Scopus

Documents

Khan, A.H., Al-Mouhamed, M., Fatayer, A., Mohammad, N.

Optimizing the Matrix Multiplication Using Strassen and Winograd Algorithms with Limited Recursions on Many-Core
(2016) *International Journal of Parallel Programming*, 44 (4), pp. 801-830. Cited 3 times.

Abstract
Many-core systems are basically designed for applications having large data parallelism. We propose an efficient hybrid matrix multiplication implementation based on Strassen and Winograd algorithms (S-MM and W-MM) on many-core. A depth first (DFS) traversal of a recursion tree is used where all cores work in parallel on computing each of the N× N sub-matrices, which are computed in sequence. DFS reduces the storage to the detriment of large data motion to gather and aggregate the results. The proposed approach uses three optimizations: (1) a small set of basic algebra functions to reduce overhead, (2) invoking efficient library (CUBLAS 5.5) for basic functions, and (3) using parameter-tuning of parametric kernel to improve resource occupancy. Evaluation of S-MM and W-MM is carried out on GPU and MIC (Xeon Phi). For GPU, W-MM and S-MM with one recursion level outperform CUBLAS 5.5 Library with up to twice as fast for arrays satisfying N≥ 2048 and N≥ 3072 , respectively. Similar trends are observed for S-MM with reordering (R-S-MM), which is used to save storage. Compared to NVIDIA SDK library, S-MM and W-MM achieved a speedup between 20× and 80× for the above arrays. For MIC, two-recursion S-MM with reordering is faster than MKL library by 14–26 % for N≥ 1024. Proposed implementations achieve 2.35 TFLOPS (67 % of peak) on GPU and 0.5 TFLOPS (21 % of peak) on MIC. Similar encouraging results are obtained for a 16-core Xeon-E5 server. We conclude that S-MM and W-MM implementations with a few recursion levels can be used to further optimize the performance of basic algebra libraries. © 2015, Springer Science+Business Media New York.

2-s2.0-84944704335
Document Type: Article
Publication Stage: Final
Source: Scopus