AES Embedded Systems Architecture

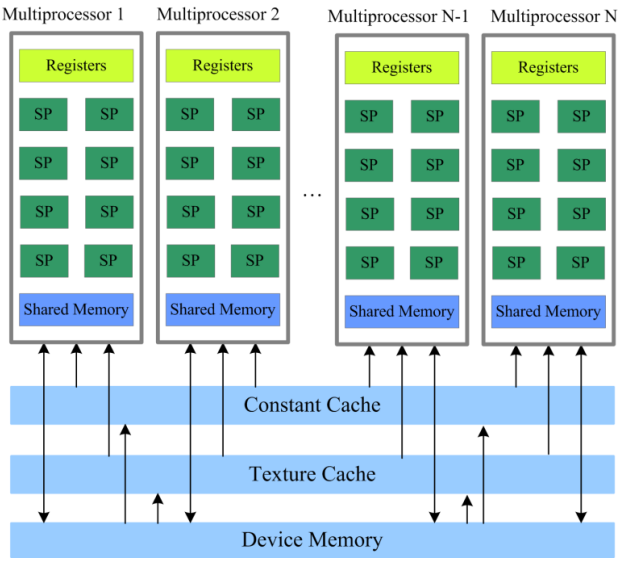
Master Thesis Project

Technische Universität Berlin

Design and Implementation of an Efficient Spatial Locality Predictor for GPUs

Background and Motivation: As GPUs computational power grows, their memory hierarchy is increasingly becoming a bottleneck. For example, there are many GPU applications which are limited by off-chip memory bandwidth. Unfortunately, the off-chip memory bandwidth is growing slower than the number of cores and has become a performance bottleneck.

Current GPU memory hierarchies' use coarse-grained memory accesses to exploit spatial locality, maximize peak bandwidth etc. The coarse-grain accesses are poor match for GPU applications with irregular control flow and memory access patterns. A dynamic granularity memory system which retains the advantages of coarse-grained accesses to exploit spatial locality and can also do fine-grained accesses to save memory bandwidth etc. will help to improve the performance of GPU applications.



Src:https://bmcresnotes.biomedcentral.com

Project Goal:

Efficient Spatial Locality Predictor (SLP) is the core of the dynamic granularity memory system. The massive multi-threading of GPUs make direct application of CPU-specific memory enhancements inefficient for improving the performance of GPUs. In this thesis we aim to design and evaluate an efficient SLP for GPUs. The design space of the SLP will be explored and based on the gathered knowledge about type of spatial locality; the SLP design space will be pruned.

Desired Skills

- Excellent C/C++
- Strong knowledge of computer architecture
- Preferably GPU architecture and CUDA

Contact Person

Sohan Lal (sohan.lal@tu-berlin.de)

References:

- D. Yoon, M. K. Jeong, M. Sullivan, M. Erez, "The Dynamic Granularity Memory System", ISCA , 2012
- M. Rhu, M. Sullivan, J. Leng, and M. Erez, " A Locality-aware Memory Hierarchy for Energy-Efficient GPU Architectures", MICRO, 2013