

AI for Processors, Processors for AI: Going New Ways for Processor Architectures

Michael Hübner

Brandenburg University of Technology Cottbus – Senftenberg
Cottbus, Germany
michael.huebner@b-tu.de

SUMMARY

The rapid evolution of artificial intelligence (AI) is driving a paradigm shift in processor architecture design. Traditional processors in the embedded computing domain, originally optimized for special-purpose computing but traditionally broad in their application domain, are increasingly sub-optimal in meeting the performance, energy efficiency, and scalability demands for applications with restricted resources. At the same time, AI techniques themselves are being used to enhance and even automate processor design, creating a reciprocal innovation cycle between AI and hardware. Additionally, AI technology can become implemented deeply into the processor hardware architecture in order to enable a adaptation of the architecture supported by inference results of the AI.

This presentation shows selected opportunities for developing architectures of embedded processors using new features of AI. First, it examines how AI accelerates innovation in processor architecture. Furthermore, first results will become presented what benefits the deployment of AI technology in processors have.

A particularly promising direction lies in run-time adaptive hardware architectures. New approaches, such as the FPGA based GPU or the high flexible iCore, illustrate how inference results from AI models can directly guide the adaptive reconfiguration of processor hardware during operation. These architectures are not static; instead, they monitor workload characteristics and dynamically tune their configuration—such

as parallelism, precision, memory hierarchy, and execution units—to operate at the most beneficial point of the power-performance curve at any given time (see figure 1).

This run-time adaptability, informed by AI-driven inference and control, enables systems to achieve higher performance while simultaneously reducing power and energy consumption. It represents a shift from the traditional worst-case provisioning approach to an intelligent, context-aware optimization that evolves with the workload in real time.

REFERENCES

- [1] S. Mahmood, M. Huebner, M. Reichenbach: “A Design-Space Exploration Framework for Application-Specific Machine Learning Targeting Reconfigurable Computing”, International Symposium on Applied Reconfigurable Computing 2023, 371-37
- [2] Hernandez, Fricke, Al Kadi, Reichenbach, Hübner: “Edge GPU based on an FPGA Overlay Architecture using PYNQ”, 2022 35th SBC/SBMicro/IEEE/ACM Symposium on Integrated Circuits and Systems Design (SBCCI)
- [3] M. M. Goncalves, F. Benevenuti, H. Munoz, M. Brandalero, M. Hubner, F. Kastensmidt, J.R. Azambuja: “Investigating Floating-Point Implementations in a Softcore GPU under Radiation-Induced Faults”, In: IEEE International Conference on Electronics, Circuits and Systems (ICECS), 2020
- [4] Michael Hübner, Diana Göhringer, C. Tradowsky, J. Henkel : “Adaptive Processor Architecture”, International Conference on Embedded Computer Systems: Architectures, Modeling, and Simulation (SAMOS XII), 2012, Samos, Greece

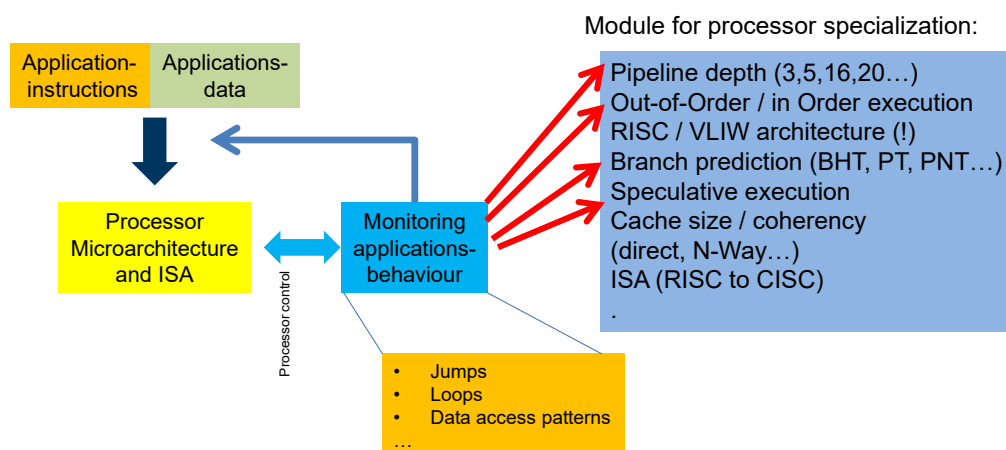


Figure 1. Monitoring and selection of processor mode