Hardware support for scratchpad memory transactions on GPU architectures

Alejandro Villegas[1], Rafael Asenjo[1], Angeles Navarro[1], Oscar Plata[1],
Rafael Ubal[2], and David Kaeli[2]

[1] Department of Computer Architecture, University of Málaga, Andalucía Tech,
29071 Málaga, Spain,
{avillegas, magonzalez, asenjo, oplata}@uma.es,

[2] Department of Electrical and Computer Engineering, Northeastern University,
Boston, MA, USA,
{ubal, kaeli}@ece.neu.edu,

Abstract. Graphics Processing Units (GPUs) have become the accelerator of choice for data-parallel applications, enabling the execution of thousands of threads in a Single Instruction - Multiple Thread (SIMT) fashion. Using OpenCL terminology, GPUs offer a global memory space shared by all the threads in the GPU, as well as a low-latency local memory space shared by a subset of the threads. The latter is used as a scratchpad to improve the performance of the applications.

We propose GPU-LocalTM, a hardware transactional memory (TM), as an alternative to data locking mechanisms in local memory. GPU-LocalTM allocates transactional metadata in the existing memory resources, minimizing the storage requirements for TM support. In addition, it ensures forward progress through an automatic serialization mechanism. In our experiments, GPU-LocalTM provides up to 100X speedup over serialized execution.

Keywords: Transactional Memory, Scratchpad Memory, GPGPU

Conclusions

In this paper we present GPU-LocalTM as a hardware TM for GPU architectures that focuses on the use of local memory. GPU-LocalTM is intended to limit the amount of additional GPU hardware needed to support TM. We propose two alternative conflict detection mechanisms targeting different types of applications. Conflict detection is performed per-bank, ensuring scalability of the solution. We find that for some applications the use of TM is not optimal and discuss how to improve our implementation for better performance. Furthermore, GPU-LocalTM introduces a serialization mechanism to ensure forward progress.

Acknowledgements

This work has been supported by projects TIN2013-42253-P and TIN2016-80920-R, from the Spanish Government, P11-TIC8144 and P12- TIC1470, from

Junta de Andalucía, and Universidad de Málaga, Campus de Excelencia Internacional, Andalucía Tech.