Thermal Management for 3D-Stacked Systems via Unified Core-Memory Power Regulation.

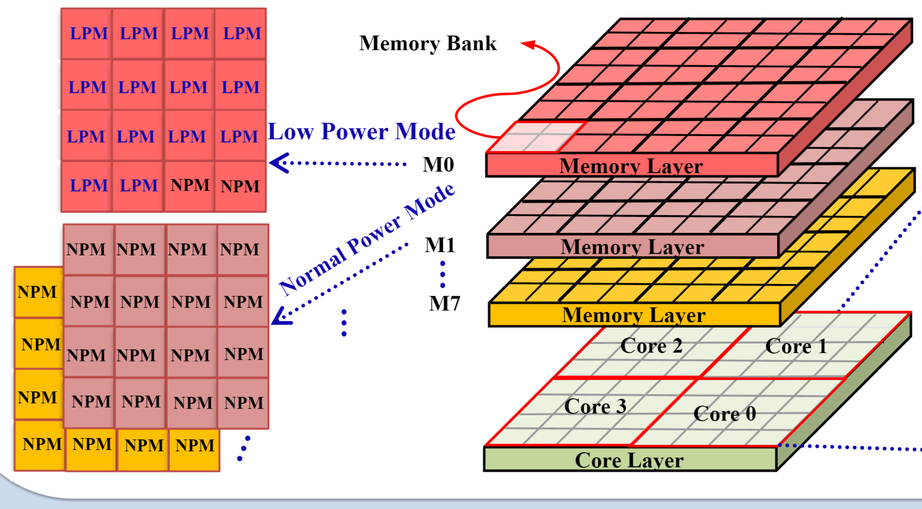
UNIVERSITY OF AMSTERDAM
Informatics Institute

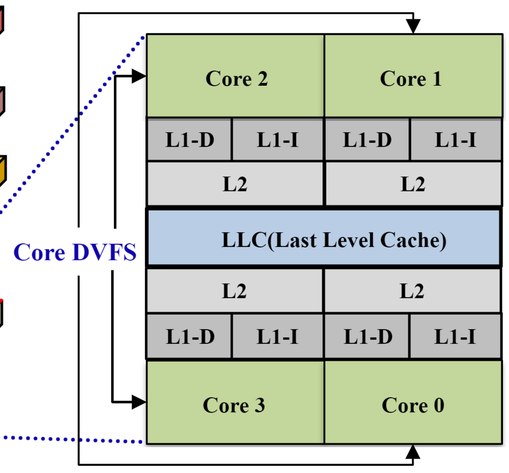
PARALLEL COMPUTING SYSTEMS

Y. Shen, L. Schreuders, A. Pathania, A.D. Pimentel

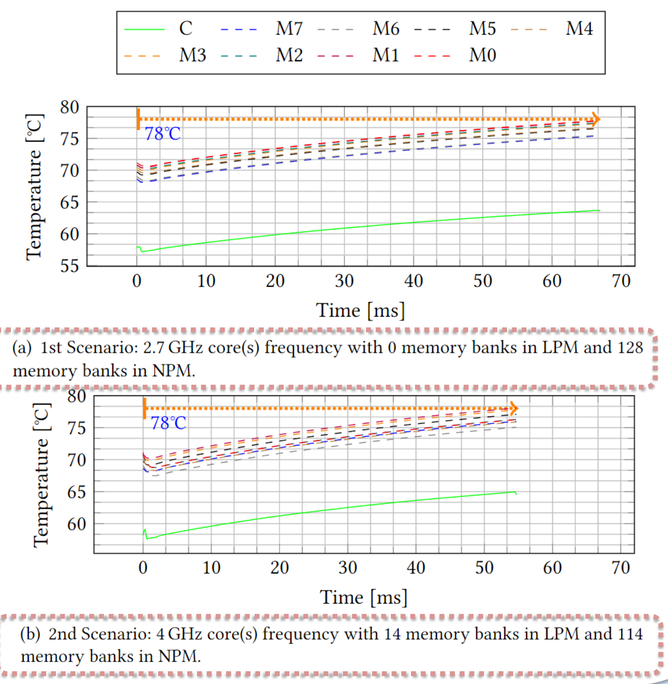
Motivational Example

- ❑ Streamcluster running in two scenarios
- ❑ Computer-intensive benchmarks **benefit more from higher core frequency** than from LPM penalties





DVFS vs LPM

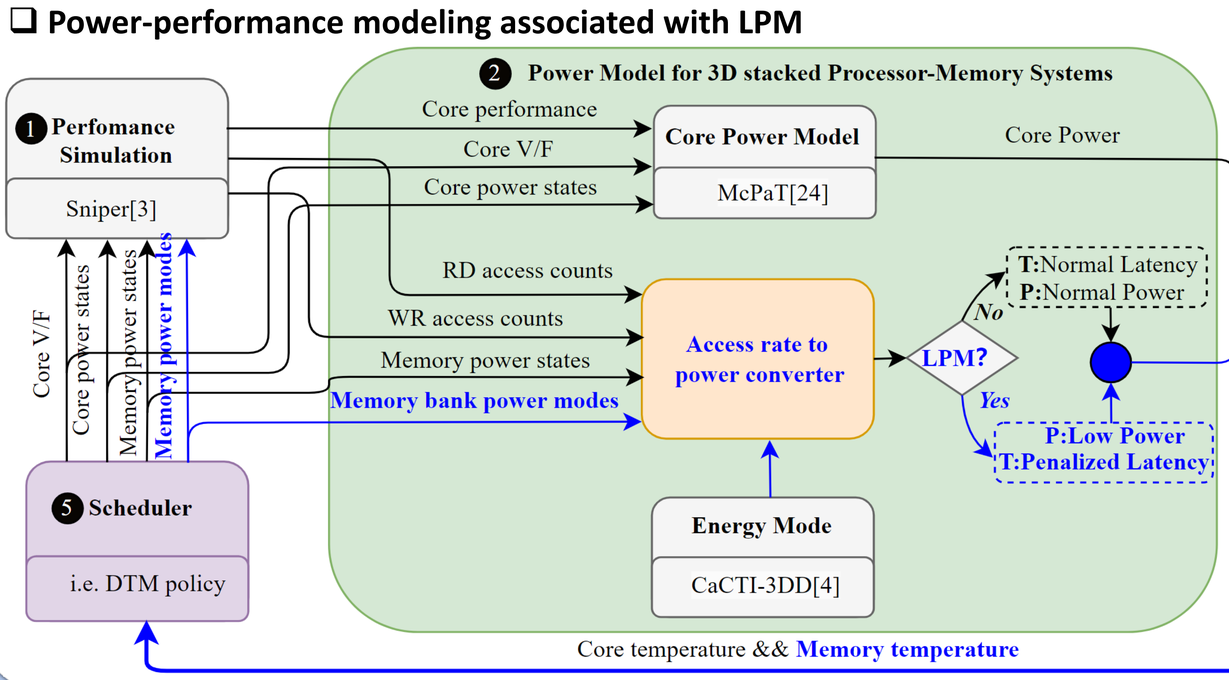


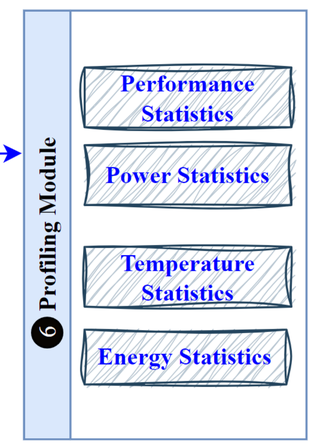
(a) 1st Scenario: 2.7 GHz core(s) frequency with 0 memory banks in LPM and 128 memory banks in NPM.

(b) 2nd Scenario: 4 GHz core(s) frequency with 14 memory banks in LPM and 114 memory banks in NPM.

LPM Implementation(the extended CoMeT tool flow)

- ❑ Power-performance modeling associated with LPM



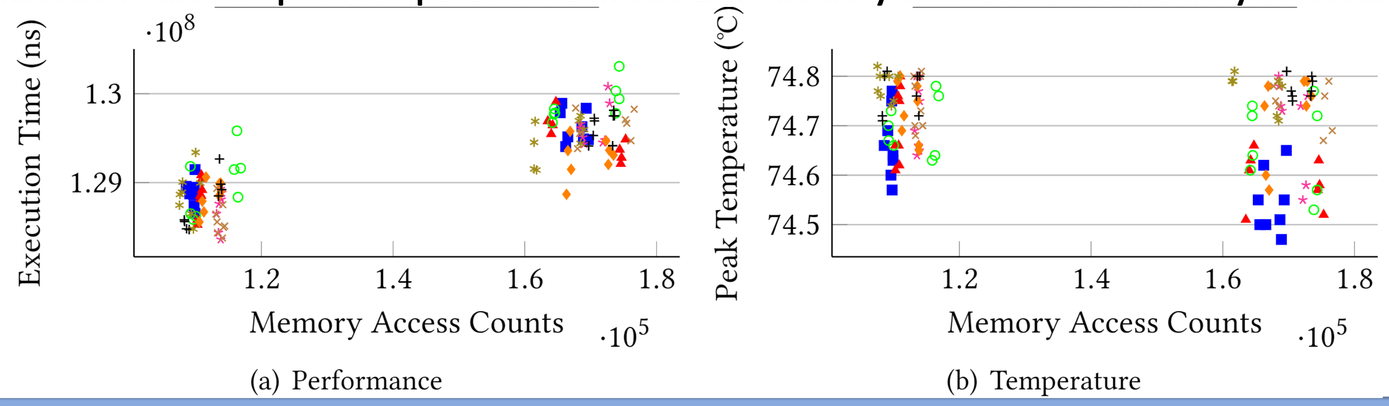


➢ ***Key features of LPM***

- ❑ Data preserved in situ in LPM, avoiding migration.
- ❑ Data inaccessible in LPM mode
- ❑ Re-accessing data requires toggling to NPM

DRAM Access Analysis for Low Power Mode

- ❑ Execution time and peak temperature with one LPM memory bank versus its memory access count



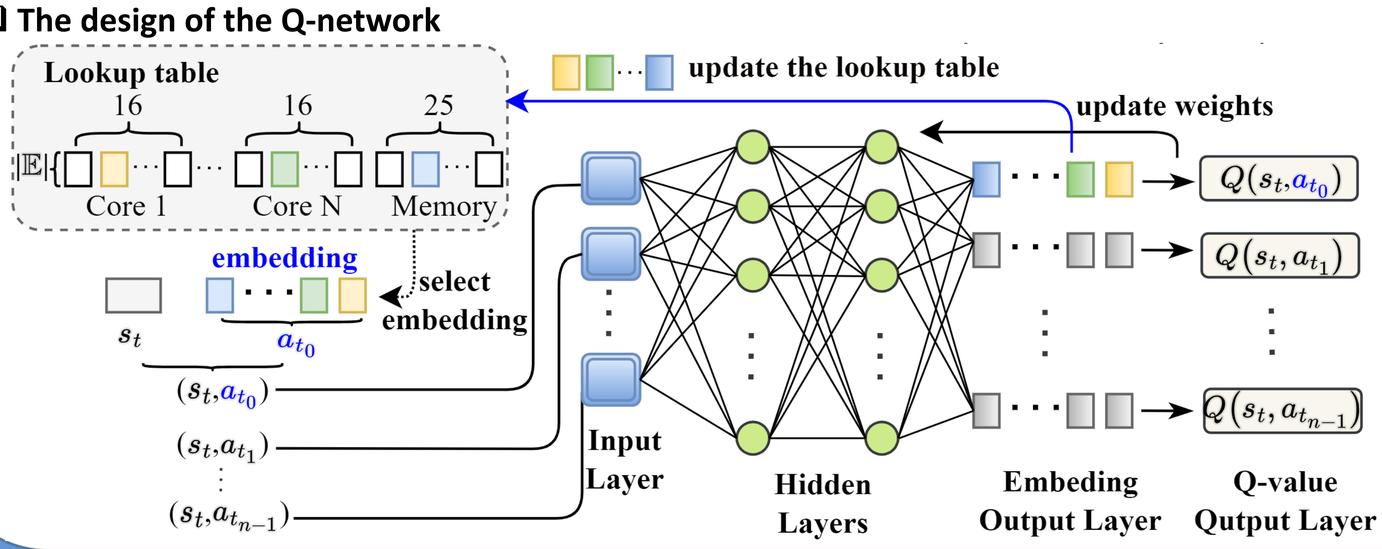
(a) Performance (b) Temperature

➢ ***Observations***

- ❑ Memory banks near the PCB layer with fewer access counts **offer better performance and thermal benefits** in LPM
- ❑ Lower layer banks with high access counts **face higher performance penalties and reduced thermal benefits** in LPM

3QUTM: A Unified DVFS and LPM Thermal Scheduler Leveraging Deep Q-learning

- ❑ The design of the Q-network



➢ ***Action Representations***

- ❑ Action embeddings
- ❑ Parameterized actions

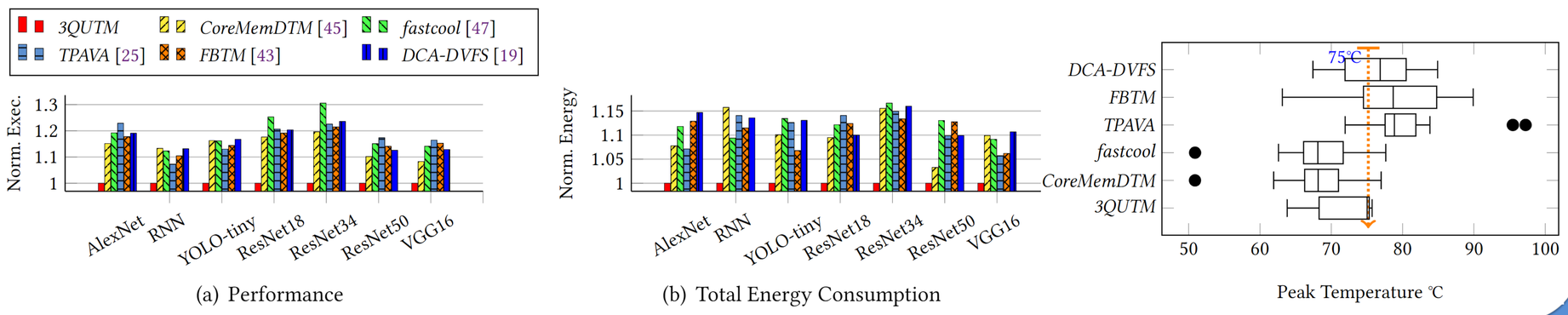
➢ ***Prioritized Experience Replay Buffer***

- ❑ Improve the sample efficiency
- ❑ Facilitate faster convergence

➢ ***Fine-Tuning for Increased Adaptability***

Evaluation

- ❑ 16 cores per layer-1 layer
- ❑ 16 memory banks per layer-8 Layers
- ❑ The arrival time of tasks follows Poisson distribution



(a) Performance (b) Total Energy Consumption