An exploration of the trade-off between GPU precision, performance and power consumption

Student: Yubo Zhi (yz39g13@soton.ac.uk)
Supervisor: Dr. Geoff Merrett (gvm@soton.ac.uk)

Nowadays, because of environmental issues and rapid growing mobile computation requirement, power consumption on electronic devices become critical, especially on platforms that are running on battery. As battery technology development is not as fast as electronic technology development, lower power consumption means longer battery life for mobile devices e.g. laptops and smart phones, also lower environmental impact. This project is purposed to reduce power consumption on applications that utilise GPU to do large scale parallelism floating point calculations.

When GPU are dealing with floating number arithmetic with a higher precision, more hardware components and computation time are required, thus consume a lot more power. The objective of this project is to dynamically control GPU calculation precision to reduce overall GPU power consumption for specific tasks. In order to achieve this, programs and benchmarks that are suitable for analysis and demonstration purpose need to be implemented and then establish a method of controlling computation precision in real-time, therefore enabling investigation of power consumption versus computation quality and performance. In addition, automatic precision feedback control based on computation quality could then be implemented. For easier power consumption analysis, the Jetson embedded development platform featuring a quad-core ARM CPU and a CUDA-enabled Kepler GPU, introduced by NVIDIA, will be used in this project.

In technical aspect, NVIDIA CUDA API will be used for controlling computation precision, which also requires the algorithms and tasks be implemented in CUDA firstly, then establish a method to do run-time precision control through things like message queue, to transfer precision control queries to real-time GPU computation manager thread. Power consumption will then be analysed by using a power analyser attached to the development board, then controlled directly from the application through on-board GPIO, automatically triggered only during GPU computation. After that, power consumption with different precision settings will be investigated, e.g. use lower precision for varies proportion of computation loops, then plot power versus quality charts.

If time allows, applications using OpenGL or OpenCL as rendering or computation interface may also be analysed by switching between different sets of code segments with different precision. By evaluating computation quality on the fly as feedback, automatic precision control may also be implemented.