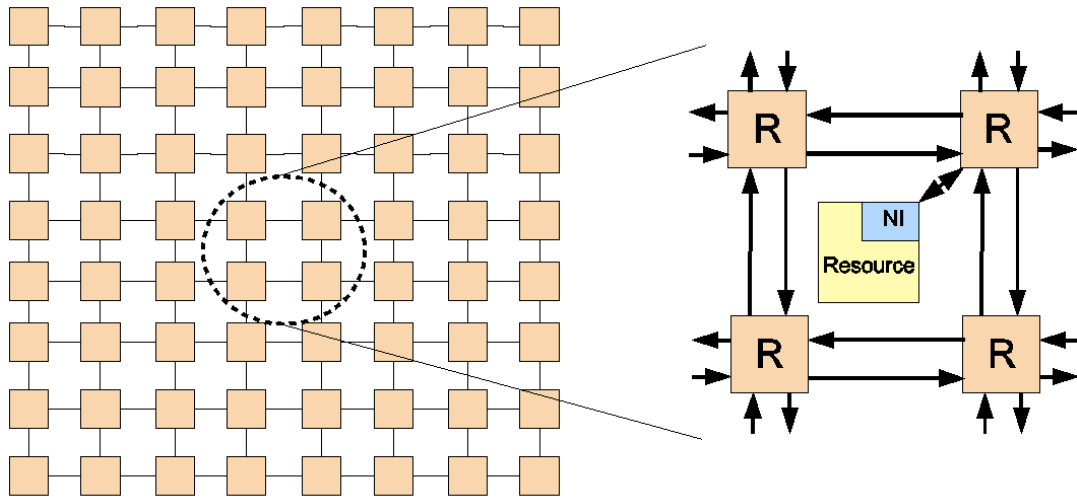
demonstrate methodologies, techniques and tools by using innovative hardware platforms provided and developed by the partners, including the “**Platform 2012**”, an early implementation of Many-core Computing Fabric provided by STMicroelectronics, and the ADRES-based many-core **COBRA platform** provided by IMEC. To ensure a wide range of application scenarios comprising the typical computation-intensive workload of a general-purpose computing system, a set of industrial high performance demanding applications will be used and customized by using the developed techniques. The selected applications are: Scalable Video Coding (HHI), Cognitive Radio (RWTH), and Multi View Image Processing (IMEC). In concrete terms, 2PARMA project has the following main goals:

2PARMA will focus on parallelisation, virtualisation, design space exploration and run-time management of many-core architectures

Programmability of Many-core Computing Fabrics: 2PARMA project tackles the issue of programmability of Multi-core Computing Fabrics at both the programming language and Operating System level. On one hand, it leverages the increasingly popular Component-Based Software Engineering and develops parallelism extraction techniques to identify opportunities for parallelisation at a high level in the design phase. Then, 2PARMA employs extensions of existing standards for parallel programming, such as OpenCL, to express data parallelism for Many-core Computing Fabrics. On the Operating System level, 2PARMA provides the means to define and deploy logic peripherals to the Many-core Computing Fabric, preserving isolation among them and efficient communication between host and Computing Fabric.

Main Objectives

The 2PARMA project focuses on the definition of suitable parallel programming models, instruction set virtualisation, run-time resource management policies and mechanisms as well as design space exploration methodologies for Many-core Computing Fabrics. The 2PARMA project will



2PARMA Many-Core Computing fabric template

Virtualisation and Continuous Adaptation:

portable bytecode representations of software are employed in 2PARMA project to provide not only portability but also the ability to adapt applications at runtime to the available resources. Instruction set virtualisation provides the means to tailor the application to the subset of the computing resources given by the coexistence of multiple applications on the computing platform, by employing dynamic compilation and optimisation techniques.

Runtime Management: given the opportunities for adaptation of the application to the available resources, 2PARMA project develops intelligent policies to manage the system resources taking into account the Quality of Service requirements imposed by the application, while optimising the resource usage for system-wide performance and energy consumption goals.

Design Space Exploration: continuous adaptation and run-time management require large amount of information on the system and the applications to take effective and timely decisions. 2PARMA goes beyond traditional design space exploration (DSE) by defining a methodology to provide synthetic information about the points of operation of each application with respect to the subsets of resources available to it. Design space exploration methodologies developed in 2PARMA provide also architectural customisation to support parallel programming models, especially communication and memory mapping.

Project Outcomes

Integrated Compiler Toolchain and OS Layer to start from a componentised application source code (C-based) to be semi-automatically parallelised into an OpenCL program and then compiled to bytecode and further dynamically translated to machine code. Then, the machine code execution and deployment will be supported by an OS layer to provide isolated logical devices efficiently communicating (device-to-device and host-to-device).

Design toolset for supporting the HW/SW co-exploration: Starting from a configurable architecture and a parallelised application description (coming from the compilation toolchain) the toolset will explore the HW/SW design space by generating:

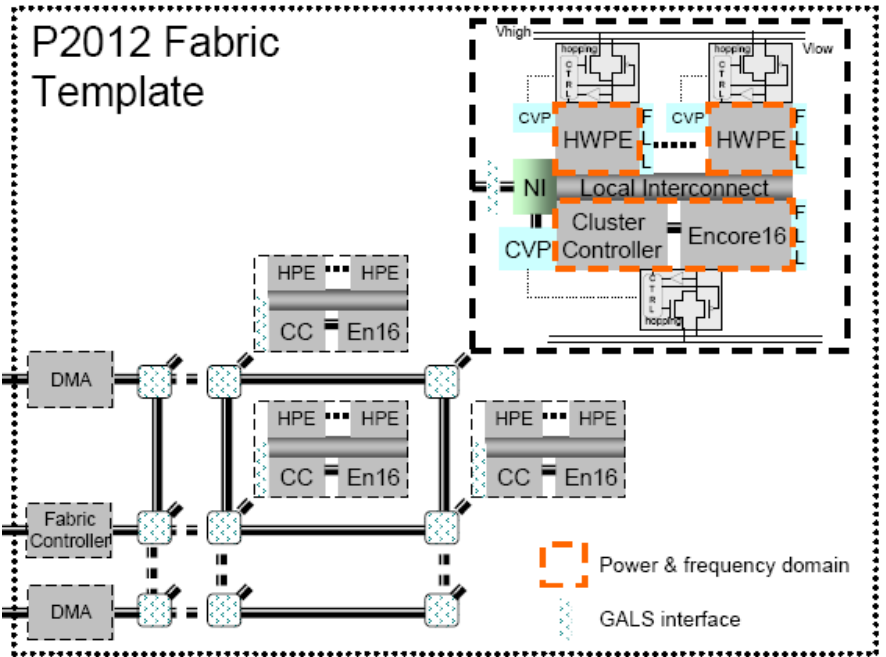
- A bottleneck analysis and optimisation of the target architecture with respect to the parallel application;
- Robust tuning of the design-time (static) configurable parameters with respect to run-time (dynamic) system behaviour;
- A set of operating points for the run-time configurable parameters to be used for dynamic resource management.

Run-time Manager: A set of techniques to manage at run-time the system resources. Based on the set of operating points given by the DSE tool at design time and the info collected at run-time on system workload and resource utilization, the run-time management techniques will optimise data allocation and data access scheduling, task mapping and scheduling and power consumption.

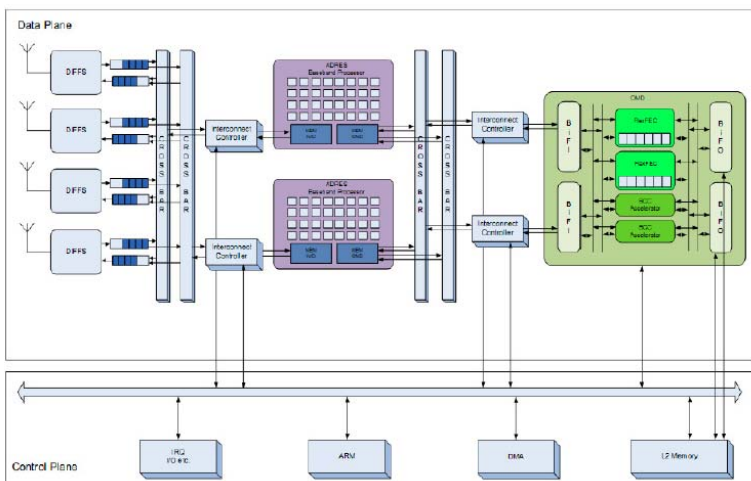
Platforms

The 2PARMA project will demonstrate methodologies, techniques and tools by using innovative hardware platforms provided and developed by STMicroelectronics and IMEC.

STM Platform 2012: The P2012 program is a cooperation between STM and CEA and aims at designing and prototyping a regular computing fabric able to improve manufacturing yield. P2012 is a high-performance programmable accelerator whose architecture meets requirements for next generation SoC products at 32nm and beyond. The goal of P2012 is twofold: from one side, it is to provide flexibility through massive programmable and scalable computing power; from the other side, to provide a solid way to deal with increasing manufacturability issues and energy constraints. Organized around an efficient Network-on-Chip communication infrastructure, P2012 allows connecting a large number of decoupled STxP70 processors SMP clusters offering flexibility, scalability and high computation density. The Platform 2012 computing fabric is composed of a variable number of 'tiles' that can be easily replicated to provide scalability. Each tile includes a computing cluster with its memory hierarchy and a communication engine. The computing fabric operation is coordinated by a fabric controller and is connected to the SoC host subsystem through a dedicated bridge, with DMA capabilities. The P2012 computing fabric is connected to a host processor such as the ARM Cortex A9, via a system bridge. In this way, the fabric is exposed to legacy OS like the GNU/Linux OS.



IMEC ADRES-based COBRA Platform: The COBRA platform is an advanced platform template targeting 4G gigabit per second wireless communication. This platform can be customised to handle very high data rates as well as low throughputs in a scalable way. This platform consists of 4 types of cores. DIFFS, an ASIP processor tuned towards sensing and synchronisation, and optimised for very low power. ADRES, a coarse-grained reconfigurable core template consisting of a number of functional units connected in a given interconnect network. FlexFEC, a flexible forward error correction ASIP capable of different outer modem processing. It is a SIMD engine template where the instruction set, bit width of the data-path and the number of SIMD slots can be chosen based on the set of requirements of the standard to be run. An ARM host processor controls the tasks on the platform (e.g. the run-time manager task). The three cores (DIFFS, ADRES and FlexFEC) can be instantiated for a mix of targeted standards that need to be supported. The communication is ensured by customised InterConnect Controller (ICC) cores that are programmable at assembly level as well. In this platform, besides the type and the size of each core, the number of each type of core can be selected based on the different standards that need to be supported on the platform.



IMEC COBRA Platform

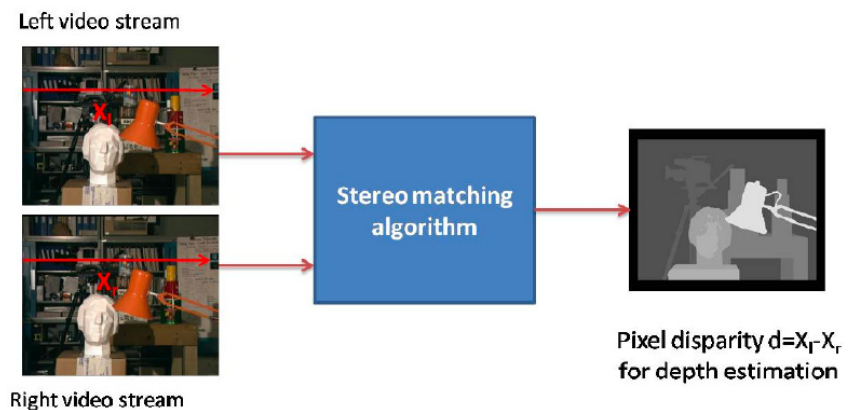
Applications

To ensure a wide range of application scenarios comprising the typical computation-intensive workload of a general-purpose computing system, a set of industrial high performance demanding applications will be used and customized by using the techniques and methodologies developed in 2PARMA project. Applications' architecture, development and integration will leverage from the acknowledged experience of three partners from the Consortium: Fraunhofer HHI for Scalable Video Coding application, RWTH Aachen University for Cognitive Radio (MAC and Physical Layer), and IMEC for Multi View Image Processing.

Scalable Video Coding: SVC also known as layered video coding has already been included in different video coding standards in the past. Scalability has always been a desirable feature of a media bit stream for different services and especially for best-effort networks that are not provisioned to provide suitable QoS and especially suffer from significantly varying throughput. Thus a service needs to dynamically adapt to the varying transmission conditions: A video encoder for example shall be capable of adapting the media rate of the video stream to the transmission conditions to provide at least acceptable quality at the clients, but shall also be able to explore the full benefits of available higher system resources. Within a typical multimedia session the video consumes the major part of the total available transmission rate compared to control and audio data. Therefore, an adaptation capability for the video bit rate is of primary interest in a multimedia session. Strong advantages of a video bit rate adaptation method relying on a scalable representation are drastically reduced processing requirements in network elements compared to approaches that require video re-encoding or transcoding. With this motivation in mind, H.264/AVC-based SVC is of major practical interest and it is therefore highly important to investigate implementation aspects of SVC.

Cognitive radio (MAC and Physical Layer): From the domain of wireless communications, this application includes both physical and MAC-layer processing. The low latency as well as high throughput and reconfiguration requirements of state-of-the-art wireless communication standards makes the cognitive radio application as a highly

appropriate use case for the 2PARMA project and its parallel programming models. The operational scenario of the physical-layer application has been defined based on cutting-edge standards, such as LTE, WiMax and WLAN. These standards have adopted OFDM transmission schemes, because of the high spectrum efficiency and low equalizer complexity. In addition, MIMO techniques are regarded as the key for improving data rates and reliability for future communication systems. The combination of MIMO and OFDM effectively enhances the achievable data rate, spectrum efficiency, and data reliability. Besides the physical-layer application, the provided cognitive MAC-layer allows re-configuration as per the changing spectrum conditions and the evolving application demands. In the component-oriented MAC design approach, special focus is set to efficiently coordinate the data and control flow at runtime. This includes flexible and efficient resource management as well as runtime modifications and exploitation of parallelism during the execution of different MAC-layer components.



The stereo-matching algorithm

Multi-View Video: With the current development of electronic, network, and computing technology, Multi-View Video (MVV) becomes a reality and allows countering the limitations of conventional single video. MVV refers to a set of N temporal synchronized video streams coming from cameras that capture the same scene from different viewpoints. The availability of multiple views of the scene broadens the application field: it extends the sensation of classical 2D video, it allows the user to freely choose a viewpoint (Free Viewpoint Video), and it provides a 3D depth impression of the scene (3D television). The application, considered in the 2PARMA project, is the cross-based stereo matching algorithm where two aligned left and right cameras are assumed.

Project Workplan

The basic idea behind the 2PARMA project is to combine the automatic extraction of parallelism to dynamic compilation in order to exploit the management of system resources at runtime. The project workplan is composed of eight workpackages (WPs): five technical WPs (WP1, WP2, WP3, WP4 and WP5), one demonstration WP (WP7), one dissemination and exploitation WP (WP6) and one management WP (WP8).

Two are the main objectives of **WP1 (Architecture/application specifications and tool requirements)**. From one side, the main goal is the definition of the requirements of the design techniques and tools to be developed in the project. From the other side, the main goal is the definition of the specifications of the industrial applications and architecture to be used for the integration and validation of the design flow in WP5 and WP7. The specifications and requirements defined in WP1 will drive the research and technical activities to be done in WP2, WP3 and WP4.

In **WP2 (Programmability of Parallel Computing Systems)**, the main goal is to define a parallel compilation toolchain and Operating System support. The compilation toolchain starts

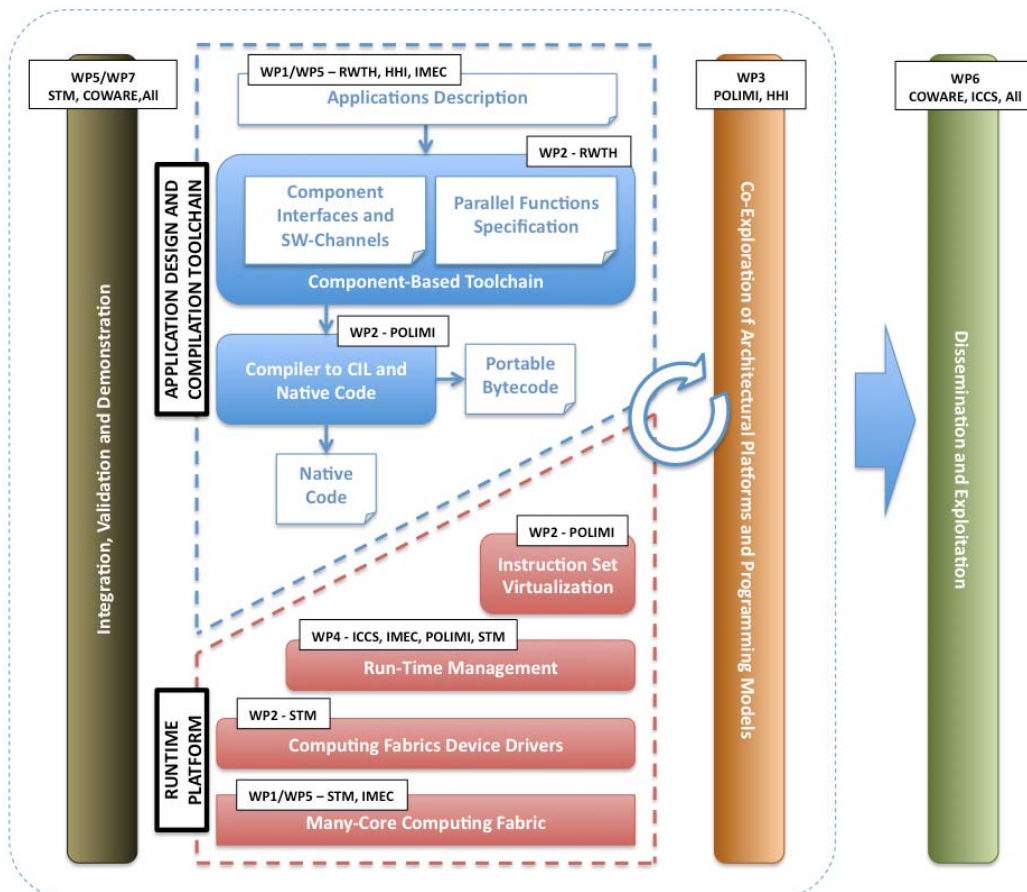
with the component-based application source code (C-based) to be assembled and compiled to bytecode and further dynamically translated to machine code. Then the machine code execution and deployment will be supported by an OS layer to provide isolated logical devices efficiently communicating (device-to-device and host-to-device). The GNU/Linux operating system will be used as the software reference common ground for what the host processor is concerned.

The main goal of **WP3 (Co-Exploration of Architectural Platforms and Programming Models)** is to develop methodologies and tools to support the HW/SW co-exploration of many-core architectures. In particular, WP3 focuses on the profiling of the parallel applications (derived from WP2) aimed at finding the bottleneck of the target platform and on the robust design space co-exploration of static and dynamic parameters (derived from the analysis done in WP1 and WP4) by considering dynamic workloads, while identifying hints/guidelines for dynamic resource management (as needed by WP4).

The partners participating in **WP4 (Run-Time Management)** will develop a run-time resource manager (RTRM) for many-core architectures. This RTRM will offer adaptive task and data allocation as well as scheduling of the different

tasks and the accesses to the data.

Furthermore, the adequate power management techniques as well as the integration to the Linux OS will be developed. The integration of the task, data and power management with the OS are performed in WP4 where the composition of the RTRM takes place. The work performed within



2PARMA Design Flow and Tools

WP4 will take into account: the requirements/specifications of many-core platforms, applications and the design techniques and tools defined in WP1; the dynamic compilation chain and OS support for resource management developed in WP2 and the design space exploration results provided by WP3.

There are also two important activities that are traversed through the project. The **Integration (WP5), Validation and Demonstration (WP7)** activities will follow the development of the 2PARMA design techniques and tools starting from the definition of the specifications to the final integration by using the architectures and applications specified in WP1. **Dissemination and exploitation (WP6)** activities represent the key factors to broaden internally and externally to the Consortium the use of the knowledge gained and the techniques and tools developed in the project, not only spreading scientific and technological achievements at the end of the 2PARMA project lifetime but also while it is in progress.

Finally, the activities in **WP8 (Project Management)** are dedicated to management, coordination and monitoring the research, demonstration, dissemination and exploitation activities carried out in the other WPs.

Main Achievements of Year 1 (2010)

1. **Requirements definition:** Thanks to the fruitful collaboration among the project partners, requirements have been collected describing the architectures and applications to be used as project use cases for integration and validation activities. Requirements on design techniques and tools have also been defined to drive the research activities in the project.
2. **Specifications of the technologies to be developed in the project:** Based on the analysis of requirements, this achievement consists of the definition of specifications of application programming interfaces and data exchange formats related to technologies and tools to be developed in the project.

This formal specification is of fundamental importance for granting the integration of design techniques and tools enabling the independent development of the modules and a seamless integration of design tools and data structures into a common design environment.

3. **Release of the initial implementation of the NoCTrace profiling tool for parallel computing platforms:** The tool has been developed to support the co-optimization of parallel programming models and architectural parameters for many core platforms. The initial version of the tool is currently downloadable from the 2PARMA project website. Based on the NoCTrace profiling tool, a bottleneck analysis of the Scalable Video Coding application running on a parallel computing platform, in particular the Platform P2012 provided by STMicroelectronics, will be done in forthcoming period.
4. **Dissemination:** Several dissemination activities started during the first period to create awareness about the project and to ensure the spreading of the project's outcomes among external industrial and academic organizations, so that the methodologies and tools developed in the project can be exploited in these sectors. The Consortium has been present at a number of international workshops and events to describe the objectives and early results of the research. Some papers have been published in international conferences to spread out the knowledge developed in the project. The 2PARMA project website (www.2parma.eu) is online since January 2010 and periodically updated to ensure visibility of the project development and results. The initial version of the Dissemination Plan is publicly available from the 2PARMA website reporting the main dissemination activities already done and the forthcoming dissemination plans for the remaining period.

