REPORTING PERFORMANCE IN THE THIRD AGE OF GPU COMPUTING - HOW TO OPTIMIZE, VERIFY AND VALIDATE GPU CODES

André R. Brodtkorb, SINTEF ICT, +4722067548, Andre.Brodtkorb@sintef.no

1. André R. Brodtkorb, SINTEF ICT

Introduction. The aim of this talk is to give an overview of best practices for reporting performance and accuracy in the third age of GPU computing. While these tasks appear trivial at first, it is still a challenging and time consuming task that is somewhat neglected in many papers. According to Owens et al. [2], the focus should be on "building real applications on which GPUs demonstrate an appreciable advantage". This is an important message, which means that we must shift the focus from reporting high "speed-up numbers" for academic toy models, to reporting a genuine advantage of using GPUs for real world problems. We will detail what this requires in this talk, by going through the required steps for verification and validation of GPU codes, with a particular emphasis on single precision versus double precision, and by going through best practices for reporting performance.

From "Looks Real" to Physically Correct. In the third age of GPU computing, it is not sufficient that GPU results look plausible: we need to thoroughly verify and validate our results. This is just as important for CPU codes, but a particular difficulty with GPUs is that double precision is often prohibitively expensive: one either has to purchase expensive "compute-GPUs", or use "gaming-GPUs" with a fraction of the double precision performance. Thus, one is often faced with having to choose double precision or performance. We go through the required steps for a thorough verification and validation of GPU codes, with a particular emphasis on the use of single precision and performance. Performance Assessment. In the first age of GPU computing, the main message to convey was that GPUs could offer a performance benefit over using the CPU for a variety of algorithms. This resulted in a speed-up race that continues even today, in which people report the GPU as being tens to hundreds of times faster than the CPU. When we examine the theoretical performance gap between the two architectures, which is currently around seven times, it becomes self-evident that something does not add up for the most extreme speed-ups. It is thus now time to start reporting a new set of metrics that give better insight into the performance than a single speed-up figure. We argue that in the third age of GPU computing, the primary metrics should be resource utilization which is tightly linked with the best practices of profile driven development. In addition, we need to start reporting good domain specific metrics, such as millions of cells per second or millions of lattice updates per second.