# 先端国際共同研究推進事業
# 2023 年度採択

# 次世代のための ASPIRE
# 半導体分野

# 2023 年～2024 年度
# 年次報告書・公開版

| | |
|---|---|
| 研究課題名 | 次世代の高効率計算基盤を実現する適応型データ圧縮ハードウェアの探求 |
| 日本側研究代表者 | 上野 知洋　理化学研究所　研究員 |
| 相手側研究代表者 | ・Franck Cappello, Senior Computer Scientist/R&D Leader, Argonne National Laboratory<br>・Jason Anderson, Professor, University of Toronto |
| 研究期間 | 2024 年 2 月 1 日～2027 年 3 月 31 日 |

1.　研究成果の概要

①　研究構想にかかる成果

＜実施したこと＞

・　適応型データ圧縮ハードウェアプラットフォームの構築と評価
・　Chisel による回路実装環境構築および共同研究・開発環境の整備
・　データ圧縮評価用データセットの収集
・　適応型・汎用データ圧縮手法に関する情報収集
・　CGRA（Coarse-Grained Reconfigurable Array、粗粒度再構成可能アーキテクチャ）フロントエンドコンパイラの開発
・　CGRA 向けプログラマブルバッファに関する研究
・　CGRA におけるレイテンシ不均衡問題に関する研究

＜得られた成果＞

適応型データ圧縮ハードウェアにおいて、可逆データ圧縮の際に発生するデータ毎の圧縮率の変動を吸収し、利用可能な出力帯域幅を使い切るためのプラットフォームシステムについて、Chisel を用いた回路の実装と設計空間探索のための量子化パラメータによる評価を行った。我々が導入した量子化パラメータを変化させると、データ圧縮アルゴリズムとは独立して圧縮率と必要計算機リソースとのトレードオフを制御できることを示した。これにより、適用環境に応じた適応型データ圧縮回路の設計空間探索が可能となり、通信帯域と計算機資源との資源分配を考慮したシステムの構築が容易となる。この成果について、国内研究会にて発表し、優秀講演賞を受賞した。また、年度内に国際会議への論文投稿を予定している。

加えて、データ圧縮性能評価のためのベンチマークデータの収集を行った。アルゴンヌ国立研究所のグループが評価に用いている公開データセットに加えて、AI 学習用のオープンソースデータセット、医療用画像や車載動画といった特定の領域向けのデータセット、参加研究機関が所有するクローズドなデータセット等、様々な領域のデータセットについて探索・収集を行った。データセット収集は来年度以降も継続して実施し、データ圧縮アルゴリズムとデータ自体の特徴との関係性を探る研究に活用する。

適応型データ圧縮の実現に向けては、最新のアルゴリズムに関する情報収集を行った。特に汎用なアルゴリズムとして、データ圧縮に用いられるアトミックな操作を組み合わせて動的に有効な圧縮手法を構築する研究や、学習済み AI による予測に基づく手法等、ハードウェア化を考慮した有望な手法について情報を集めた。現在、調査結果をまとめたサーベイ論文を執筆中であり、来年度に投稿予定である。

CGRA に関する研究では、トロント大学が開発中の CGRA-ME 向けコンパイラの技術を利用し、我々が開発中の Riken CGRA のアーキテクチャに適したフロントエンドコンパイラを開発した。これにより、C または C++で記述されたソースコードから、データフローグラフを経由して CGRA の配置配線の設定ファイル出力までのフローが完成する。この研究は、トロント大学から招聘した修士課程の学生を中心に行われた。

CGRA のアーキテクチャに関する研究では、CGRA 上でデータを効率的に活用しメモリアクセス回数を減らすために、プログラマブルバッファの開発を行った。これは任意のロケーションからデータを出力できる FIFO バッファであり、例としてステンシル計算等の数値計算時に、データの再利用によるメモリアクセス削減効果が期待できる。この研究は、熊本大学からトロント大学に渡航した修士課程の学生を中心に行われた。コンパイラおよびプログラマブルバッファに関する研究は来年度以降に論文としてまとめ、成果発表を行う。

最後に、Riken CGRA アーキテクチャ上でのレイテンシの不均衡に関する研究を行った。これは演算エレメントの複数の入力について、それぞれのデータパスにおけるレイテンシの違いから、待ち合わせのために大きな FIFO が必要となる問題に対処するための研究である。明治大学を中心に、CGRA 上でのレイテンシ不均衡問題を

モデル化し、それに基づく適切な FIFO サイズの調査を実施した。予備調査の結果は国内会議で発表済であり、より広範な研究結果を国際ワークショップに投稿中である。

② 国際頭脳循環の促進にかかる成果

＜実施したこと＞

- 国際会議等を利用した日本、米国、カナダの研究者の顔合わせ
- 学生 2 名のトロント大学への長期派遣
- トロント大学の院生 1 名の 2 か月間の招聘
- 研究計画および渡航計画打ち合わせのための複数回のアメリカ・カナダへの出張
- 海外機関からの参加者を含むクローズドな研究ミーティングの開催

＜得られた成果＞

サンフランシスコで開催された国際会議 IPDPS2024 において、Dr. Cappello、Prof. Anderson を含むプロジェクト参加者の顔合わせのためのランチミーティングを実施した。また、LBNL （ローレンス・バークレー国立研究所）において開催された ARCHIDE ワークショップ等において、既存の交流の枠組みを活用したコネクションを拡大する取り組みを行った。

また、国内の修士課程学生 2 名がトロント大学へ渡航した。1 名は 2024 年 6 月下旬から 12 月下旬までの半年間の渡航の中で、CGRA アーキテクチャにおけるプログラマブルバッファの共同研究を行った。もう 1 名は 2025 年 1 月にトロント大学に渡航し、1 年間 CGRA におけるレイテンシ不均衡問題について共同研究を行う予定である。他方、トロント大学の修士課程学生を 2024 年 6 月から 2 か月間、理化学研究所に招聘し、CGRA フロントエンドコンパイラの開発およびメモリアクセス時のアドレス演算の効率化について共同研究を実施した。

さらに、研究代表者がアルゴンヌ国立研究所とトロント大学に複数回出張し、渡航・招聘の計画、研究内容のすり合わせ、研究計画の立案等を目的とする打ち合わせを行った。加えて、適応型データ圧縮ハードウェアの先行研究調査報告、ベンチマークデータセットの情報共有、圧縮アルゴリズムの探求、ハードウェア設計の具体化などを目的として、2024 年 12 月に熊本大学にて研究参加者によるクローズドなミーティングを実施した。2025 年度にミラノにて予定されているワークショップ CGRA4HPC については本 ASPIRE プロジェクトの協賛として開催し、これまでの研究成果の報告や関連研究の講演などを予定している。

2. 研究実施体制

| 研究テーマ | 中心となる研究者氏名 | 所属機関・部署・役職名 |
|---|---|---|
| 研究テーマ1:<br>適応型データ圧縮HWの実現 | 上野 知洋<br>Franck Cappello<br><br>吉井 一友 | 理化学研究所・計算科学研究センター・研究員<br>Argonne National Laboratory・MCS・Senior Computer Scientist/R&D Leader<br>Argonne National Laboratory・MCS・Principal Specialist |
| 研究テーマ2:<br>CGRAアーキテクチャと自動再構成 | Boma Adhi<br>Jason Anderson | 理化学研究所・計算科学研究センター・研究員<br>University of Toronto・ECE・Professor |

3. 代表的な業績（原著論文、プレスリリース、表彰など）

＜原著論文＞

[1]. 北爪 開人、上野 知洋、吉井 一友、木山 真人、藤田 典久、小林 諒平、佐野 健太郎、朴 泰祐、“適応型帯域圧縮ハードウェアプラットフォームの Chisel 実装と評価” 信学技報,、vol.124、no.188、RECONF2024-50、pp.41-46、2024 年
[2]. Franck Cappello, Mario Acosta, Emmanuel Agullo, Hartwig Anzt, Jon Calhoun, Sheng Di, Luc Giraud, Thomas Grützmacher, Sian Jin, Kentaro Sano, Kento Sato, Amarjit Singh, Dingwen Tao, Jiannan Tian, Tomohiro Ueno, Robert Underwood, Frédéric Vivien, Xavier Yepes, Yoshii Kazutomo, Boyuan Zhang, Multifacets of lossy compression for scientific data in the Joint-Laboratory of Extreme Scale Computing, Future Generation Computer Systems, Volume 163, 2025
[3]. 岡田拓実、長名保範、飯田全広、Boma Adhi、佐野健太郎、Omar Ragheb、Jason Anderson、“RIKEN CGRA 向けステンシル計算用プログラマブルバッファの導入と評価”電子情報通信学会技術研究報告、RECONF2024-125、2025 年

＜表彰＞

[1]. 北爪 開人、“適応型帯域圧縮ハードウェアプラットフォームの Chisel 実装と評価” 電子情報通信学会リコンフィギュラブルシステム研究会優秀講演賞、2024 年 11 月
[2]. Tomohiro Ueno、“Promotion of International Collaborative Research Project to Explore Adaptive Bandwidth Compression Hardware” RIKEN Ohbu Award （RIKEN Research and Technology Incentive Award）、2025 年 3 月