論文 / 著書情報
Article / Book Information

題目(和文)	
Title(English)	Cache Blocking and Parallel Runtime Scheduling of Hierarchical Matrices
著者(和文)	Deshmukh Sameer Satish
Author(English)	Sameer Satish Deshmukh
出典(和文)	学位:博士(工学), 学位授与機関:東京工業大学, 報告番号:甲第12558号, 授与年月日:2023年9月22日, 学位の種別:課程博士, 審査員:横田 理央,吉瀬 謙二,宮﨑 純,DEFAGO XAVIER,小野 峻佑
Citation(English)	Degree:Doctor (Engineering), Conferring organization: Tokyo Institute of Technology, Report number:甲第12558号, Conferred date:2023/9/22, Degree Type:Course doctor, Examiner:,,,,
学位種別(和文)	博士論文
Category(English)	Doctoral Thesis
種別(和文)	論文要旨
Type(English)	Summary

（博士課程）
Doctoral Program

論 文 要 旨

THESIS SUMMARY

系・コース： Department of, Graduate major in	Computer Science Computer Science	系 コース	申請学位（専攻分野）： Academic Degree Requested	博士 Doctor of	(Engineering)
学生氏名： Student's Name	DESHMUKH Sameer Satish		審査員主査： Chief Examiner	Rio Yokota	

要旨（英文 800 語程度）
Thesis Summary (approx.800 English Words)

Important scientific problems in electrostatics, acoustics and statistics can be solved using the Boundary Element Method (BEM). The coefficient matrix from a BEM discretization results in a dense matrix, leading to $O(N^2)$ and $O(N^3)$ time complexity for the matrix-vector product and direct factorization algorithms, respectively. The underlying geometry from which the dense matrix is generated can be exploited to approximate the interactions between far points. These points are expressed in the dense matrix as off-diagonals blocks. Low rank approximation of the off-diagonal blocks of such a structured dense matrix can reduce the time complexity of the matrix-vector product and direct factorization algorithms to $O(N)$ by trading off accuracy for time. The accuracy of the algorithm can be tuned to match the required accuracy of the application.

Algorithmic developments of the compression, multiplication and factorization routines of such low rank approximated matrices have led to reduction in the time complexity and better accuracy for a wide variety of problems. However, the implementation of such routines on modern, highly parallel computer architectures for such routines still requires the use of numerical software that was originally written for computation of dense linear algebra. The small, irregular kernels that are prevalent in the algorithms involving low rank approximation result in inefficient execution on modern hardware.

In this thesis, we propose improvements to the efficiency of the matrix-vector product and direct factorization algorithms of low rank approximated dense matrices on shared and distributed memory machines. We show that our techniques are applicable to 2D problems for modeling acoustic waves, electrostatics and statistics with acceptable accuracy for the application in question. We target the block low rank and hierarchically semi-separable low rank matrix formats for this study since they have been shown to work well with 2D problems.

The first part of this improves the efficiency of the matrix-vector product. The matrix-vector product is useful for problems arising from the wave equation. The low rank matrix multiplication algorithm is a key computational kernel of the matrix vector product. We design efficient cache blocking algorithms by leveraging the Execution-Cache-Memory performance model. Our portable implementation making use of parallel loops written in a high level language with architecture-specific assembly micro-kernels. This preserves the portability of our method and allows us outperform vendor optimized BLAS libraries on the Fujitsu A64FX, Intel Xeon 6148 and AMD EPYC 7502 CPUs. We improve the matrix vector product of low rank matrices by obtaining a 2x performance improvement in the low rank matrix multiplication algorithm on all CPU architectures.

The second part of this thesis improves the distributed memory factorization of hierarchical matrices arising out of 2D problems from electrostatics and statistics. Specifically, we shown that we achieve up to 2x improvement in weak scaling performance over the state-of-the-art in dense direct factorization on up to 128 nodes of Fugaku. We show that this improvement in speed can be achieved with similar or better accuracy than other state-of-the-art implementations making use of low rank approximation of structured dense matrices. The main reasons behind the improvement can be attributed to the use of the HSS-ULV algorithm with a distributed asynchronous runtime system. The combination of an algorithm $O(N)$ time and communication complexity with an asynchronous distributed runtime system results in better overlap of computation with communication and results in better performance.

備考 ： 論文要旨は、和文 2000 字と英文 300 語を 1 部ずつ提出するか、もしくは英文 800 語を 1 部提出してください。
Note : Thesis Summary should be submitted in either a copy of 2000 Japanese Characters and 300 Words (English) or 1copy of 800

Words (English).

注意：論文要旨は、東工大リサーチリポジトリ（T2R2）にてインターネット公表されますので、公表可能な範囲の内容で作成してください。
Attention: Thesis Summary will be published on Tokyo Tech Research Repository Website (T2R2).