Performance/energy aware optimization of parallel applications on GPUs under power capping

Adam Krzywaniak, Pawel Czarnul
Faculty of Electronics, Telecommunications and Informatics
Gdansk University of Technology, Poland
adam.krzywaniak@pg.edu.pl, pczarnul@eti.pg.edu.pl

In the paper, we present results of research on performance and energy aware optimization of parallel applications run on modern GPUs under power capping. Power capping has been introduced as a feature available for modern server, desktop and mobile CPUs as well as GPUs through tools such as Intel's RAPL, AMD's APM, IBM's Energyscale for CPUs and NVIDIA's NVML/nvidia-smi for NVIDIA GPUs [2,3]. Our recent review [2] of energy-aware high performance computing surveys and reveals open areas that still need to be addressed in this field of study. While there have been several works addressing performance and energy awareness of CPU-based systems, the number of papers related to finding optimal energy-aware configurations using GPUs is still very limited. For instance, paper [4] looks into finding an optimal GPU configuration in terms of the number of threads per block and the number of blocks. Paper [7] finds best GPU architectures in terms of performance/energy usage. In this work, we provide a follow up of our previous research [5,6] for CPU-based systems aimed at finding interesting performance/energy configurations obtained by setting various power caps. Within this paper, we provide results of running selected and widely considered benchmarks such as: NPB-CUDA which is an implementation of the NAS Parallel Benchmarks (NPB) for Nvidia GPUs in CUDA 1 as well as cublasgemm-benchmark 2 . Specifically, preliminary tests indicate that using power capping on an NVIDIA GTX 1070 GPU installed within a workstation with an Intel(R) Core(TM) i7-7700 CPU @ 3.60GH (4 cores, 8 logical processors) and 16 GBs of RAM:

1. for NAS SP class C benchmark we were able to save 13.5% of total energy consumption at the cost of execution time increased by less than 3%,
2. for NAS BT class B benchmark we were not able to reach any measurable energy gains,
3. for GEMM operations (10 iterations for square matrices of size 4096 up to 16384) we were able to save 16.2% of total energy consumption at the cost of execution time increased by 29%.

This indicates that potential gains depend very much on the application and problem size (NAS Parallel Benchmark class). In the paper, we provide extended data from these tests on several modern GPUs available at the Faculty of Electronics, Telecommunications and Informatics, Gdansk University of Technology, Poland. Following work [1], we could also observe gradual and reasonably slow drop of power consumption on a GPU right after application has finished.

Keywords: performance/energy optimization, power capping, GPU, NAS Parallel Benchmarks.