Backgrounder

European Processor Initiative (EPI, https://www.european-processor-initiative.eu/) project is one of the cornerstones of European strategic HPC agenda.

Below we give executive overview of Europe's recent HPC activities and show how EPI is one of the premier examples of strategic unification of efforts and resources towards European exascale goals.

The idea behind EPI initiative was rooted when Horizon2020, biggest EU Research and Innovation programme, was planned by the European Commission with the specific ICT Call in the Work Programme (H2020-ICT-2017-2, under the topic ICT-42-2017 Framework Partnership Agreement in European low-power microprocessor technologies) with the aim to fund the design and implementation of a roadmap for a new family of low-power European processors for extreme scale computing, high-performance Big-Data and a range of emerging applications. This Call was a strategically envisioned spark that would ignite further industry and R&I activities to follow.

A group of key European industrial and scientific organizations joined to form a consortium under the name European Processor Initiative. The EPI's proposal was chosen as the winning one under the stringent selection criteria and the first stage of development, under the Specific Grant Agreement of the European Processor Initiative (EPI-SGA1: 826647), was signed. The European Processor Initiative project formally kicked off and started its activities in December 2018.

In parallel, in March 2017, the European Commission initiated another strategic activity with the formal signature of the EuroHPC declaration – an agreement in which the signatory countries commit to work together with each other and with the European Commission to acquire, build and deploy an integrated world-class High Performance Computing infrastructure. With 7 member states signing the initial declaration, the number has grown to 29 member states as of Oct 2019.

Following the declaration, the European Commission proposed to setup the EuroHPC Joint Undertaking, a 1-billion-Euro joint initiative between the EU and European countries to develop a World Class Supercomputing Ecosystem in Europe. The proposal has been formally adopted by the Council of the European Union on 28 September 2018. Located in Luxembourg, the Joint Undertaking started operating in November 2018.

EuroHPC will permit the EU and participating countries to coordinate their efforts and share resources with the objective of deploying in Europe a world-class supercomputing infrastructure and a competitive innovation ecosystem in supercomputing technologies, applications and skills.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826647

EUROPEAN PROCESSOR INITIATIVE

FRAMEWORK PARTNERSHIP AGREEMENT IN EUROPEAN LOW-POWER MICROPROCESSOR TECHNOLOGIES

The Joint Undertaking will pool EU and national resources in High-Performance Computing with the aim of:

- acquiring and providing a world-class petascale and pre-exascale supercomputing and data infrastructure for Europe's scientific, industrial and public users, matching their demanding application requirements by 2020. This would be widely available to users from the public and private sector, to be used primarily for research purposes;
- supporting an ambitious research and innovation agenda to develop and maintain in the EU a world-class High-Performance Computing ecosystem, exascale and beyond, covering all scientific and industrial value chain segments, including low-power processor and middleware technologies, algorithms and code design, applications and systems, services and engineering, interconnections, know-how and skills for the next generation supercomputing era.

Supercomputing is additionally one of the five key digital topics where the EU's investment should significantly increase: under the next long-term EU budget for 2021-2027 the Commission proposes under Digital Europe programme to invest €2.7 billion in projects to build-up and strengthen supercomputing and data processing in Europe.

Nowadays, European research institutes, universities and industry consume 30% of global supercomputing capacities, but Europe is able to provide just 5% of those capacities – it is clear Europe needs to build its own supercomputers.

The EPI project is recognized and established as one of the cornerstones of this strategic plan. The EuroHPC Joint Undertaking announced it would support future activities of the European Processor Initiative to develop, using European technologies, the low-power microprocessors needed to power supercomputers. This will make Europe less reliant on foreign technology in a field that is essential for many areas of the digital economy in high-performance computing and beyond, such as connected and autonomous vehicles, and big data servers.

Sources

- "EPI." European Processor Initiative, https://www.european-processor-initiative.eu/project/epi/. *Accessed 17 Oct. 2019.*
- "Horizon 2020." Horizon 2020 - European Commission, https://ec.europa.eu/programmes/horizon2020/en. *Accessed 22 Oct. 2019.*
- EuroHPC Joint Undertaking - Documents https://eurohpc-ju.europa.eu/documents.html. *Accessed 17 Oct. 2019.*
- European Commission - PRESS RELEASES - Press Release - High-Performance Computing and the EuroHPC Joint Undertaking. https://europa.eu/rapid/press-release_MEMO-18-5901_en.htm. *Accessed 17 Oct. 2019.*
- YouTube - European Supercomputers - https://www.youtube.com/watch?v=5qDrPhIdgBU&feature=youtu.be. *Accessed 17 Oct. 2019.*
- European Commission - PRESS RELEASES - Press Release - EU Budget: Commission Proposes €9.2 Billion Investment in First Ever Digital Programme. https://europa.eu/rapid/press-release_IP-18-4043_en.htm. *Accessed 22 Oct. 2019.*

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826647

EUROPEAN PROCESSOR INITIATIVE CONSORTIUM

Bayerische Motoren Werke Aktiengesellschaft

Today, the BMW Group, with its 31 production and assembly facilities in 15 countries as well as a global sales network, is the world's leading manufacturer of premium automobiles and motorcycles, and provider of premium financial and mobility services.

w www.bmwgroup.com

Bull SAS (Atos group)

Every day our 110,000 people in 73 countries are developing and implementing innovative digital solutions that support the business transformation of clients and address the environmental and social challenges we all face.

w www.atos.net

Infineon Technologies AG

Infineon Technologies AG is a world leader in semiconductor solutions that make life easier, safer and greener. Microelectronics from Infineon is the key to a better future.

w www.infineon.com

Barcelona Supercomputing Center – Centro Nacional de Supercomputacion

BSC-CNS is the national supercomputing centre in Spain. We specialise in high performance computing (HPC) and manage MareNostrum, one of the most powerful supercomputers in Europe, located in the Torre Girona chapel.

w www.bsc.es

Kalray SA

Kalray develops manycore processors for new intelligent systems that have the capability to analyze on the fly a very large amount of information, make decisions and interact with the world. Those processors are based on patented MPPA architecture.

w www.kalrayinc.com

Forschungszentrum Jülich GmbH

We conduct research to provide comprehensive solutions to the grand challenges facing society in the fields of energy and environment, information and brain research. Our aim is to lay the foundation for the key technologies of tomorrow.

w www.fz-juelich.de

Foundation for Research and Technology Hellas

FORTH, founded in 1983, is one of the largest research centers in Greece, with highly qualified personnel, well-organized facilities, and a reputation as a top-level research institution worldwide. FORTH-ICS focuses on ICT research and on commercializing its results.

w www.forth.gr

Commissariat à l'Energie Atomique et aux énergies alternatives

The French Alternative Energies and Atomic Energy Commission (CEA) is a key player in research, development and innovation in defence and security, low carbon energies, technological research for industry and fundamental research.

w www.cea.fr

Eidgenössische Technische Hochschule Zürich

As a technical and scientific university, ETH Zürich is committed to a broad range of expertise that enables novel combinations of knowledge. ETHZ regularly appears at the top of international rankings as one of the best universities in the world.

w www.ethz.ch

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826647

EUROPEAN PROCESSOR INITIATIVE CONSORTIUM

SemiDynamics technology services SL

SemiDynamics specializes in microprocessor architecture, front-end design and verification services, with a special focus in RISC-V development, both for ASICs and FPGA solutions.

W www.semidynamics.com

Instituto Superior Technico

Técnico is the largest school of Architecture, Engineering, Science and Technology in Portugal, involving a community of over 10,000 people.

W www.tecnico.ulisboa.pt

Fraunhofer Gesellschaaft zur Förderung der angewandten Forschung E.V.

Fraunhofer is Europe's leading application-oriented research organization. Our research efforts are geared entirely to people's needs: health, security, communication, energy and the environment.

W www.fraunhofer.de

Alma mater studiorum - Universita di Bologna

Almost 1000 years old, the University of Bologna is known as the oldest University of the western world. It is one of the most important institutions of higher education across Europe and the second largest university in Italy.

W www.unibo.it

Chalmers tekniska högskola

Chalmers University of Technology conducts research and offers education in technology, science, shipping and architecture with a sustainable future as its global vision.

W www.chalmers.se

Universita di Pisa

The University of Pisa is a public institution with twenty Departments, and high-level research centers in these HPC sectors: engineering, physics, computer science. Research activities in EPI are carried out by DII (Dept. Information Engineering).

W www.unipi.it

Sveučilište u Zagrebu, Fakultet elektrotehnike i računarstva

UNIZG-FER is Croatia's leading academic and research institution in the field of electrical engineering, computing and ICT. Its HPC Architecture and Application Research Center has long-standing experience in performance-optimized HPC architecture.

W www.fer.unizg.hr

E4 Computer Engineering SPA

E4 Computer Engineering is an innovative and agile provider of integrated engineering solutions, technology projects and advanced scalable products for HPC, private & hybrid cloud, high-performance storage, HPDA and AI.

W www.e4company.com

Cineca Consorzio Interuniversitario

Cineca is a not-for-profit Consortium, made up of 67 Italian universities*, 9 Italian Research Institutions, 1 Polyclinic and the Italian Ministry of Education. Today it is the largest Italian computing centre, one of the most important worldwide.

W www.cineca.it

W www.european-processor-initiative.eu

@EuProcessor

European Processor Initiative

European Processor Initiative

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826647

EUROPEAN PROCESSOR INITIATIVE CONSORTIUM

SURFsara BV

SURF is the collaborative ICT organisation for Dutch research and education. We offer researchers in the Netherlands access to the best possible facilities and expertise.

w www.surf.nl

Grand Equipement National de Calcul Intensif

Since its creation in 2007 by the Public authorities, GENCI is working to increase the use of numerical simulation and high performance computing (HPC) for boosting competitiveness within the French economy, across all scientific and industrial fields.

w www.genci.fr

STMICROELECTRONICS SRL

STMicroelectronics is a world leader in providing the semiconductor solutions that make a positive contribution to people's lives, today and into the future.

w www.st.com

Extoll GmbH

EXTOLL, a German high-tech SME, is dedicated to provide leading-edge semiconductor IP for the high-speed interconnect market.

w www.extoll.de

Karlsruher Institut für Technologie

The Institute for Information Processing Technologies, one laboratory in the department of Electrical Engineering at KIT, focuses on VLSI integration, embedded HW/SW Co-Design and accelerator solutions, as well as dependable microsystems.

w www.itiv.kit.edu

Prove & Run

Prove & Run's mission is to help its customers resolve the security challenges of large-scale deployments of connected devices by providing cost-effective, highly secure, certified, off-the-shelf operating systems and hypervisors.

w www.provenrun.com

Elektrobit Automotive GmbH

Elektrobit (EB) is an award-winning and visionary global supplier of embedded and connected software products and services for the automotive industry.

w www.elektrobit.com

Menta SAS

Menta is a privately held company in France. The company delivers 100% third party standard-cells embedded FPGA IPs for SoC, ASIC or ASSP designs.

w www.menta-efpga.com

SIPEARL SAS

SiPearl is a fabless chip company, that came into being from EPI and is now newly added to the Initiative. It will serve as its industrial hand. SiPearl is tasked with building RHEA and subsequent generations of chips, market, support and sales.

w www.sipearl.com

w www.european-processor-initiative.eu
European Processor Initiative
@EuProcessor
European Processor Initiative

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826647

EUROPEAN PROCESSOR INITIATIVE

FRAMEWORK PARTNERSHIP AGREEMENT IN EUROPEAN LOW-POWER MICROPROCESSOR TECHNOLOGIES

EPI embedded FPGA (eFPGA) core

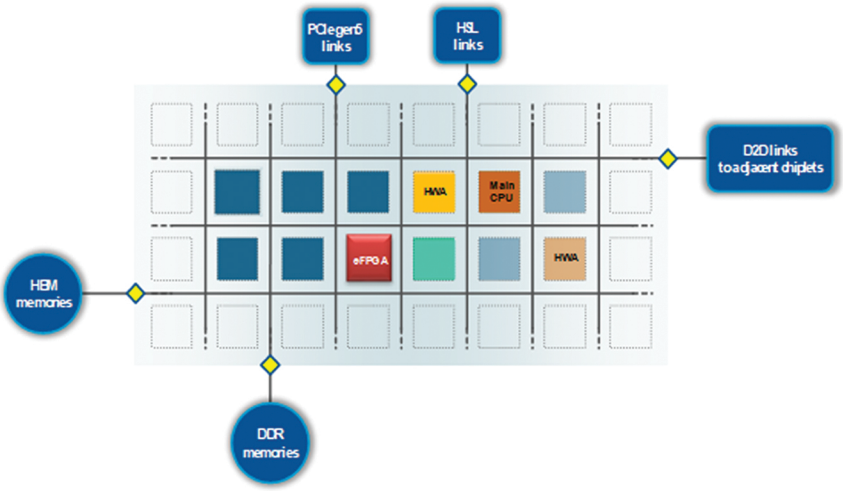
- European embedded field programmable gate arrays “eFPGA” is provided by Menta now in the fifth architecture generation.
- Pure digital IP guarantees a very fast delivery; Menta’s standard-cell based approach enables a rapid port of the eFPGA IP to a new process geometry/variant, while using the standardized and automated tool flow for SoCs.
- It is a key differentiator IP of the GPP Chip integrating high-performance computing requirements of exascale machines with dynamic configurability.
- It enables hardware accelerators improving density, intensive computing and programmability of SoCs for high-performance computing applications as well as automotive applications.
- The technology is provided with an ASIC-like design for test and verification methodology that is crucial to enabling the first success of IPs in the most advanced process nodes.
- The IP is supported by Menta’s unique eFPGA IP specification software, Origami Designer which enables silicon architects to fine-tune the eFPGA fabrics to the requirements of high-performance computing and automotive control unit applications.

The eFPGA tile, which is integrated into the General Purpose Processor Chip (GPP), contributes to an energy-efficient allocation of the necessary performance by an optimal interaction with the main CPU and the dedicated Hardware Accelerators (HWA).

Menta eFPGA IP is optimized for general purpose HPC and automotive applications such as image-processing using machine-learning (ML). It allows post-production functions like customer customization and proprietary elements. In addition, it can consider emerging security aspects, like run-time reconfigurable crypto and post quantum public crypto accelerators.

The eFPGA core plays a key role in an optimal hardware/software codesign system, enabling reconfiguration options for the next generation of European HPC and automotive industry.

Hardware acceleration features are moved on-chip, without the limitations or overhead due I/O pad-count or chip-to-chip communication interfaces. The eFPGA core is provided to EPI customers with the corresponding programming software-tool, Origami Programmer, which generates the bitstream that targets and optimizes RTL to the needs of Menta eFPGA architecture. The technology does not rely on third party software tools, which target generic FPGA architectures and thus deliver suboptimal results.



www.european-processor-initiative.eu

European Processor Initiative

@EuProcessor

European Processor Initiative

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826647

EUROPEAN PROCESSOR INITIATIVE

FRAMEWORK PARTNERSHIP AGREEMENT IN EUROPEAN LOW-POWER MICROPROCESSOR TECHNOLOGIES

Approach

The favoured fundamental approach is a master-slave constellation with one or several automotive-qualified processors operating as masters and one or several HPC-GPP (High Performance Computing General Purpose Processors) and accelerators operating as slaves.

Steps

In order to achieve this goal, the following research work will be done:
Extracting functional and non-functional requirements for the automotive eHPC platform. Apart from aspects like safety, security or energy efficiency, automotive application specific demands such as connectivity are taken into consideration. Moreover, use case scenarios, focusing on how much embedded performance is really needed for future automotive applications will be defined.

- A detailed roadmap for the implementation for the eHPC platform and the eHPC-MCU including arising business cases and revenue models.
- Automotive eHPC platform architecture including all related methods, processes and tools.
- Development of the automotive SDK for the eHPC platform.
- Testing, demonstrating and evaluating the automotive eHPC platform, in order to validate the project findings.
- There is dedicated effort on the partial qualification of the HPC GPP in order to ensure it is adapted sufficiently for the automotive domain.

All in all

The following basic characteristics of an automotive system must be ensured:

- Functional safety
- Data security
- Real-time capability
- Availability and reliability for the entire life cycle
- Lowest possible energy consumption.

Automotive eHPC Compute Platform

Without innovative solutions, the digital progress in the automotive sector will end in a deadlock because of insufficient computing power for new and increasing fields of application like 360-degree environment recognition and other real-time systems.
In order to continue on the road of success, the basic approach is the use of suitable high performance processors from the HPC sector. The specific challenge is the integration of those high performance consumer processors that have to be compliant with the relevant application and environmental requirements in the real-time domain of automotive.
From the system viewpoint, adapting and using consumer processors must under no circumstances have a negative impact on the function, the real-time behaviour, the availability and the reliability of the automotive compute platform.

In the EPI project, the Automotive stream aims at a scalable ECU platform intending to come up with a pioneering automotive processor (eHPC MCU). The latter will be expanded in view of architecture and performance ability so that it is able to act as master and control and survey one or more number crunchers. In the master-slave constellation, EPI partners will provide a number cruncher, in the form of a specific automotive version of the HPC-GPP.
Both parts will be matched concerning their architecture, profile and performance. As to the eHPC MCU, aspects like safety, security, fall back or redundancy for reduced application will be taken into consideration. At the same time, it is of main relevance to achieve the top Automotive Safety Integrity Level D (ASIL-D) at system level which might be expected and necessary for autonomous driving application. All of this will help to open the gates to autonomous driving, this way preparing the way even for use at level5.

www.european-processor-initiative.eu
European Processor Initiative
@EuProcessor
European Processor Initiative

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826647

EUROPEAN PROCESSOR INITIATIVE

FRAMEWORK PARTNERSHIP AGREEMENT IN EUROPEAN LOW-POWER MICROPROCESSOR TECHNOLOGIES

Furthermore

In the extremely cost-driven automotive sector, specific standards must be observed, for example for data communication, interfaces or networks. In light of this, the use of complex and expensive special solutions like active cooling should be avoided if possible.

With respect to the HPC domain, one of the main incentives to adapt the HPC GPP is the expected improved exploitation of the highly expensive development of such a complex processor - not only in the context of the original purpose with local servers and super computers but also embedded domains like automotive. Additional computing power should not exceed given cost and energy budgets.

Since an integration of HPC GPP and HPC accelerators into the automotive eHPC has to be technologically, functionally and economically successful, fundamental questions must be determined and clarified right from the beginning. Some of the central issues in question could be:

- The profile of requirements on the automotive side concerning development process and implementation of the HPC GPP processor, for instance with regard to the documentation.
- Which interfaces or peripherals must be additionally considered in the HPC GPP architecture?
- What kind of automotive proven chip assembly method for GPP chip has to be chosen?
- Which technological adjustments of the HPC GPP can be realised without depleting the targeted economic synergies between the HPC and the automotive type of the GPP?

Other challenges

The system “Automotive eHPC Compute Platform“ must operate in the required automotive temperature range from -40°C to 155°C. Defined usage profiles must be observed. The task of the automotive eHPC MCU is to monitor correct function of the slave processors. They must safeguard real-time capability, protect the system for security reasons and perform a reduced application in parallel in order to take over in case of unavailability or degradation of the HPC GPP. The eHPC compute platform has to achieve fail-operational status in order to be suitable for automated driving level 4 and 5.

Highly complex

The complexity of different rule sets to be taken into account for the qualification process in automotive underlines the challenge of the research activities. Although a full chip qualification of an automotive version of the GPP will be impossible, it is intended to test and if necessary, prepare for qualifications of dedicated IP-blocks and interfaces under automotive conditions. The qualification determines not only methods and process flows to be deployed but also relevant responsibilities. After all, the general approach is to define an adequate qualification flow that describes the basic elements of every qualification process. Each qualification covers the items:

- Reliability qualification
- Characterisation of object
- Manufacturability including testability

An efficient qualification methodology is essential to meet optimum turnaround times and optimised costs. The qualification strategy has to be knowledge- and application-based rather than stress-test-based in order to make Q&R concerns, risks and opportunities transparent. Possible failure mechanisms must be explored in advance and studied with utmost effectiveness. The qualification of a processor shall cover not only the product itself, but all materials, used components and processes needed to produce and deliver the device. The production process of the processor has to be qualified together with the processor or a processor-like test vehicle, if the technology is described by key characteristics as having a major influence on the reliability, performance or functionality of the processor.

Complementary to a stress-test-driven qualification, a failure-mechanism-driven reliability qualification is recommended. During the qualification, environmental conditions as well as thermal, electrical or mechanical stress factors have to be respected.

Beyond Automotive

The automotive processor units have to fulfil relevant application requirement specifications. They are characterised by the widely varying use conditions and loads which all have in common: long lifetime expectations and stringent failure rate targets. In the end, customer expectations as well as requirements concerning production quality/stability and reliability performance must be fulfilled.

Taking a general view beyond the horizon, Automotive could be treated as the appropriate touchstone for the deployment of computation technology in a lot of further embedded domains like machinery, medical, avionics and others. The emergence of HPC processor capability in Europe will strongly reinforce manufacturing and service-provision value chains and will enable new synergies that will benefit an important part of European industry.

www.european-processor-initiative.eu
European Processor Initiative
@EuProcessor
European Processor Initiative

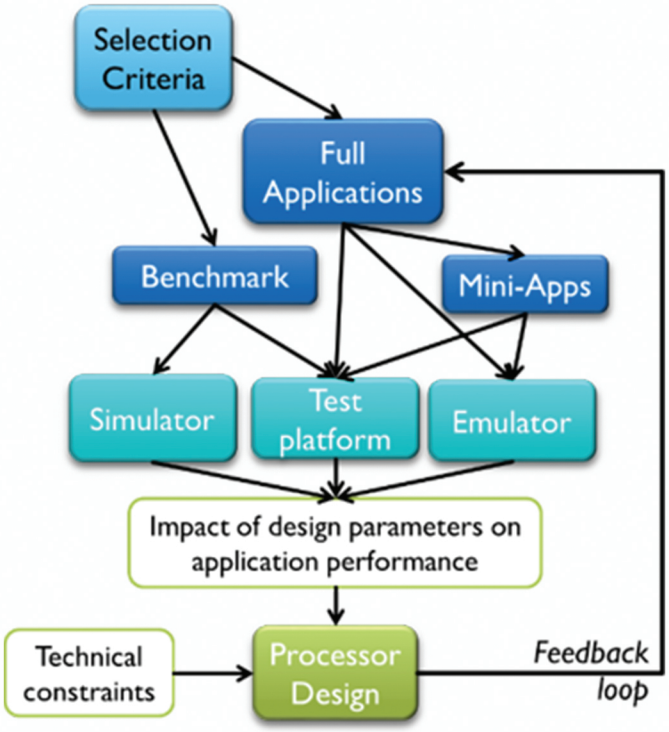
This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826647

EUROPEAN PROCESSOR INITIATIVE

FRAMEWORK PARTNERSHIP AGREEMENT IN EUROPEAN LOW-POWER MICROPROCESSOR TECHNOLOGIES

EPI Co-design approach

- Bi-directional and iterative interaction process between:
  - o application experts and
  - o hardware (HW) and system-software (SW) developers
- Multi-level suite of benchmarks
  - o from very low-level synthetic benchmarks to high-level applications
- Methodology with multi-level models & simulators
  - o analytical models, high level
  - o simulation based (e.g. gem5 simulation engine)
  - o reference platform (e.g. Marvell ThunderX2)
- Node-level co-design parameters (e.g.):
  - o GPP: SVE length, number of SVE pipelines per core
  - o Accelerator: vector registers length, ratio accelerator-vs-GPP cores
  - o Memory (size and BW): Cache, HBM, DDR



The success of the processor technologies developed in EPI depends to a large extent on how well they fulfil the needs of the end-users: scientific and industrial application developers. Co-design is understood as a bi-directional and iterative interaction process between application owners and hardware (HW) and system-software (SW) developers, in which benchmarks and applications are employed to identify the impact of design decisions onto the performance of the codes.

www.european-processor-initiative.eu
European Processor Initiative
@EuProcessor
European Processor Initiative

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826647

EUROPEAN PROCESSOR INITIATIVE

FRAMEWORK PARTNERSHIP AGREEMENT IN EUROPEAN LOW-POWER MICROPROCESSOR TECHNOLOGIES

EPI benchmark suite

The EPI benchmark suite is based on a multi-level set of benchmarks with increasing code complexity: from very low-level test codes (simple flow of instructions to test functional units), passing by synthetic benchmarks, and reaching up to mini-apps (simplified versions of an application preserving its main characteristics but with much fewer lines of code), or even full application codes (for performance evaluation at later stages of the project). All disclosable elements of the EPI benchmark suite are available in a public code repository: https://gitlab.version.fz-juelich.de/epi-wp1-public

Covered application fields

The selection criteria applied to choose the codes in the EPI benchmark suite ensures that all components of the EPI design can be tested, and that fields with large weight on the current and predicted user portfolio on HPC systems are covered. The selection includes not only "traditional" HPC applications but also emerging fields such as deep learning (e.g. convolutional neural networks) and high performance data analytics:

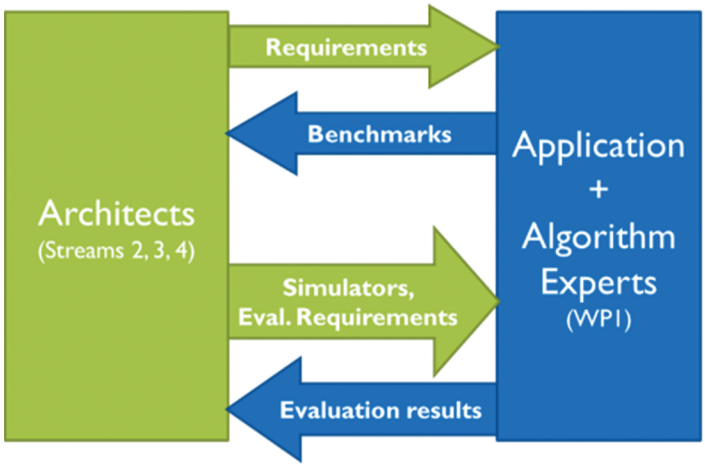
- Biophysics, Biology, Medicine
- Earth Sciences, Climate
- High energy physics, fusion
- Material Sciences
- CFD, hydrodynamics
- PDE
- Image / Media
- Automotive
- Cryptography
- HPDA
- Machine Learning, Deep learning
- Cloud, Data Base
- Reference benchmarks (HPL, HPCG, Stream, DGEMM...)

The multi-level benchmark suite contains synthetic benchmarks, mini-applications, and full-fledged codes. This shall allow a wide range of studies, from detailed design analysis and simulations in the early stages of the project, up to platform evaluation and comparison when EPI evaluation vehicles are available. In this context it is important to highlight that the benchmark suite is by no means "frozen" or "closed". The current selection of applications is to be considered as a pool from which the most suitable workloads are selected, depending on the study that needs to be performed. If specific needs are identified later on in the project, which cannot be addressed with the current codes, new codes might be added.

Interaction with architects

A constructive and efficient interaction between WP1 and the hardware and software architects is crucial in order to enable real co-design, i.e. identifying and giving answer to the most crucial design questions based on the analysis of application requirements. This interaction has been established through regular face-to-face technical meetings. In them, a methodology has been developed based on running a small set of benchmarks and use cases with the simulation tools available to the project: Gem5, SESAM, and MUSA. This has allowed to give indications on the preferred balance between various design parameters, such as number of cores per chip, size of caches, length of vector units, or memory capacity.

Once the first test systems are available, the EPI benchmark suite will be used to evaluate and compare performance and usability with other technologies, providing further co-design input for next generation EPI processors chips. The results of these co-design activities will have an impact on the final design of the EPI processor technologies, ensuring that they fulfil the requirements of their targeted markets.



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826647

EUROPEAN PROCESSOR INITIATIVE

FRAMEWORK PARTNERSHIP AGREEMENT IN EUROPEAN LOW-POWER MICROPROCESSOR TECHNOLOGIES

Crypto-Accelerators for EPI Security

In the framework of strategic EU markets (e.g. HPC for automotive, aerospace, AI, industry4.0, robotics, biomedical), robustness and safety issues play a major role, requiring layered and itemized architecture with specific HW and dedicated private resources for security responsibility. For this reason, in close collaboration between University of Pisa (I) and Provenrun (F), the General Purpose Processor (GPP) ecosystem integrates a Security Subsystem (Fig. 1), that relies on an isolated blocks architecture, with specific HW and dedicated private resources. Each isolated block (Security Domain) guarantees security service availability and full independence of each security scheme that is accelerated in HW by an embedded cryptographic co-processor (Crypto-Tile).

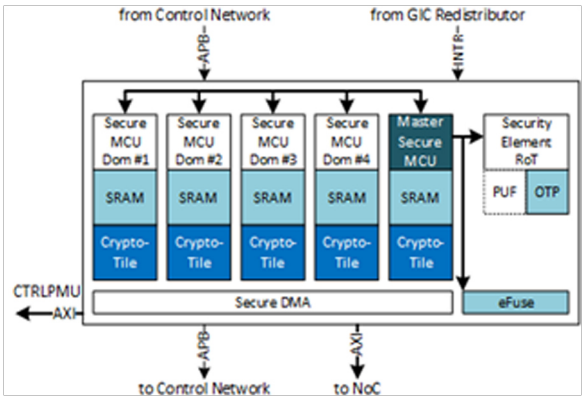


Fig. 1. GPP Security Subsystem

Crypto-Tile co-processor features:

- Implementation of Security Recommendations addressing threats, vulnerabilities and errors
- Secure MCU interface for installation and management of cryptographic keys and configuration, control and status of cryptographic operations
- Secure DMA interface for high-bandwidth data transfer
- HW acceleration of symmetric-key cryptographic algorithms (AES and AES modes of operation: ECB, CBC, CFB, OFB, CTR, CMAC, CCM, GCM, XTS), for 128b and 256b keys
- HW acceleration of public-key cryptography arithmetic on 256b and 521b elliptic curves (ECC)
- Supporting SW aided ECC schemes (ECDSA, ECDH, ECIES, ECMQV)
- HW acceleration of hash functions SHA2 and SHA-3 for computation of digests on 224, 256, 384 and 512 bits and supporting SW aided high-level hash schemes (HMAC)
- HW acceleration for random numbers generation, supporting both external seeds and internal seeds (by means of an embedded fully digital entropy generator)
- Evolution towards Post-Quantum Cryptography
- Support to security protocols and security standards such as TLS, SSH, MACsec, IPsec, WAVE, ESI, ITS

The Crypto-Tile architecture (Figure 2) consists in global registers for management of internal keys slots dedicated to each engine, and for secure configuration and handling of the tile and 4 crypto-processors (CP) for symmetric-key algorithms (AES), elliptic curve cryptography operations (ECC), hash functions (SHA) and random numbers generation (RNG). Each CP integrates a digital engine and dedicated registers for controlling and handling cryptographic operations and transferring data. Any registers of the Crypto-Tile can be accessed by the Secure MCU (Master Control Unit, e.g. a low-power RISCV or CortexM core) and, in addition, data registers of crypto-processors can be accessed also by a specific interface to Secure DMA (Direct Memory Access) controller. Dedicated SW drivers will be developed to integrate the HW acceleration offered by the Crypto-Tile in the secure GPP Operating System Kernel. The Crypto-Tile IP can be connected through an AXI full memory-mapped interface. Synthesis results on Artisan 7nm TSMC technology demonstrate that the Crypto-Tile IP can run up to several GHz (e.g. ECC@1.8GHz, SHA@5GHz, AES@3GHz) ensuring a throughput in data encryption and signature verification orders of magnitude higher than what achievable via SW on ARMv8 64b architecture.

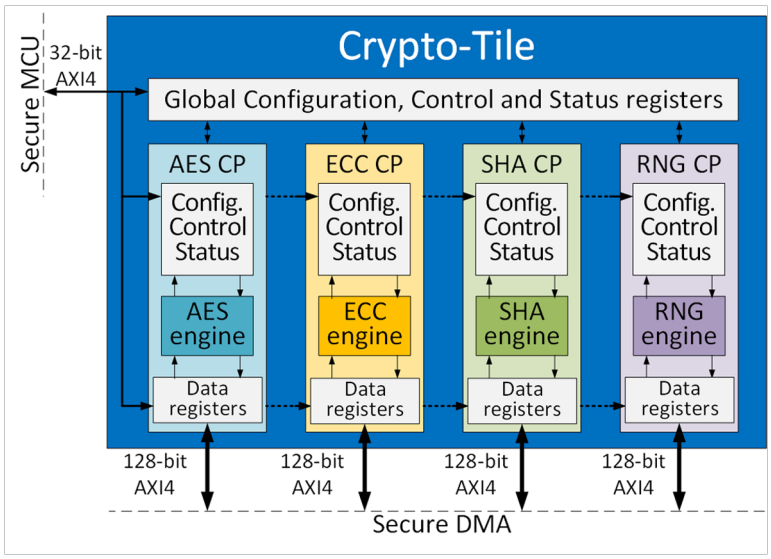


Fig. 2. Outline of Crypto-Tile architecture

www.european-processor-initiative.eu
European Processor Initiative
@EuProcessor
European Processor Initiative

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826647

EUROPEAN PROCESSOR INITIATIVE

FRAMEWORK PARTNERSHIP AGREEMENT IN EUROPEAN LOW-POWER MICROPROCESSOR TECHNOLOGIES

Accelerator for the Automotive Stream

The objective of the EPI automotive stream is to support Autonomous Driving Systems (ADS) based on a mainstream automotive safety/security MCU and two or more 'number crunchers' derived from the General Purpose Processor (GPP) (Figure 1). Each 'number cruncher' will implement the perception and path planning functions. As the GPP is based on ARM cores, it provides high performances for 64-bit and 32-bit floating-point computations. These are applicable to the path planning functions of ADS. The GPP also embeds compute units dedicated to acceleration, in particular the MPPA tiles for vehicle perception. Key vehicle perception functions include: sensor data segmentation and fusion; object detection and tracking; coupling with the localization functions. These functions require an integrated mix of CPU, GPU, and NPU (neural processing unit) capabilities, while meeting the soft real-time constraints of the AUTOSAR Adaptive standard.

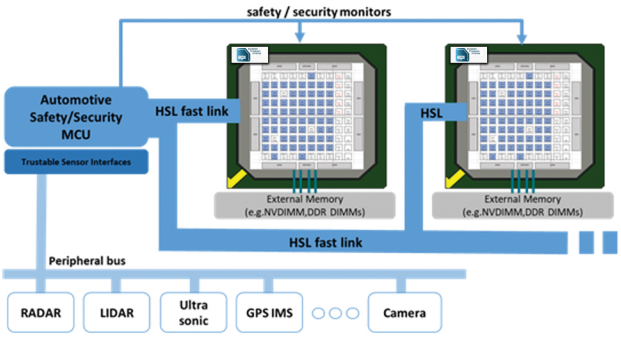


Figure 1 Automotive Stream embedded HPC platform

The MPPA accelerator tile for EPI features:

- Four Processing Engine (PE) cores.
- Resource Management (RM) core.
- Full crossbar local interconnect.
- 2MB local memory (SPM or L2$).
- Micro-programmable DMA engine.
- Debug/trace Support Unit (DSU).
- Connection to the GPP AMBA 5 NoC.

Each PE or RM core features:

- 6-issue 64-bit VLIW architecture.
- 256-bit wide Load/Store Unit (LSU).
- 128-bit wide high-performance FPU.
- Tightly-coupled tensor coprocessor.

The Kalray software development framework is based on the C/C++/OpenMP/POSIX standards. It provides optimized application libraries (OpenCV, CNN, FFT, BLAS) and implements offloading API supported in GCC or LLVM. Offloading from ARM CPUs is based on three key developments:

- Retargeting of the LLVM compiler to the Kalray VLIW core architecture, in order to compile the OpenCL-C dialect.
- Extending the POCL open-source project with the ability to launch GCC C/C++/OpenMP 'native kernels' for execution on MPPA clusters under the lightweight POSIX ClusterOS. POCL provides the OpenCL 1.2 Embedded Profile host API.
- Implementing the OpenAMP open standard from the Multicore association, which allows standard Linux running on the ARM cores to manage the life cycle and the system call remoting requests from the lightweight POSIX ClusterOS.

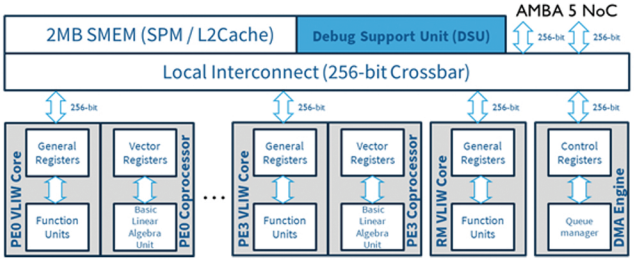


Figure 2 Overview of the MPPA accelerator tile for EPI

The MPPA accelerator tile (Figure 2) is developed for vehicle perception in the EPI Automotive Stream. This tile is based on the compute unit ('compute cluster') of the Kalray MPPA3 processor. Indeed, the MPPA processors already excel at vehicle perception, thanks to their architecture designed for high-performance signal processing, image processing, bit-level processing, and deep learning inference. They are fully programmable in C/C++/OpenMP under GCC or LLVM, and are able to host RTOS, SMP POSIX and Linux operating systems. Other tools include a deep learning compiler that produces highly optimized code from trained neural networks provided under the standard Caffe, TensorFlow and ONNX formats. In the EPI Automotive Stream, the MPPA3 processor will be used in the Intermediate Reference Platform (IRP) to support software development and performance analysis of ADS functions until the first iteration of the GPP is available.

www.european-processor-initiative.eu
European Processor Initiative
@EuProcessor
European Processor Initiative

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826647

EUROPEAN PROCESSOR INITIATIVE

FRAMEWORK PARTNERSHIP AGREEMENT IN EUROPEAN LOW-POWER MICROPROCESSOR TECHNOLOGIES

Posit-based ML & DNN Acceleration for AI in EPI

The adoption of Machine Learning (ML) and Deep Neural Networks (DNN) to enable Artificial Intelligence (AI) techniques on edge devices in strategic EU markets (e.g. automotive, aerospace, industry4.0, biomedical, robots), is requiring HW and SW acceleration to achieve High Performance Computing with high energy efficiency. To this aim, in collaboration between University of Pisa and Kalray, the use of Posits in the EPI project aims at exploiting a new type of arithmetic that allows the reduction of power consumption and circuit complexity for data processing and storage. The fundamental application of the Posit research topic is the acceleration of Deep Neural Network and the acceleration of autonomous driving applications. Posits are an effective alternative to classic integer and floating-point (IEEE 754) arithmetic to preserve the accuracy of floats but with just half of the bits. This means that it is possible to use this novel format for data compression and for computation, i.e. doubling the bandwidth and/or halving the memory complexity, without losing accuracy.

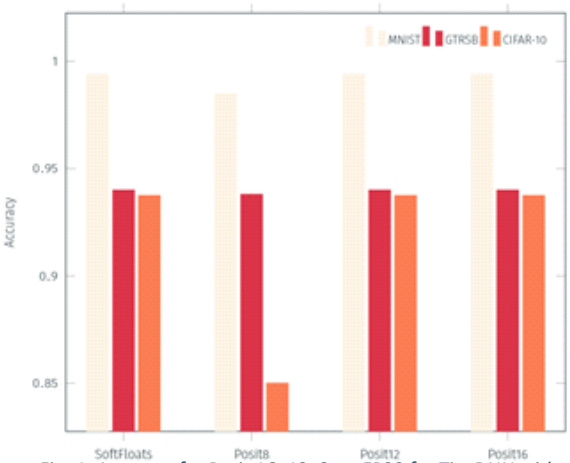


Fig. 1. Accuracy for Posit 16, 12, 8 vs. FP32 for TinyDNN with MNIST, GTRSB and CIFAR data sets

Posit C++ library developed by UniPisa provides:

- Custom definition of any Posit type
- Different back-ends for accelerated emulation (floating point, fixed point, tabulation)
- Support for FPGA synthesis to provide HW acceleration of a Posit Processing Unit
- Support for different vector processors used in EPI: ARM v8.2 Scalable Vector Extension; RISC-V “V” 0.8 extension

Posits in Kalray MPPA (Massively Parallel Processor Array):

- Posit8 numbers identified as an effective compressed representation for the Float32 parameters: the results of rounding can be restricted to Posit8,0 or Posit8,1 numbers, with the benefit of reducing by half the memory capacity and bandwidth

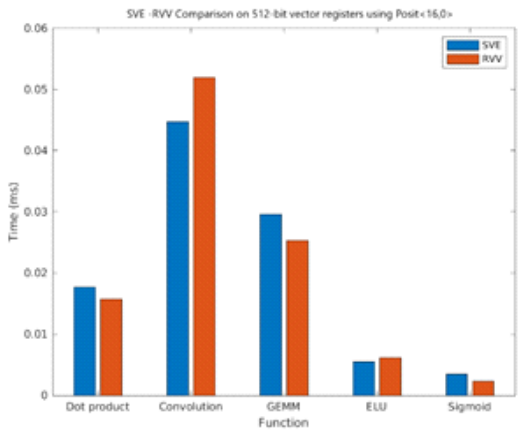


Fig. 2. Execution time of Posit C++ library using ARM SVE and RISC-V V backends with 512-b vector registers for several AI functions: dot product, convolution, Exponential Linear Unit (ELU), sigmoid, General Matrix-Multiply (GEMM)

Moreover, Posits can be further squeezed down to a quarter of the size of floats, losing little-to-none in terms of accuracy. In EPI, Posits have been implemented:
i) through a SW library (Pisa CppPosit) and the SW framework is completed by the open source tinyDNN Deep Neural Network library extended to support the new cppPosit features;
ii) in Kalray MPPA (Massively Parallel Processing Array).
Posit8 numbers have been identified by Kalray as an effective compressed representation for the Float32 network parameters: instead of rounding Float32 parameter values to Float16, the results of rounding can be restricted to Posit8,0 or Posit8,1 numbers, with the primary benefit of reducing by half the memory capacity and bandwidth required by the network parameters. Kalray focuses on the Posit8,0 and Posit8,1 numbers because they are exactly represented as Float16 numbers, and thus can benefit from the exact Float16.32 dot-product operator of the MPPA3 co-processors. This evaluation should lead to the inclusion of new arithmetic instructions to expand Posit8 to Float16 in the MPPA IP delivered to the H2020 European Processor Initiative.

www.european-processor-initiative.eu
European Processor Initiative
@EuProcessor
European Processor Initiative

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826647

EUROPEAN PROCESSOR INITIATIVE

FRAMEWORK PARTNERSHIP AGREEMENT IN EUROPEAN LOW-POWER MICROPROCESSOR TECHNOLOGIES

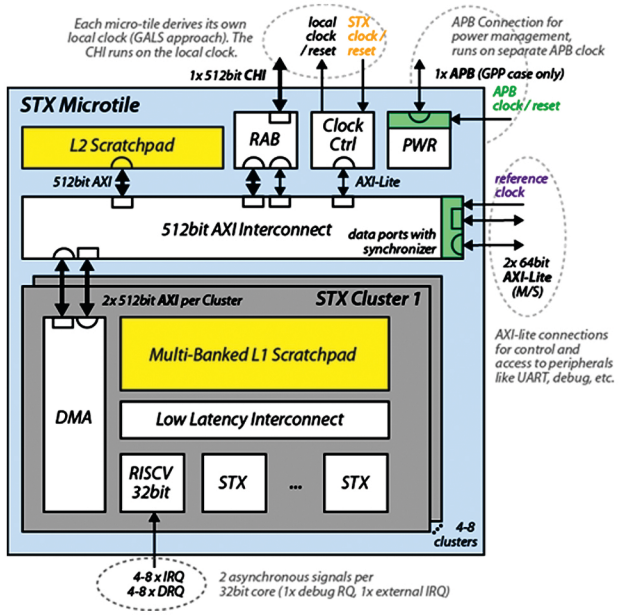
Domain specific accelerators in EPI
STX (stencil/tensor accelerator)

Architecture

- Domain specific accelerator for Machine Learning and Stencil operations within EPAC tile
- Designed to provide at least 10x more energy efficient computing for these workloads
- A typical STX accelerator has 4-8 clusters where each has
  - multiple specialized compute units
  - one 32bit RISC-V processor for support
  - energy efficient local scratchpad memory
  - DMA to transfer data from/to the accelerator

Goal: at least 5x more energy efficiency (TFLOPS/W)

Description



From the beginning EPI explicitly considered “specialised blocks for stencil and deep learning (DL) acceleration. The vector and stencil capabilities will address workloads in HPC centres, while the DL block will target learning acceleration” as part of the acceleration stream motivated by “optimised performance and energy efficiency” for “specialised computations”. In the initial DoA, two different domain specific accelerators (NTX for machine learning, and a stencil accelerator) were suggested. During the first few months of the project, researchers from Fraunhofer Institute, ETH Zürich and University of Bologna were able to merge the functionality of both units into a very efficient computation engine that has been named STX (stencil/tensor accelerator).

Such ”domain-specific accelerators” are now a major trend in industry, as can be seen by multiple new announcements in the 2019 Hot-Chips symposium and AI Summit by industry heavyweights as a multitude of startups that have presented acceleration engines that were based on specialized datapaths and not general purpose vector units, confirming the significant differentiation in architectures needed for achieving top efficiency and performance in the machine learning domain.

The main goal of STX is to achieve a significantly higher (at least 5x-10x) energy efficiency over general purpose/vector units. The efficiency tells us how many computations can be performed with the unit, and the early target for the STX unit was to achieve at least 5x more energy efficiency (TFLOPS/W) than the vector unit on deep learning applications. In the first few months of the project, it became clear that these estimations are rather conservative, and the effective efficiency within EPI chips will be significantly higher. For applications that require only inference using quantized networks, this efficiency will be another 10x higher.

STX has been designed as a modular building block with several parametrization options. Each STX accelerator consists of several clusters of computing units, a typical instance would have four such clusters. Each cluster in turn consists of specialized computing engines as well as up to two RISC-V cores that are used to control the computing engines and perform additional operations. All these units will access a local scratchpad memory, which will be filled using a centralized DMA unit. This configuration allows for 64 GFLOPS (single precision FP), and multiple instances of STX can be instantiated in an EPAC tile.

STX is programmed using OpenMP, there are solutions that allow regular operations to be offloaded to the STX unit from an ARM system (in the GPP) or the 64-bit RISC-V core (in the EPAC tile) using both GCC and LLVM based flows that will be further refined as part of the project.

www.european-processor-initiative.eu
European Processor Initiative
@EuProcessor
European Processor Initiative

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826647

EUROPEAN PROCESSOR INITIATIVE

FRAMEWORK PARTNERSHIP AGREEMENT IN EUROPEAN LOW-POWER MICROPROCESSOR TECHNOLOGIES

Description

- Domain specific accelerator for scientific computing, designed for the resolution of large ill-conditioned systems of equations
- Designed for the accurate computation (up to 256 bit fractional parts) of large systems
- Targets 10x to 100x acceleration of variable precision computation (compared to software solutions)
- A single VRP processor features:
  - Fully functional branch control for efficient convergence
  - 32 internal registers for up to 256 bit arithmetic operations
  - Transparent cache access for arbitrary size of variables

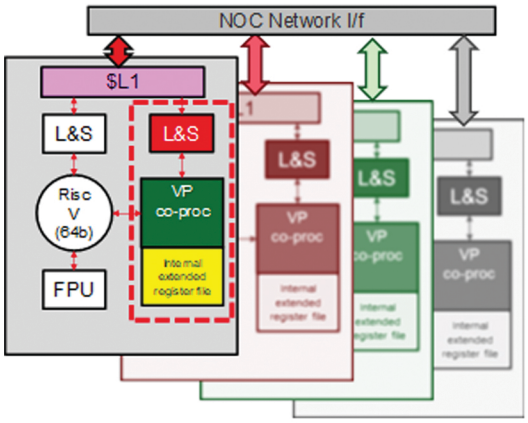


Figure 1 Variable precision tile structure

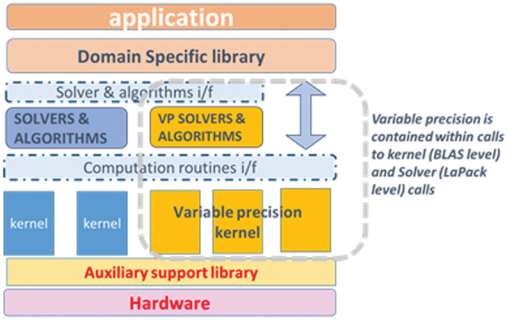


Figure 2 Layered programming model for variable precision

Domain Specific Accelerator in EPI – VRP (VaRiable Precision Processor)

The VaRiable-Precision Unit (VRP) enables efficient computation in scientific domains with extensive use of iterative linear algebra kernels, such as physics and chemistry. Augmenting accuracy inside the kernel reduces rounding errors and therefore improves computation's stability. Usual solutions for this problem have a very high impact in memory and computation time (e.g. use double precision in the intermediate calculations).

The hardware support of variable precision, byte-aligned data format for intermediate data optimizes both memory usage and computing efficiency. When the standard precision unit cannot reach the expected accuracy with standard precision (aka double), the variable precision unit takes the relay and continues with gradually augmenting precision until the tolerance error constraint is met. The offloading from the host processor (GPP) to the VRP unit is ensured with zero-copy handover thanks to IO-coherency between EPAC and GPP.

The VRP is embedded as a functional unit in a 64-bits RISC-V processor pipeline. The unit extends the standard RISC-V Instruction with hardwired arithmetic basic operations in variable precision for scalars: add, subtract, multiply and type conversions. It implements other additional specific instructions for comparisons, type conversion and transfers to cache.

The unit features a dedicated register file for storing up to 32 scalars with up to 256 bits of mantissa precision. Its architecture is pipelined for performance, and it has an internal parallelism of 64-bits. Thus, internal operations with higher precisions multiple of 64 bits are executed by iterating on the existing hardware.

The VRP's programming model is meant for a smooth integration with legacy scientific libraries such as BLAS, MAGMA and linear solver libraries. The integration in the host memory hierarchy is transparent for avoiding the need of data copy, and the accelerator offers a standard support for C programs. The libraries are organized in order to expose the variable precision kernels as compatible replacements of their usual counterparts in the BLAS and solver libraries. The complexity of arithmetic operations is confined as much as possible within the lower level library routines (BLAS), as represented below. Consistently, the explicit control of precision is exclusively handled at solver level.

www.european-processor-initiative.eu
@EuProcessor
European Processor Initiative
European Processor Initiative

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826647