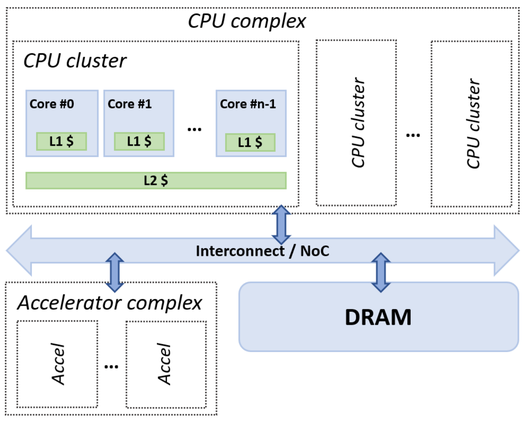
The Importance of Worst-Case Memory Contention Analysis for Heterogeneous SoCs

Lorenzo Carletti, Gianluca Brilli, Alessandro Capotondi, Paolo Valente, and Andrea Marongiu

UNIMORE, University of Modena and Reggio Emilia, 41125 Modena, Italy

Motivation

- **Memory contention problem.** Shared memory is an hardware bottleneck in modern multicore architectures. Bursts of memory requests submitted by co-scheduled applications can exceed the available memory bandwidth, causing disproportional slowdowns.

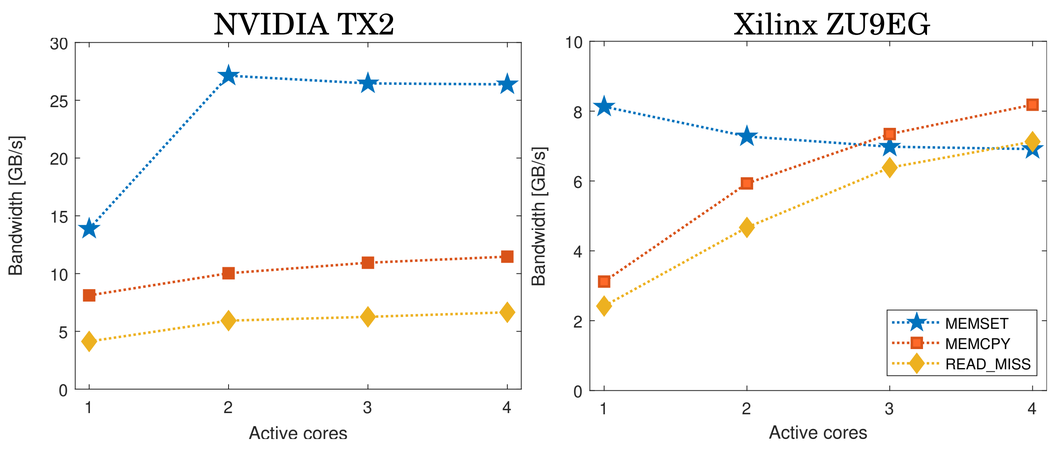


- **Worst case interference.** When trying to mitigate the amount of interference that a system is subject to, having a proper understanding of the worst case is needed to prove the validity of a solution. Certain published works [1]-[3] make the assumption that Read-only synthetic benchmarks should be both:

1. The tasks most subject to DRAM interference caused from other programs.
2. The tasks causing the highest amount of DRAM interference.

Bandwidth Analysis

- Reference HeSoCs: the GPU-based **NVIDIA TX2** and the FPGA-based **Xilinx ZU9EG**.
- **Memory bandwidth saturation** is a key component of DRAM interference. For that reason, we conducted an analysis on the bandwidth that various types of traffic generate on the reference HeSoCs. 100 MB large synthetic benchmarks: **READ_MISS**, **MEMCPY** and **MEMSET**.



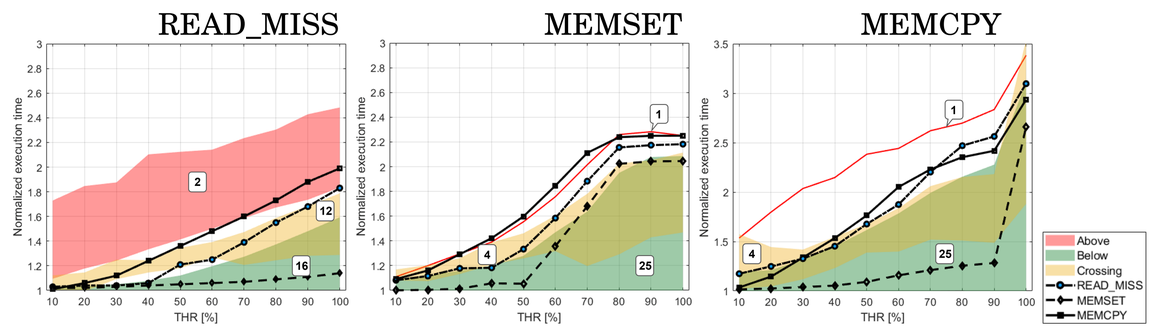
- **Read-only traffic generates low bandwidth** for the reference HeSoCs. This made us question the previous papers' claims, and we decided to investigate the actual amount of DRAM interference which the traffic types can cause.

References

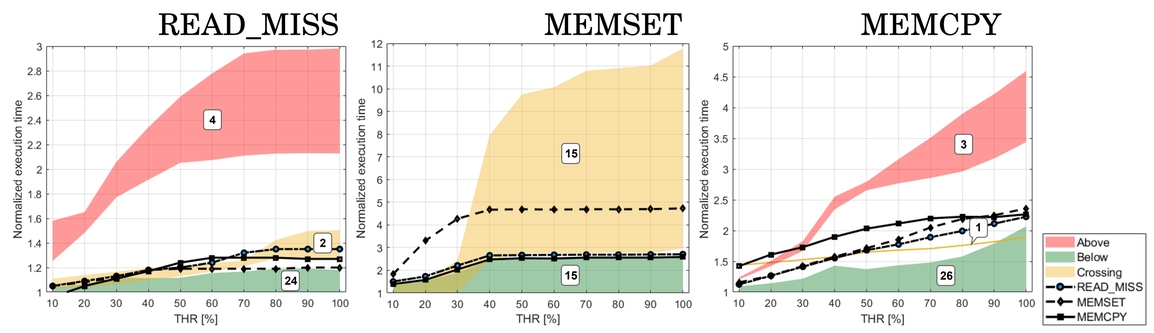
- [1] P. Radojkovi, S. Girbal, A. Grasset, E. Qui Nones, S. Yehia, and F. J. Cazorla, **On the Evaluation of the Impact of Shared Resources in Multithreaded COTS Processors in Time-Critical Environments.**
- [2] N. Capodieci, R. Cavicchioli, I. S. Olmedo, M. Solieri, and M. Bertogna, **Contending memory in heterogeneous SoCs: Evolu- tion in NVIDIA Tegra embedded platforms**, in 2020 IEEE 26th International Conference on Embedded and Real-Time Computing Systems and Applications (RTCSA), 2020, pp. 110.
- [3] G. Brilli, R. Cavicchioli, M. Solieri, P. Valente, and A. Marongiu, **Evaluating Controlled Memory Request Injection for Efficient Bandwidth Utilization and Predictable Execution in Heterogeneous SoCs**, ACM Trans. Embed. Comput. Syst., sep 2022, just Accepted.

Interference Analysis

- Executed both **Synthetic benchmarks** and **Polybench** against **READ_MISS**, **MEMSET** and **MEMCPY** running on the other cores, with increasing throttling factor (**THR%**) for the reference HeSoCs.



Platform: NVIDIA TX2



Platform: Xilinx ZU9EG

- **Results:**

1. Due to the effects of Cache thrashing, certain Polybench get more slowed down than the read-only synthetic benchmarks (Red and Yellow regions). Up to **12** × slowdown!
2. Causing the highest DRAM interference: **NVIDIA TX2** - MEMCPY, **Xilinx ZU9EG** - MEMSET.
3. Most subject to DRAM interference: **NVIDIA TX2** - READ_MISS, **Xilinx ZU9EG** - MEMSET.

Future work

- **Full-on paper** on the matter, with a deep dive on the cause for such high latency increase due to cache thrashing.
- **Exploration** of the effects of using both **CPU and FPGA** cores to cause interference (graph below).
- Long term goal: **Study of novel QOS guaranteeing techniques** to handle accelerator-based DRAM interference on these kinds of HeSoCs.