EINDHOVEN UNIVERSITY OF TECHNOLOGY ETH Zürich IBM Research | Zurich

NAPEL: Near-Memory Computing Application Performance Prediction via Ensemble Learning

Gagandeep Singh, Juan Gomez-Luna, Giovanni Mariani, Geraldo F. Oliveira, Stefano Corda, Sander Stuijk, Onur Mutlu, Henk Corporaal

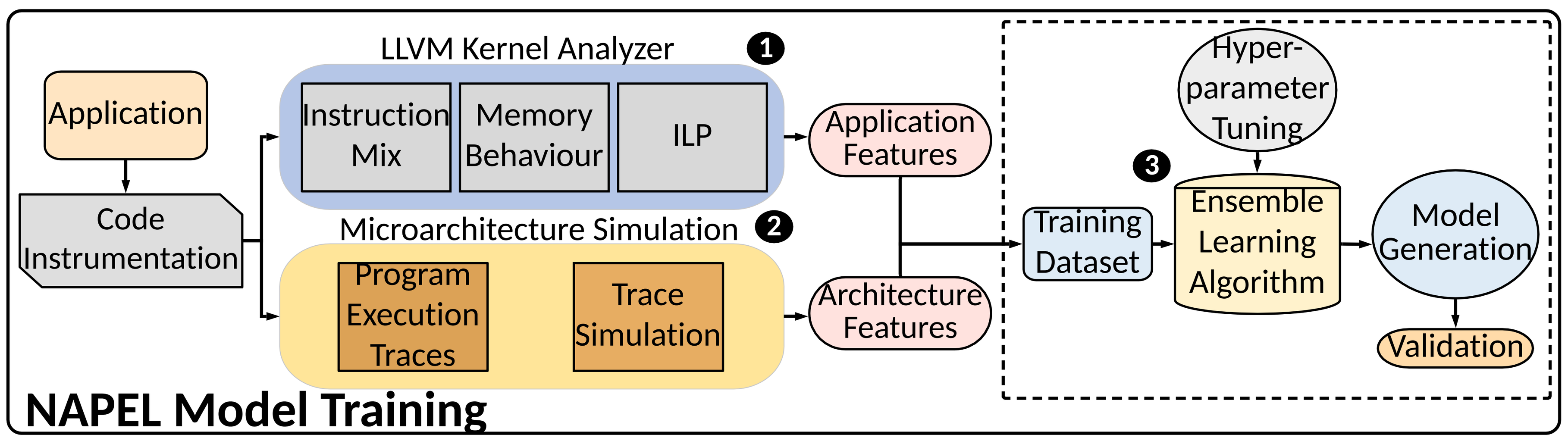
Funded by the Horizon 2020 Framework Programme of the European Union MSCA-ITN-EID

Motivation

- Exorbitant amount of data
- The high cost of energy for data movement
- A paradigm shift towards processing close to the data i.e., near-memory computing (NMC)
- However in early design-stage, simulation are extremely slow, imposing long run-time

NAPEL: Performance Prediction via Ensemble Machine Learning

- Fast and accurate performance and energy prediction for a previously-unseen application
- Microarchitecture-independent characterization with architectural simulation responses to train an ensemble algorithm
- Intelligent statistical techniques to extract meaningful data with minimum experimental runs



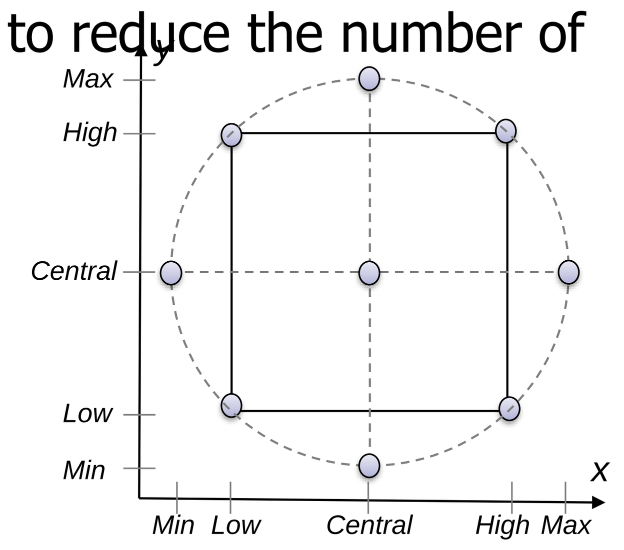
Phase 1: LLVM Kernel Analysis

- Microarchitecture-independent kernel analysis to generate an application profile independent of the NMC architecture

Application Feature	Description
Instruction Mix	The fraction of instruction types (integer, floating point, memory, etc.)
ILP	Instruction-level parallelism on an ideal machine
Data/Instruction reuse distance	For a given distance δ, probability of reusing one data element/instruction (in a certain memory location) before accessing δ other unique data elements/instructions (in different memory locations)
Memory traffic	Percentage of memory reads/writes that need to access the main memory, assuming a cache of size equal to the maximum reuse distance
Register traffic	An average number of registers per instruction
Memory footprint	Total memory size used by the application

Phase 2: Central Composite Design

- Design of experiment techniques [1] are used to reduce the number of experiments to train NAPEL
- Central composite design (CCD) is applied to minimize the uncertainty of a nonlinear polynomial model that accounts for parameter interactions
- In CCD, each input parameter can have five levels: *min, low, central, high, maximum*

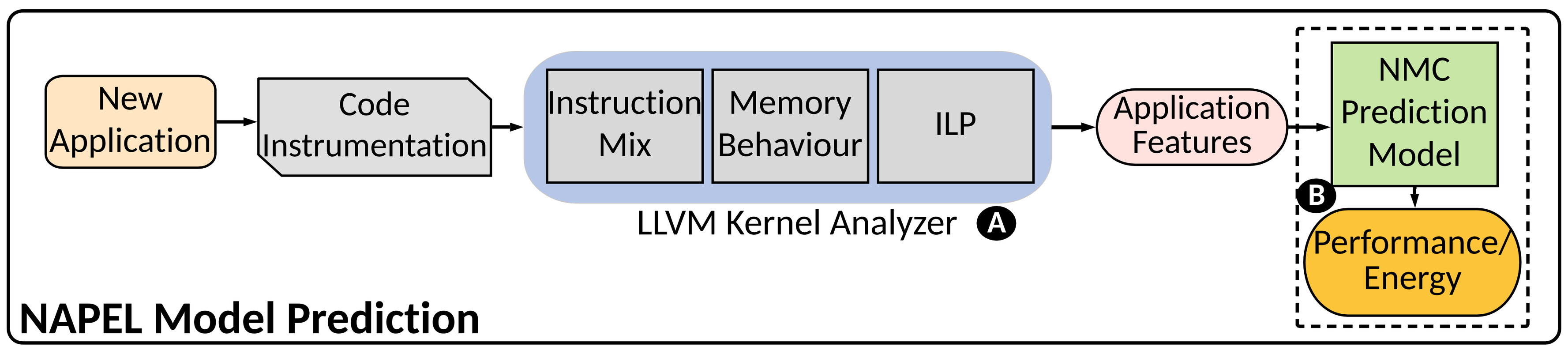


Phase 3: Ensemble Machine Learning

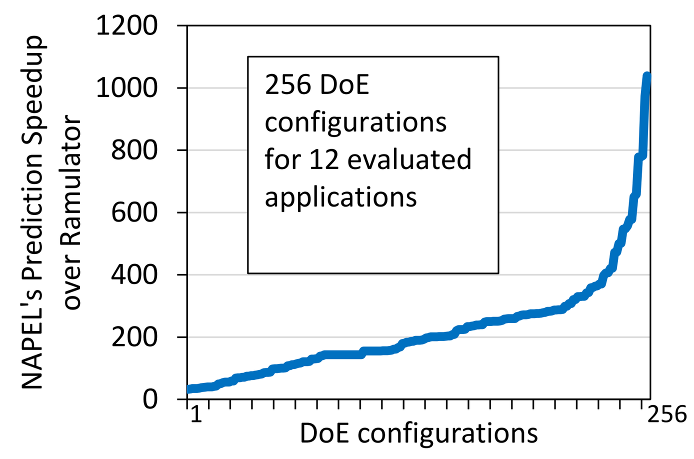
- We employ random forest (RF) as our ML algorithm, which embeds procedures to screen input features
- With hyper-parameters tuning to optimize the accuracy of ML algorithm

NAPEL Prediction

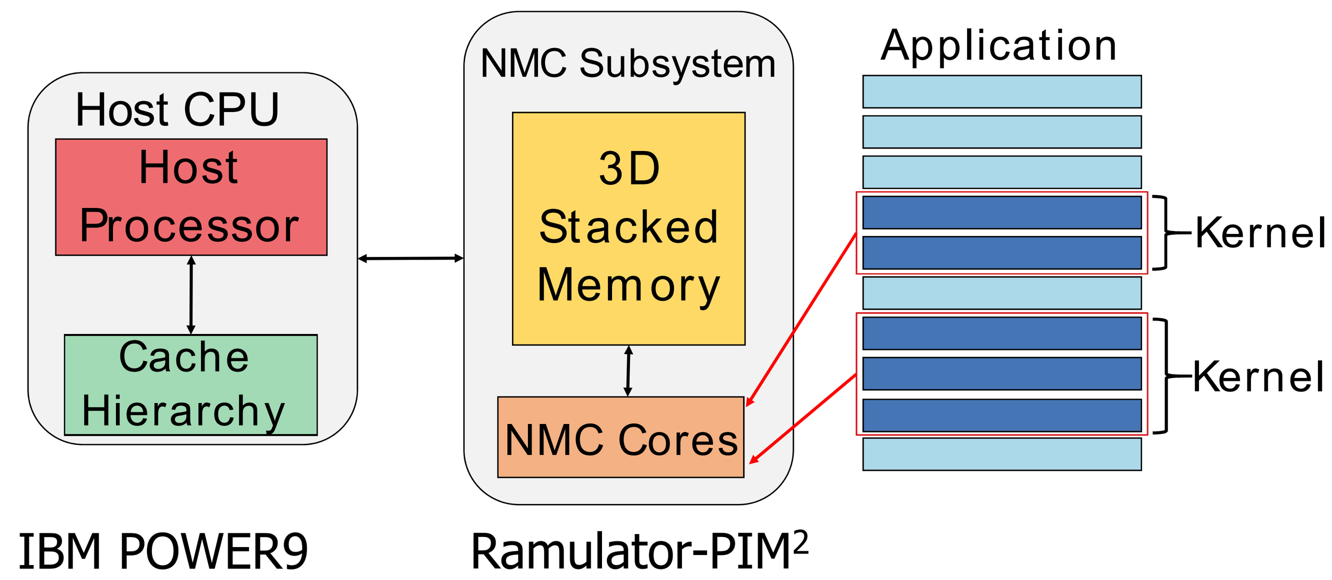
- Cross-platform prediction of a completely unseen application by only using micro-architectural independent application features



- 220x faster, on average, than our NMC simulator (min. 33x, max. 1039x)

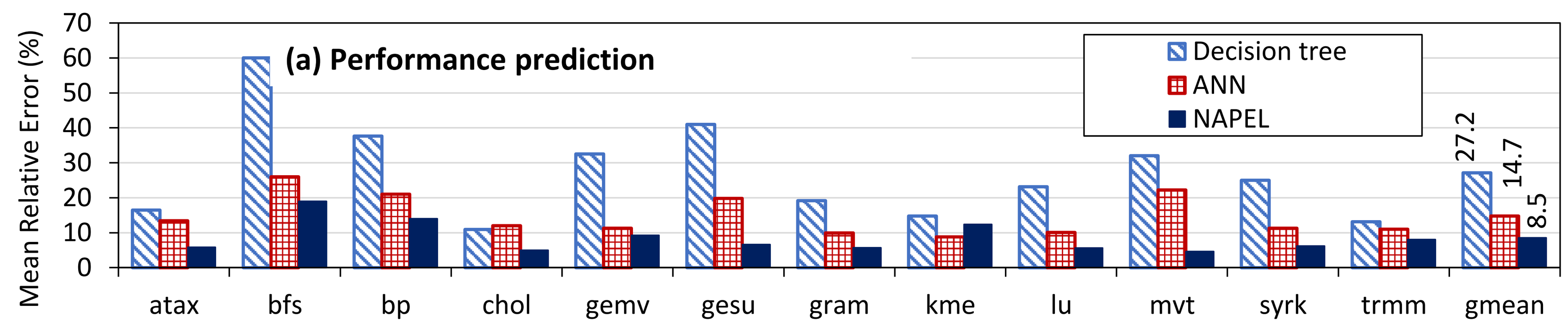


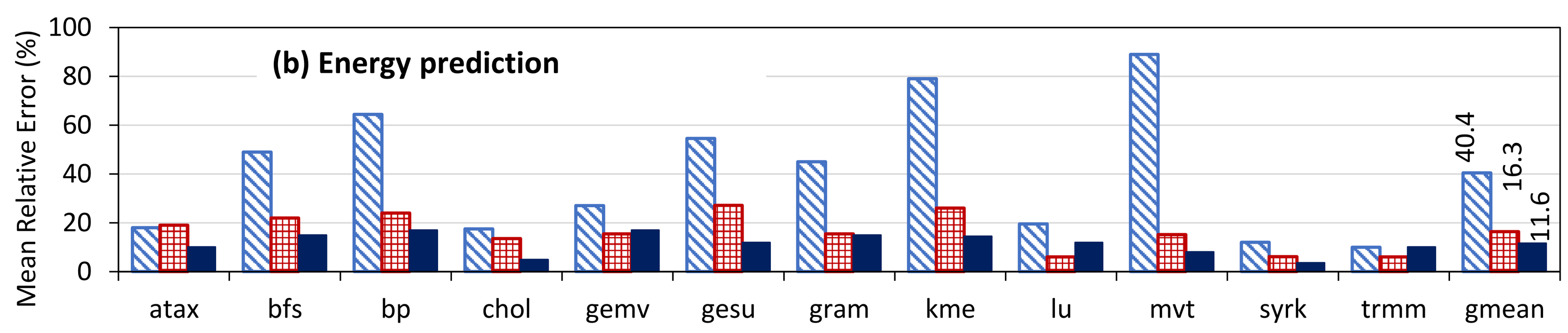
NMC Architecture



Evaluation

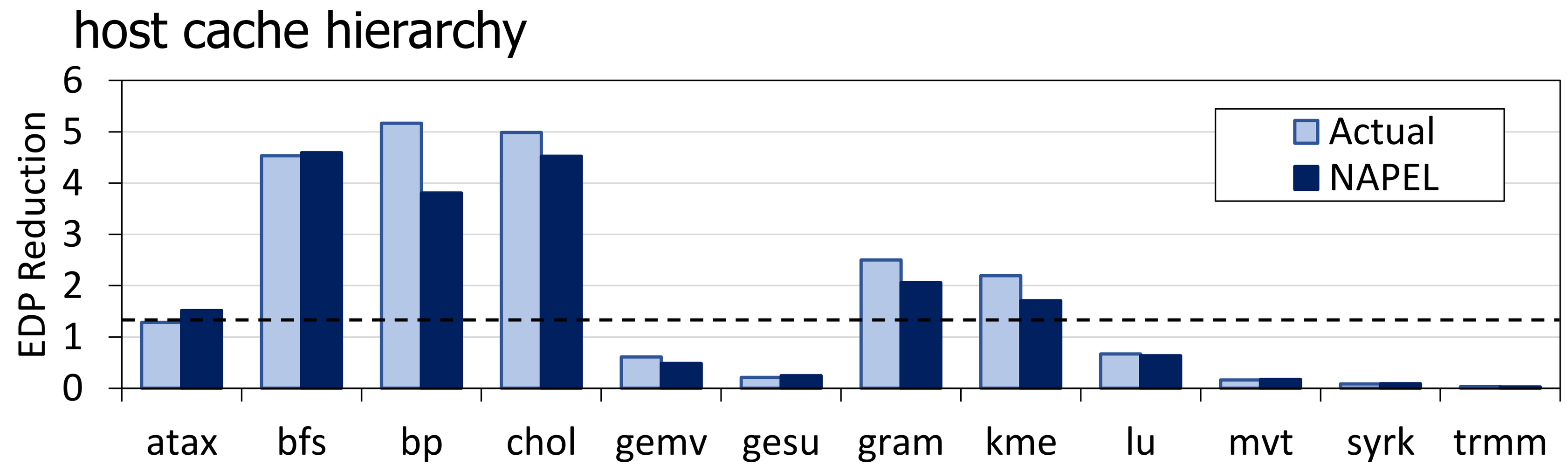
- MRE of 8.5% and 11.6% for performance and energy prediction
- NAPEL is 1.7x (1.4x) and 3.2x (3.5x) better in terms of performance (energy) estimation than ANN and decision tree





NMC Suitability Analysis

- NAPEL provides an accurate prediction of NMC suitability
- MRE between 1.3% to 26.3% (average 14.1) for EDP prediction
- Workloads with EDP<1, are not suitable for NMC and can leverage the host cache hierarchy



References

[1] D. C. Montgomery, Design and anlysis of experiments, (2017)

[2] https://github.com/CMU-SAFARI/ramulator-pim/