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

令和　６ 年　 ６ 月 ２６ 日現在

機関番号：１２６０８
研究種目：若手研究
研究期間：2021～2023
課題番号：２１Ｋ１７７２０

研究課題名（和文）大規模疎行列処理のためのインストレージアクセラレータの創出

研究課題名（英文）In-Storage Accelerator Architectures for Large-Scale Sparse Matrix Processing

研究代表者
ＣＨＵ　ＴｈｉｅｍＶａｎ（Chu, Thiem Van）

東京工業大学・科学技術創成研究院・助教

研究者番号：８０８３８２３５

交付決定額（研究期間全体）：（直接経費）　　3,600,000円

研究成果の概要（和文）：本研究では，大規模疎行列処理を高速かつ高効率に行うためのインストレージアクセラレータアーキテクチャを含む包括的な疎行列処理アーキテクチャの開発を目指している．その第一ステップとして，疎行列と疎行列の積という基本演算に焦点を当て，高速かつ高効率なアーキテクチャの研究を進め，FPGA（Field-Programmable Gate Array）によるハードウェアプロトタイプの実装および評価を行った．主な成果として，VLSI（Very-Large-Scale Integration）とシステム分野の国際会議ASP-DAC 2024での論文発表，3件の招待講演，および2件の受賞が挙げられる．

研究成果の学術的意義や社会的意義
本研究の成果は，疎行列処理の高速化と高効率化を実現することで，ビッグデータ解析，機械学習，科学計算の複雑なシミュレーションなど多くのアプリケーションにおいて重要な計算カーネルの性能向上および計算資源の節約に寄与する．本研究によって提案された手法は，学術的にはアーキテクチャおよびハードウェア設計に新たな知見を提供し，社会的にはデータ分析や人工知能などの発展に大きな影響を与えると期待できる．

研究成果の概要（英文）：This study aims to develop a comprehensive sparse matrix processing architecture, including an in-storage accelerator architecture, to perform large-scale sparse matrix processing with high performance and efficiency. As the first step, the research focuses on the basic operation of sparse-sparse matrix multiplication, advancing the study of a high-performance and efficient architecture, and implementing and evaluating a hardware prototype using FPGA (Field-Programmable Gate Array). Major achievements include the presentation of a paper at the Asia and South Pacific Design Automation Conference (ASP-DAC'24), three invited talks, and two awards.

研究分野：計算機システム

キーワード：疎行列処理　疎行列疎行列積　データフロー　アーキテクチャ　FPGA

１．研究開始当初の背景

膨大なデータを処理するグラフ処理や機械学習等の多くのアルゴリズムは，大規模疎行列に対する処理として定式化できる．疎行列処理の基本演算として，疎行列と疎行列の積（SpMSpM: sparse-sparse matrix multiplication），疎行列と密行列の積（SpMM: sparse-dense matrix multiplication），疎行列と密ベクトルの積（SpMV: sparse matrix-vector multiplication），疎行列転置（SpMT: sparse matrix transposition）が挙げられる．

疎行列処理は密行列処理のように並列性を持つが，その並列性が不規則であるため，ランダムなメモリアクセスや低い時間的・空間的局所性などの問題を引き起こし，汎用計算用のメニーコア CPU 及び GPU での処理を苦手とする．ムーアの法則の終焉により，疎行列処理の高速化は更に困難となっていて，革新的な処理アーキテクチャが求められている．

近年，疎行列処理を高速化するためのいくつかの領域特化型アクセラレータが開発されている[Dorrance+ FPGA'14，Fowers+ FCCM'14，Pal+ HPCA'18，Sadi+ MICRO'19，Zhang+ HPCA'20，Srivastava+ MICRO'20]．これらのアクセラレータは，Intel 社の Sparse BLAS や Nvidia 社の cuSPARSE などのメニーコア CPU および GPU に最適化された実装と比較して，数倍から十数倍までの優れた性能を達成している．しかし，以下の 2 つの解決すべき課題が残っている．

- 第一に，Society 5.0 時代のビッグデータ処理アプリケーションで求められる TB スケールの大規模疎行列処理に対応する必要がある．既存のアクセラレータは，疎行列処理の致命的な問題であるメモリアクセスボトルネックを緩和するために，HBM のような高速メモリを使用し，全てのデータがその高速メモリに格納できることを前提としている．しかし，HBM の容量は多くても数 GB であり，対応できる行列サイズは限られている．例えば，最先端のアクセラレータ[Zhang+ HPCA'20，Srivastava+ MICRO'20]が対応できる最大入力疎行列サイズは数百 MB である．TB スケールの大規模疎行列の処理では，低スループットおよび高レイテンシのストレージアクセスが頻繁に発生し，処理性能が大幅に低下する．
- 第二に，実用性の高いアクセラレータを実現するためには，複数の処理をサポートする必要がある．既存のアクセラレータは SpMV や SpMSpM のいずれかの処理にしか着目していなかった．しかし，実際のアプリケーションでは，現実世界の複雑な疎データを取り扱うために，複数の処理を同時に求める場合が多い．

２．研究の目的

前述の背景を踏まえ，本研究では，大規模疎行列処理を高速化・高効率化するために，インストレージアクセラレータアーキテクチャを含む，包括的な疎行列処理アーキテクチャの開発を目指している．その第一ステップとして，SpMSpM に焦点を当て，高速かつ高効率なアーキテクチャの研究を進めており，FPGA によるハードウェアプロトタイプの実装および評価を行った．

３．研究の方法

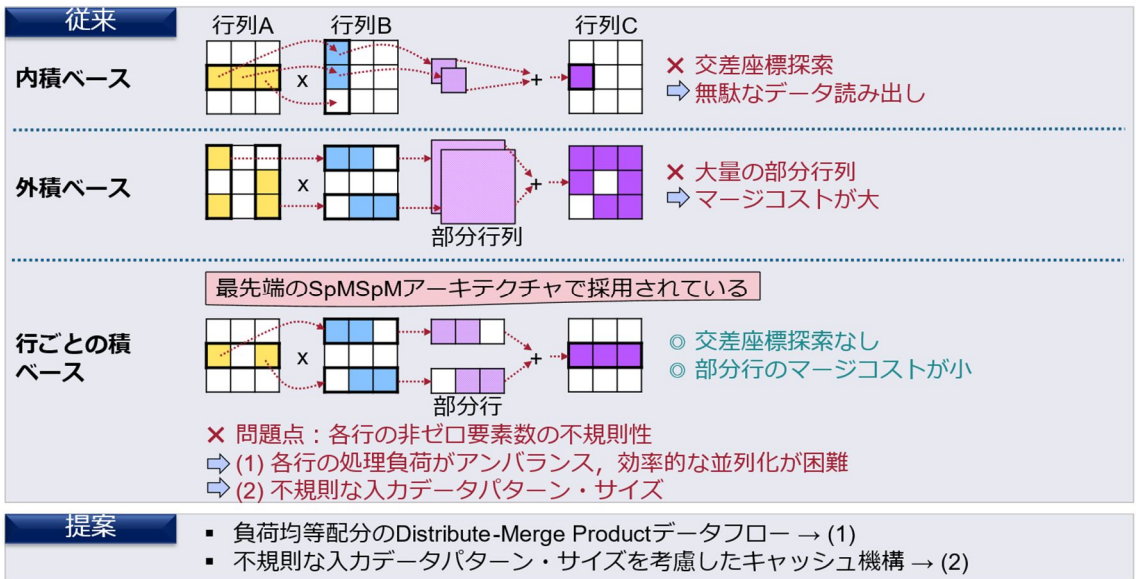


図 1. 従来の SpMSpM 処理データフローと本研究の提案の概要.

SpMSpM の基本的な処理データフローとして，内積（Inner Product），外積（Outer Product），行／列ごとの積（Row/Column-wise Product．Gustavson's Product とも呼ばれている）が挙げられる（図 1）．内積ベースのデータフローでは，入力行列を繰り返し何度も読み込む必要があ

り，インデックスがマッチしている要素同士のみの乗算を行うため，メモリアクセス量が膨大になるわりに，多くの無効なアクセスが発生する．この問題は，入力行列の密度が低いほど深刻になる．外積ベースのデータフローでは，内積ベースのデータフローにおける無駄なデータ読み出しの問題が発生しないが，途中結果の部分行列の大きなマージコストが課題となる．そこで，近年提案された SpMSpM アーキテクチャは行／列ごとの積ベースのデータフローを採用している（図 1 での説明は行ごとの積ベースのデータフローになっている．列ごとの積ベースのデータフローも同様である）．しかし，このデータフローでも，各行／列の非ゼロ要素数の不規則性に起因して，各行／列の処理負荷がアンバランスになり，効率的な並列化が困難である．また，不規則な入力データパターンやサイズの問題もある．

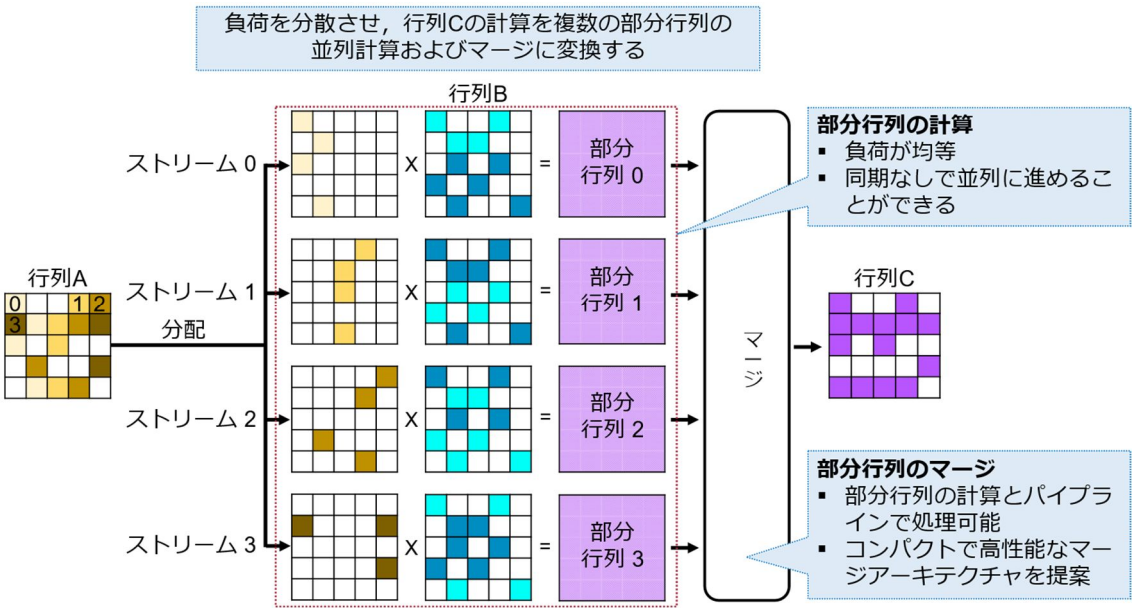
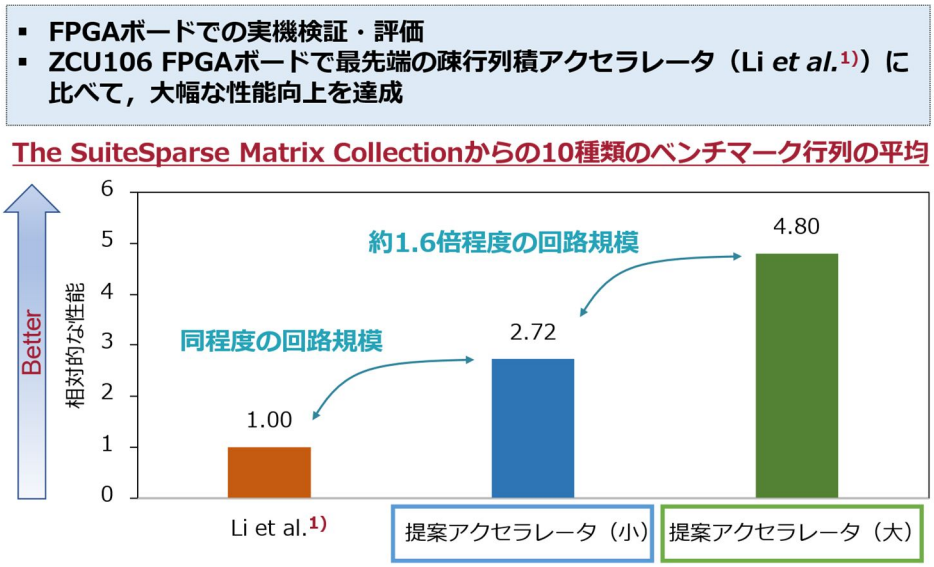


図 2. 提案の SpMSpM 処理データフロー（Distribute-Merge Product）の概要.

本研究は，既存の SpMSpM 処理データフローの問題の解決を目指し，処理負荷を均等に分配する Distribute-Merge Product データフロー（図 2）とその特徴にあった入力データキャッシュ機構を提案した．提案の Distribute-Merge Product データフローでは，SpMSpM 処理は複数（図 2 では 4 つ）の部分行列の計算とマージのパイプライン処理で行われ，各部分行列の計算負荷が均等で，同期なしで並列に進めることができるため，高速で効率的である．このデータフローに基づくアーキテクチャを設計し，FPGA ボードでの実機検証・評価を行った．



[1] Li et al., “An Efficient Gustavson-based Sparse Matrix-matrix Multiplication Accelerator on Embedded FPGAs,” IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 2023.

図 3. 提案 SpMSpM アクセラレータの性能評価結果の概要.

4．研究成果

ZCU106 FPGA ボードで実施した，提案の SpMSpM アーキテクチャと最先端の SpMSpM アーキテク

チャ [Li *et al.*, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 2023] との比較実験において，同程度の回路規模で 2.72 倍，約 1.6 倍の回路規模で 4.80 倍の性能向上を確認した（図 3）．これにより，提案した SpMSpM 処理データフローとアーキテクチャの有効性が示された．この研究成果は，VLSI とシステム分野の国際会議 ASP-DAC'24 (Asia and South Pacific Design Automation Conference, 2024) で発表され，さらに 3 件の招待講演及び 2 件の受賞を果たしている．SpMSpM は，ビッグデータ解析，機械学習，科学計算の複雑なシミュレーションなど，多岐にわたるアプリケーションで重要な計算カーネルとして位置づけられており，本研究はこれらの分野への大きな貢献が期待される．

今後の展望としては，提案した SpMSpM 処理データフロー・アーキテクチャに基づき，SpMM や SpMV，SpMT などの他の疎行列の基本演算への拡張を図る．これにより，より包括的な疎行列処理アーキテクチャの確立を目指す．また，疎データ処理を基盤とした融合・分野横断的な研究の推進も計画している．

５．主な発表論文等

〔雑誌論文〕　計4件（うち査読付論文　4件／うち国際共著　1件／うちオープンアクセス　0件）

１．著者名 Yuta Nagahara, Jiale Yan, Kazushi Kawamura, Masato Motomura, Thiem Van Chu	４．巻 1
２．論文標題 Sparse-Sparse Matrix Multiplication Accelerator on FPGA featuring Distribute-Merge Product Dataflow	５．発行年 2024年
３．雑誌名 Asia and South Pacific Design Automation Conference (ASP-DAC)	６．最初と最後の頁 785-791
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/ASP-DAC58780.2024.10473865	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Yuta Nagahara, Jiale Yan, Kazushi Kawamura, Masato Motomura, Thiem Van Chu	４．巻 1
２．論文標題 Efficient COO to CSR Conversion for Accelerating Sparse Matrix Processing on FPGA	５．発行年 2024年
３．雑誌名 International Conference on Consumer Electronics (ICCE)	６．最初と最後の頁 1-2
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/ICCE59016.2024.10444348	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Philippos Papaphilippou, Thiem Van Chu	４．巻 73
２．論文標題 Efficient Deadlock Avoidance for 2-D Mesh NoCs That Use OQ or VOQ Routers	５．発行年 2024年
３．雑誌名 IEEE Transactions on Computers	６．最初と最後の頁 1414-1426
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/TC.2024.3365954	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 該当する

１．著者名 Thiem Van Chu, Ryuichi Kitajima, Kazushi Kawamura, Jaehoon Yu, Masato Motomura	４．巻 1
２．論文標題 A High-Performance and Flexible FPGA Inference Accelerator for Decision Forests Based on Prior Feature Space Partitioning	５．発行年 2021年
３．雑誌名 International Conference on Field-Programmable Technology (FPT)	６．最初と最後の頁 1-10
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/ICFPT52863.2021.9609699	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計4件（うち招待講演　3件／うち国際学会　1件）

1．発表者名 永原 雄大, Jiale Yan, 川村 一志, 本村 真人, Thiem Van Chu
2．発表標題 効率的な負荷分配による高並列疎行列積アーキテクチャの研究
3．学会等名 第23回情報科学技術フォーラム（FIT2024） トップコンファレンスセッション（招待講演）
4．発表年 2024年

1．発表者名 Thiem Van Chu
2．発表標題 Sparse-Sparse Matrix Multiplication Accelerator on FPGA featuring Distribute-Merge Product Dataflow
3．学会等名 Workshop on FPGA Technologies for Adaptive Computing (FTAC), International Conference on Supercomputing (ICS)（招待講演）（国際学会）
4．発表年 2024年

1．発表者名 永原 雄大, Jiale Yan, 川村 一志, 本村 真人, Thiem Van Chu
2．発表標題 ［記念講演］分散マージ乗算手法に基づく疎行列疎行列積アクセラレータ
3．学会等名 電子情報通信学会 VLSI設計技術研究会（VLD）（招待講演）
4．発表年 2024年

1．発表者名 永原雄大, 安藤洸太, 川村一志, 劉載勲, 本村真人, Thiem Van Chu
2．発表標題 外部メモリアクセス抑制による高効率疎行列積アクセラレータの研究
3．学会等名 電子情報通信学 コンピュータシステム研究会（CPSY）, 並列／分散／協調システムとディペンダブルコンピューティングおよび一般（SWoPP）
4．発表年 2022年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関