Institute of Informatics
Parallel Computing Systems

Stephen Swatman - s.n.swatman@uva.nl
Ana-Lucia Varbanescu - a.l.varbanescu@uva.nl
April 4th, 2022

Performance Engineering

Assignment 1: The Roofline Model

The goal of this assignment is to practice with the Roofline model for a small application kernel. To this end, several models - on different architectures and different versions of the code - need to be built. For all models, we require to build the model manually and compare it against a version generated by a tool of your choice (examples to be provided).

You are provided a sequential implementation of a matrix multiplication algorithm (see Canvas). The application reads matrices A and B from two different files and writes the output to a third file. The format in which the matrices are stored is the "matrix market" format (`https://math.nist.gov/MatrixMarket/formats.html#MMformat`). We use `mmio.c` and `mmio.h` for handing the I/O of these matrices (`https://math.nist.gov/MatrixMarket/mmio-c.html`).

We provide a set of 5 matrices for your tests (arc130, ash958, bcsstk13, G6, tols1090). All performance results and model validation should be reported for all five matrices. As you have to multiply $A \times B$, you can assume $B = A$ for square matrices and $B = transpose(A)$ for rectangular ones. More matrices are available at `https://sparse.tamu.edu/` (where the selected five are also originating from), and you can also generate your own input data. Please assume matrices are dense, even if the number of non-zero elements is low. Reporting results for more matrices is appreciated, but not mandatory.

Assignment 1.1: The Roofline model - sequential - 1p

Build the Roofline model for the reference version of your matrix multiplication.

Assignment 1.2: Matrix multiplication optimization - 2p

Improve the performance of matrix multiplication, based on the information from the Roofline model. Consider techniques to improve the memory behavior (i.e., improve caching) and the computational efficiency (i.e., consider making use of SIMD, possibly using compiler pragma's and/or flags) of the application. Measure the performance of the original and improved application version(s); report performance as improvement over the reference version.

Assignment 1.3: The Roofline model - sequential + optimized - 2p

Add the new Roofline model for your optimized matrix multiplication version(s) to the original. Comment on the findings from comparing these models.

Assignment 1.4: Parallel matrix multiplication - 3p

Please build a (simple) model to determine what is the best performance to be achieved, for the full application, through parallelization. Parallelize the matrix multiplication kernel for the CPU (we recommend using OpenMP) and for the GPU (we recommend using CUDA or OpenACC). Build a Roofline model for each implementation and, using it as an indicator, improve the performance further (if possible) or explain why that is no longer feasible.

Assignment 1.5: Model validation/assessment and reflection - 2p

Please assess your models (Roofline sequential, estimation for the maximum parallel performance, Roofline parallel) for all 5 provided datasets. Are your models correct? How useful are they? What do you need to further improve them, to be more useful and/or accurate?

Submission due date: April 13, 2022, at 20:00