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

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

56th Design Automation Conference (DAC), Las Vegas
4th-June-2019

TU/e EINDHOVEN UNIVERSITY OF TECHNOLOGY
ETH zürich
IBM Research | Zurich

Executive Summary

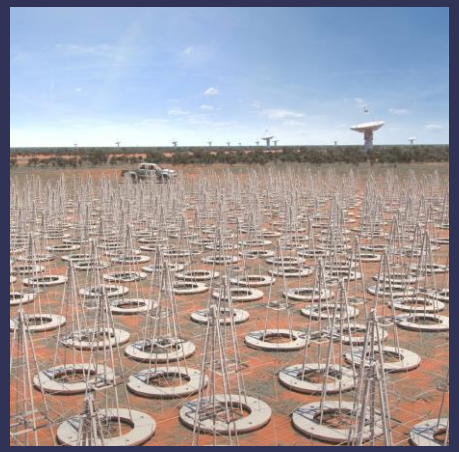
- **Motivation:** A promising paradigm to alleviate data movement bottleneck is *near-memory computing (NMC)*, which consists of placing compute units close to the memory subsystem
- **Problem:** Simulation times are extremely slow, imposing long run-time especially in the early-stage design space exploration
- **Goal:** A quick high-level performance and energy estimation framework for NMC architectures
- **Our contribution: NAPEL**
  - Fast and accurate performance and energy prediction for previously-unseen applications using ensemble learning
  - Use intelligent statistical techniques and micro-architecture-independent application features to minimize experimental runs
- **Evaluation**
  - NAPEL is, on average, 220x faster than state-of-the-art NMC simulator
  - Error rates (average) of 8.5% and 11.5% for performance and energy estimation

We open source Ramulator-PIM: https://github.com/CMU-SAFARI/ramulator-pim/

SKA
300PB

uploads on facebook
180PB

searches on Google
98PB



YouTube
15PB

CERN
15PB

NASDAQ
3PB

SKA
300PB

uploads on facebook

Massive amounts of data

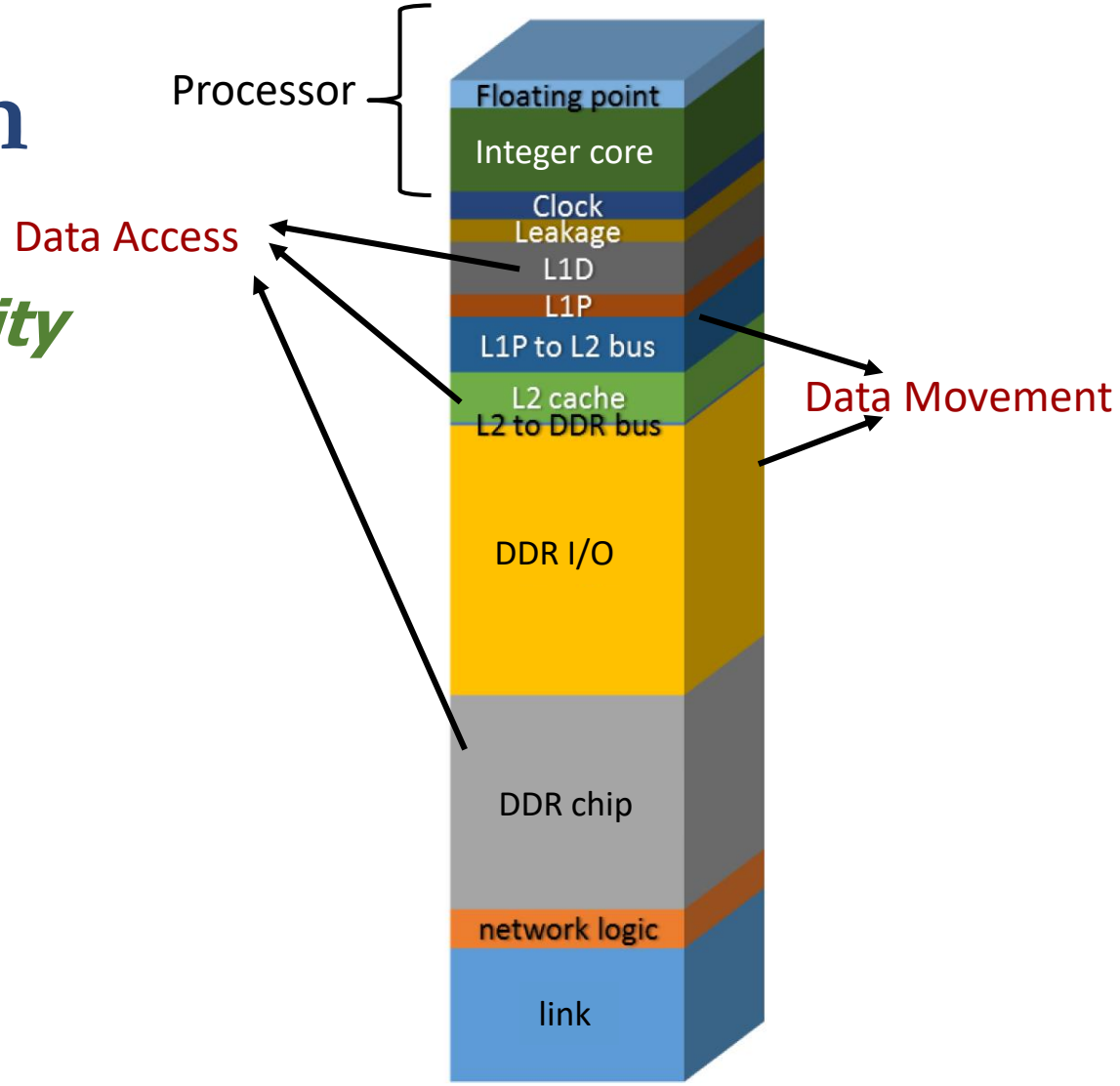
YouTube
15PB

CERN
15PB

NASDAQ
3PB

Compute Centric Approach

- Memory hierarchies take advantage of ***locality***
  - Spatial locality
  - Temporal locality
- Not suitable for all workloads
  - Graph processing
  - Neural networks
- Data access consumes a major part
  - Applications are increasingly data hungry
- Data movement energy dominates compute
  - Especially true for off-chip movement

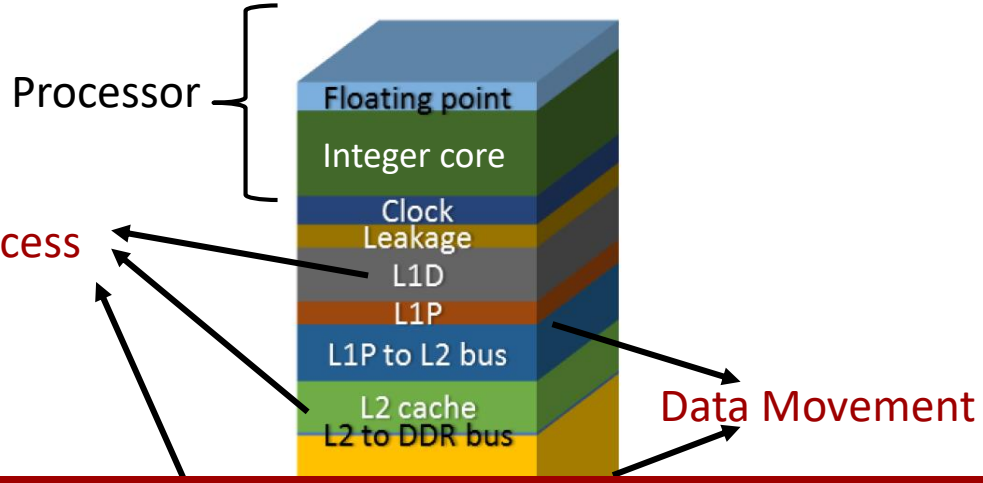


System-level power break down*

** R. Nair et al., "Active memory cube: A processing-in memory architecture for exascale systems", IBM J. Research Develop., vol. 59, no. 2/3, 2015*

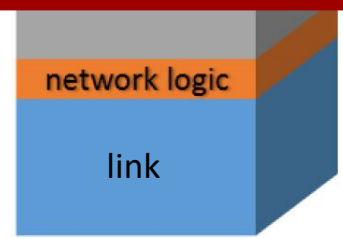
Compute Centric Approach

- Memory hierarchies take advantage of ***locality***
  - Spatial locality



Data movement bottleneck

- Data access consumes a major part
  - Applications are increasingly data hungry
- Data movement energy dominates compute
  - Especially true for off-chip movement

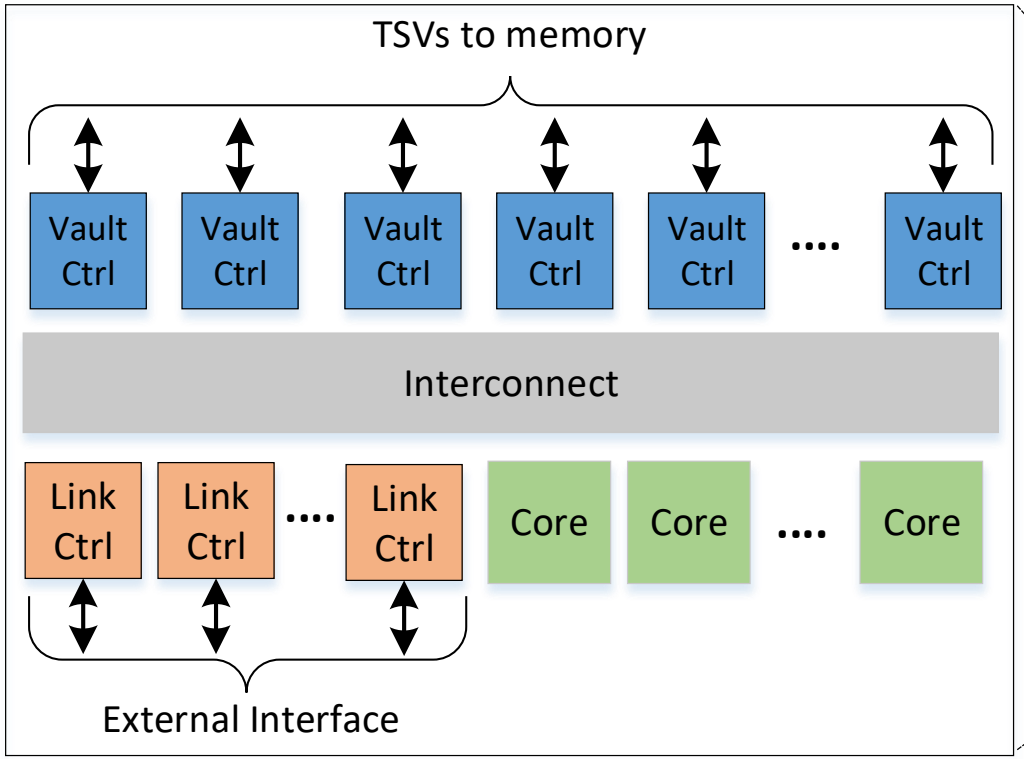


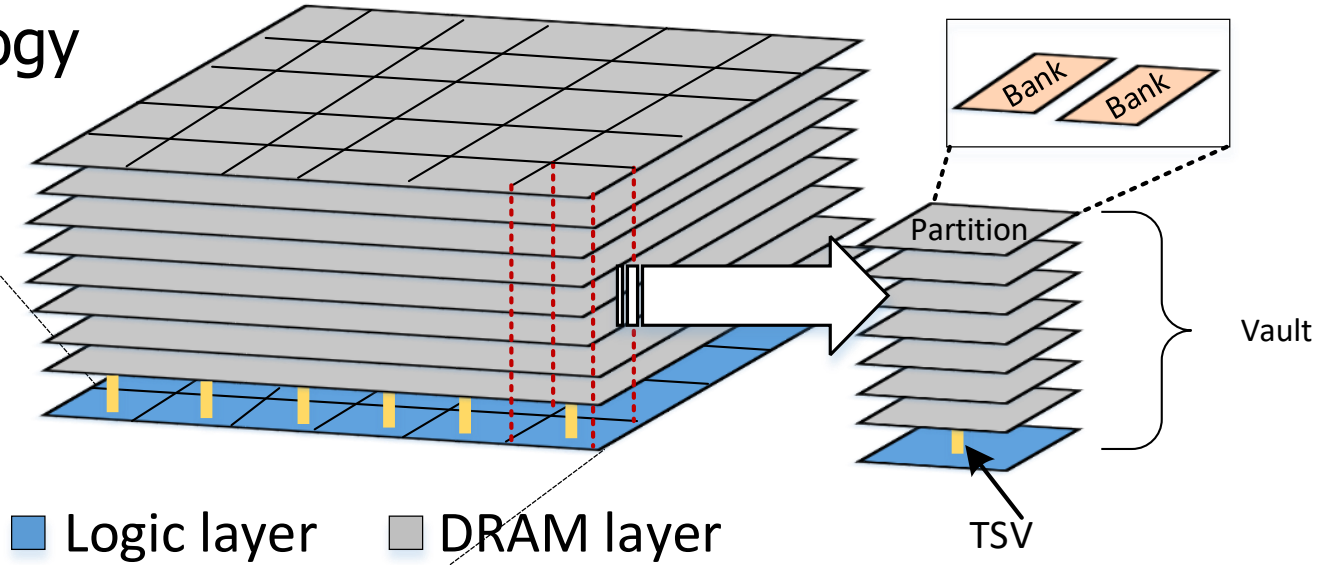
System-level power break down*

** R. Nair et al., "Active memory cube: A processing-in memory architecture for exascale systems", IBM J. Research Develop., vol. 59, no. 2/3, 2015*

Paradigm Shift - NMC

- Compute-centric to a data-centric approach
- Biggest enabler – stacking technology





NMC Simulators

- Simulation for:
  - Design space exploration (DSE)
  - Workload suitability analysis
- NMC Simulators:
  - Sinuca, 2015
  - HMC-SIM, 2016
  - CasHMC, 2016
  - Smart Memory Cube (SMC), 2016
  - CLAPPS, 2017
  - Gem5+HMC, 2017
  - Ramulator-PIM[1], 2019

[1]Ramulator-PIM: https://github.com/CMU-SAFARI/ramulator-pim/

NMC Simulators

- Simulation for:
  - Design space exploration (DSE)
  - Workload suitability analysis

Simulation of real workloads can be 10000x slower than native-execution!!!

  - Smart Memory Cube (SMC), 2016
  - CLAPPS, 2017
  - Gem5+HMC, 2017
  - Ramulator-PIM[1], 2019

[1]Ramulator-PIM: https://github.com/CMU-SAFARI/ramulator-pim/

NMC Simulators

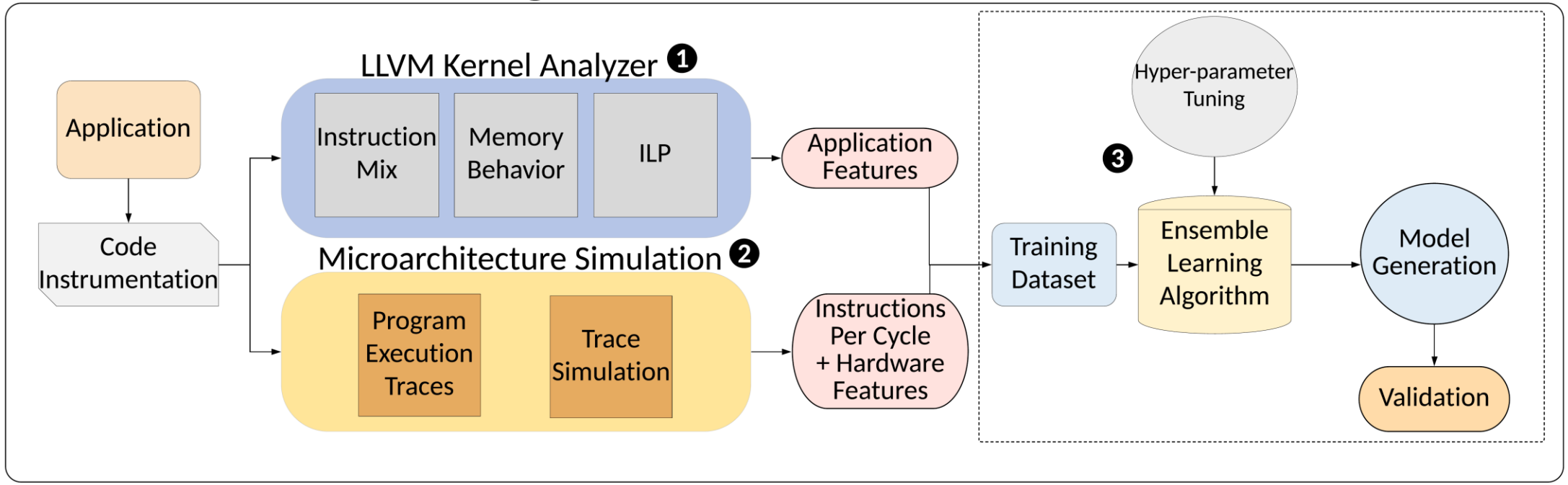
- Simulation for:
  - Design space exploration (DSE)
  - Workload suitability analysis
- Smart Memory Cube (SMC), 2016
- CLAPPS, 2017
- Gem5+HMC, 2017
- Ramulator-PIM[1], 2019

Idea: Leverage ML with statistical techniques for quick NMC performance/energy prediction

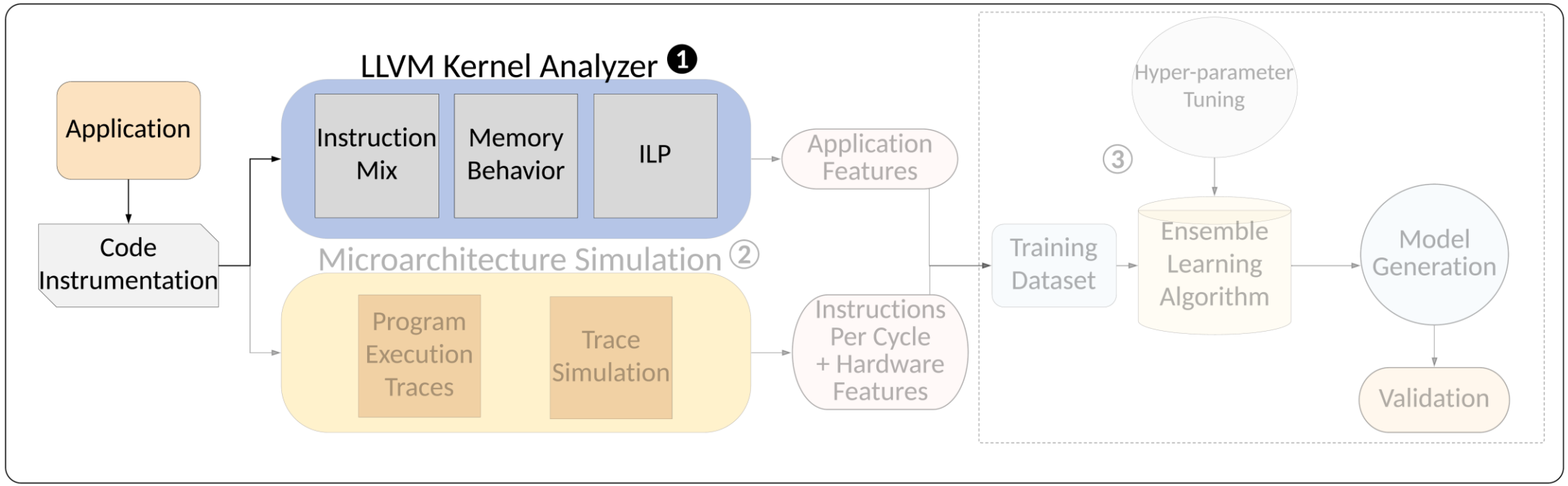
[1]Ramulator-PIM: https://github.com/CMU-SAFARI/ramulator-pim/

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

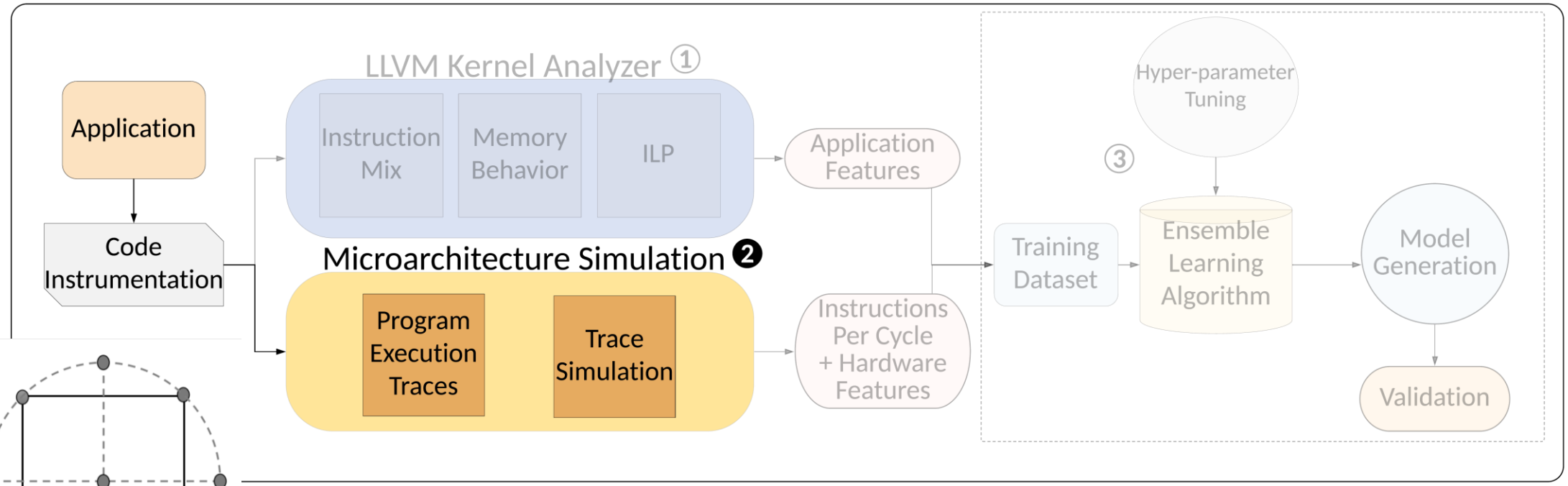
NAPEL Model Training

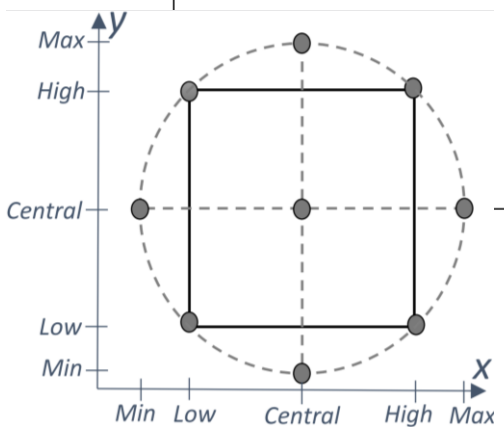


Phase 1: LLVM Analyzer



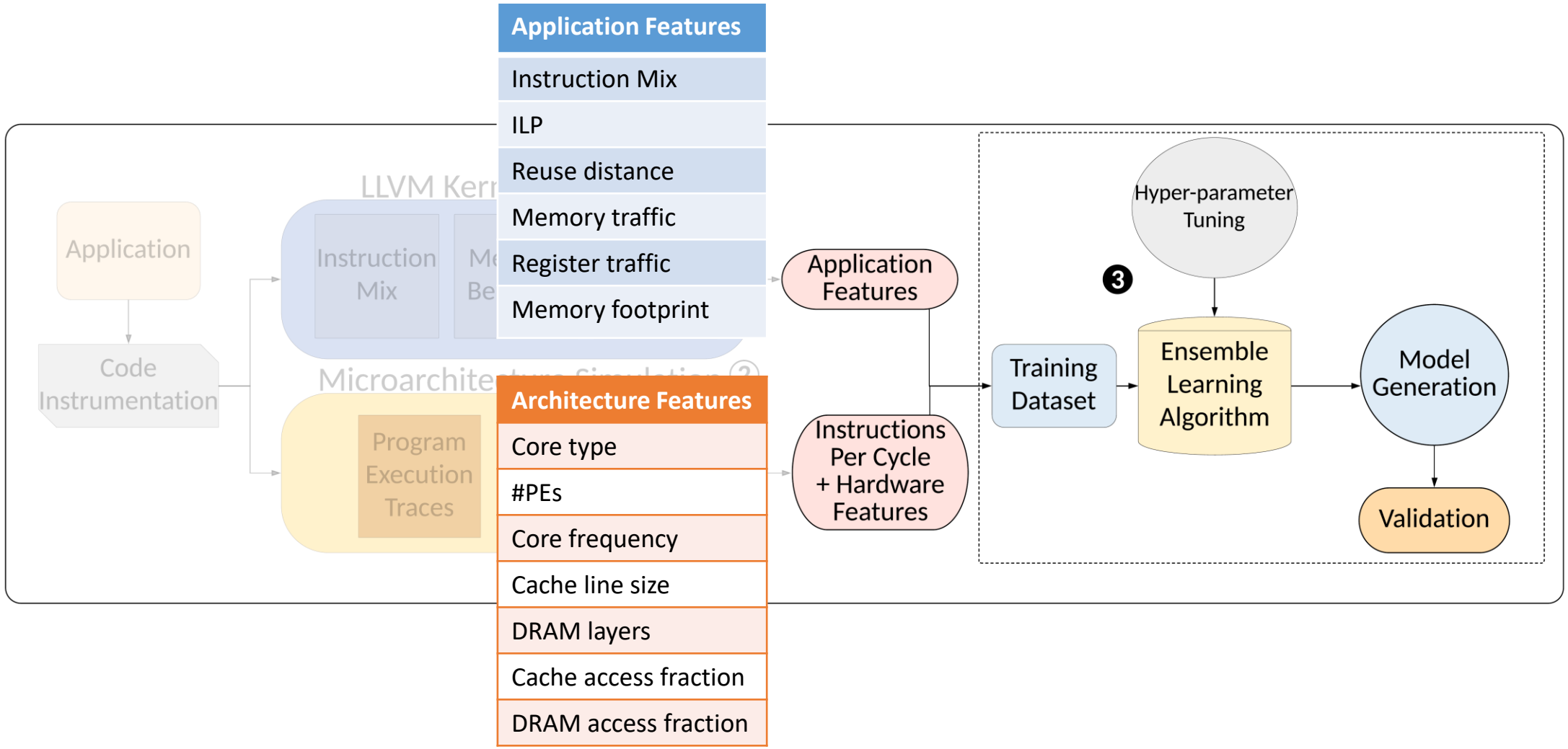
Phase 2: Microarchitecture Simulation



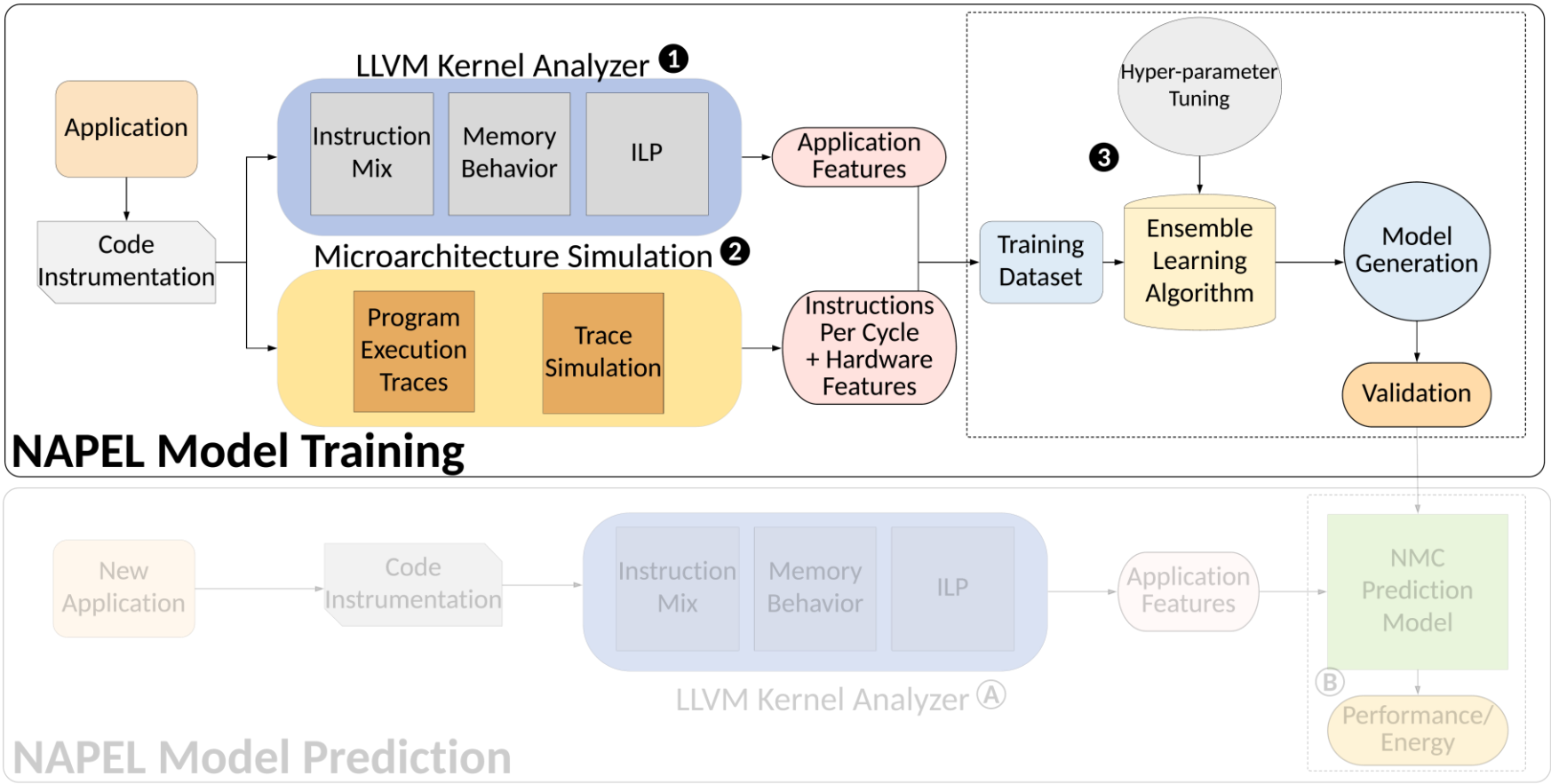


Central composite design of experiments technique to minimize the number of experiments while data collection

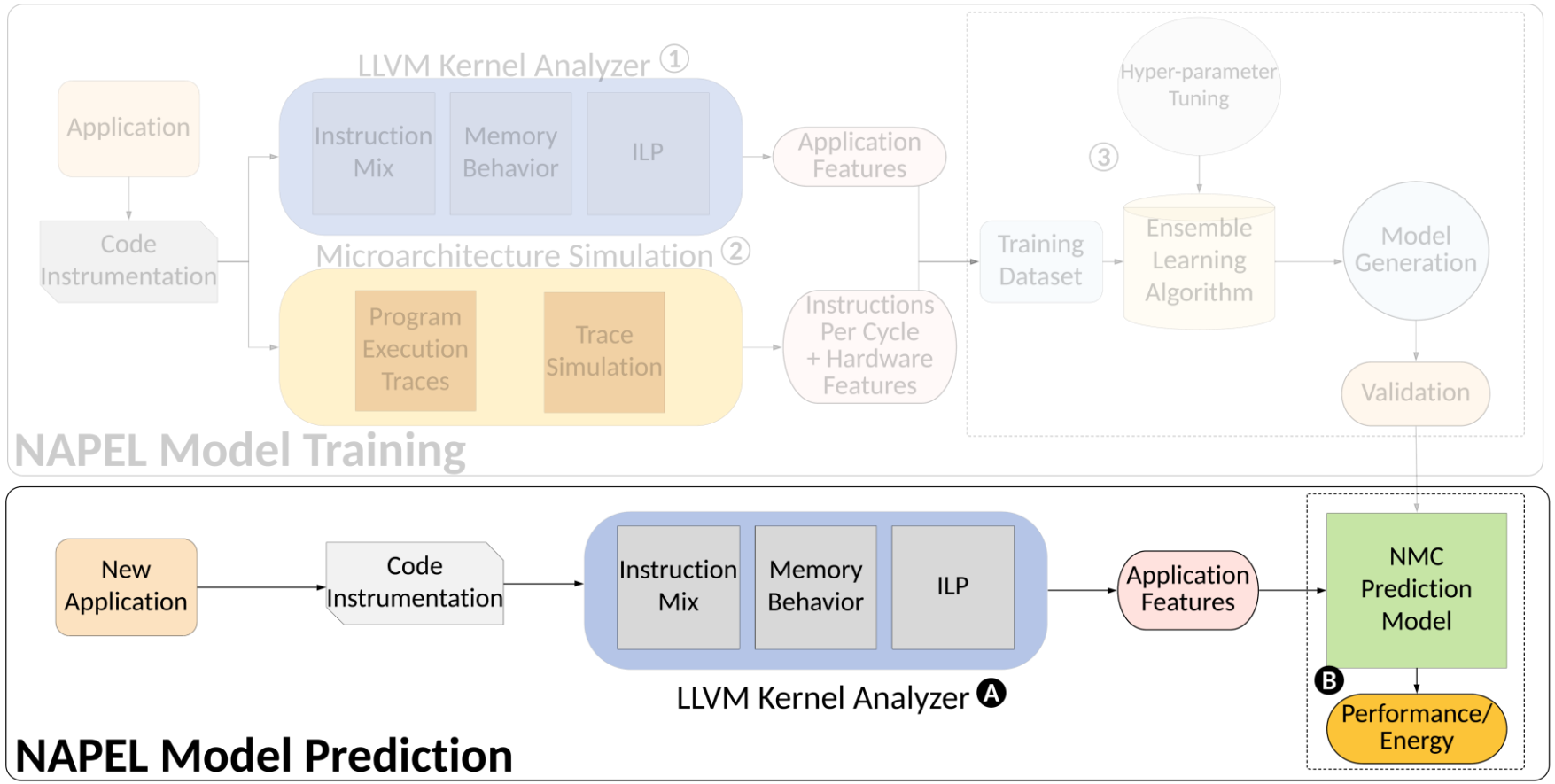
Phase 3: Ensemble ML Training



NAPEL Framework

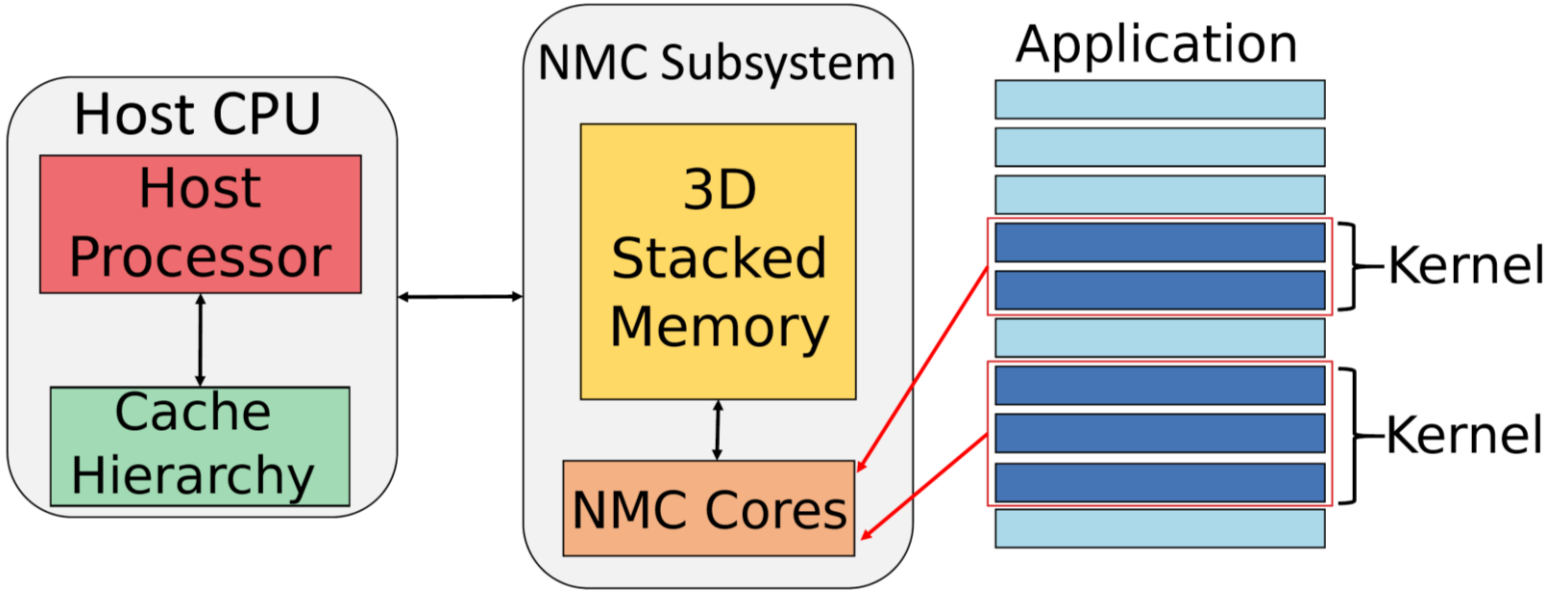


NAPEL Prediction



Experimental Setup

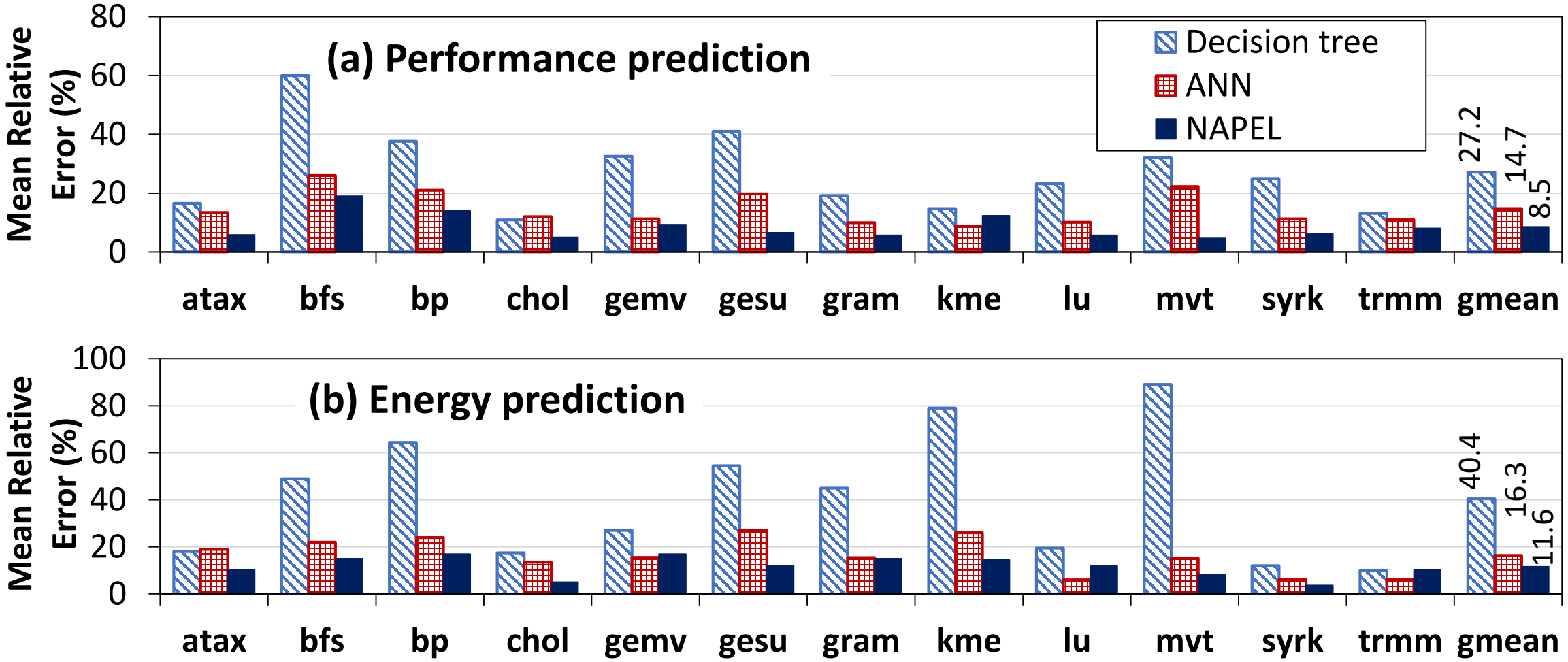
- Host System
  - **IBM POWER9**
  - Power: AMESTER
- NMC Subsystem
  - **Ramulator-PIM[1]**



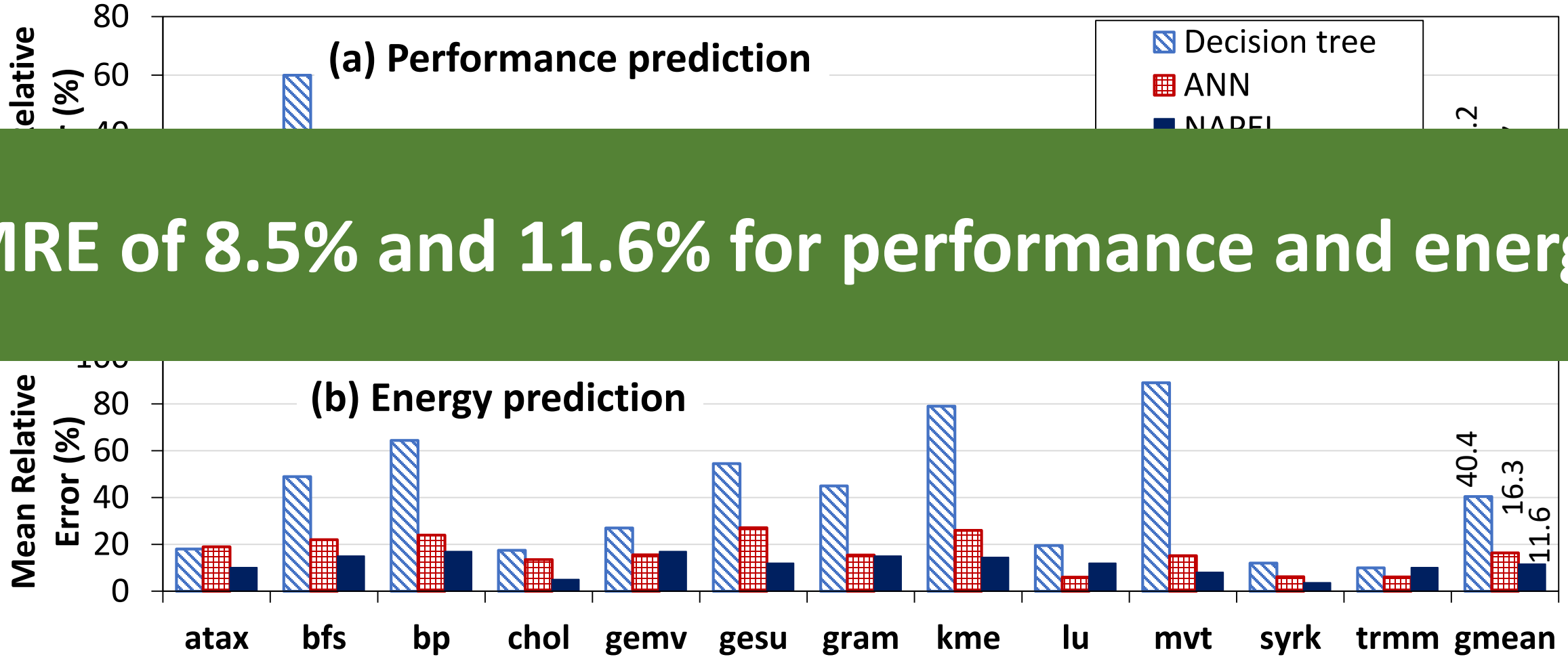
- Workloads
  - **PolyBench** and **Rodinia**
  - Heterogeneous workloads such as image processing, machine learning, graph processing etc.
- Accuracy in terms of mean relative error (MRE)

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

NAPEL Accuracy: Performance and Energy Estimates



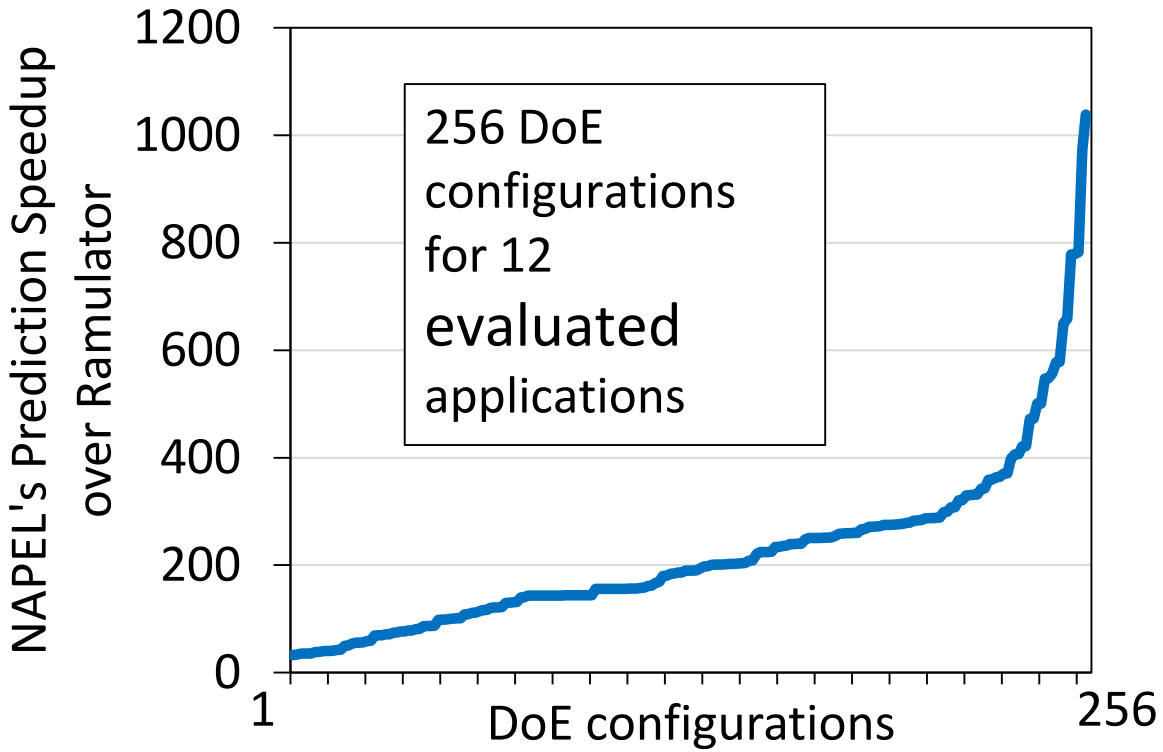
NAPEL Accuracy: Performance and Energy Estimates



MRE of 8.5% and 11.6% for performance and energy

Speed of Evaluation

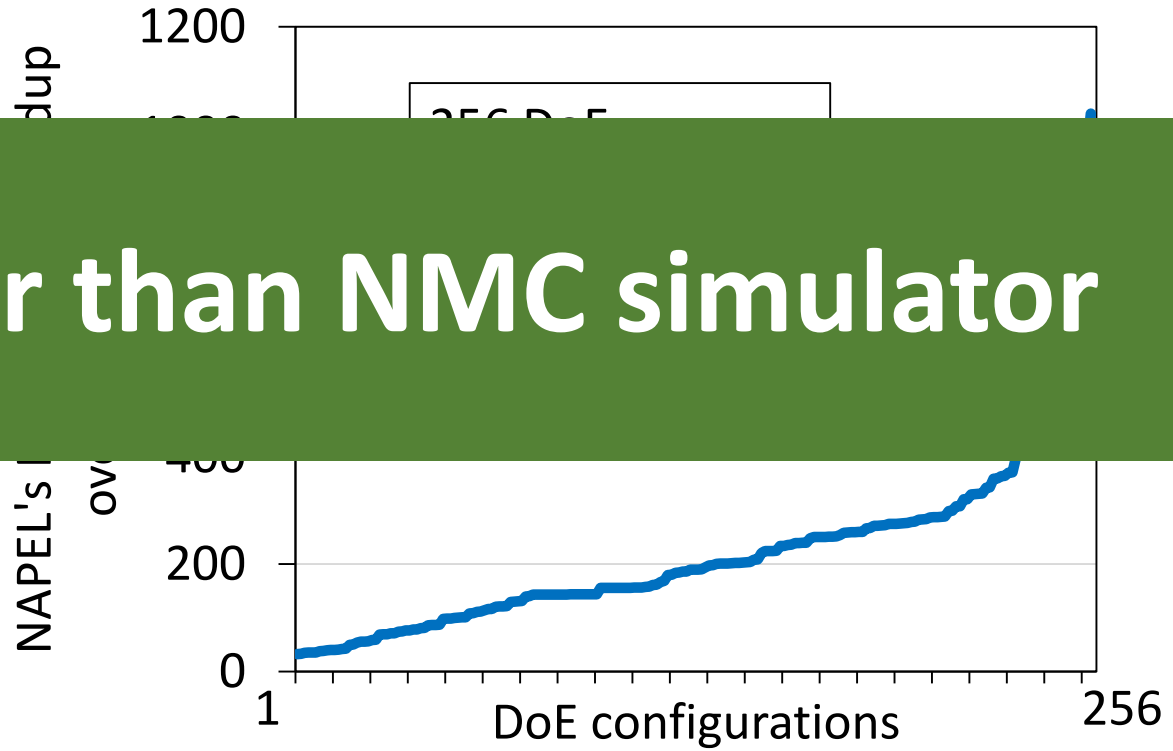
Application	Training/Prediction Time			
Name	#DoE conf.	DoE run (mins)	Train+Tune (mins)	Pred. (mins)
atax	11	522	34.9	0.49
bfs	31	1084	34.2	0.48
bp	31	1073	43.8	0.47
chol	19	741	34.9	0.49
gemv	19	741	24.4	0.51
gesu	19	731	36.1	0.51
gram	19	773	36.5	0.52
kme	31	742	36.9	0.55
lu	19	633	37.9	0.51
mvt	19	955	38.0	0.54
syrk	19	928	35.7	0.51
trmm	19	898	37.6	0.48



Speed of Evaluation

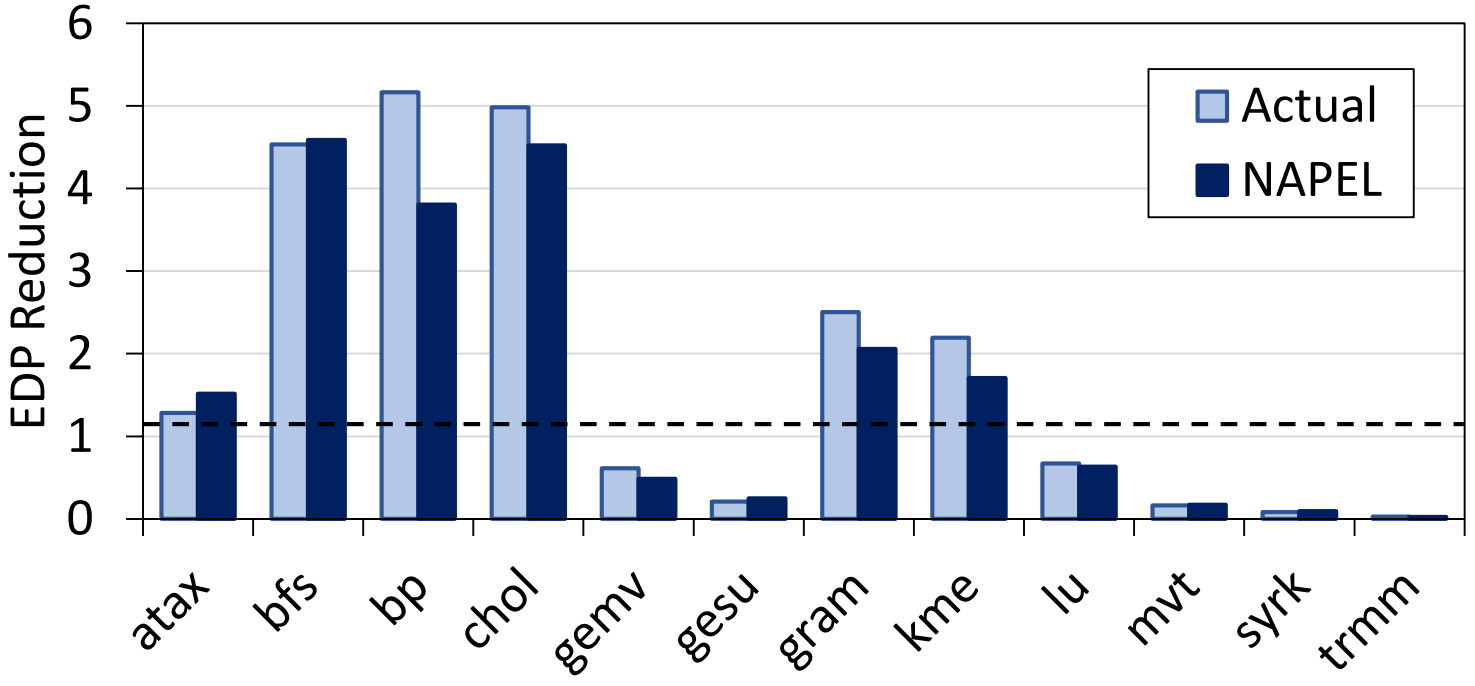
Application	Training/Prediction Time			
kme	31	742	36.9	0.55
lu	19	633	37.9	0.51
mvt	19	955	38.0	0.54
syrk	19	928	35.7	0.51
trmm	19	898	37.6	0.48

220x (up to 1039x) faster than NMC simulator



Use Case: NMC Suitability Analysis

- Assess the potential of offloading a workload to NMC
- NAPEL provides accurate prediction of NMC suitability
- MRE between 1.3% to 26.3% (average 14.1%)



Conclusion and Summary

- **Motivation:** A promising paradigm to alleviate data movement bottleneck is *near-memory computing (NMC)*, which consists of placing compute units close to the memory subsystem
- **Problem:** Simulation times are extremely slow, imposing long run-time especially in the early-stage design space exploration
- **Goal:** A quick high-level performance and energy estimation framework for NMC architectures
- **Our contribution: NAPEL**
  - Fast and accurate performance and energy prediction for previously-unseen applications using ensemble learning
  - Use intelligent statistical techniques and micro-architecture-independent application features to minimize experimental runs
- **Evaluation**
  - NAPEL is, on average, 220x faster than state-of-the-art NMC simulator
  - Error rates (average) of 8.5% and 11.5% for performance and energy estimation

We open source Ramulator-PIM: https://github.com/CMU-SAFARI/ramulator-pim/

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

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

56th Design Automation Conference (DAC), Las Vegas
4th-June-2019

TU/e EINDHOVEN UNIVERSITY OF TECHNOLOGY
ETH zürich
IBM Research | Zurich