Programming abstractions and optimizing compilers for energy-efficient computing

Jeronimo Castrillon
Technische Universität Dresden
Dresden, Germany
jeronimo.castrillon@tu-dresden.de

1st Workshop on NetZero Carbon Computing (NetZero'23)
Montreal, Canada
February 25, 2023

abstract: The demise of scaling laws in micro-electronics has led to an era of innovation in software and hardware architectures aimed at improving the energy efficiency of computing systems. Albeit still highly relevant, software optimizations for mainstream systems, which make the bulk of today's computing systems, provide ever-decreasing returns in the range of single-digit percentages. This is why lots of attention has rightfully turn to domain-specific architectures and emerging technologies which promise improvements of one to several orders of magnitude. Software development for these novel systems is still characterized by low-level expert coding and brittle toolchains, preventing hardware innovations from reaching a broader impact. In this talk, we discuss ongoing efforts on providing high-level programming abstractions and optimizing compilers to automatically target emerging computing systems. We do this by looking at three ongoing projects.

First, we describe a collaborative HW-SW effort to reduce the energy footprint of baseband processing in upcoming cellular networks, predicted to surpass the 1500 TWh mark in 2030 [1]. We discuss an ongoing effort to model and simulate different 5G and user profiles [7] in the context of the award-winning BMBF E4C project[1]. We show how the semantic information provided by dataflow models can be leveraged for domain-specific resource allocation and scheduling at the stringent latency constraints of 5G baseband processing [4].

We then discuss ongoing efforts in the EU EVEREST Project[2] to build a *system development kit* for heterogeneous and reconfigurable systems [5]. Concretely, we describe an MLIR-based end-to-end compilation flow from tensor abstractions [10, 6] onto state-of-the-art reconfigurable systems [9].

Finally, we describe an extensible compilation flow for linear algebra abstractions that transparently generates code for novel in-memory and near-memory computing systems [8, 2]. This is ongoing work in the context of the DFG HetCIM project within the Priority Program on Disruptive Memory Technologies (SPP 2377)[3]. In this context, we discuss recently proposed in-memory computing using racetrack memory [3] as a promising direction for highly dense, robust and energy-efficient systems.

[1]`bit.ly/3LXdkSo`
[2]`https://everest-h2020.eu`
[3]`https://spp2377.uos.de`

References

[1] Anders SG Andrae and Tomas Edler. On global electricity usage of communication technology: trends to 2030. *Challenges*, 6(1):117–157, 2015.

[2] Asif Ali Khan, Hamid Farzaneh, Karl F. A. Friebel, Lorenzo Chelini, and Jeronimo Castrillon. Cinm (cinnamon): A compilation infrastructure for heterogeneous compute in-memory and compute near-memory paradigms. January 2023.

[3] Asif Ali Khan, Sebastien Ollivier, Stephen Longofono, Gerald Hempel, Jeronimo Castrillon, and Alex K. Jones. Brain-inspired cognition in next generation racetrack memories. *ACM Transactions on Embedded Computing Systems (TECS)*, 21(6):79:1–79:28, March 2022.

[4] Robert Khasanov, Julian Robledo, Christian Menard, Andres Goens, and Jeronimo Castrillon. Domain-specific hybrid mapping for energy-efficient baseband processing in wireless networks. *ACM Transactions on Embedded Computing Systems (TECS), special issue of the 2021 International Conference on Compilers, Architecture, and Synthesis of Embedded Systems (CASES)*, 20(5s), September 2021.

[5] Christian Pilato, Stanislav Bohm, Fabien Brocheton, Jeronimo Castrillon, Riccardo Cevasco, Vojtech Cima, Radim Cmar, Dionysios Diamantopoulos, Fabrizio Ferrandi, Jan Martinovic, Gianluca Palermo, Michele Paolino, Antonio Parodi, Lorenzo Pittaluga, Daniel Raho, Francesco Regazzoni, Katerina Slaninova, and Christoph Hagleitner. EVEREST: A design environment for extreme-scale big data analytics on heterogeneous platforms. In *Proceedings of the 2021 Design, Automation and Test in Europe Conference (DATE)*, DATE'21, pages 1320–1325, February 2021.

[6] Norman A. Rink and Jeronimo Castrillon. TeIL: a type-safe imperative Tensor Intermediate Language. In *Proceedings of the 6th ACM SIGPLAN International Workshop on Libraries, Languages, and Compilers for Array Programming (ARRAY)*, ARRAY 2019, pages 57–68, New York, NY, USA, June 2019. ACM.

[7] Julian Robledo and Jeronimo Castrillon. Parameterizable mobile workloads for adaptable base station optimizations. In *Proceedings of the IEEE 15th International Symposium on Embedded Multicore/Many-core Systems-on-Chip (MCSoC-22)*, pages 381–386, December 2022.

[8] Adam Siemieniuk, Lorenzo Chelini, Asif Ali Khan, Jeronimo Castrillon, Andi Drebes, Henk Corporaal, Tobias Grosser, and Martin Kong. OCC: An automated end-to-end machine learning optimizing compiler for computing-in-memory. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD)*, 41(6):1674–1686, August 2021.

[9] Stephanie Soldavini, Karl F. A. Friebel, Mattia Tibaldi, Gerald Hempel, Jeronimo Castrillon, and Christian Pilato. Automatic creation of high-bandwidth memory architectures from domain-specific languages: The case of computational fluid dynamics. *ACM Transactions on Reconfigurable Technology and Systems (TRETS)*, September 2022.

[10] Adilla Susungi, Norman A. Rink, Albert Cohen, Jeronimo Castrillon, and Claude Tadonki. Meta-programming for cross-domain tensor optimizations. In *Proceedings of 17th ACM SIGPLAN International Conference on Generative Programming: Concepts and Experiences (GPCE'18)*, GPCE 2018, pages 79–92, New York, NY, USA, November 2018. ACM.