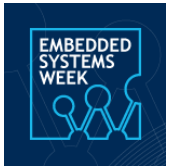
UNIVERSITY OF AMSTERDAM
Informatics Institute



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

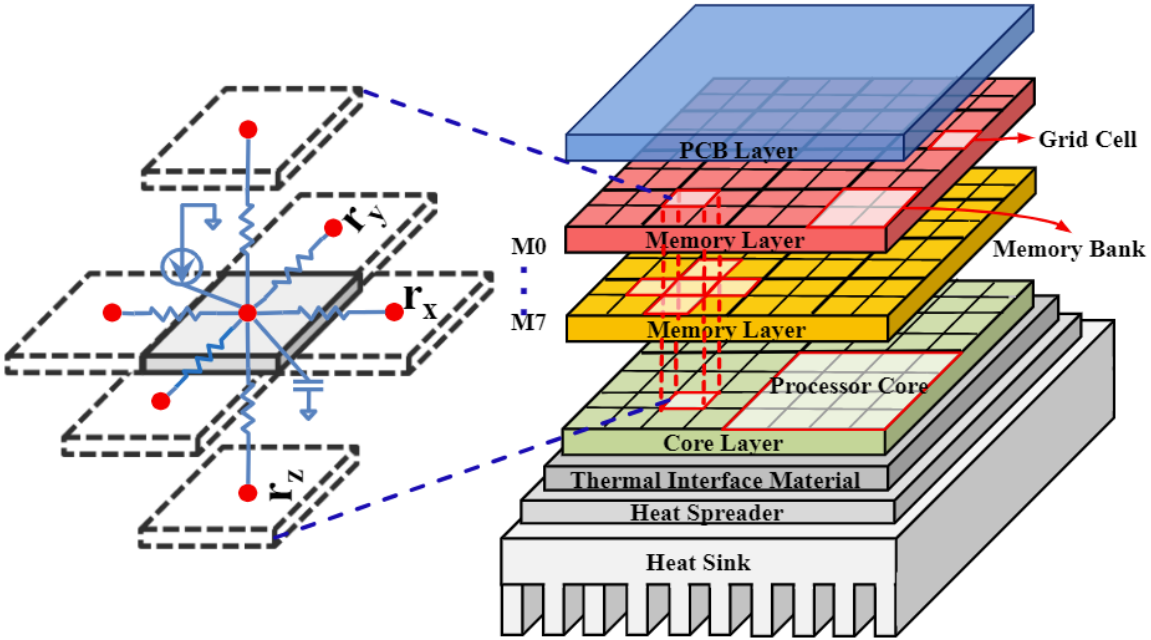
Y. Shen, L. Schreuders, A. Pathania, A.D. Pimentel

Outline

- ***Research background***
- A motivational example
- LPM implementation
- DRAM access profiling
- 3QUTM: a unified thermal scheduler via Deep Q-learning
- Experimental results
- Conclusion

Research background→3D-stacked systems

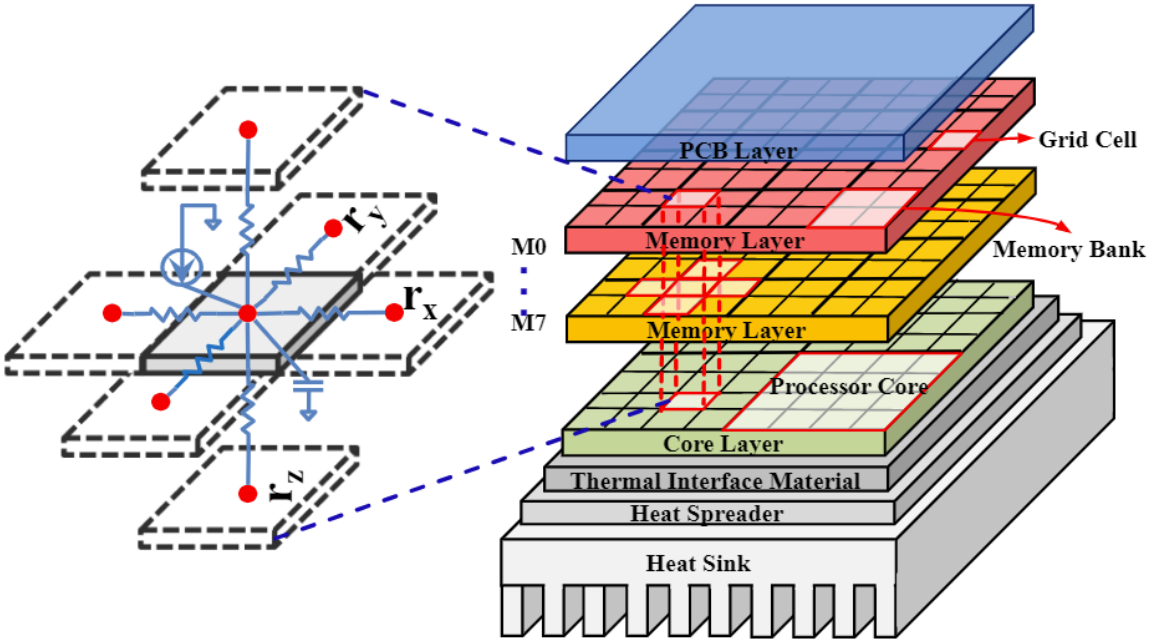
❑ 3D-stacked Processor-Memory Architecture



- Allow Heterogenous Integration
- Stacked (Logic/DARM) Dies
  - TSV(Through Silicon Via) interconnection
  - More compact material
- Higher Bandwidth via TSV
- Lower Latency

Research background→3D-stacked systems

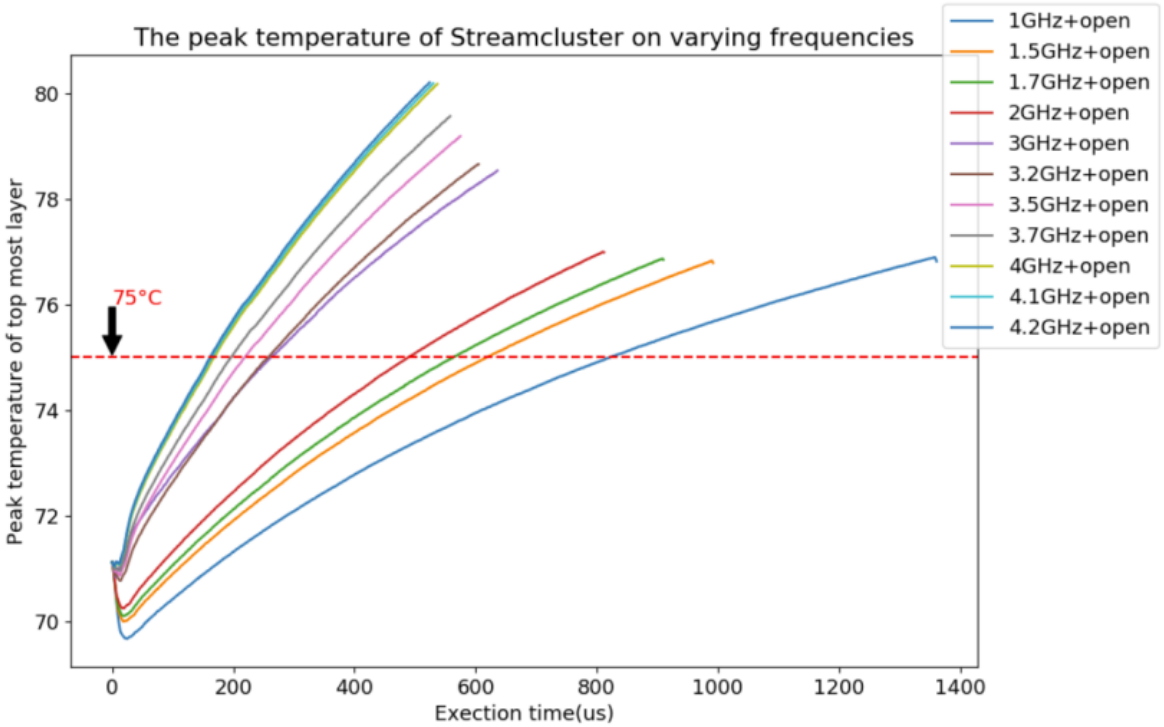
❑ 3D-stacked Processor-Memory Architecture



- Higher cost
  - Simulation needed
  - CoMeT
- Higher power density

Research background→3D-stacked systems

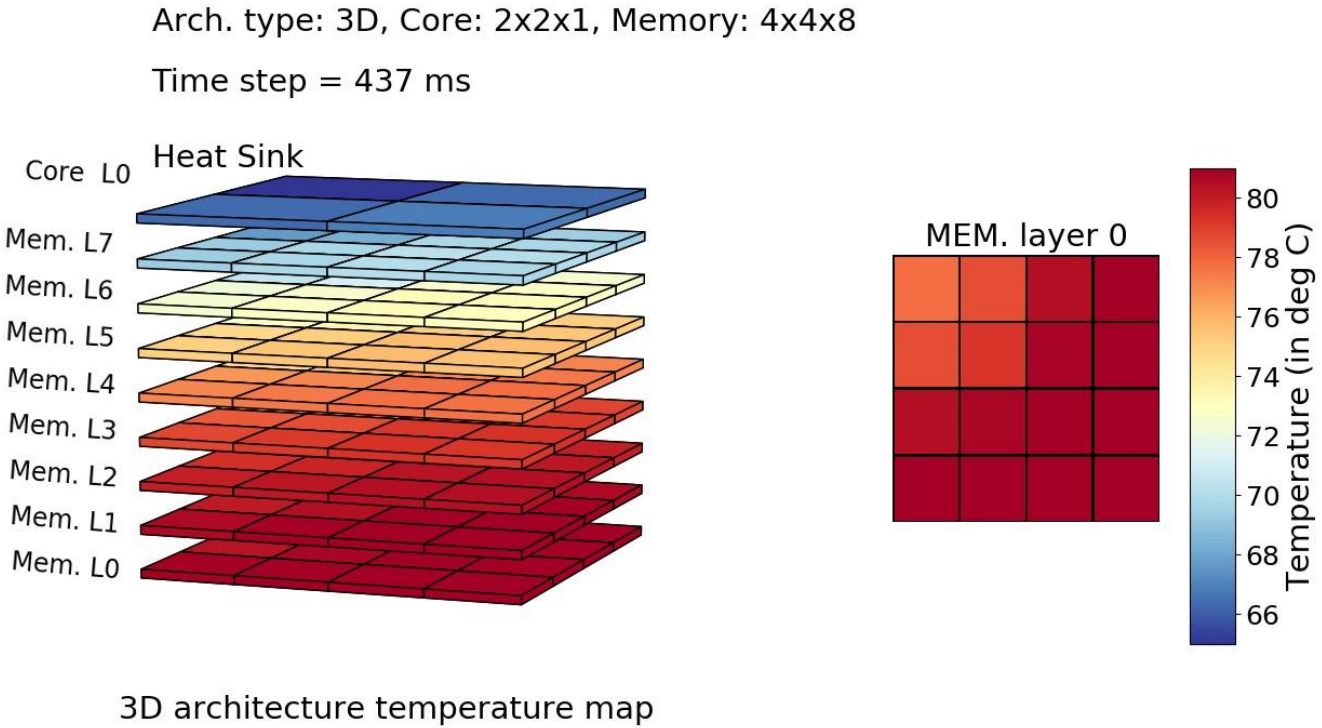
❑ 3D-stacked Processor-Memory Architecture



- Higher cost
  - Simulation needed
  - CoMeT
- Higher power density

Research background→3D-stacked systems

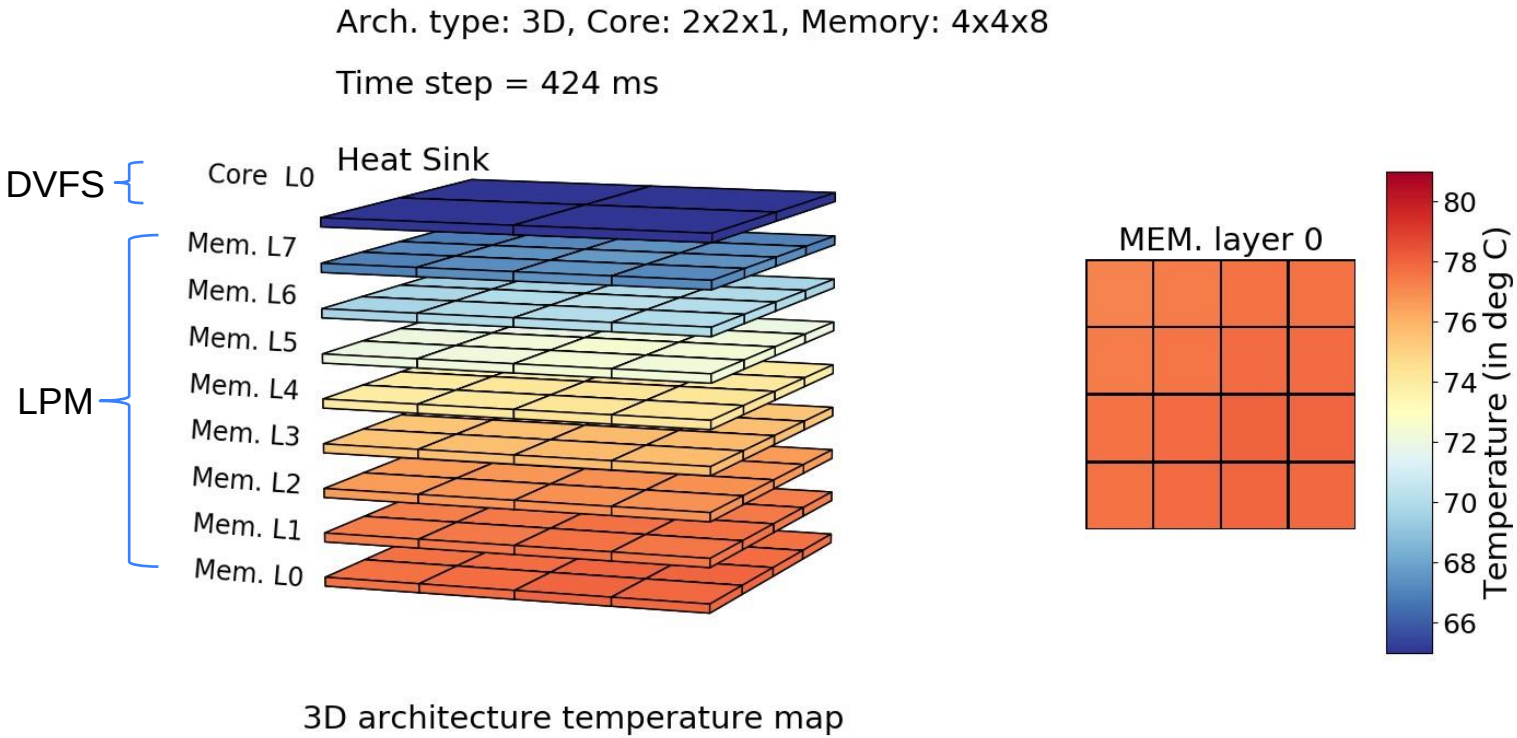
- Thermal behavior in 3D-stacked systems(running blacksholes)



3D architecture temperature map

Research background→3D-stacked systems

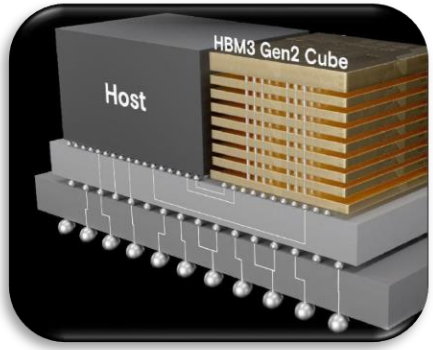
- Thermal behavior in 3D-stacked systems(running blacksholes)



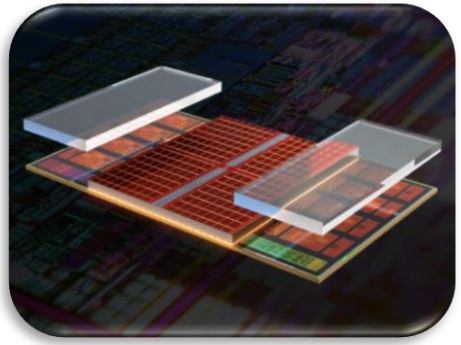
3D architecture temperature map

Research background→3D-stacked systems

- 3D-stacked chips in industry

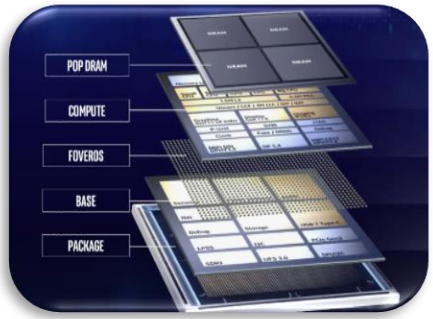


Micron HBM

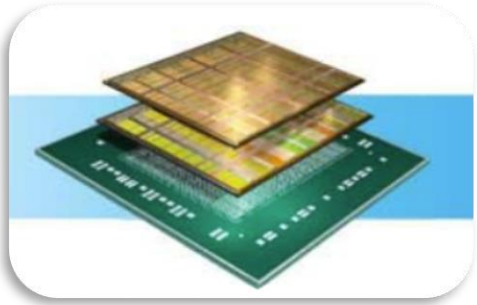


AMD Zen 3 Ryzen

Apple M2



Intel Lakefield



Xilinx Virtex Ultrascale

Samsung HBM2E Flashbolt

Outline

- Research background
- ***A motivational example***
- LPM implementation
- DRAM access profiling
- 3QUTM: a unified thermal scheduler via Deep Q-learning
- Experimental results
- Conclusion

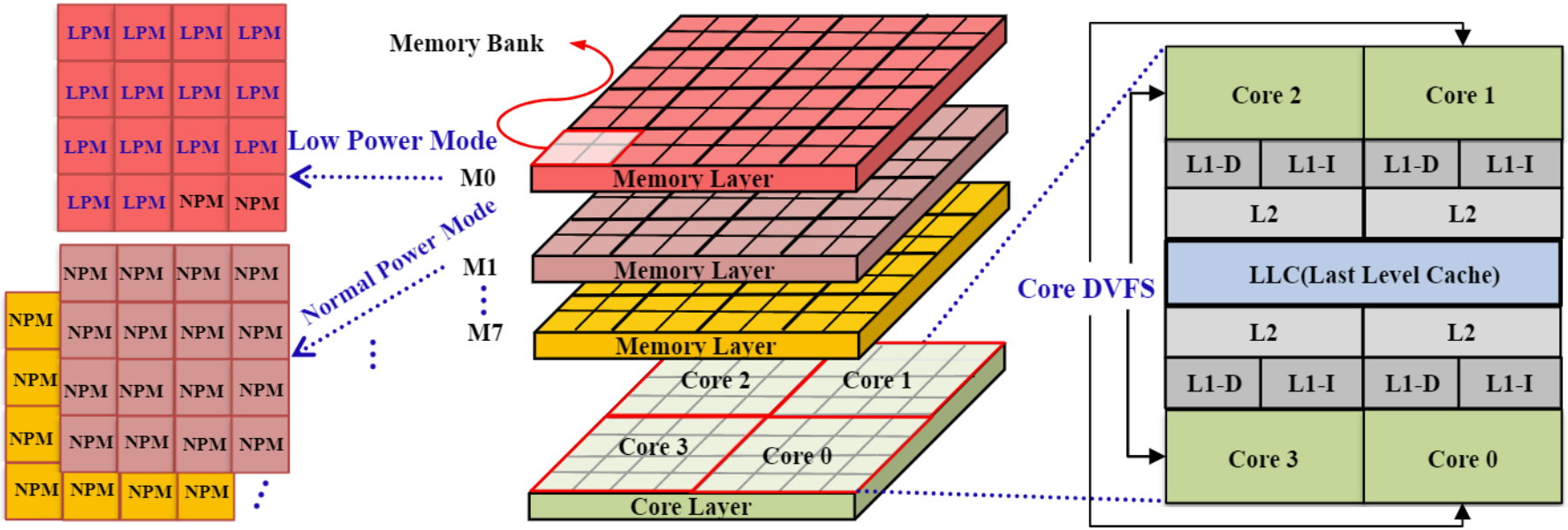
A motivational example

- Experimental platform: simulated 3D-stacked system
  - 4 core per layer-1 layer
  - 16 memory banks per layer-8 layers
- Thermal threshold 78(℃)
- Experimental configuration

	Benchmark	Thread	DTM method	Execute at
Scenario 1	streamcluster	1 master, 3 slaves	DVFS	Core 0,1,2,3
Scenario 2	streamcluster	1 master, 3 slaves	LPM	Core 0,1,2,3

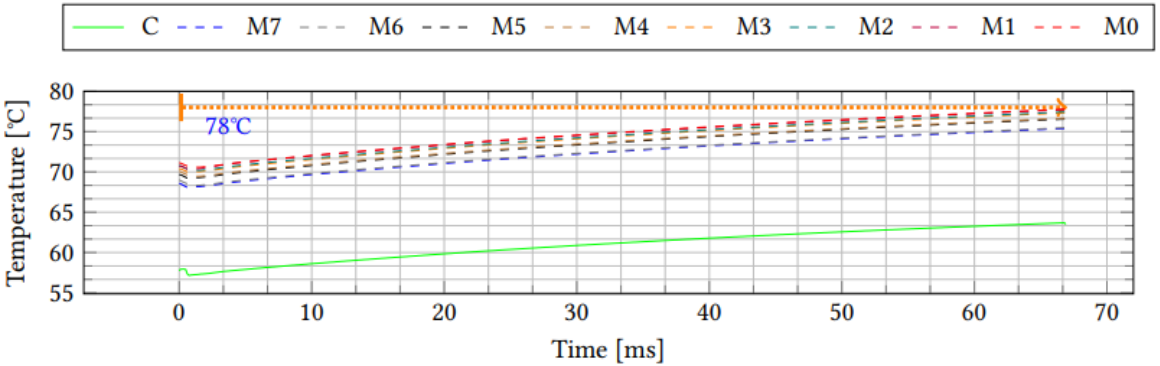
A motivational example

- 3D-stacked processor-memory systems with DVFS and LPM

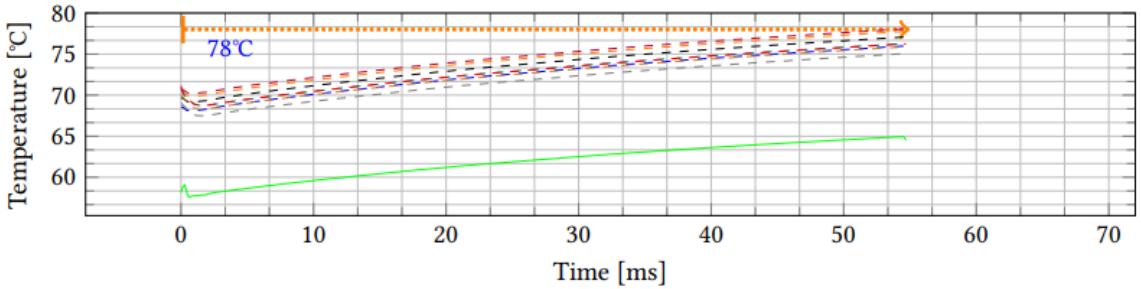


A motivational example

- Layer-wised peak temperature for different layers



(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.

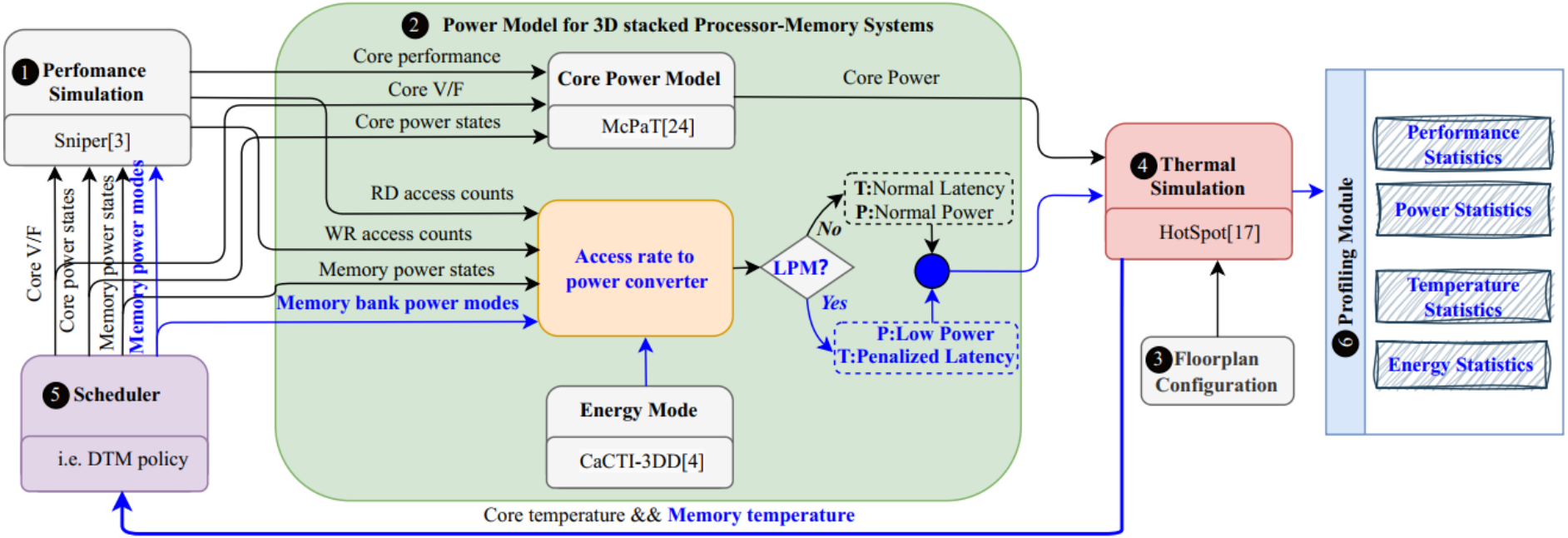
- LPM+4GHz is 18.33% faster than No LPM+2.7GHz
- Computer-intensive benchmarks benefit more from higher core frequency than from LPM penalties

Outline

- Research background
- A motivational example
- ***LPM implementation***
- DRAM access profiling
- 3QUTM: a unified thermal scheduler via Deep Q-learning
- Experimental results
- Conclusion

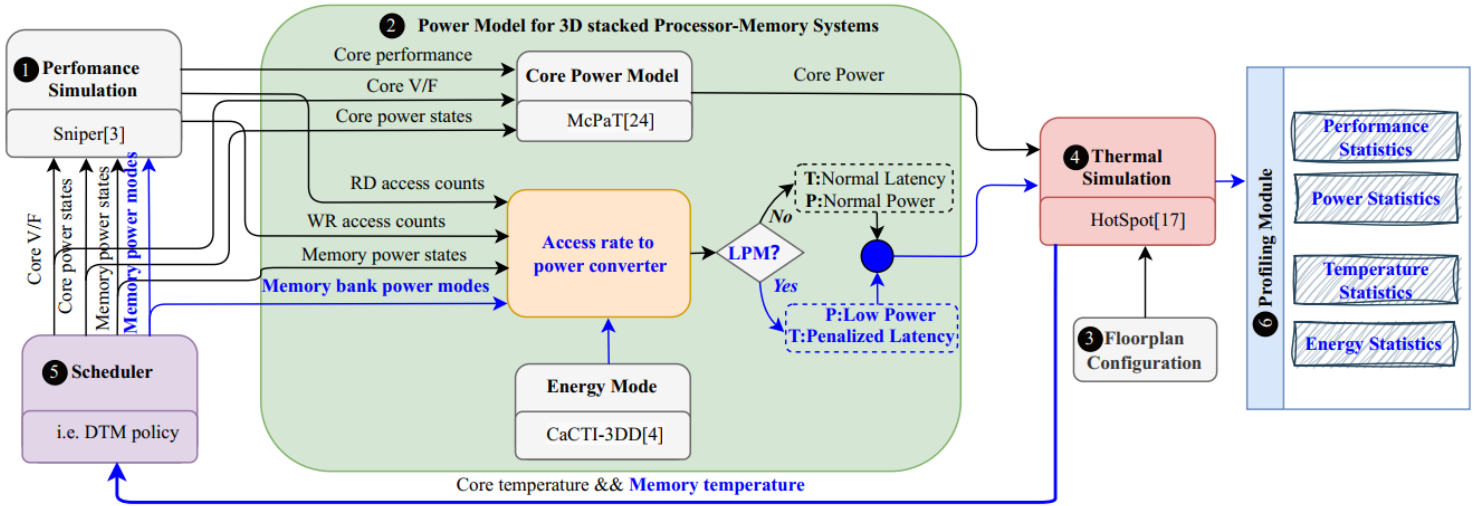
Low power mode(LPM) implementation

- Power-performance modeling associated with LPM



Low power mode(LPM) implementation

- Key features of LPM
  - Data preserved in situ in LPM, avoiding migration
  - Data inaccessible in LPM mode
  - Re-accessing data requires toggling to NPM

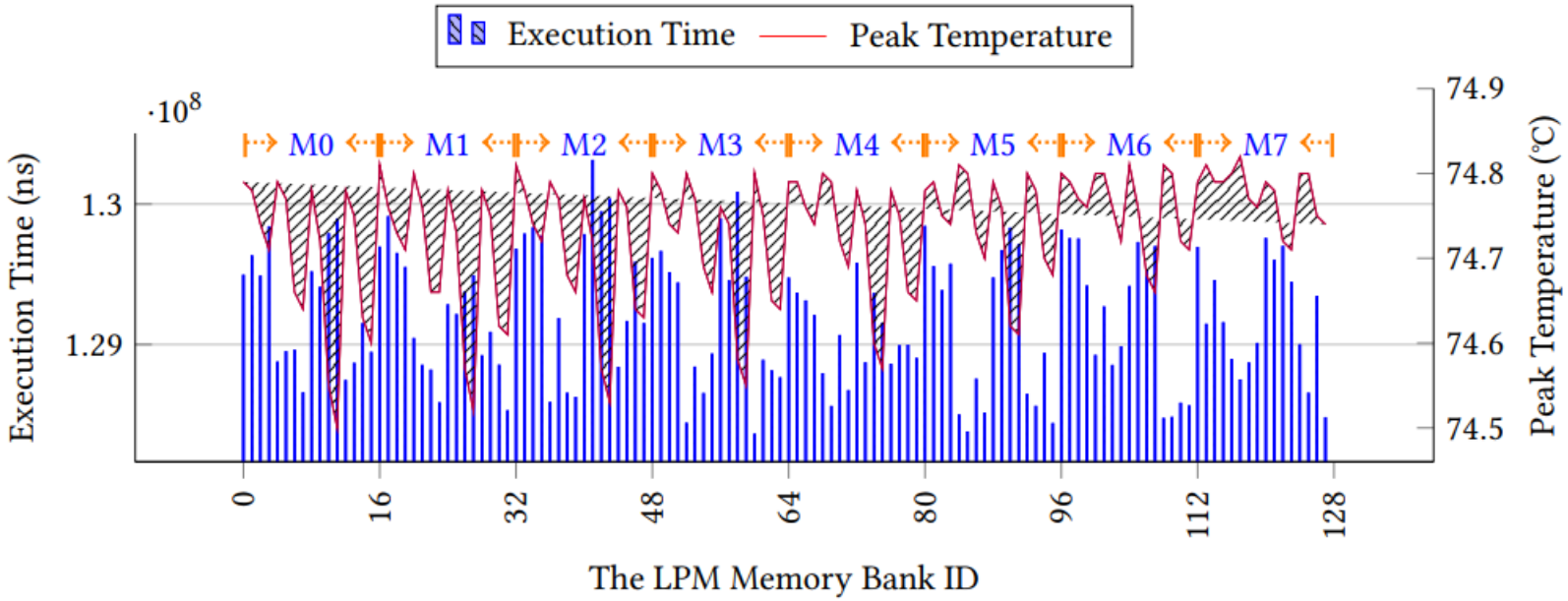


Outline

- Research background
- A motivational example
- LPM implementation
- ***DRAM access profiling***
- 3QUTM: a unified thermal scheduler via Deep Q-learning
- Experimental results
- Conclusion

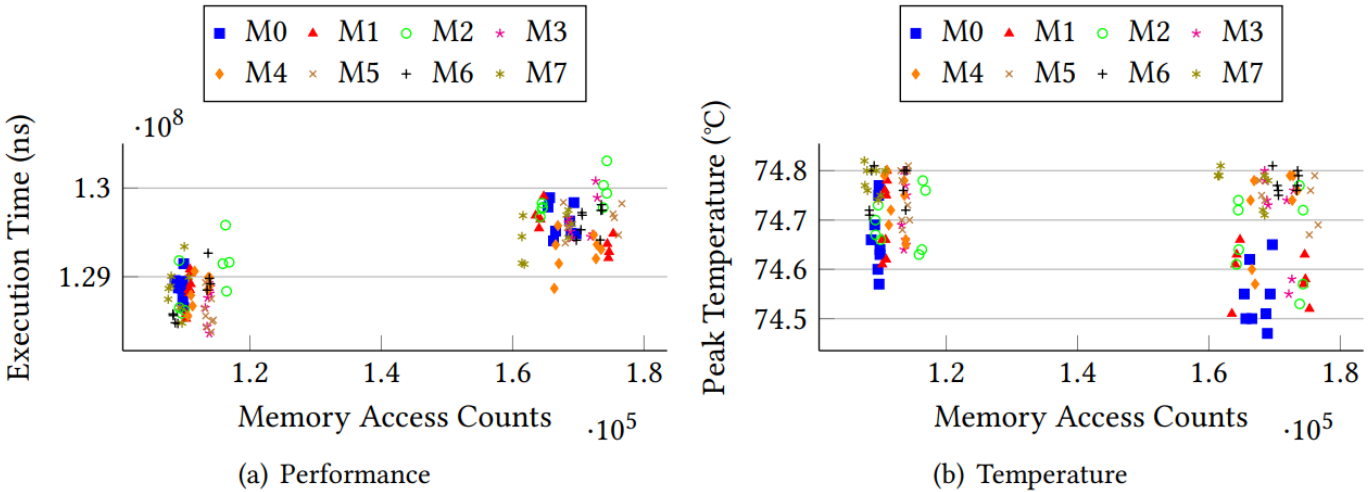
DRAM Access Analysis for Low Power Mode

- Performance and peak Temperature Variations in 3D-stacked Systems with Varied LPM
  - Turn off 1 memory bank per test



DRAM Access Analysis for Low Power Mode

- Performance and peak Temperature Variations in 3D-stacked Systems with Varied LPM
  - 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

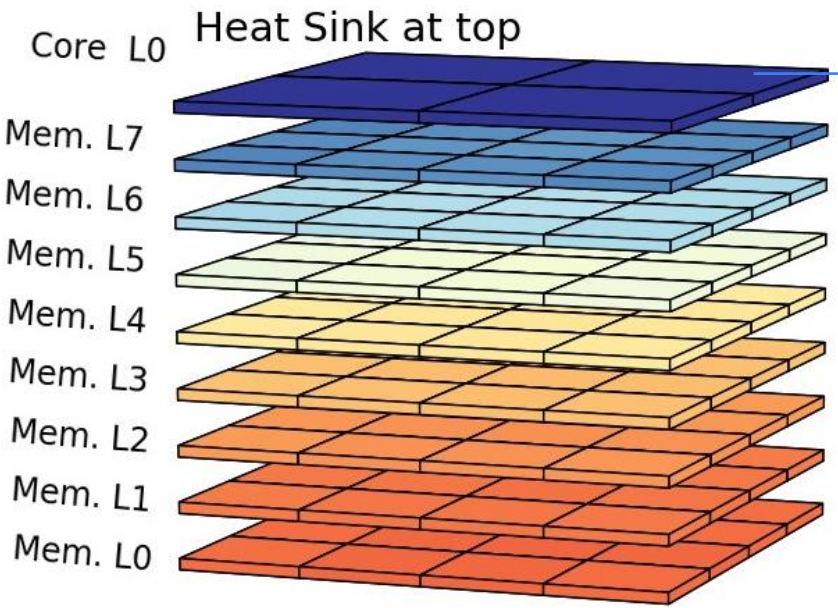


Outline

- Research background
- A motivational example
- LPM implementation
- DRAM access profiling
- ***3QUTM: a unified thermal scheduler via Deep Q-learning***
- Experimental results
- Conclusion

3QUTM: a unified thermal scheduler via Deep Q-learning

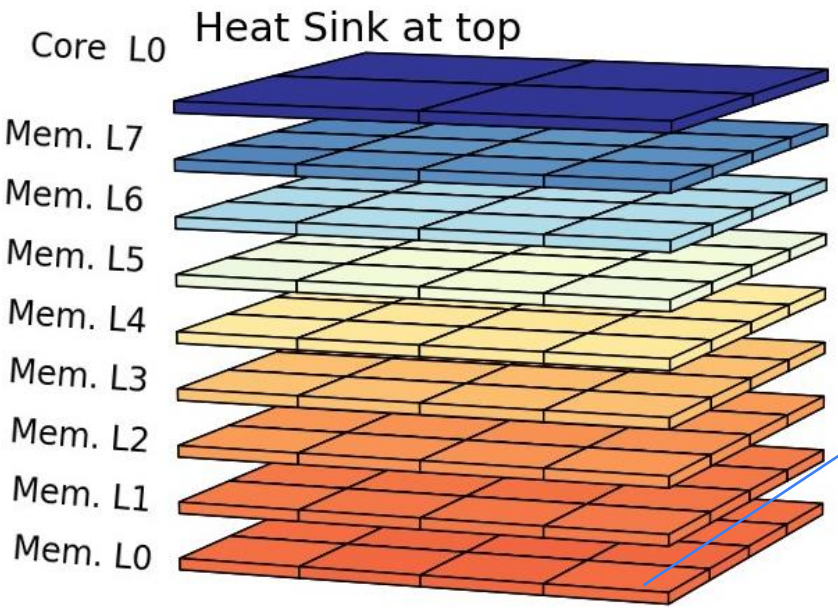
- Action space
  - DVFS for core(16^4)
  - LPM for memory bank



V(v)	F(GHz)	V(v)	F(GHz)
0.648	1.0	0.907	2.6
0.680	1.2	0.940	2.8
0.713	1.4	0.972	3.0
0.745	1.6	1.004	3.2
0.778	1.8	1.037	3.4
0.810	2.0	1.069	3.6
0.842	2.2	1.102	3.8
0.875	2.4	1.134	4.0

3QUTM: a unified thermal scheduler via Deep Q-learning

- Action space($16^4 * 2^{128} = 2^{144}$)
  - DVFS for core(16^4)
  - LPM for memory bank(2^{128})



LPM(Low power mode)

NPM(Normal power mode)

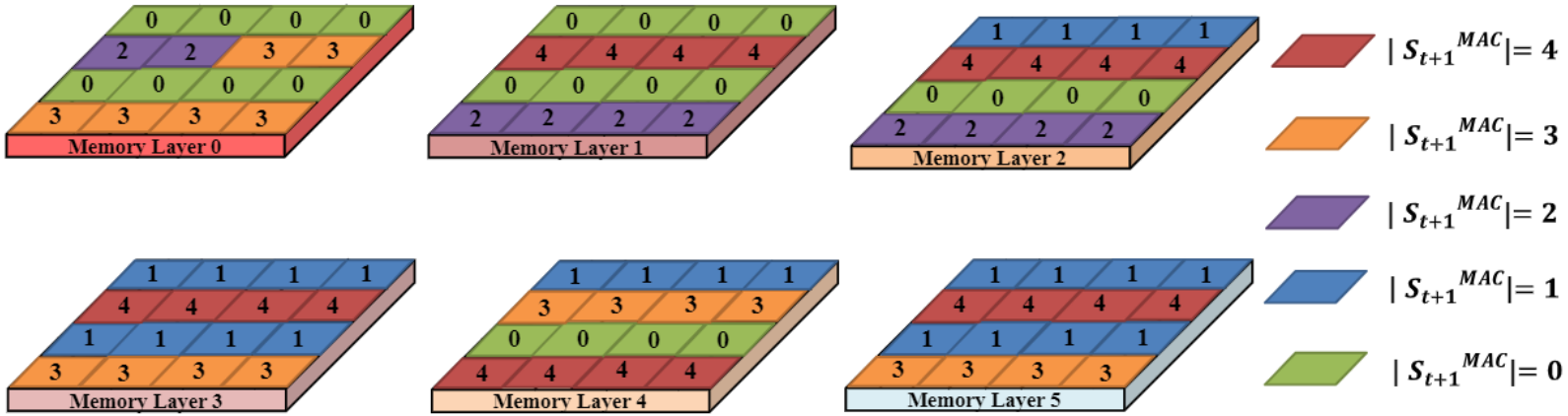
3QUTM: a unified thermal scheduler via Deep Q-learning

- State space
  - Discrete state: s^{IPC}, s^{f}, s^{U}, s^{MAC}
  - Continuous state: s^{hotL}, $s^{T_{peak}}$, $s^{P_{tot}}$

State	Description	Discrete values
s^{IPC}	Billion Instructions per Cycle per Core	0,1
s^{f}	Frequency per Core	1.0,1.2,$\cdots$,4.0
s^{U}	Utilization per Core	0,1,2,3,4,5,6
$s^{T_{peak}}$	Peak temperature of 3D-stacked Systems	-
$s^{T_{hotL}}$	Average Temperature of the Hottest Memory Layer	-
$s^{P_{tot}}$	Total Power Consumption	-
s^{MAC}	Memory Access Count per Memory Bank	0,1,2,3,4

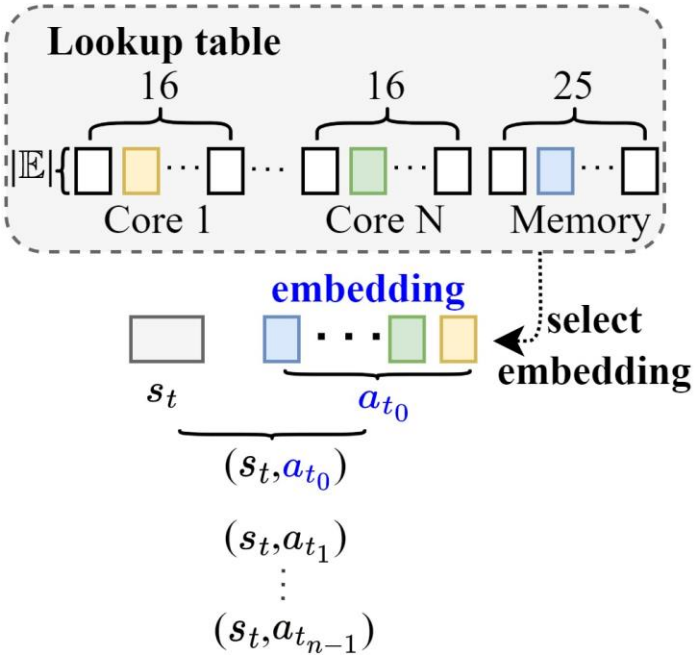
3QUTM: a unified thermal scheduler via Deep Q-learning

- DRAM access profiling(s^{MAC})
  - 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 thermal scheduler via Deep Q-learning

- Action Embedding
  - Parameterize each subaction(each core and each DRAM action)
  - Embedding the parameterized subaction



3QUTM: a unified thermal scheduler via Deep Q-learning

- Reward function
  - r_0 represents system performance
  - r_1 represents temperature
  - r_2 represents the impact of LPM memory banks

$$r_{t+1}(s_t, s_{t+1}) = \zeta r0_{t+1} + \phi r1_{t+1} + \varphi r2_{t+1} + \frac{\Phi}{|s_{t+1}^{P_{tot}}|}$$

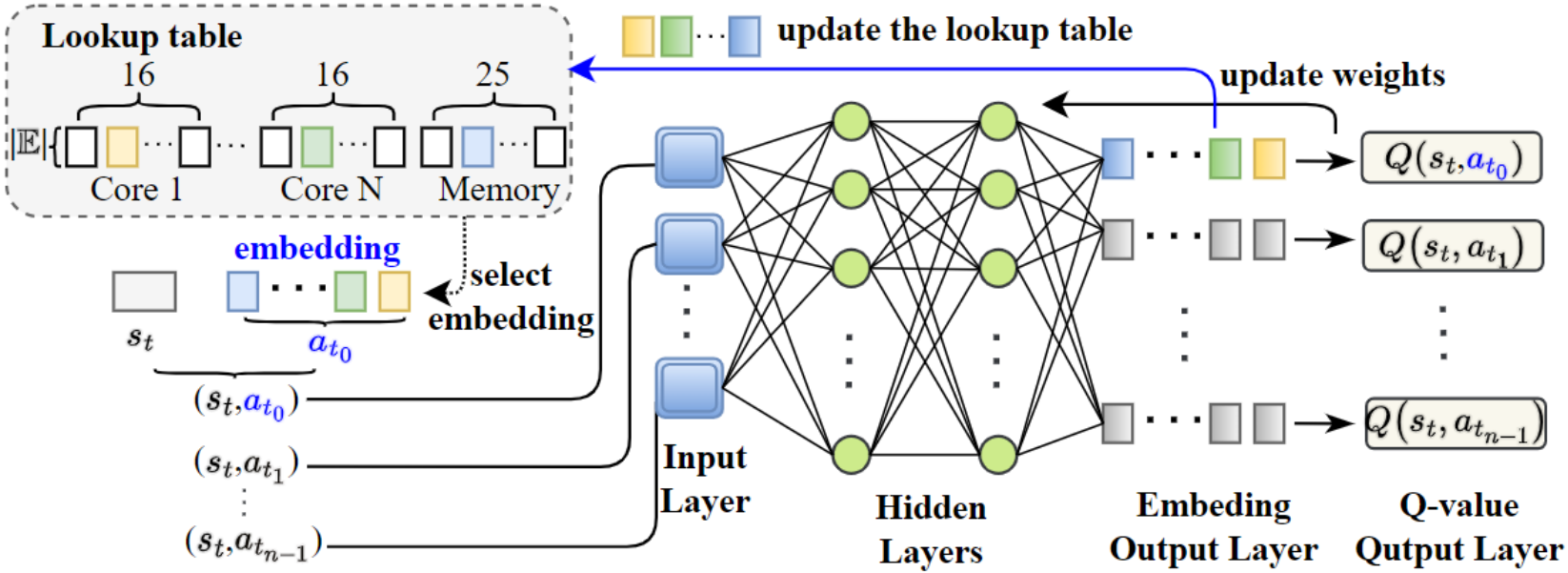
Average system throughput (core frequency, IPC, utilization)

The gap between T_{peak}, and T_{op}, T_{cr} (an exponential law)

Switching frequently accessed memory banks to LPM increases penalties.

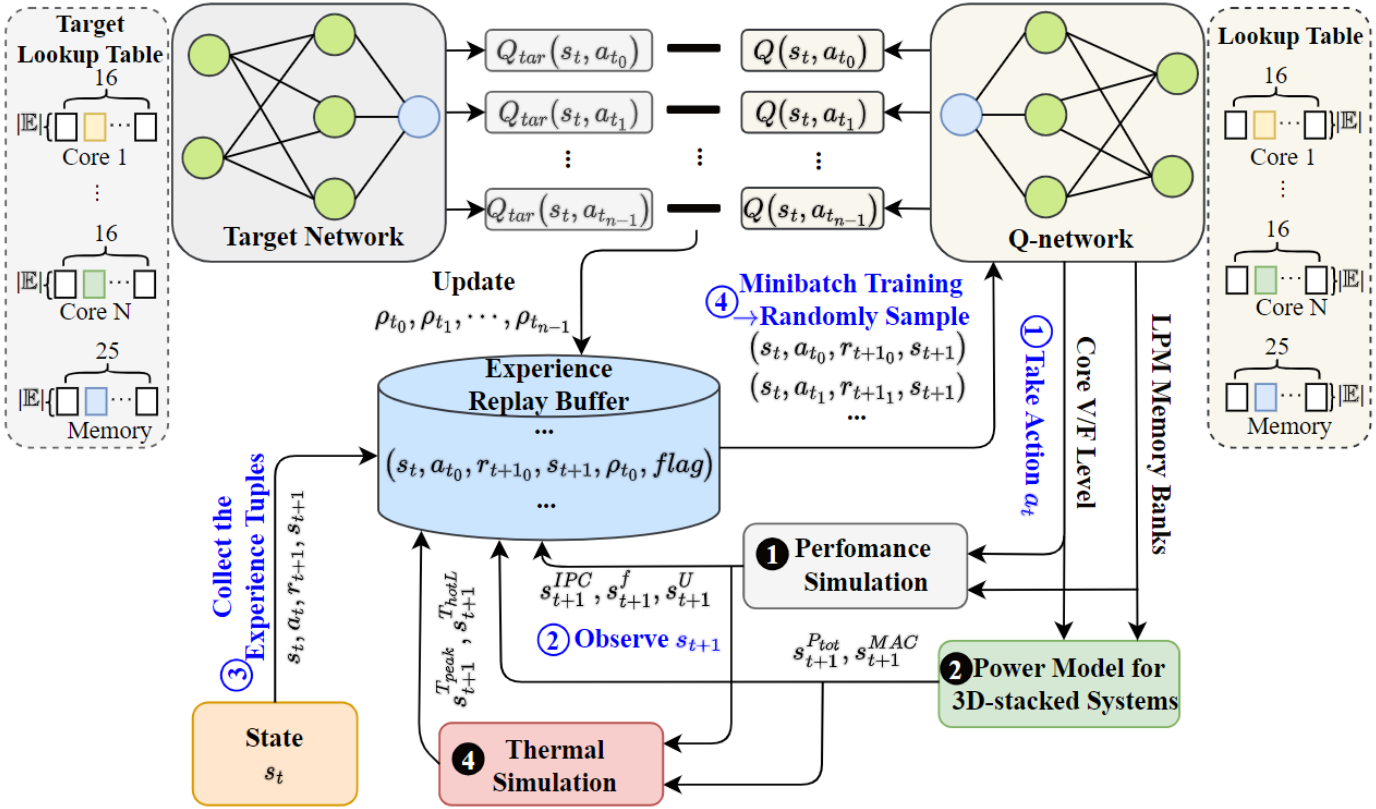
3QUTM: a unified thermal scheduler via Deep Q-learning

- The design of the Q-network
  - Parameterize each subaction(each core and each DRAM action)
  - Embedding the parameterized subaction



3QUTM: a unified thermal scheduler via Deep Q-learning

- The training of 3QUTM



3QUTM: a unified thermal scheduler via Deep Q-learning

- The inference of 3QUTM in simulated 3D-stacked systems
  - **C++ Conversion**:
    - Well-trained Q-network transition from Python to C++
    - Preparation the package for CoMeT simulator integration
  - **Binary Code Compilation**:
    - Q-network compiled into binary post C++ conversion
    - Enhanced execution speed and memory efficiency
    - Suitable for real-world hardware deployment
  - **Execution on CoMeT**:
    - Binary file loaded onto 3D-stacked processor-memory system

Outline

- Research background
- A motivational example
- LPM implementation
- DRAM access profiling
- 3QUTM: a unified thermal scheduler via Deep Q-learning
- ***Experimental results***
- Conclusion

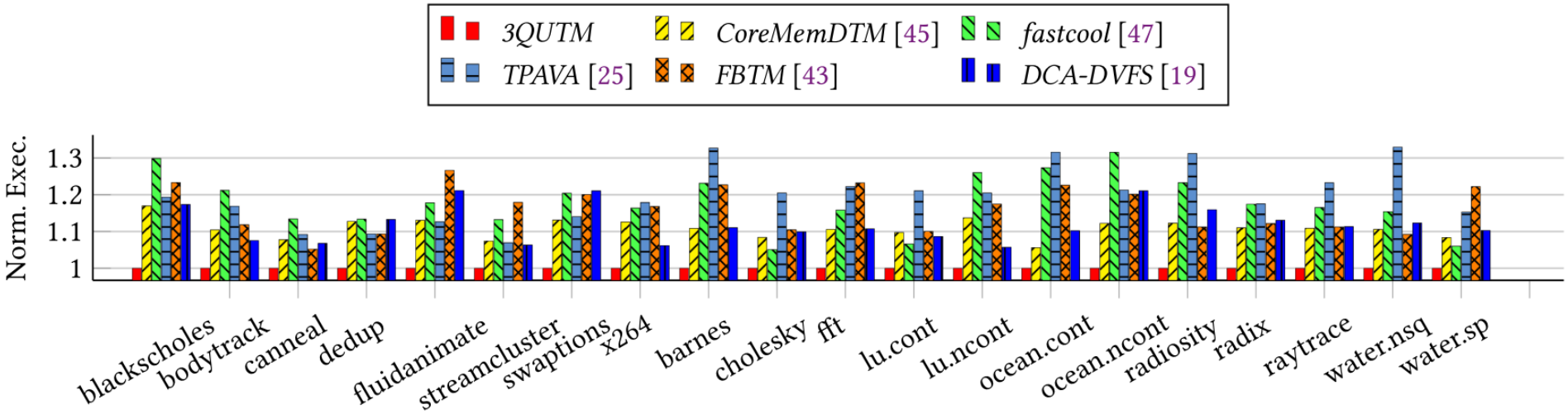
Experimental results

- Configuration
  - Training using PARSEC benchmarks
  - In the inference phrase, using PARSEC + Splash-2

Core Parameters	
Number of Cores	4/16, 1 layer
Core Model	x86, 4.0 GHz, 22 nm, out-of-order
L1 I/D cache	32/32 KB, 4/4-way,64 B-block
L2 cache	private,512 KB, 8-way, 64 B-block
L3 cache	512 KB, 16-way, 64 B-block
Memory Parameters	
3D-stacked Memory	8 GB, 8 layers, 16 channels, 128 banks
Memory Bandwidth	25.6 GB/s

Experimental results

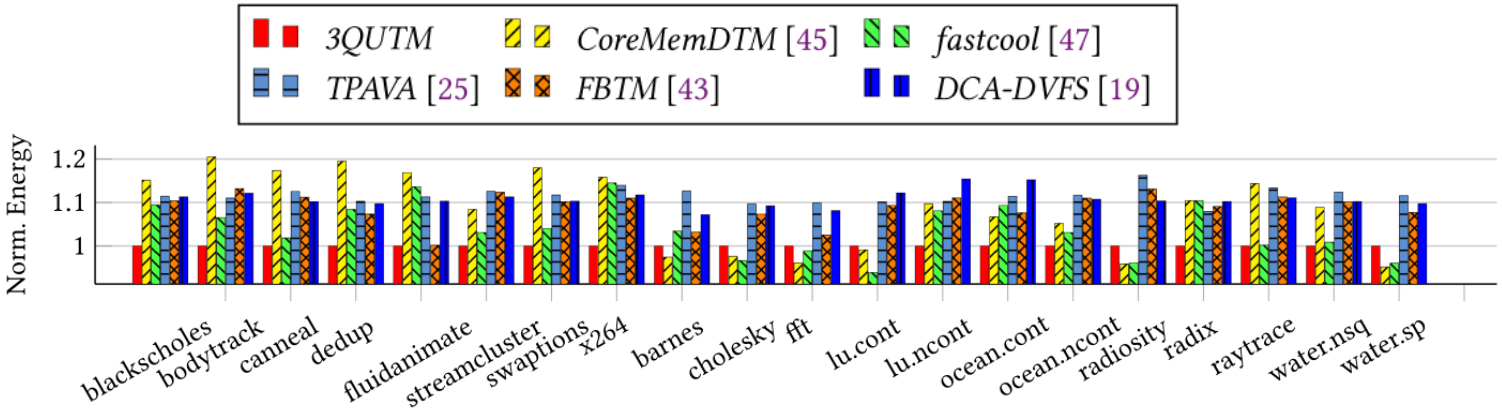
- Performance improvements on a 4-core 3D-stacked systems
  - Training using PARSEC benchmarks
  - In the inference phrase(thermal management), using PARSEC+Splash-2



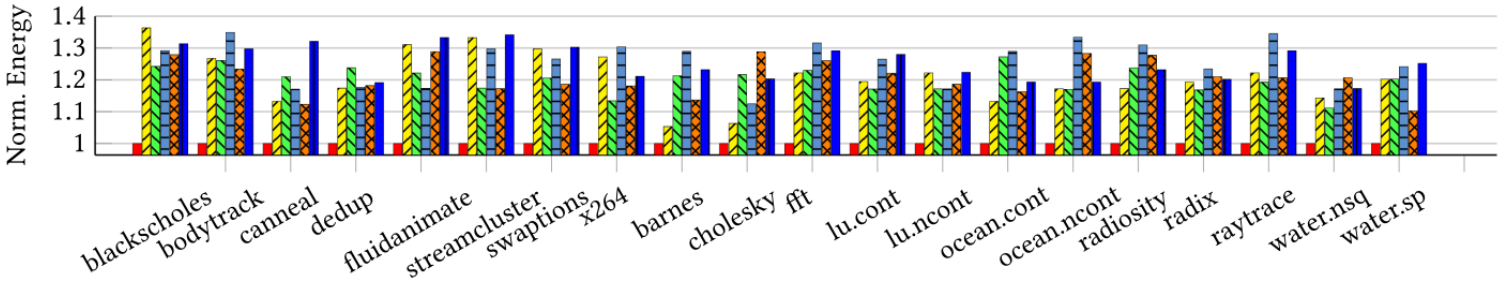
(a) Performance

Experimental results

- Energy improvements on a 4-core 3D-stacked systems



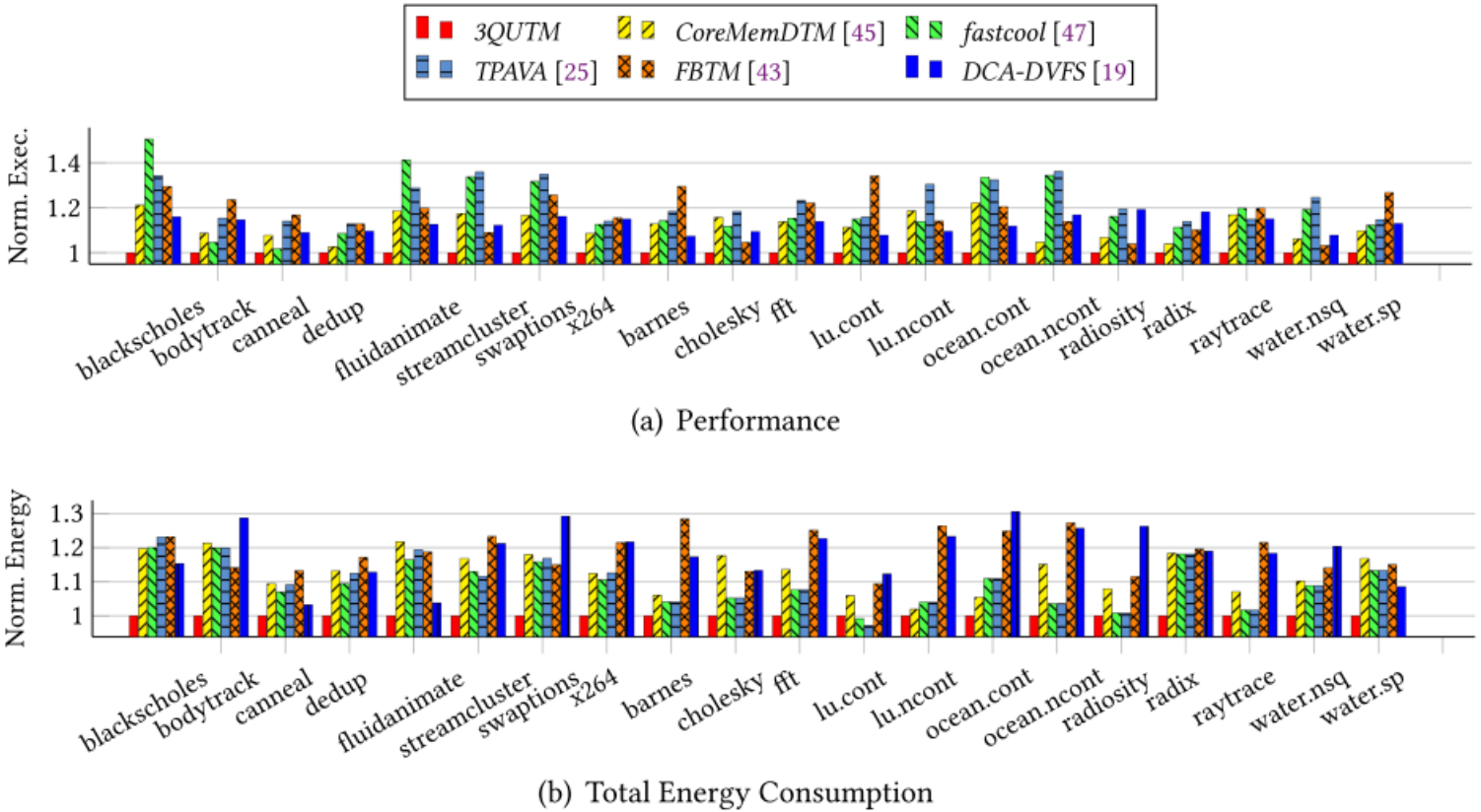
(b) Core Energy Consumption



(c) Memory Energy Consumption

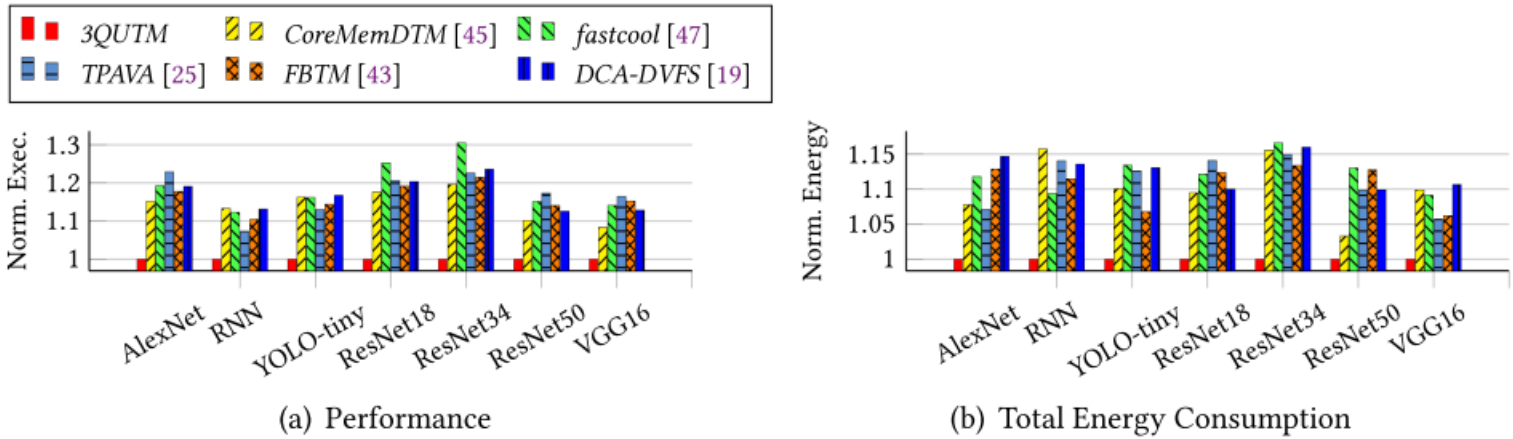
Experimental results

- Performance and energy improvements on a 16-core 3D-stacked systems



Experimental results

- Performance and Energy Consumption for DNN Workloads in Inference

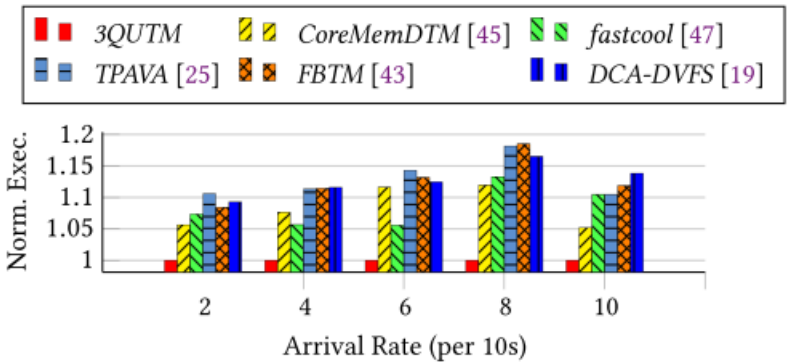


(a) Performance

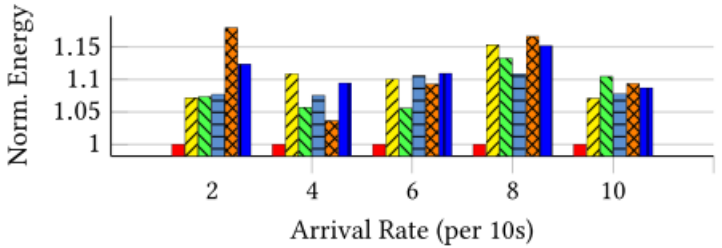
(b) Total Energy Consumption

Experimental results

- Performance and Energy Consumption for DNN Workloads in Inference
  - Open systems: The arrival time follows Possion Distribution



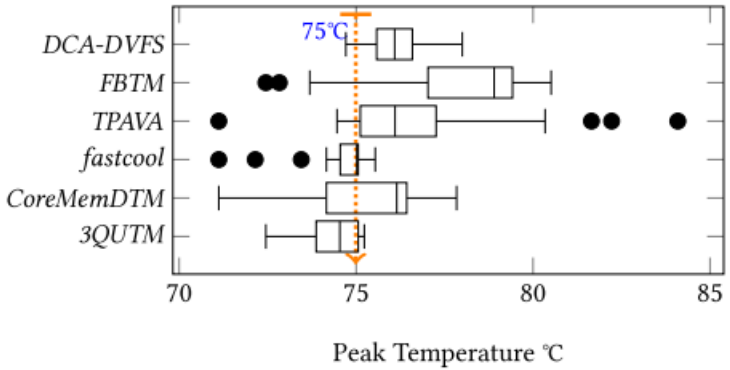
(a) Performance



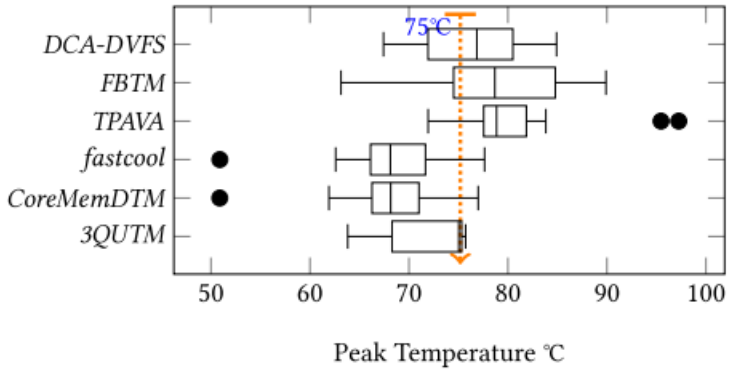
(b) Total Energy Consumption

Experimental results

- Peak Temperature Analysis



(a) Temperature Comparison on a 4-core 3D-stacked System



(b) Temperature Comparison on a 16-core 3D-stacked System

Outline

- Research background
- A motivational example
- LPM implementation
- DRAM access profiling
- 3QUTM: a unified thermal scheduler via Deep Q-learning
- Experimental results
- ***Conclusion***

Conclusion

- We developed a Low Power Mode (LPM) optimized for 3D-stacked systems.
- We introduce 3QUTM, a DQN-based unified thermal scheduler for intelligent thermal management.
- We conducted comparative experiments, yielding notable performance and energy savings across various benchmarks and DNN workloads.
- We confirmed 3QUTM's efficacy in temperature regulation within 3D-stacked systems, suggesting promising avenues for future research..

Thanks for your attention

Questions?