# 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) | DeshmukhSameer Satish |
| Author(English) | Sameer Satish Deshmukh |
| 出典(和文) | 学位:博士(工学),<br>学位授与機関:東京工業大学,<br>報告番号:甲第12558号,<br>授与年月日:2023年9月22日,<br>学位の種別:課程博士,<br>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),<br>Conferring organization: Tokyo Institute of Technology,<br>Report number:甲第12558号,<br>Conferred date:2023/9/22,<br>Degree Type:Course doctor,<br>Examiner:,,,, |
| 学位種別(和文) | 博士論文 |
| Category(English) | Doctoral Thesis |
| 種別(和文) | 審査の要旨 |
| Type(English) | Exam Summary |

（博士課程）

# 論文審査の要旨及び審査員

| 報告番号 | 甲第　　号 | | 学位申請者氏名 | | Sameer Deshmukh | |
|---|---|---|---|---|---|---|
| | | 氏　名 | 職　名 | | 氏　名 | 職　名 |
| 論文審査<br>審査員 | 主査 | 横田 理央 | 教授 | 審査員 | 小野 峻佑 | 准教授 |
| | 審査員 | 宮崎 純 | 教授 | | | |
| | | 吉瀬 謙二 | 教授 | | | |
| | | Defago Xavier | 教授 | | | |

論文審査の要旨（2000字程度）

本論文は「Cache Blocking and Parallel Runtime Scheduling of Hierarchical Matrices」と題し、N×Nの密行列の行列分解の計算量を$O(N^3)$から$O(N)$に低減できる階層的低ランク近似法において有効なキャッシュブロッキング及びランタイムを用いた並列実装に関するもので、英文で書かれており全6章からなる。

１章「Introduction」では、高性能計算分野において中心的な役割を果たしてきた線形代数ライブラリの発展の歴史を振り返りながら近年注目されている階層的低ランク近似法の利点と問題点について述べている。具体的には、階層的低ランク近似法が多くのアプリケーションにおいて密行列分解の計算量を$O(N^3)$から$O(N)$に低減できる利点について触れ、その内部で発生する小さな行列積の演算性能が既存のライブラリでは低いことと、分散並列化を行う際に生じるデータの依存関係が性能低下をもたらしていることを問題点として挙げている。その上で、本論文の目的として複数の行列に跨るキャッシュブロッキングによる小さな行列積の性能向上と、ランタイムを用いることによる分散並列化による性能の向上を挙げている。

２章「Low Rank Approximation and Factorization of Dense Matrices」では、本論文の前提となる密行列分解について、それが科学技術計算においてどのような位置づけにあるのか、これまでどのような高速解法が提案されてきたかを詳細に解説している。有限差分法や有限要素法などから生じる疎行列とは異なり、境界要素法から生じる密行列は全く別の高速解法が必要になることを述べている。特に、密行列を階層的に分割し、その部分行列を低ランク近似することにより生成されるHSS行列やH行列の有用性について解説している。また、当該分野の研究が主に応用数学の研究者を中心に行われてきたため、高性能計算分野で頻繁に用いられている高速化・並列化の技法を適用した例が少ないことを問題として投げかけている。

３章「Background of Parallel Numerical Libraries and Computer Architecture」では、行列演算の高速化における定石であるキャッシュブロッキングについてSIMDやNUMAの概念を交えながら解説し、行列の次元が小さい場合に最先端の線形代数ライブラリでもプロセッサの理論演算性能を引き出すことが困難であることを述べている。また、小さい行列積を複数まとめて行う場合にバッチ処理により演算性能が向上することを挙げながらも、行列が細長い場合にはバッチ処理をしたとしてもプロセッサの理論演算性能を引き出すことが困難であることを述べている。さらに、Intel、AMD、Fujitsuのプロセッサにおいて性能の可搬性を実現するために、パフォーマンスモデルを用いて性能最適化を行う方法についても紹介している。次に、分散並列環境下において行列分解を行う際の処理の依存関係が非有向循環グラフで表せることを示し、それを実行時に解析することで処理の並列度を最大化するランタイムについて解説している。

４章「Cache Blocking for Batched Low-Rank Matrix Multiplication」では、階層的低ランク近似法を適用した行列―ベクトル積において生じる多数の小さな行列積の演算効率を上げるための新たなキャッシュブロッキング手法を提案している。Intel、AMD、Fujitsuのプロセッサの命令ごとのレイテンシを調べた上でパフォーマンスモデルにそれぞれの値を代入し、小さな行列積を行う際の演算性能の理論的な上限値を算出している。これを目標値として複数の行列に跨るキャッシュブロッキングを適用し、実装のチューニングを行うことでIntel、AMD、Fujitsuの各社が提供している線形代数ライブラリを大幅に超える演算性能を実現している。また、階層的低ランク近似のランク、ブロックの大きさなどを変化させて同様の比較を行い、全てのランクやブロックサイズにおいてバッチサイズが十分大きい場合は、提案の実装が既存のライブラリに対して優位であることを示している。

５章「HSS Matrix Factorization with PaRSEC」では、階層的低ランク近似によって生成される

HSS 行列に対して分散並列環境下で行列分解を行う際に、PaRSEC ランタイムを用いることの有効性を検証している。既存研究には PaRSEC を密行列の行列分解に適用したものはあるが、HSS 行列に適用する場合には計算負荷が不均一になることを明らかにした上で、クリティカルパスの負荷が均一になるように工夫することでこの問題を軽減できることを示している。その上で、STRUMPACK や LORAPO などの既存研究との比較実験を行い、ParSEC がノード数を増やした場合にも優位であることを示している。

６章「Conclusion」では、得られた知見と結論についてまとめている。主要な結論としては、階層的低ランク行列の行列―ベクトル積の際に生じる多数の小さな行列積に対して、本論文で提案するキャッシュブロッキング手法を適用することで、既存の線形代数ライブラリに対して約 2 倍の高速化を実現できること、階層的低ランク行列の行列分解を分散並列環境下で行う際に本論文で提案する手法をランタイムと併用することで、既存研究よりも計算量が少なく並列化効率の良い手法が実現できることが挙げられる。

以上より、本論文で提案している手法を用いることで高性能計算分野において中心的な役割を果たしてきた線形代数ライブラリの演算量を $O(N^3)$ から $O(N)$ に低減しながらも、プロセッサの演算性能を最大限に引き出すキャッシュブロッキング、分散並列化の効果を最大限に引き出すランタイムの実装を行い、それにより大幅な高速化を実現できることを示している。これは高性能計算分野において工学的に重要な貢献であるため博士（工学）の学位に相応しいと判断できる。