Square Partitioning for Parallel Matrices Multiplication

Thomas Lambert*[1]

[1]RealOpt (INRIA Bordeaux - Sud-Ouest) – CNRS : UMR5251, Université de Bordeaux, INRIA – France

Abstract

The problem of partitioning a matrix into a set of sub matrices has received a lot of attention in the last few years. This operation is indeed crucial when considering dense linear algebra kernels on heterogeneous platforms. Let us for instance consider dense matrix multiplication based on Canon's-like algorithm, restricted for the sake of simplicity to the multiplication C=AB of two square n x n matrices A and B. Let us further assume that the matrices are partitioned into blocks, whose size is chosen so as to be well adapted to all types of resources (typically CPUs and GPUs). Then, at step k of the algorithm, the outer product of the k-th column of blocks of A and the k-th row of blocks of B is computed. Let us assume that processor P holds a set of s blocks whose projections along the different axis have respective size h and w. Then, the volume of computations P needs to perform is proportional to s and the volume of communications is proportional to h+w. In order to balance the computing load, each processor should receive a number of blocks proportional to its relative speed. In turn, the overall volume of communications is proportional to the sum of the projections of the areas owned by the different processors along the axes. Therefore, in order to minimize the processing time while minimizing the overall volume of communication, the optimization problem is amenable to the problem of partitioning a square into a set of zones of prescribed area (in order to balance the load) such that the sum of the projections along the two axes is minimized (in order to minimize the communications).

*Speaker