PCS5120 Homework

Giuliano A. F. Belinassi[1]

[1]Institute of Mathematics and Statistics (IME) – University of São Paulo (USP)
Rua do Matão, 1010 – São Paulo – SP – Brazil

In various research fields where computational linear algebra is required eighter because of its facility or inexisence of analytical solutions, can take advantage of numerical linear algebra packages such as Lapack or BLAS. One major concern about these libraries is its performance, subject discussed in this report. Here we target the routine designed to multiply two matrices in single precision floating point (`SGEMM`).

Althrough implementing a matrix-matrix multiplication seems trivial by its concept, it is fairly difficult to provide an efficient code because of various reasons, such as cache usage. The use of techniques such as block matrix-matrix multiplication can yield better results due to a better cache usage, but a question that can be asked from this approach is what is the block size that maximizes performance?

1. The database

We used the database "*SGEMM GPU kernel performance Data Set*" provided by *UCI machine learning repository*[1]. Briefly, it has timings of multiplication of two matrices in *ms*, each one of size 2048×2048, using a combination of 14 parameters, totalizing 241601 lines in the database. Since the database provides four run times per line, we added a new column "mean" that is the average of these values.

2. Data analysis

We used Orange in our analysis. Our first objective was to find the distribuition of the average of the four executions per sample to check possible improvements or deterioration of the performance. Figure 1 shows such distribution. Notice that most of the averages concentrate under 200ms, and there are some data aroung 2400ms. Such high timing can be caused by the parameters itself or be an outlier because there were other programs running on the computer.

3. Conclusions

[1]`https://archive.ics.uci.edu/ml/datasets/SGEMM+GPU+kernel+performance`

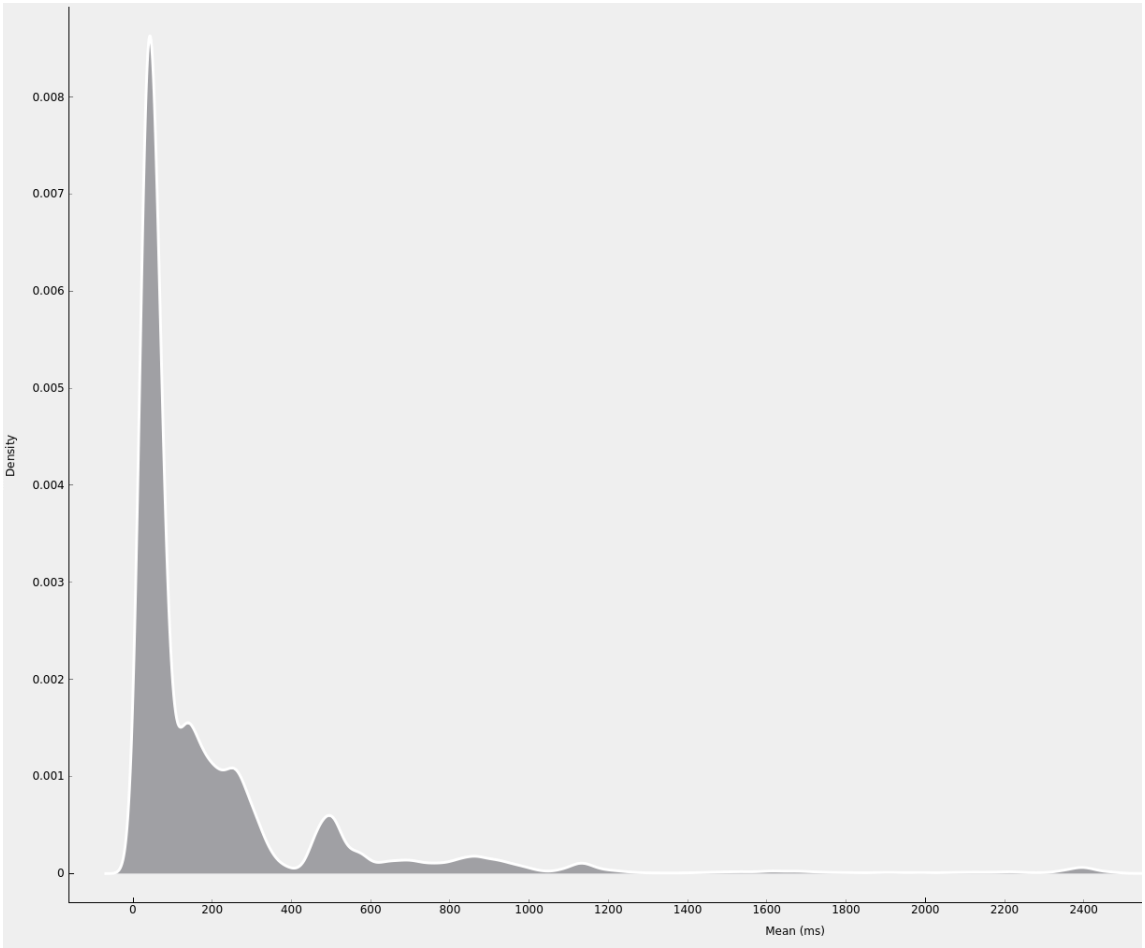

Figure 1. Distribution of the average of four samples per parameter.