A New Direction to Parallelize Winograd's Algorithm on Distributed Memory Computers

D. K. Nguyen[1], **I. Lavallee**[2], **M. Bui**[3], **and Q. T. Ha**[4]

Abstract: Winograd's algorithm to multiply two $n \times n$ matrices reduces the asymptotic operation count from $O(n^3)$ of the traditional algorithm to $O(n^{2.81})$, thus on distributed memory computers, the association of Winograd's algorithm and the parallel matrix multiplication algorithms always gives remarkable results. Within this association, the application of Winograd's algorithm at the inter-processor level requires us to solve more difficult problems in designing but it forms the most effective algorithms.

To use Winograd's algorithm at the inter-processor level, the most significant point is to determine the sub matrices after having recursively executed r time the Winograd's formula (these sub matrices correspond to the nodes of level r in the execution tree of Winograd's algorithm) and then to find the result matrix from these sub matrices (corresponding to the process of backtracking the execution tree). It is simple to solve this problem for a sequential machine, but it's much harder for a parallel machine. With a definite value of r, we can manually solve this problem like in the previous papers with $r = 1$, 2, but the solution for the general case has not been found. In this paper, we present our method to determine all the nodes at the unspecified level r in the execution tree of Winograd's algorithm, and to show the expression representing the relation between the result matrix and the sub matrices at the level recursion r; this expression allows us to calculate directly the result matrix from the sub matrices calculated by parallel matrix multiplication algorithms at the bottom level. By combining this result with a good storage map of sub matrices to processor, and with the parallel matrix multiplication algorithms based on traditional algorithm (1D-systolic, 2D-systolic, Fox (BMR), Cannon, PUMMA, BiMMeR, SUMMA, DIMMA, ...) we have a general scalable parallelizing of Winograd's algorithms on distributed memory computers.

From a different view, we generalize the Winograd's formulas for the case where the matrices are divided into 2^k parts (the case $k = 2$ gives us original formulas) thus we have a whole new direction to parallelize Winograd's algorithm.

[1,2,3] LaISC - Ecole Pratique des Hautes Etudes et Universite de Paris 8
No. 41, Gay Lussac Street, 75005 Paris, France
Kien.Duc-Nguyen@univ-paris8.fr, Ivan.Lavallee@univ-paris8.fr, Marc.Bui@univ-paris8.fr

[4] Faculty of Information Technology
Hanoi University of Technology
No. 1, Dai Co Viet Road, 10307 Hanoi, Vietnam
Trung.Ha-Quoc@it-hut.edu.vn