

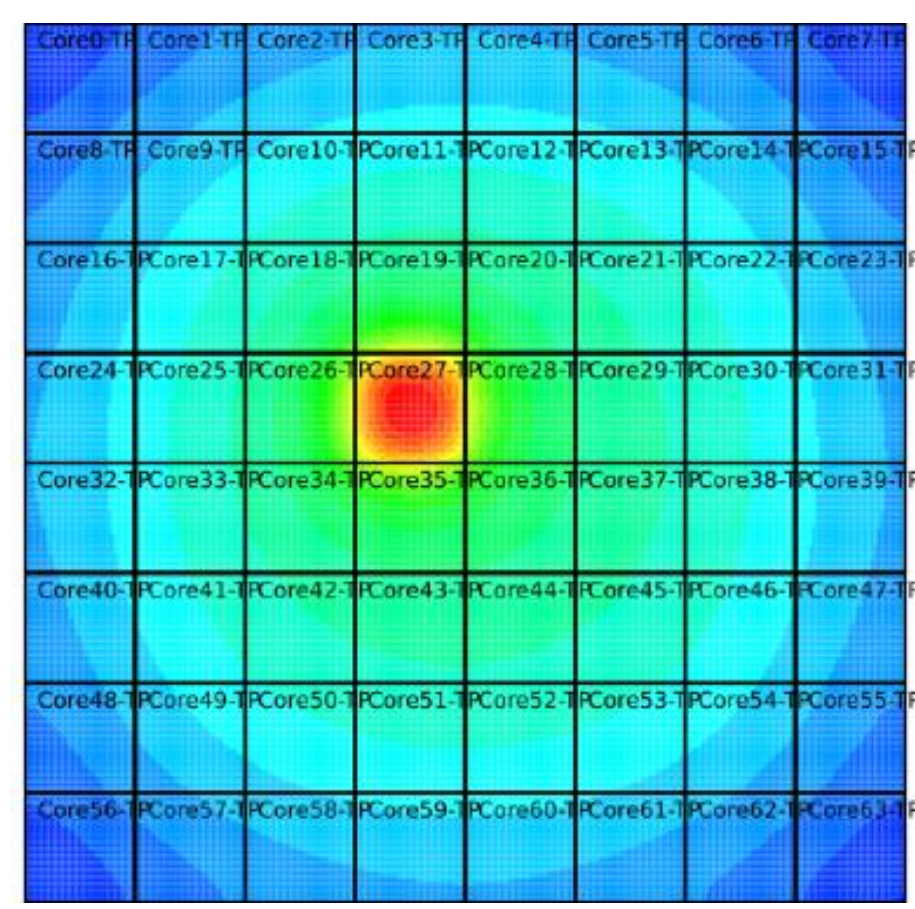
Thermal Management for S-NUCA Many-Cores via Synchronous Thread Rotations

Research Background (S-NUCA Architecture + Task Rotations)

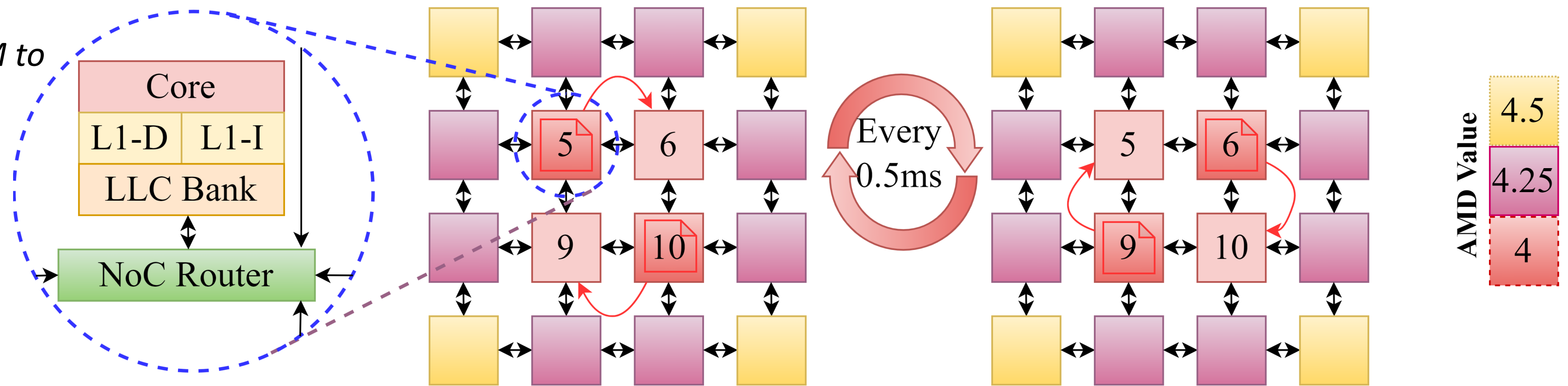
Higher power density + Dark silicon

Logically shared but physically distributed cache + Inherent Heterogeneity (Cache latency is non-uniform)

Thread rotations can balance the hot and cold core temperatures

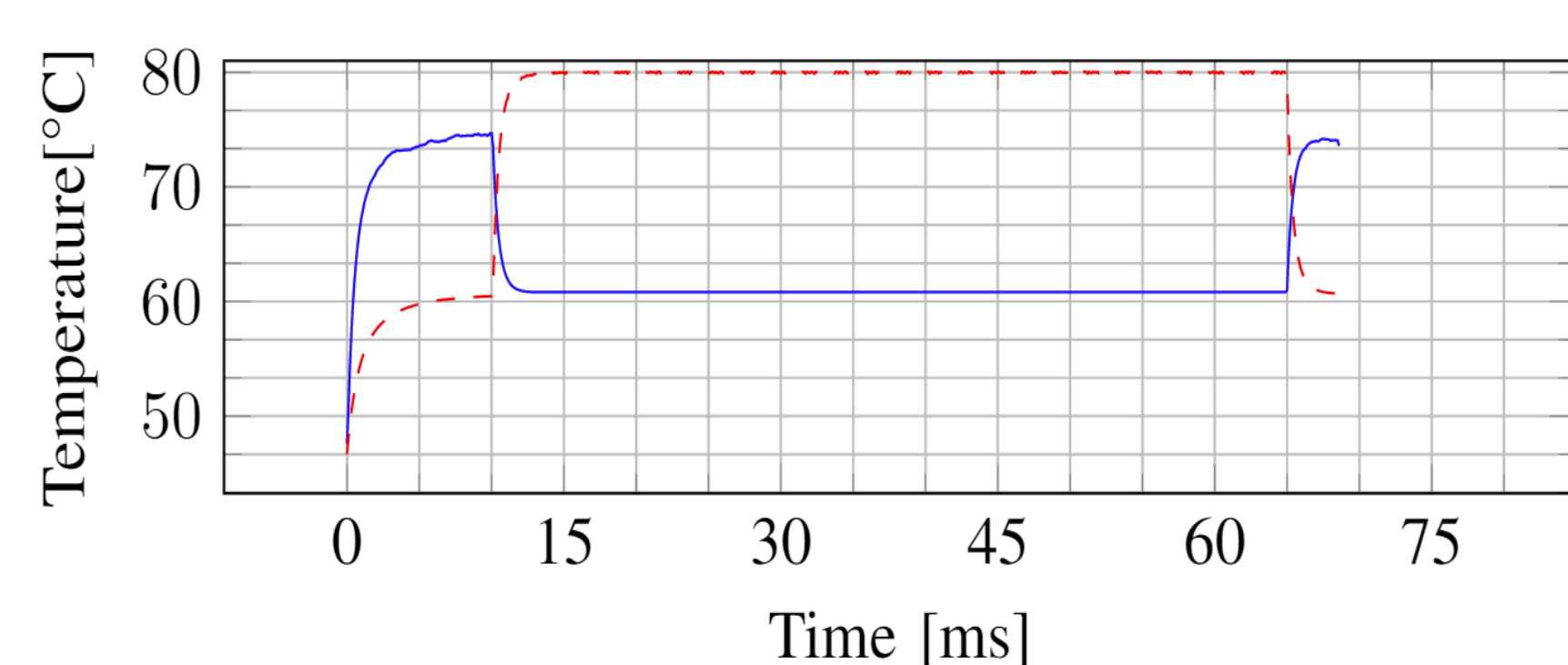


Requires intricate DTM to unlock performance potential

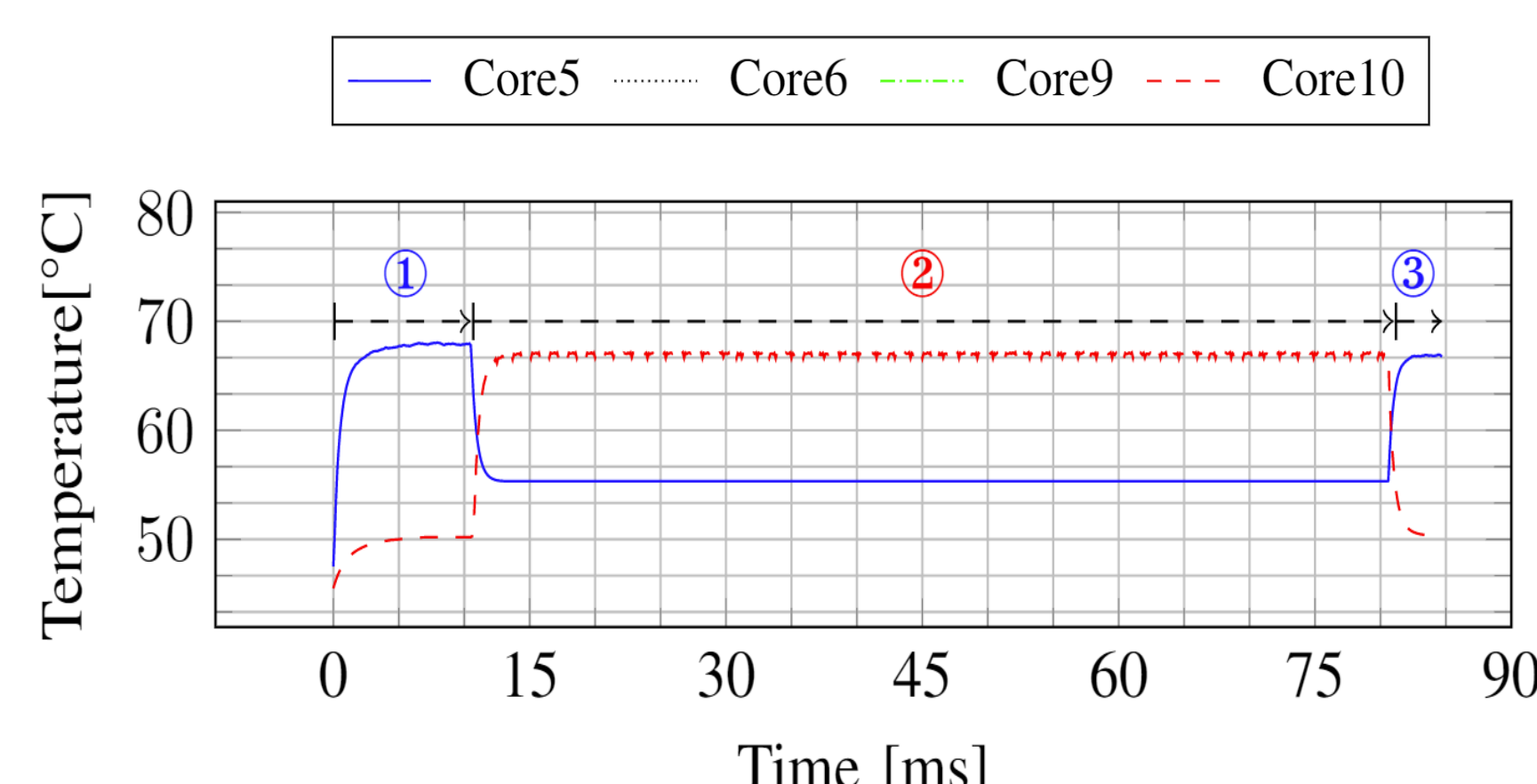


Motivational Example

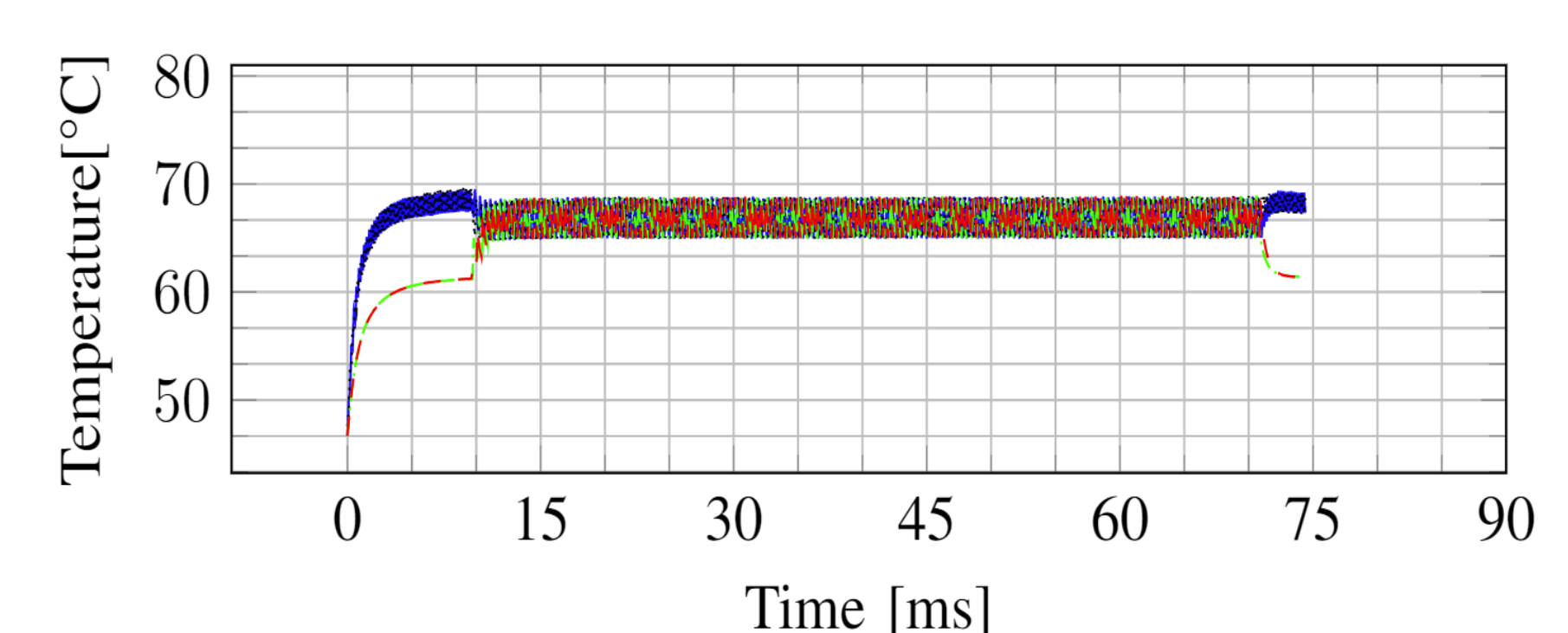
Thread rotations penalty vs DVFS-based penalty



(a) Fixed Peak Frequency



(b) Thermally Safe Frequency provided by TSP



(c) Synchronous Thread Rotation at Peak Frequency

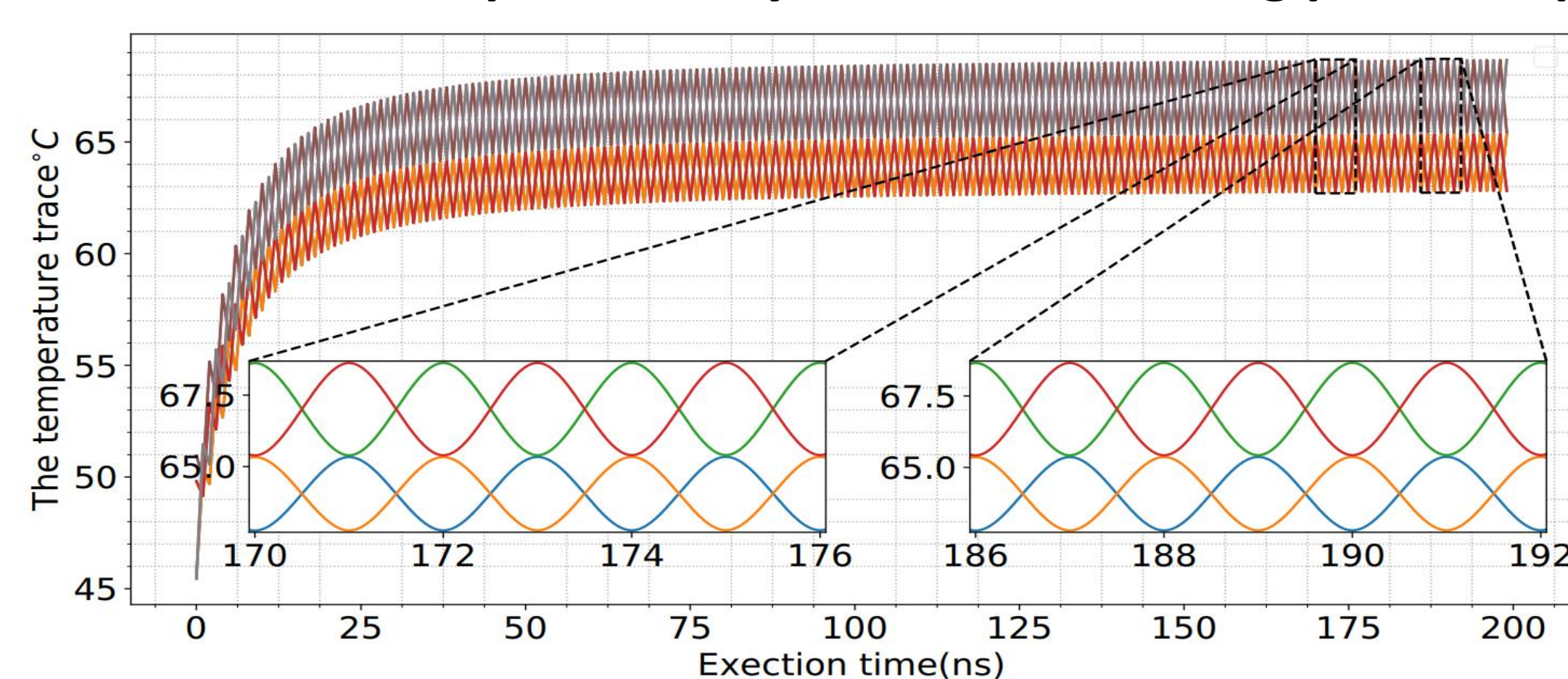
	Benchmark	Threads	DTM method	Execute at	Peak temp(°C)	Exec time(ms)	Penalty(%)
Case (a)	blackscholes	1 master, 1 slave	No	Core 5,10	80.03	67.97	-
Case (b)	blackscholes	1 master, 1 slave	TSP	Core 5,10	67.94	84.49	19.55
Case (c)	blackscholes	1 master, 1 slave	Thread rotations	Core 5,6,9,10	69.32	74.47	8.72

Case (c) is 10.83% faster than Case (b) \Rightarrow Thread rotations penalty < DVFS-based penalty

One-shot Peak Temperature Calculation

Thread rotations periodically exhibit a recurring peak temperature pattern

Theoretically calculate the peak temperature



recurring pattern

heat transfer properties

$$T_{peak} = f(v, w, \delta, \tau, P)$$

Auxiliary matrix using floorplan-based constants

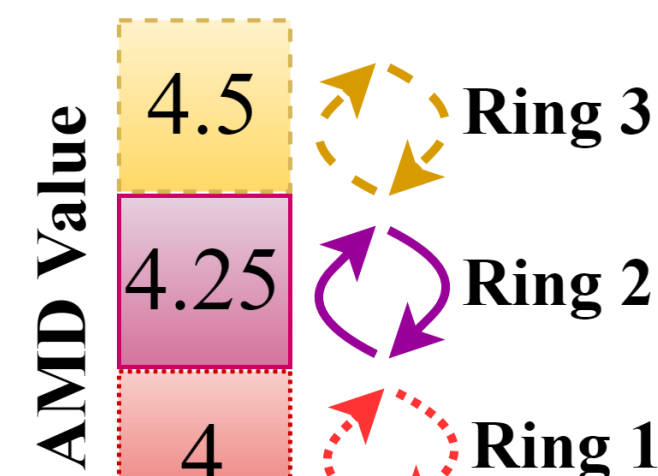
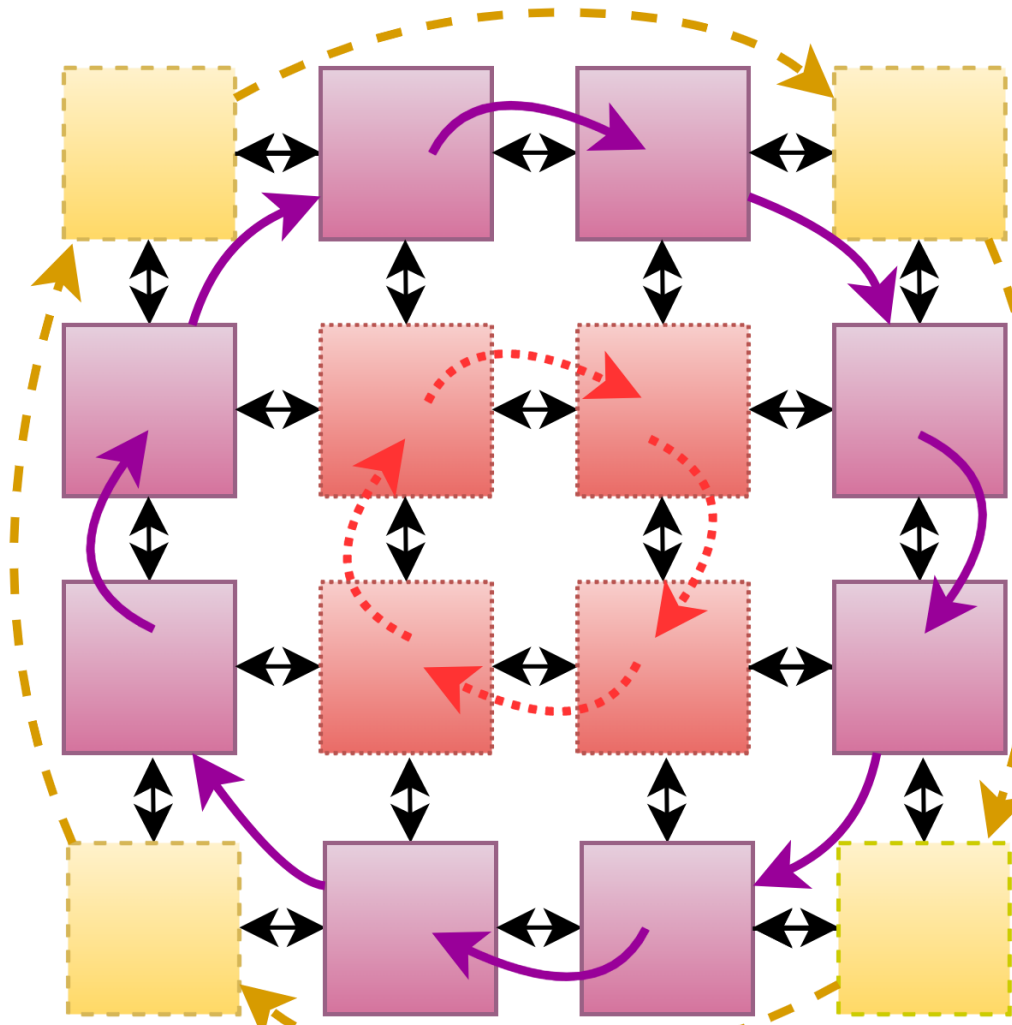
Rotation speed

Rotation period

Power traces

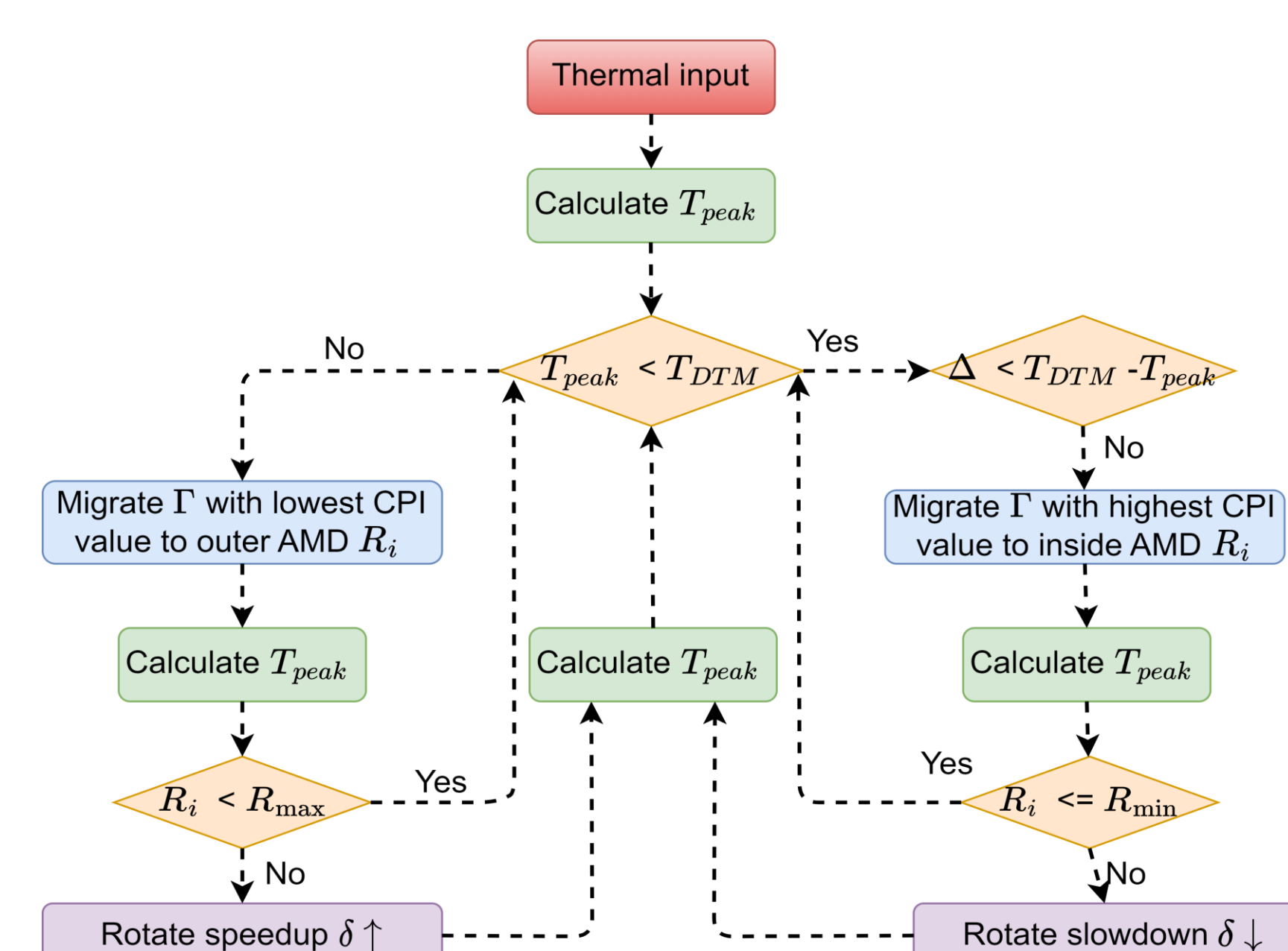
Hot-potato Scheduling

Thermal and architecture-aware synchronous thread rotations



Thermal dissipation condition
Ring1 < Ring2 < Ring3

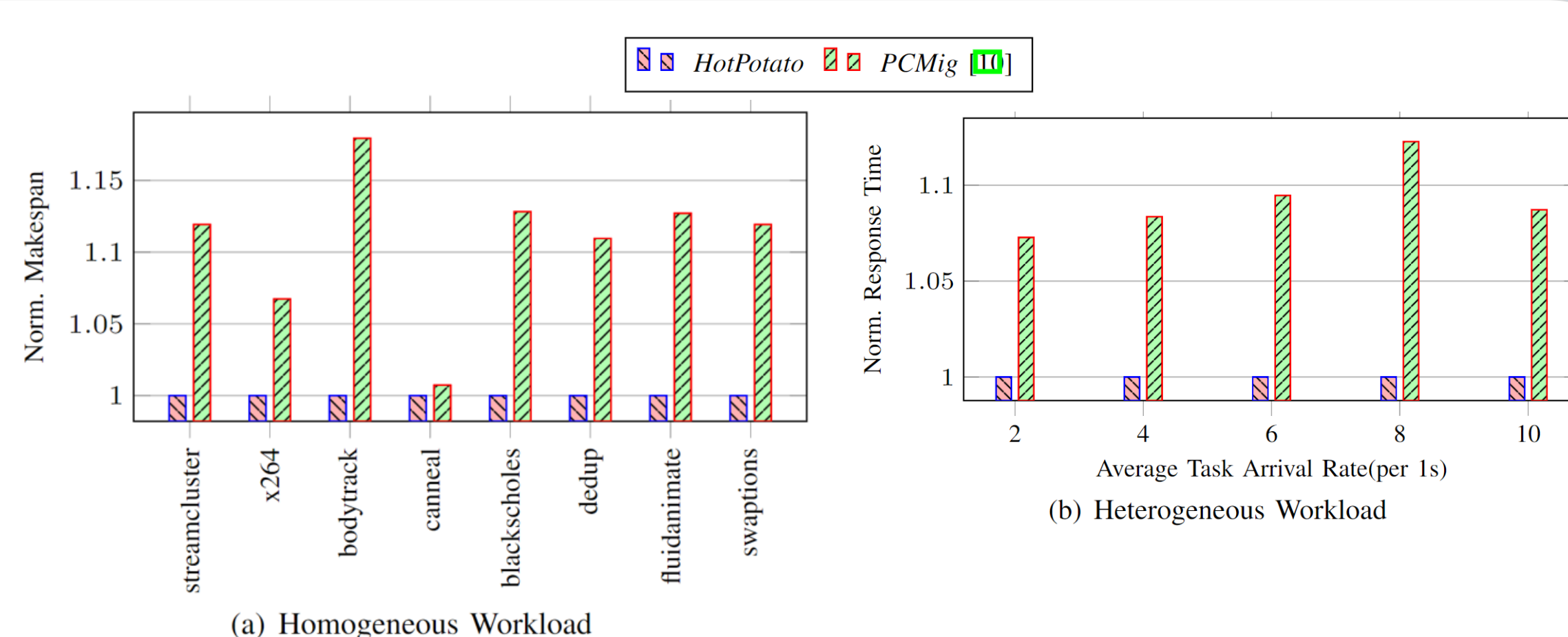
Access cache latency
Ring1 < Ring2 < Ring3



Evaluation

Simulated open systems using HotSniper

The arrival time of threads follows Poisson distribution



Authors



Yixian Shen



Sobhan Niknam



Anuj Pathania



Andy Pimentel

Email: y.shen@uva.nl