Optimization of Pattern Matching
Compression Algorithms for GPUs

The OPAC project studied the optimization of high performance image coding algorithms, for many-core (GPUs) and multicore platforms (CPUs), relying on several parallel standards such as OpenMP and OpenCL. A broad range of platforms were studied, namely, high end servers, laptops, embedded systems and mobile devices from both the raw performance and energy consumption point of view.

GENERAL MOTIVATION AND OBJECTIVES

Compression of digital images and videos is mandatory to contain the data deluge and to preserve network bandwidth. However, modern image/video compression algorithms achieve high compression ratio at the cost of computational complexity, requiring massive amounts of computational power and energy, and long encoding time. This is also the case for the Multimedia Multiscale Pattern algorithm, which has been developed by the members of the IT research team in the past decade. One of the main goals of the OPAC project was to optimize the raw performance of the algorithms for a broad range of platforms – high end servers, laptops, embedded systems and mobile devices –, while assessing the energy consumption. OPAC also focused on implementing CPU optimized versions, resorting to the OpenMP and OpenCL standards, creating CPU/multicore versions that can be fairly compared to GPU/multicore ones.

CHALLENGE

OPAC tackled several challenges. First, pattern-based compression algorithms have limited parallelism, since the compression of an input block is linked to the outcome of the previous block and so on. Therefore, opportunities for parallelism need to be explored in the inner-block compression. Second, although parallel standards allow for code portability across platforms, performance is not portable. For instance, an OpenCL code tuned for GPUs but it will not deliver the top performance that can be achieved with a GPU-optimized OpenCL code.

First challenge was addressed by optimizing well known algorithms used in image/video encoding. This was the case for the Least Square Prediction (LSP), and the Walsh-Hadamard transform. The second set of challenges was tackled by building many fine-tuned versions of the algorithms for all the available platforms.

WORK DESCRIPTION AND ACHIEVEMENTS

OpenMP and OpenCL versions of MMP tuned for CPU were developed in the case of the OpenCL standards, existing GPU-optimized versions were adapted and later optimized for multicore/multiple CPU machines. Highly optimized versions of OpenMP were built, leveraging the parallelism available within the intra-block compression. All of the versions were assessed for raw performance and energy consumption on a wide range of hardware platforms, namely high end and multiple CPUs servers, laptops, and System on a Chip (Raspberry Pi 2 and NVIDIA’s Jetson TK1). Experience and knowledge were learnt in collecting and processing simultaneously performance and energy consumption measurements. Another important achievement was the use of the challenging SIMD instructions to optimize GPU-versions of image encoding software. This was done within the OpenCL platforms, boosting the performance of CPU-based code, while maintaining code portability.

Main Project Team


Journal Publications


The final code yields fast execution times, while consuming a fraction of the energy of other platforms.

In conclusion, the OPAC project achieved several meaningful results: use of OpenMP and OpenCL for optimizing CPU code; Use of CPU’s SIMD instructions within the context of OpenCL study and OpenCL’s Optimized version of LSP and the Walsh-Hadamard algorithms; coupling of energy consumption measurements with execution time performance measurements.