様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　３０　年　　６　月　１３　日現在

機関番号：　１２６０８
研究種目：　若手研究(A)
研究期間：　2016～2017
課題番号：　１６Ｈ０５８５９

研究課題名（和文）FMMとH行列を組み合わせた大規模連立一次方程式の反復解法

研究課題名（英文）Large scale iterative solvers by combining FMM and H-matrices

研究代表者
横田　理央（Yokota, Rio）

東京工業大学・学術国際情報センター・准教授

研究者番号：２０７６０５７３

交付決定額（研究期間全体）：（直接経費）　　5,100,000円

研究成果の概要（和文）：平成28年度にはFMMのH行列への拡張とH行列によるLU分解コードの開発を行った。この際、exaFMMで採用しているdual tree traversalをH行列のadmissibilityの判定に用いることでタスクベースの並列化を容易に行うことができた。平成29年度には内部カーネルのチューニングと実アプリケーションによるmultigrid法との比較を行った。Batched MAGMAを採用することにより小規模な行列をGPU上で高速に処理することができた。Multigrid法に対する優位性は行列の条件数に依存し、並列度がますとH行列が優位になることが分かった。

研究成果の概要（英文）：In FY2016, we extended the FMM to H-matrices and developed a LU decomposition code using H-matrices. The dual tree traversal of exaFMM was used to determine the block cluster tree for arbitrary admissibility conditions, which allowed tasked based parallelization of the compression part of the H-matrix code. In FY2017, we further optimized inner kernels of the H-matrix code and compared H-matrices with multigrid for real applications. The use of batched MAGMA enabled us to maximize the performance of GPUs even for small matrices. The advantage of H-matrices over multigrid depends on the condition number of the matrix, while the H-matrix becomes advantageous as the degree of parallelism increases.

研究分野：　高性能計算

キーワード：　H行列　FMM　GPU　LU分解

１．研究開始当初の背景

大規模連立一次方程式の高速解法は流体、構造、電磁界、音響などの数値解析に幅広く用いられており、最近では機械学習の分野でも盛んに用いられている。多くのアプリケーションでは計算時間の９割近くが連立一次方程式の解法に費やされるため、その部分を高速化する効果は大きい反面、大規模並列計算機上でその高速化を行うことは容易ではない。

連立一次方程式の解法として見た場合の FMM に内在する並列性は高く、並列数が増えるにしたがって multigrid 法に対して優位になる傾向があるため、次世代の大規模計算機で広く用いられることが予想される。最近では FMM の代数学的拡張である H 行列、H^2 行列、HSS 行列などが盛んに研究されている。国際的には H 行列に関連する論文が急激に増えており、応用数学の分野で最も注目されている研究領域の一つである。しかし、国内からは H 行列に関する論文はほとんど出ておらず、当該研究領域の国際水準に大きく遅れをとっている。

２．研究の目的

研究期間内には H 行列と FMM を組み合わせた手法を構築し、両手法の長所を兼ね備えた実装を実現する。連立一次方程式の直接解法のみならず、反復法の前処理にも適用する。現在主流の前処理手法である multigrid 法と比較し、その定量的優位性を明らかにする。性能比較は流体解析、電磁界解析、機械学習の３つの性質の異なるアプリケーションについて同等の計算条件、計算機環境の下で行う。また、これを通して第３者のアプリケーションコードとの標準インターフェイスを構築し使いやすさを評価する。

３．研究の方法

平成２８年度には、限られた偏微分方程式しか解くことのできない現在の FMM を一般的な偏微分方程式に適用できる H 行列による連立一次方程式の解法へと拡張する。これは FMM の多重極展開の部分を代数学的な低ランク近似に置き換えることで行うことができる。この際に FMM が適用可能な問題については H 行列を FMM に変換することでメモリの消費量を低減する。これは FMM の観点から H 行列を見た場合に冗長な情報を保存していることが明らかになることから実現できる工夫である。

平成２９年度には、線形代数カーネルのチューニングの専門家と協力し H 行列の内部カーネルのチューニングを行う。また、流体解析、電磁界解析、機械学習の各アプリケーションごとにパラメータのチューニングを行い、FFTW のような使いやすい標準インターフェイスを構築する。上記の３つの性質の異なるアプリケーションについて H 行列と multigrid 法の比較を行うことで H 行列が優位になるパラメータのレンジを各アプリケーションごとに示す。

４．研究成果

FMM と H 行列は表裏一体であることに本研究は着目した。つまり、同じ密行列の階層的低ランク近似手法でありながら、FMM は球面調和関数などを用いることで解析的に低ランク近似を行うことで行列を直接生成することなくその圧縮を行うことができるのに対して H 行列は randomized SVD などを用いることで代数的に低ランク近似を行い行列を陽的に生成する。FMM は行列を保存しないためメモリ消費量を削減できるが、H 行列は行列を保存するため演算量は減るがメモリ消費量は増加する。

本研究では、まず FMM と H 行列を同一の軸上でメモリ消費量と演算量の観点から比較した。これは行列を保存しない FMM と保存する H 行列のトレードオフをメモリ消費量と計算時間の観点から厳密に比較するためである。

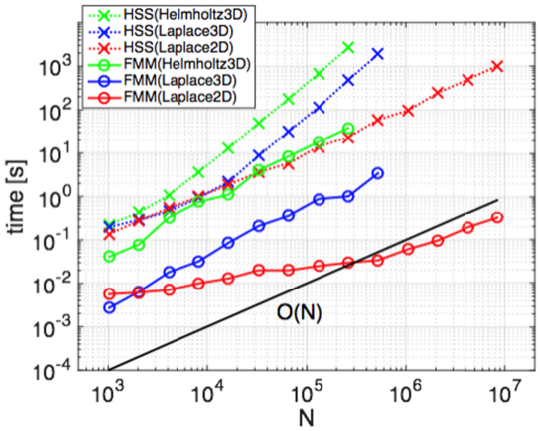

図１　FMM と HSS の計算時間の比較

図１に FMM と HSS の計算時間の比較を示す。ここで比較しているのは 2 次元の Laplace 方程式、3 次元の Laplace 方程式、3 次元の Helmholtz 方程式の Green 関数解から生成される行列を HSS で行列―ベクトル積を計算したものと、それらの方程式の Green 関数解を FMM で計算したものである。計算に用いたのは Intel Ivy Bridge E5-2695v2 の 12 core の CPU で、計算はいずれも OpenMP や AVX で並列化され１ソケットを用いて実行した。横軸の N は未知ベクトルの大きさを表している。FMM はいずれの方程式についても O(N)の計算時間を有していることが分かる。一方、HSS は 2 次元の方程式では O(N)の挙動を示すが、3 次元の方程式では HSS のランクが増大することで O(N)に漸近しない挙動が見られた。これは FMM の優位性というよりは weak admissibility を用いている HSS の弱点と

解釈できる。H 行列のような strong admissibility を用いる手法ではこのような挙動は見られない。

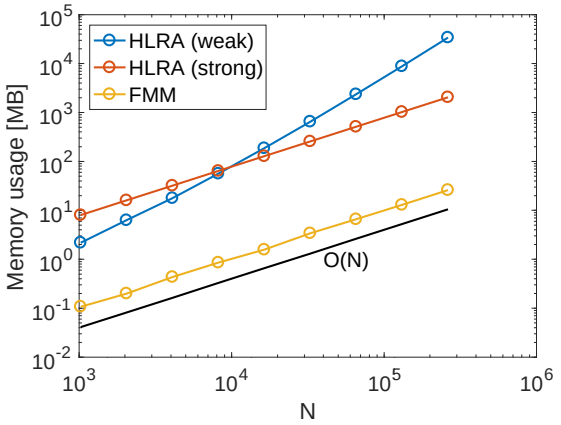

図 2 FMM と HLRA のメモリ消費量の比較

図 2 には 3 次元の Laplace 方程式をそれぞれ FMM と HLRA で解いたときのメモリ消費量を示す。HLRA は HSS のような weak admissibility と H 行列のような strong admissibility の両方が適用可能な汎用な手法である。図 1 と同様に FMM は O(N)の挙動を示しているのに対して、weak admissibility の HLRA では O(N)になっていない。ただし、strong admissibility の HLRA では O(N)の挙動を示している。また、FMM と HLRA(strong)はともに O(N)であるものの、FMM の方が 2 桁ほど小さいことが見てとれる。

FMM が HLRA に比べて 2 桁少ないメモリ消費量で済むのは、HLRA が行列の要素を圧縮しながらも保存するのに対して、FMM は行列―ベクトル積の際に要素を生成するため、マトリックスフリーとなるからである。FMM でも M2L の変換行列を保存することで高速化する手法が用いられるが、この場合でも M2L のセルの相対的な位置関係 316 通りについてのみ保存すれば良いため、O(1)のメモリ消費量で済む。HLRA は FMM の M2L 変換行列を球面調和関数ではなく randomized SVD などで代替しているに過ぎないので、実質的には FMM の変換行列の並進対称性は HLRA にも当てはまる。この性質を利用すると HLRA の低ランクブロックのほとんどが同じ値を持っているはずであり、N によらず 316 通りの可能性しかないことが分かる。本研究ではこの性質を利用することで HLRA のメモリ消費量を FMM とほぼ同等まで低減することができた。

本研究で用いている FMM と HLRA のハイブリッド化は ExaFMM というコードに実装した。ExaFMM は github 上にオープンソースで公開されており、流体解析、電磁場解析、分子動力学のシミュレーションなどで広く使われている。AVX, OpenMP, CUDA, MPI などあらゆる並列化を用いており、TBB や Cilk などのタスク並列化、QUARK などのランタイムも実験的に実装されている。本研究では FMM を連立一次方程式の直接解法として用いることができるように ExaFMM を拡張した。

図 1,2 に示されるような Laplace, Helmholtz 方程式は偏微分方程式を離散化すると疎行列が得られる。これを疎行列の直接解法を用いて解くこともできれば、反復法の前処理に multigrid 法を用いて解くこともでき、ここでは FMM を Laplace, Helmholtz 方程式の Green 関数解の行列―ベクトル積の高速解法として用いることで反復法の前処理に用いるができることを示した。この場合、FMM の精度を落としても前処理としては十分な効果が得られることが分かった。FMM の精度を落としすぎると反復数が増えるため、最適な精度で FMM を計算することが必要になる。ここで示す結果は、最適な精度で計算された場合の FMM の結果である。これを踏まえて、FMM を前処理に用いた場合と multigrid 法を前処理に用いた場合と元々の疎行列に直接解法を適用した場合の比較を示す。

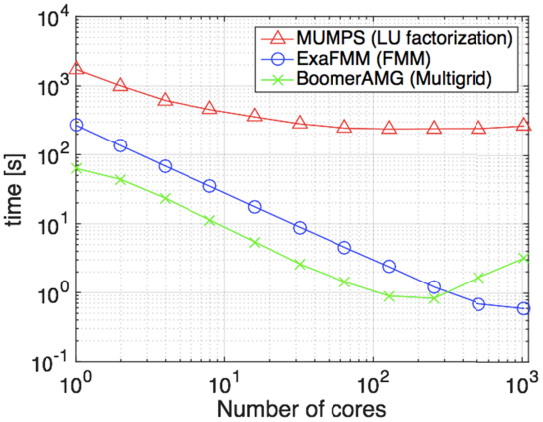

図 3 FMM と multigrid 法と直接法の比較

図 3 にはこのときの FMM と multigrid 法と直接法との比較を示す。また、ここではそれぞれの手法の並列化効率を比較するために横軸はコア数をとって比較した。縦軸は反復法全体が収束するのにかかった時間もしくは直接法の計算時間を表す。凡例にある MUMPS は直接法、ExaFMM は FMM、BoomerAMG が multigrid 法の結果を示している。ここで解いているのは図 1,2 にあったものと同様な 3 次元の Laplace 方程式である。MUMPS は逐次の計算時間も他の 2 手法に比べて大きくなっており、並列数を増やしたときの性能向上もあまり見られなかった。FMM と multigrid 法を比較すると、コア数の少ない時は multigrid 法の方が高速であるが、コア数が増えるにつれて FMM が優位になることが分かる。これにより、本研究の主な目的であった FMM を連立一次方程式の解法として見た場合の multigrid

法との直接比較を通して定量的な優位性を検証することができた。

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計6件）

① R. Yokota, H. Ibeid, and D. E. Keyes, Fast Multipole Method as a Matrix-Free Hierarchical Low-Rank Approximation, Lecture Notes in Computational Science and Engineering, Springer, 査読有, Vol. 117, 2017, pp. 227-244 DOI:10.1007/978-3-319-62426-6_17

② H. Ibeid, R. Yokota, J. Pestana, D. Keyes, Fast Multipole Preconditioners for Sparse Matrices Arising from Elliptic Equations, Computing and Visualization in Science, 査読有, Vol. 18, No. 6, 2017, pp. 213–229 DOI:10.1007/s00791-017-0287-5

③ J. E. Castrillon-Candas, M. G. Genton, R. Yokota, Multi-level Restricted Maximum Likelihood Covariance Estimation and Kriging for Large Non-gridded Spatial Datasets, Spatial Statistics, 査読有, Vol. 18, 2016, pp. 105–124 DOI:10.1016/j.spasta.2015.10.006

④ H. Ibeid, R. Yokota, D. Keyes, A Performance Model for the Communication in Fast Multipole Methods on HPC Platforms, International Journal of High Performance Computing Applications, 査読有, Vol. 30, No. 4, 2016, pp. 423–437 DOI:10.1177/1094342016634819

⑤ **横田理央，階層構造を有する低ランク行列近似手法の数値計算とデータ構造，シミュレーション，日本シミュレーション学会，査読有，Vol.35，No.3，2016，pp．147-153**

⑥ **横田理央，FMM と H^2(HSS) 行列のトレードオフについて，計算工学，査読有，Vol.21，No.4，2016，pp．3498-3501**

〔学会発表〕（計22件）

① I. Yamazaki, A. Abdelfattah, A. Ida, S. Ohshima, S. Tomov, R. Yokota, J. Dongarra. Analyzing Performance of BiCGStab with Hierarchical Matrix on GPU clusters, 32nd IEEE International Parallel & Distributed Processing Symposium, 査読有, Vancouver, Canada, May. (2018).

② S. Ohshima, I. Yamazaki, A. Ida, R. Yokota. Optimization of Hierarchical Matrix Computation on GPU, SC Asia, 査読有, Singapore, Mar. (2018).

③ H. Naganuma, R. Yokota. Accelerating Convolutional Neural Networks Using Low Precision Arithmetic, HPC Asia, 査読有, Tokyo, Japan, Jan. (2018).

④ **桑村祐二，大沢和樹，横田理央．自然勾配法の近似手法における学習パラメータの調整，情報処理学会全国大会，査読有，Mar．(2018)．**

⑤ **大友広幸, 大沢和樹, 横田理央. フィッシャー情報行列の階層的低ランク近似を用いた深層学習，情報処理学会 全国大会，査読有，Mar.(2018)．**

⑥ **大友広幸, 大沢和樹, 横田理央. フィッシャー情報行列のクロネッカー因子分解を用いた深層ニューラルネットワークの分散学習，第163回ハイパフォーマンスコンピューティング研究発表会，査読有，Mar．(2018)．**

⑦ K. Oosawa, R. Yokota, Evaluating the Compression Efficiency of the Filters in Convolutional Neural Networks, The 26th International Conference on Artificial Neural Networks, 査読有, Sardinia, Italy, 11-14 Sep. (2017).

⑧ M. AbdulJabbar, M. Al Farhan, R. Yokota, D. Keyes, Performance Evaluation of Computation and Communication Kernels of the Fast Multipole Method on Intel Manycore Architecture, 23rd EUROPAR, 査読有, Galicia, Spain, 29 Aug. - 1 Sept. (2017).

⑨ K. Oosawa, A. Sekiya, H. Naganuma, R. Yokota, Accelerating Matrix Multiplication in Deep Learning by Using Low-Rank Approximation, The 2017 International Conference on High Performance Computing & Simulation, 査読有, Genoa, Italy, 17-21 Jul. (2017).

⑩ M. AbdulJabbar, G. Markomanolis, H. Ibeid, R. Yokota, D. Keyes, Communication Reducing Algorithms for Distributed Heirarchical N-Body Methods, ISC High Performance, Lecture Notes in Computer Science, 査読有, Vol. 10266, pp. 79–96, Frankfurt, Germany, 18-22, Jun. (2017).

⑪ **長沼大樹, 横田理央. 畳み込みニューラル ネットワークにおける低精度演算を用いた高速化の検証，査読有，GTC Japan，Dec.(2017)．**

⑫ **大沢和樹, 関谷翠, 長沼大樹, 横田理央. 低ランクテンソル分解を用いた畳み込みニューラルネットワークの高速化，パターン認識･メディア理解研究会，査読有，Oct.(2017)．**

⑬ **長沼大樹, 関谷翠, 大沢和樹, 大友広幸, 桑村裕二, 横田理央. 深層学習における低精度演算を用いた高速化及びアクセラレーターの性能評価，パターン認識・メディア理解研究会，査読有，Oct.(2017)．**

⑭ 長沼大樹, 大沢和樹, 関谷翠, 横田理央. 深層学習における半精度演算を用いた圧縮モデルの高速化, 日本応用数理学会年会, 査読有, Sep.(2017).
⑮ 大島 聡史, 山崎 市太郎, 伊田 明弘, 横田理央. GPU クラスタ上における階層型行列計算の最適化, SWOPP, 査読有, Jul. (2017).
⑯ 大沢和樹, 関谷翠, 長沼大樹, 横田理央. 畳み込みニューラルネットワークの低ランク近似を用いた高速化, 第 22 回計算工学講演会, 計算工学講演会論文集, 査読有, Vol.22, May.(2017).
⑰ 本山義史, 遠藤敏夫, 松岡聡, 横田理央, 福田圭祐, 佐藤育郎. 低ランク近似行列による CNN における畳み込み演算の最適化, 第 158 回ハイパフォーマンスコンピューティング研究発表会, 2017-HPC-158 No.25, Mar.(2017).
⑱ 関谷翠, 大沢和樹, 長沼大樹, 横田理央. 低ランク近似を用いた深層学習の行列積の高速化, 第 158 回ハイパフォーマンスコンピューティング研究発表会, Mar.(2017).
⑲ R. Yokota, Hierarchical Low-Rank Approximations at Extreme Scale, ISC High Performance, 招待講演 Frankfurt, Germany, 18 - 22, June, (2017).
⑳ R. Yokota, Compute-Memory Tradeoff in Hierarchical Low-Rank Approximation Methods, SIAM Conference on Computational Science and Engineering, 招待講演 Atlanta, USA, 27 February - 3 March, (2017).

６．研究組織

(1)研究代表者

横田　理央　（YOKOTA, Rio）
東京工業大学・学術国際情報センター・准教授
研究者番号：20760573

(2)研究分担者

なし

(3)連携研究者

なし

(4)研究協力者

シャオイエ　リー (Xiaoye S. Li)
Lawrence Berkeley National Laboratory

デイビッド　キース　(David E. Keyes)
King Abdullah University of Science and Technology