An Approach to Automatic Tuning for Parallel Householder QR Decomposition

Takeshi Fukaya[1)], Yusaku Yamamoto[2)], Shao-Liang Zhang[3)]

1),2),3) Graduate School of Engineering, Nagoya University

1) t-fukaya@na.cse.nagoya-u.ac.jp

2) yamamoto@na.cse.nagoya-u.ac.jp

3) zhang@na.cse.nagoya-u.ac.jp

We consider parallel computing of the Householder QR decomposition on SMP machines. This decomposition is one of the basic tools in matrix computations and is used in various problems such as the least square problem and the singular value decomposition of a rectangular matrix.

Since this algorithm consists almost entirely of BLAS routines such as matrix-vector multiplications, the simplest way of parallelization is to parallelize each BLAS routine. Moreover, using the blocking technique [1], we can use matrix-matrix multiplications, which can be efficiently parallelized.

On the other hand, the TSQR algorithm has been proposed in 2007 [2]. Thanks to this algorithm, coarse-grain parallelization of the QR decomposition for a Tall Skinny matrix has become possible. Additionally, using the blocking technique, we can apply this algorithm to the QR decomposition of not necessarily tall skinny matrices [3].

For efficient parallel computing, we need to consider two points at the same time; how to combine the BLAS-level parallelism and the TSQR algorithm, and how to partition a matrix into blocks. In our poster, we aim for automatic determination of these two things depending on the target machine and the size of the target matrix.

In our approach, we first identify parameters to optimize. Next we define an objective function based on the hierarchical structure of the computation. We show that the optimization problem can be transformed into the Bellman equation, which can be solved using dynamic programming [4]. Finally we propose a practical solution based on the performance model.

Performance evaluation on Xeon processor with 8 cores shows that performance of computation tuned with our approach is about as high as that tuned by hand.

References

[1] Dongarra, J., Duff, I., Sorensen, D. and van der Vorst, H.: *Numerical Linear Algebra for High-Performance Computers*, SIAM, 2 edition (1998).

[2] Langou, J.: AllReduce algorithms: application to Householder QR factorization, *Proceedings of the 2007 International Conference on Preconditioning Techniques for Large Sparse Matrix Problems in Scientific and Industrial Applicatipns*, pp. 103–106 (2007).

[3] Demmel, J., Grigori, L., Hoemmen, M. and Langou, J.: Communication-avoiding parallel and sequential QR factorizations, Technical report, Electrical Engineering and Computer Sciences, University of California Berkeley (2008).

[4] Bellman, R.: *Dynamic Programming*, Dover Publications (2003).