IBM Research | Zurich

TU/e EINDHOVEN UNIVERSITY OF TECHNOLOGY

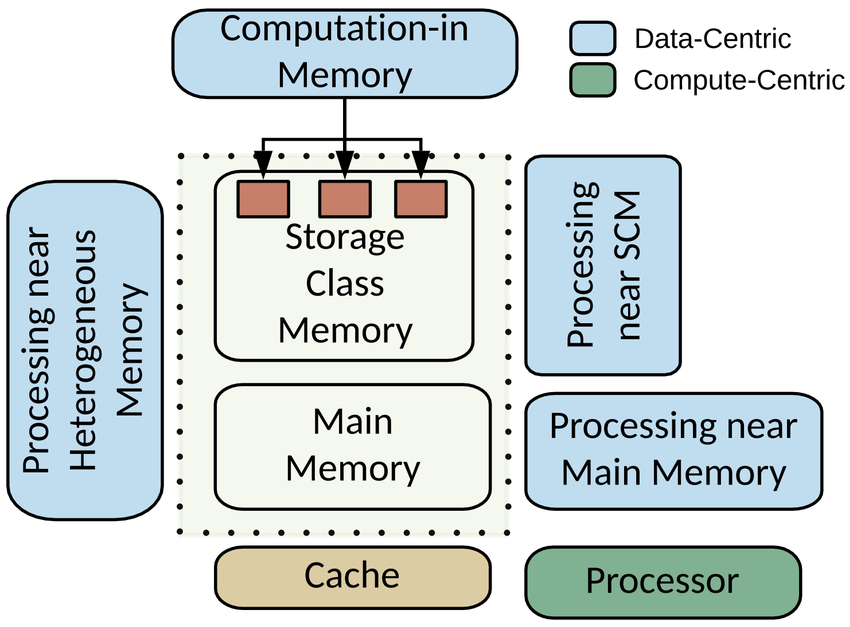
Scaling Stencil Computation on OpenPOWER Near-Memory Architecture

Gagandeep Singh, Dionysios Diamantopoulos, Sander Stuijk,
Henk Corporaal, Christoph Hagleitner

Funded by the Horizon 2020 Framework Programme of the European Union MSCA-ITN-EID

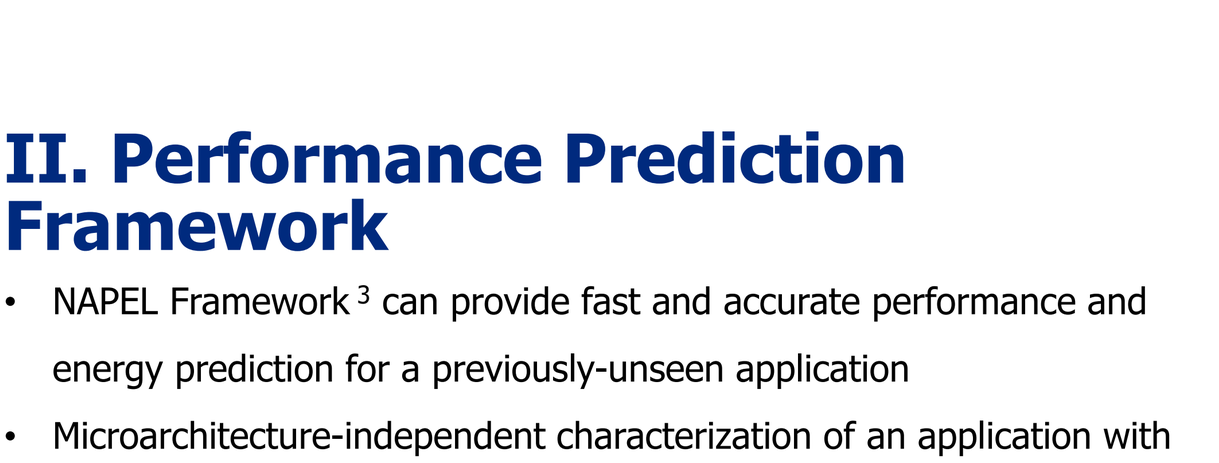
I. Motivation

- An exorbitant amount of data [1]
- The high cost of energy for data movement [1]
- A paradigm shift towards processing close to the data, i.e., near-memory computing (NMC) [2]
- In early design-stage, simulations are extremely slow, imposing long run-time [3]



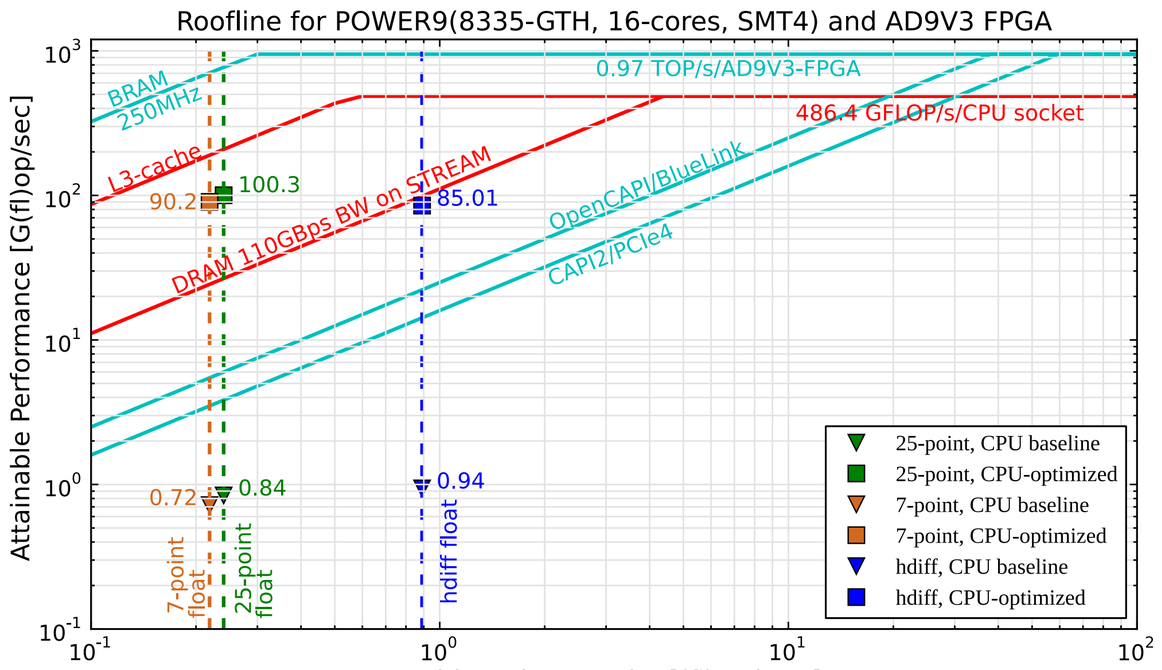
II. Performance Prediction Framework

- NAPEL Framework [3] can provide fast and accurate performance and energy prediction for a previously-unseen application
- Microarchitecture-independent characterization of an application with architectural simulation responses to train an ensemble algorithm
- Intelligent statistical techniques [4] to extract meaningful data with minimum experimental runs



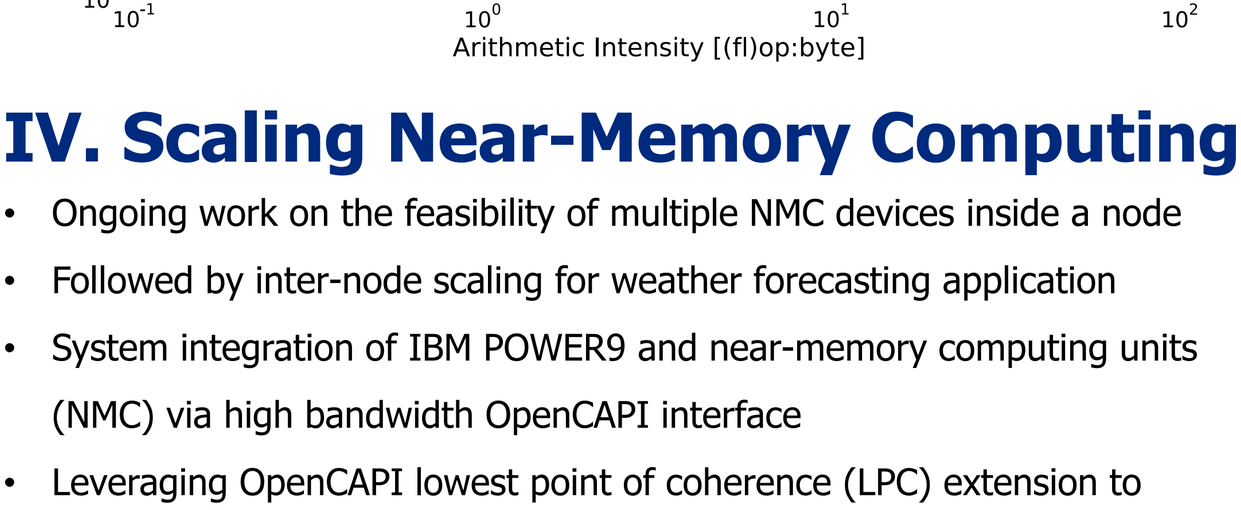
III. Stencil Computation on a CPU-Centric Architecture

- Implementing 3D stencils from weather application on IBM POWER9 CPU
- Performance gap due to a combination of memory hierarchy under-utilization and application's complex access patterns
- FPGA-aware roofline, which consists of AD9V3 FPGA board with CAPI2 and OpenCAPI links, indicates that FPGA is a promising device for near-memory platform



IV. Scaling Near-Memory Computing

- Ongoing work on the feasibility of multiple NMC devices inside a node
- Followed by inter-node scaling for weather forecasting application
- System integration of IBM POWER9 and near-memory computing units (NMC) via high bandwidth OpenCAPI interface
- Leveraging OpenCAPI lowest point of coherence (LPC) extension to attach an accelerator functional unit (AFU)
- LPC allows AFU memory to act as system memory. Hence, the AFUs can be used to for acceleration near system memory.



References

[1] Mutlu, O., et al. "Enabling Practical Processing in and near Memory for Data-Intensive Computing" *Proceedings of the 56th Design Automation Conference (DAC), 2019*

[2] Singh, G., et al. "A Review of Near-Memory Computing Architectures: Opportunities and Challenges." *21st Euromicro Conference on Digital System Design (DSD),* IEEE, 2018

[3] Singh, G., et al. "NAPEL: Near-Memory Computing Application Performance Prediction via Ensemble Learning." *Proceedings of the 56th Design Automation Conference (DAC), 2019*

[4] D. C. Montgomery, Design and anlysis of experiments, 2017