修士学位論文要約（令和 4 年 3 月）

大規模行列に対するコレスキー分解のための スケーラブルアーキテクチャと FPGA 実装に関する研究

菅野 航平

指導教員：張山 昌論，学位論文指導教員：Hasitha Muthumala Waidyasooriya

A Study on Scalable Architecture and FPGA Implementation for Cholesky Decomposition of Large Matrix

Kouhei KANNO

Supervisor: Masanori HARIYAMA，Research Advisor: Hasitha Muthumala WAIDYASOORIYA

Cholesky decomposition is a matrix decomposition with a wide range of applications, such as solving simultaneous linear equations. Conventional speed-up methods using FPGAs mainly use internal memory. Therefore, the size of the matrix that can be supported is greatly limited. In this paper, we propose a scalable architecture that supports large-scale matrices by using a systolic array, and formulate the required external memory bandwidth and the number of logic units. The evaluation results showed a 6.23 times speedup over the CPU in a 32768×32768 matrix.

1. はじめに

コレスキー分解は，連立一次方程式の求解に代表される幅広い活用範囲を持つ行列分解である．FPGA を使用した従来の高速化においては，処理速度との兼ね合いから主な記憶を内部メモリに頼ることで，対応可能な行列サイズが強く制限されるという問題が存在した．本研究においては，効率的なデータ使用と並列処理によって大規模行列にも対応可能なスケーラブルアーキテクチャを提案し，その実装と評価，設計のための定式化を行った．

2. 提案アーキテクチャ

コレスキー分解は行列 A を下三角行列 L とその転置の積に分解する操作である．この計算には平方根を求める工程があり，処理速度と精度の観点から，実用上は代替として，対角行列 D を加えた以下の式 (1) に示す修正コレスキー分解が使用される．

$$A = LDL^T \tag{1}$$

ここで，L^T は L の転置行列である．修正コレスキー分解の計算は図 1 のフローチャートに示すとおり 3 重のループ構造をしている．これは最も内側のループが要素 1 つ分の計算を表し，外側に向かって列方向，行方向の繰り返しを表している．

この手順を反映した提案アーキテクチャを図 2 に示す．アーキテクチャにはシストリックアレイを取り入れた．データが横方向の PE 1 行を伝搬する中で要素 1 つが求められ，これが列数分パイプライン処理として繰り返されている．PE の縦方向の接続は次の行を求める際に再使用可能なデータを送っている．これにより外部メモリへのアクセスを削減し，アクセス時間による速度低下を回避している．

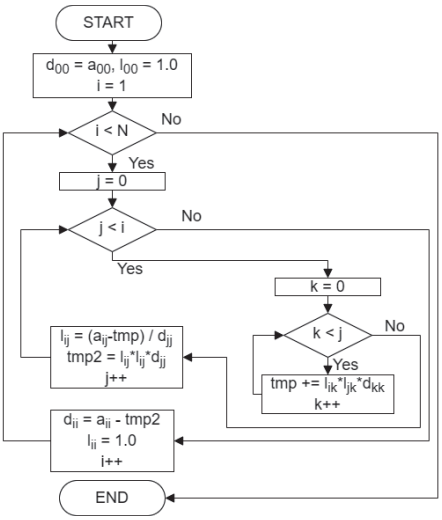

図 1. 修正コレスキー分解のフローチャート

3. 最適設計

提案アーキテクチャにおける処理時間は，PE 構成が R 行 C 列の場合，入力となる $N \times N$ の大規模行列に対しておよそ $\frac{N^3}{6RC}$ となる．

図 3 に LAPACK 用いた CPU による処理時間と，

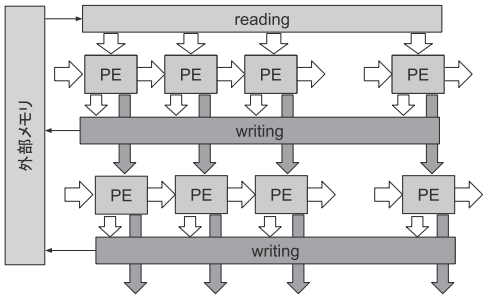

図 2. 提案アーキテクチャ

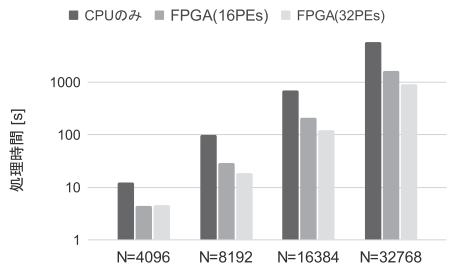

図 3. $N \times N$ 行列に対する CPU と FPGA それぞれの処理時間

Intel FPGA SDK for OpenCL を用いて提案アーキテクチャを実装した FPGA を用いた場合における処理時間を比較したグラフを示す．黒が CPU，灰色と薄い灰色がそれぞれ PE 数 16，32 の処理時間を表している．入力行列のサイズが大きくなるほど，処理のオーバーヘッド等による影響の割合が小さくなる分，CPU に対する FPGA による処理時間のスピードアップが増加し，PE 数 16 では最大 3.49 倍，PE 数 32 では最大 6.23 倍となった．処理性能と対応サイズについて，従来との比較を表 1 に示す．

また実装に際して制約条件となる要素としては，(1) 外部メモリのバンド幅，(2) リソース使用量の 2 つがある．メモリバンド幅に関して，要求される最低限のバンド幅 B_w[Byte/s] を，処理全体で外部メモリに読み書きされるデータの総量を処理時間で割ることで求めた．以下にその式を示す．

$$B_w = C \times ds \times F \tag{2}$$

ここで C は PE の列数，ds は各データのサイズ [Byte]，F は FPGA の動作周波数 [Hz] である．

またリソースの使用量に関しては，最も使用量が多かったロジックユニットに着目した．各 PE 構成でのロジック使用量をグラフにプロットすることで，PE 構成との間に以下の式で表されるような関係を持つことがわかった．

$$f(R, C) = aRC + bR + c \tag{3}$$

ここで a, b, c は実装環境によって決定される定数である．式 (2) と (3) から，実装環境の最大メモリバンド幅 B_{max} と使用可能なロジックユニット数 L_{max} を用いて，設計の際に考慮すべき制約条件を表す式は以下のように表せる．

$$B_{max} > C \times ds \times F \tag{4}$$

$$L_{max} > aRC + bR + c \tag{5}$$

したがって提案アーキテクチャにおける最適な設計は，式 (4)，(5) によって求められる領域のうち，総 PE 数を表す $R \times C$ が最大となる格子点 (R, C) を求めることに相当する．

表 1. 従来との比較

手法	使用 FPGA	最大行列サイズ	入出力格納メモリ	処理性能 [GFLOPS]
文献 [1]	Intel Stratix II EP2S130F1508C3	25×25	内部メモリ	0.21
文献 [2]	Xilinx Virtex-5 XC5VSX95T-2	256×256	内部メモリ	510
文献 [3]	Xilinx Virtex-6 XC6VSX475T	1024×1024	内部メモリ	21.0
本研究	Intel Stratix 10 GX 2800	32768×32768	外部メモリ	38.3

4. まとめ

大規模行列における修正コレスキー分解のためのスケーラビリティを備えたアーキテクチャを提案し，FPGA 上での実装と評価を行った．また最適設計のための制約式として，要求メモリバンド幅とリソース使用量について定式化を行った．

文献

1) J. Luo, Q. Huang, S. Chang, X. Song, and Y. Shang, “High throughput cholesky decomposition based on FPGA,” in 2013 6th International Congress on Image and Signal Processing (CISP), vol. 3. IEEE, 2013, pp. 1649–1653.
2) D. Yang, G. D. Peterson, and H. Li, “High performance reconfigurable computing for cholesky decomposition,” in Proceedings of the Sym- posium on Application Accelerators in High Performance Computing (UIUC ’09), 2009.
3) D. Yang, J. Sun, J. Lee, G. Liang, D. D. Jenkins, G. D. Peterson, and H. Li, “Performance comparison of cholesky decomposition on gpus and fpgas,” in Symposium on Application Accelerators in High Performance Computing, 2010, pp. 1–3.