Devish Mundra, CSCI 5593

Report #1, February 3, 2019

Report on "CPU-Assisted GPGPU on Fused CPU-GPU Architectures"

This paper is based upon a technique that takes the advantage of the fused architecture that allows graphical processing units (GPU) and central processing unit (CPU) on a single chip to collaborate boosting processor performance by sharing the same on-chip L3 cache and off-chip memory like the intel sandy bridge and AMD accelerated processing unit. The proposed technique uses a program that has been compiled to leverage the architecture to allow the CPU and GPU to collaborate on the computational tasks.

The approach developed by the authors leverages the computational power of the GPU, while taking the advantage of the CPU's more flexible data retrieval and better handling of the complex tasks. The current generation CPU/GPU systems has helped in creating an energy efficient system. But still the CPU cores and GPU cores still work almost exclusively on separate functions. They rarely collaborate with each other to execute a given program, so we are unable use them as efficiently as they could be.

GPUs are designed for handling graphics, but they are also very good at handling large amount of parallel processing, particularly in those applications where the same process needs to be applied to a large amount of the data. The biggest problem for using GPUs for general purpose computations is that they don't handle complex, branches heavy code very well, but at the other side these are the strengths of the CPUs. The authors make use of the advantage of GPUs to move the data to and from the memory. To avoid GPU starving for data the authors proposed to keep the level 3 cache filled with data. This is more efficient because it allows CPUs and GPUs what they are good at. GPUs are good at performing computations. CPUs are good at making decisions and flexible data retrieval

The authors proposed method uses the CPUs faster L3 cache to pre-fetching to feed the data to GPU, by cutting out performance drags that comes with GPU code accessing memory. The authors designed a program for CPU-assisted GPGPU launches a pre-execution program at startup on the CPU to pre-fetch data to be processed by GPU code and load it into the L3 cache onboard chip, which allows process threads running in the GPU to hit the L3 cache directly, rather fetching from memory, reducing latency and significantly boosting performance.

In other words, CPUs and GPUs fetch data from off-chip main memory at approximately same speed, but GPUs can execute the functions that use that data more quickly. So, if a CPU determines what data a GPU will need in advance, and fetches it from off-chip main memory, that allows GPU to focus on executing the functions themselves and overall process less time. In the preliminary testing authors found that this approached improved fused processor performance by an average of 21.4 percent.

Reference

[1] **Huiyang** Zhou, Mike Mantor, Ping Xiang, Yi Yang, "CPU-Assist GPGPU on Fused CPU-GPU Architecture" Special Issue on: 18th International Symposium on High Performance Computer Architecture, 27 February 2012, New Orleans, doi: 10.1109/HPCA.2012.3168948