As effect of the of the Dennard’s scaling and with the steaming out of the Moore’s law computer architectures in wide range of market segments (i.e. embedded systems, cyber physical systems and high performance computing) are becoming heterogenous. Examples of such architectures are the APPLE A12X (8-Core CPU, 7-Core GPU and a NPU), Xilinx ZYNQ UltraSCALE+ (Cortex A53, Cortex R5, GPU ARM Mali-400 and one FPGA), NVidia Tegra X2 (6-Core CPU and 256-CUDA Cores), as well as top 10 Top500 supercomputing systems. Moreover, heterogeneous SoCs can be found also inside high performance computing systems for controlling their power consumption and security.

At the same time, as a support for the programmers to develop complex applications, the last 10 years have seen the birth of an important set of programming models/libraries, compilers and IRs (Intermediate Representations) that provide the programmer with an abstracted view of the underlying hardware (i.e. OpenACC, LLVM, OpenCL, TensorFlow, etc.). However, such abstractions may hinder the efficient exploitation of architectural resources. For this reason, there is a need for techniques that, by analyzing the source code, allow to determine the most suitable compute unit for the execution of computational kernels. Recently, deep learning approaches to code analysis and allocation has gained attention and revealed as a promising solution to this problem [1,2]. However, its applicability to complex and maybe reconfigurable heterogeneous architectures is still to be proved.

2. RESEARCH ACTIVITY (Attività di ricerca)

(ENG)

The research activity aims at developing: i) automatic code analysis techniques at IR level and ii) automatic allocation depending on the resulting code analysis. State-of-art techniques such as the ones based on deep learning applied to source code will be used as starting point [1]. Promising innovative techniques based on IR-level source code analysis and classification will be explored and improved [2]. As deep learning techniques require a rich data set for training and test, a challenge of this research will be to automatically collect and clean-up source code kernels from online repositories. The techniques will be applied to two prototype heterogeneous embedded platforms, namely Xilinx ZYNQ UltraSCALE+ [3] and PULP/RISC-V [4] platforms which are used as power controller of the EPI GPP processor. The research activity will also evaluate how these training and testing of deep learning models can be integrated with current HPC infrastructure, like the one of CINECA.

This research activity is line with the objective of the WP4 and WP7 of the EPI EuroHPC project and CINECA collaboration.

(ITA)

L’attività di ricerca mira a sviluppare: i) tecniche di analisi automatica del codice a livello IR e ii) allocazione automatica in base ai risultati dell’analisi del codice. Tecniche state-of-art come quelle basate sul Deep Learning applicate al codice sorgente saranno usate come punto di partenza [1]. Verranno poi esplorate ottimizzate tecniche innovative basate sull’analisi e la classificazione del codice sorgente a livello di IR [2]. Poiché le tecniche di deep learning richiedono un set di dati ricco per le fasi di training e test, una sfida di

Questa attività di ricerca è parte del WP4 e WP7 del progetto EuroHPC EPI e della collaborazione con CINECA.

3. ACTIVITY PLAN

The researcher will acquire or consolidate, in parallel with the research activity, the knowledge of: i) advanced tool chains, compilation tools, intermediate representation languages and their back-ends for embedded targets; ii) multicore heterogeneous architectures, their performance characteristics and their reconfiguration opportunities; iii) deep learning techniques such as LSTM and (variational) autoencoders and their optimization strategies. The research activity will be done in the context the WP4 and WP7 of the EPI EuroHPC project and CINECA collaboration.

The research will encompass the following phases:

- Survey of the state-of-art IR-level optimization tools
- Development of source code collection tools for training and test dataset creation
- Characterization of embedded heterogeneous platform performance and reconfiguration options
- Design and development of automatic code analysis and allocation strategy (tokenization, deep learning network architecture hyperparameter tuning)
- Evaluation on target architectures using training and test datasets on HPC systems
- Dissemination of the work through international conferences and journals

4. REFERENCES


