Benchmarking Fermi Microarchitecture

Tristan Overney, Clément Humbert

November 9, 2014

1 Goals

The goals of this research is to expose the microarchitecture implemented by Nvidia Fermi cards such as: pipeline length, instructions latency, scheduling patterns.

2 Methods

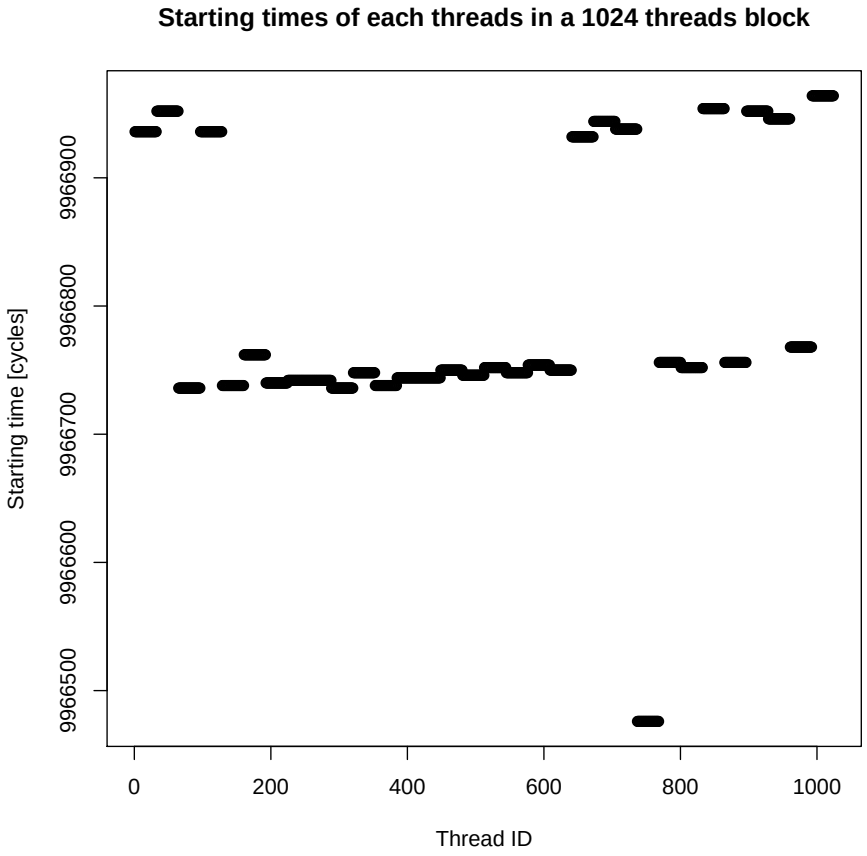
To achieve the aforementionned goals, a serie of specially crafted CUDA kernels were used. These usually contain large batches of dependent instructions that were timed with the assistance of the clock64() function offered by the CUDA API.

The benchmark programs have been ran on a machine equipped with a: Nvidia GeForce GTX 580.

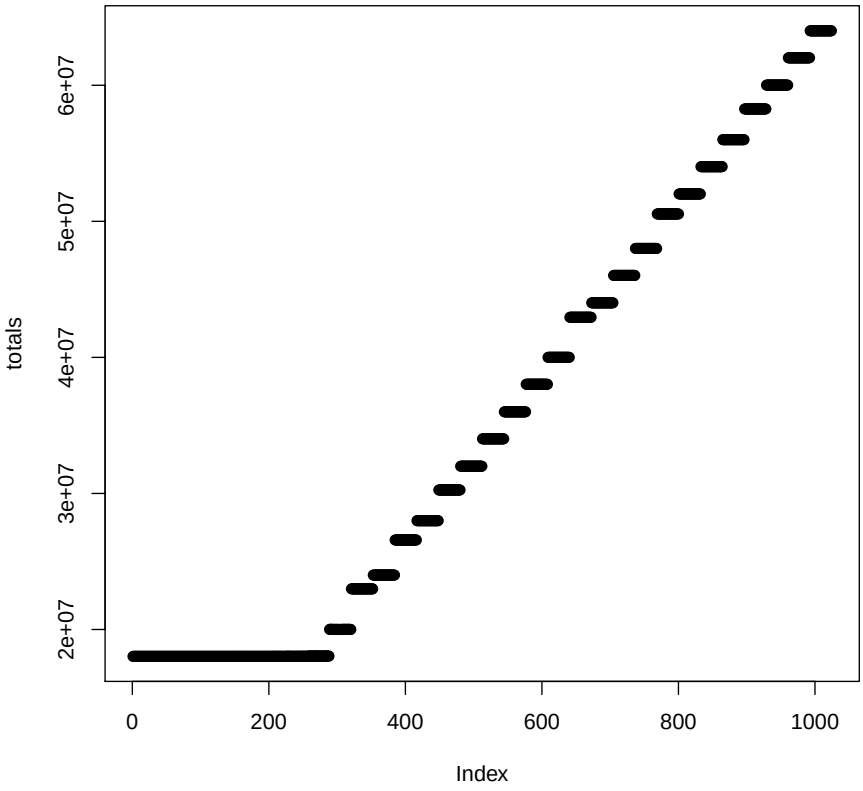
3 Integers multiplication

This section contains the results obtained through the previously described methods using large batches of integer multiplications.

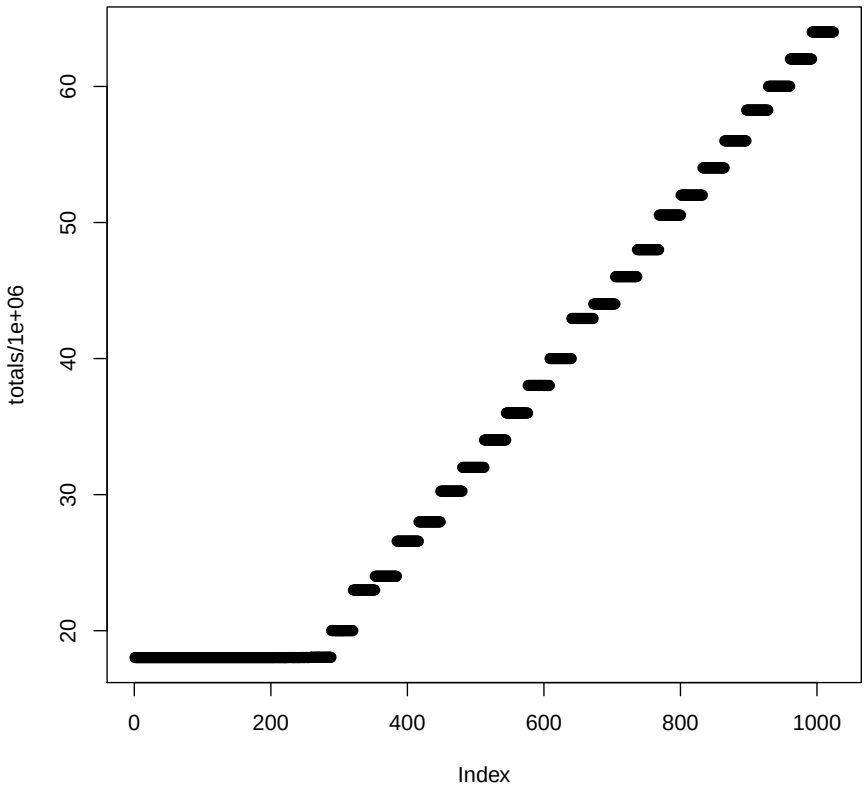
3.1 Integer multiplication: 1024 threads starting times



3.2 Benchmark running time against number of threads

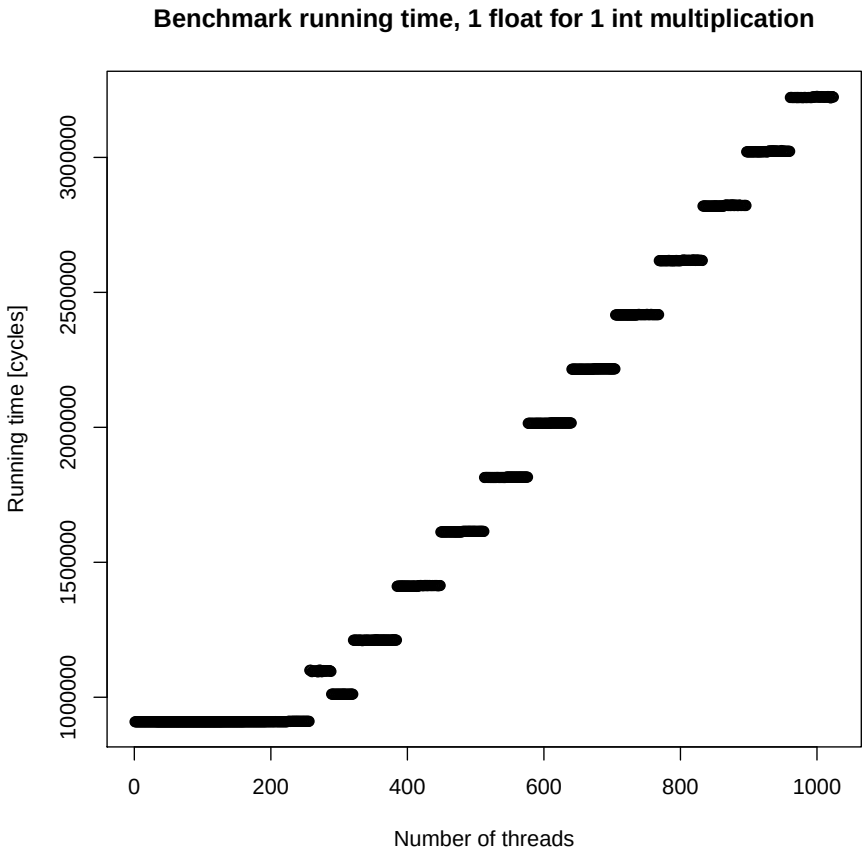


3.3 Benchmark running times divided by number of multplications

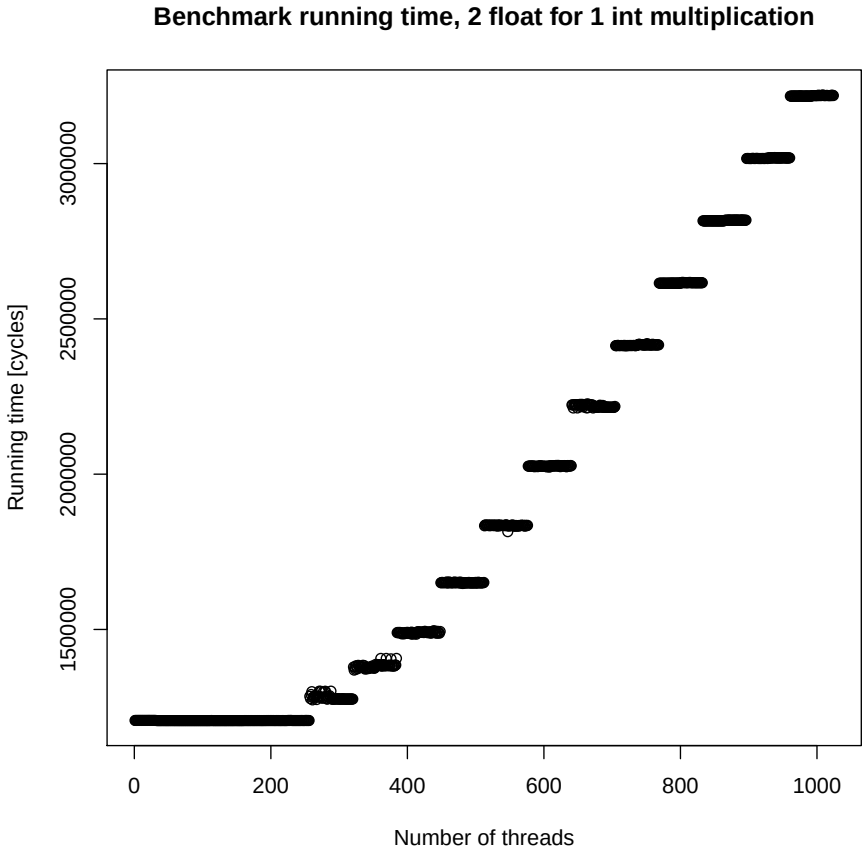


4 Mixing floating points and integer multiplication

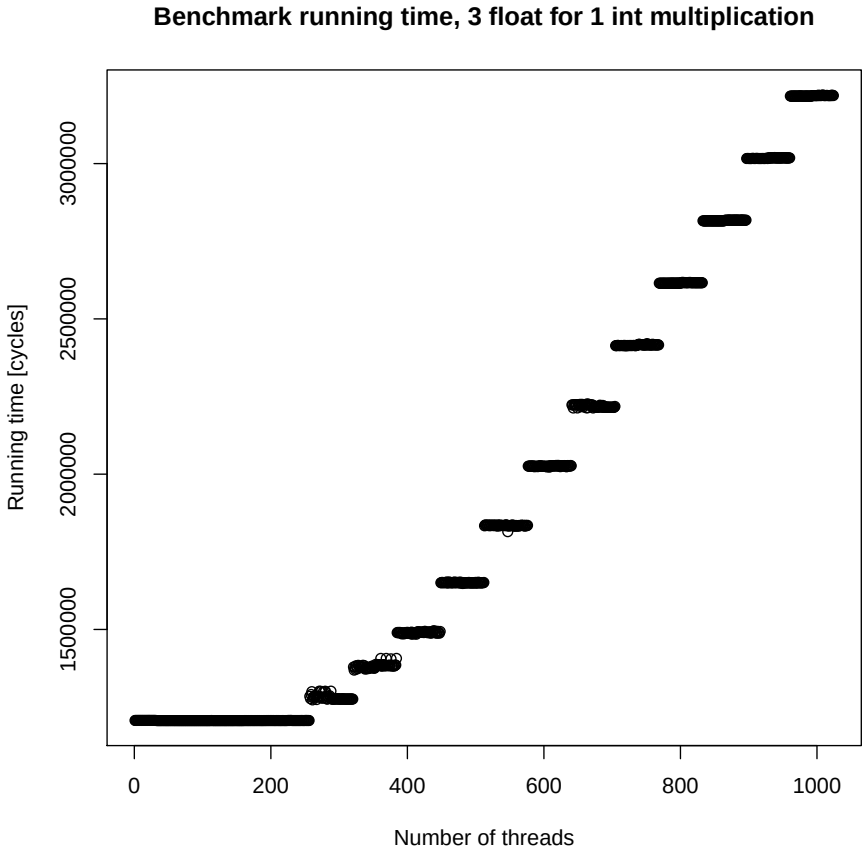
4.1 Benchmark running times, 1 floating points for 1 integer multiplication



4.2 Benchmark running times, 2 floating points for 1 integer multiplication



4.3 Benchmark running times, 3 floating points for 1 integer multiplication



5 Interpretation