SORS: "Advancing the State of the Art of Directive-Based Programming for GPUs: Runtime and Compilation"

Abstract

In this talk we will review new possibilities for optimizing directive-based code from both runtime and compilation perspectives. First, we will introduce a runtime framework for OpenACC to facilitate dynamic analysis and compilation. Especially, our framework realizes automatic asynchronous execution and multi-GPU use based on the status of kernel execution and data availability while taking advantage of an on-the-fly mechanism for compilation and program optimization. We add a versatile code-translation method for multi-device utilization by which manually-optimized applications can be distributed automatically while keeping original code structure and parallelism.

Second, we implement a novel flexible optimization technique that operates by inserting a code emulator phase to the tail-end of the compilation pipeline. Our tool emulates the generated code using symbolic analysis by substituting dynamic information and thus allowing for further low-level code optimizations to be applied. We implement our tool to support both CUDA and OpenACC directives as the frontend of the compilation pipeline, thus enabling low-level GPU optimizations for OpenACC that were not previously possible.

Third, we propose the use of a modern optimization technique, equality saturation, to optimize sequential code utilized in directive-based programming for GPUs. Our approach realizes less computation, less memory access, and high memory throughput simultaneously. Our fully-automated framework constructs single-assignment forms from inputs to be entirely rewritten while keeping dependencies and extracts optimal cases. Overall, we cover runtime techniques and optimization methods based on dynamic information, low-level operations, and user-level opportunities.

Short Bio

Kazuaki Matsumura is a Sr. Software Engineer on the NVIDIA HPC Compiler team. His research interests lie in compiler design and program optimization for high-performance computing. He is currently working on many aspects of nvc/nvc++/nvfortran from the front end to the back end.

Speakers

Speaker: Kazuaki Matsumura. Sr. Software Engineer on the NVIDIA HPC Compiler team.
Host: Antonio Peña. Accelerators and Communications for High Performance Computing. Computer Sciences, BSC.
Barcelona Supercomputing Center - Centro Nacional de Supercomputación