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 １４ 日現在

機関番号：１２１０２
研究種目：基盤研究(B)（一般）
研究期間：2020～2023
課題番号：２０Ｈ０４１５２

研究課題名（和文）大局的エントロピー予測によるデータ圧縮の最適化技法の開発

研究課題名（英文）Development of optimization technique for data compression by predicting global data entropy

研究代表者
山際　伸一（Yamagiwa, Shinichi）

筑波大学・システム情報系・准教授

研究者番号：１０５７４７２５

交付決定額（研究期間全体）：（直接経費）　　8,900,000円

研究成果の概要（和文）：無限遠に続くデータストリームに対し、その局所のエントロピーを圧縮データを管理する変換テーブルの利用率を元に予測できる原理を発見し、その知見をもとにデータ単位を最少で1ビットにロスレス圧縮できる手法Adaptive Stream-based Entropy Codingを開発できた。このデータエントロピーはシャノンの平均情報量に追従し、大域的なデータエントロピーが表され、局所で最大限の圧縮が行えることを示した。リアルタイムにデータストリームを圧縮解凍できるハードウェア指向のアルゴリズムを開発した。この圧縮器・解凍器はハードウェアにコンパクトに実装できるだけなく高速に動作することを実証した。

研究成果の学術的意義や社会的意義
従来からのデータストリームを扱う圧縮器の内部動作を元にして、局所のエントロピーの変化を監視することで、大域的なエントロピーを求められる原理を解明した。圧縮器に、未来に入力されるデータの傾向を予測して、圧縮器での符号の決定ができれば、局所的なデータの出現傾向に動的に従い、最適な圧縮率を得られるのではないか？という学術的な疑問に対し、その方法を解明した。ハードウェア実装できるアルゴリズムを開発し、ロスレス圧縮方式を開発した。IoTやAIにおける、通信データ量の増大やストレージの小型化といった今後発展していくビッグデータ時代の産業に応用できる。

研究成果の概要（英文）：This project found a principle that predicts realtime entropy of a continuous data stream based on the occupation ratio of the lookup table in the compressor. The table manages original data to generate the compressed ones. Based on the finding, this project developed a lossless data compressed that shrinks a data unit to a single bit at least. We developed a method called Adaptive Stream-based Entropy Coding. The data entropy predicted by our mechanism follows Shannon's entropy and represents the global data entropy. The experimental results show that ASE Coding invokes the most effective realtime compression. We also developed a hardware-oriented algorithm that compresses/decompresses any continuous data stream in real time. We also demonstrated that the compressor/decompressor is implemented compactly in hardware and works in high speed.

研究分野：計算機アーキテクチャ

キーワード：データ圧縮

様　式　Ｃ－１９、Ｆ－１９－１（共通）

１．研究開始当初の背景

復号後に元のデータに完全に戻すロスレスデータ圧縮技術は 1950 年代のシャノンの平均情報量による符号化技術から始まり、エントロピー符号化による算術符号、そして、ハフマン符号化といった現代でも使われる重要な方式が生まれた。そして、LZW に代表される頻出するデータをテーブルに登録し、そのテーブルのインデックスで、より小さな符号を割り当てるテーブル変換方式が発明された。これらの従来法は、データの先頭から順に、決められたデータ幅（シンボル幅と呼ぶ）で頻出パターンの検索を行い、再度現れるシンボルパターンを小さい符号に割り当てていく。しかし、この符号化過程では時系列で圧縮器が読み出す順番でのデータの傾向（すなわち、エントロピー）に従うため、全体のエントロピーを把握しているわけではない。ある圧縮後のデータの縮小率（圧縮率）は、符号化に用いたデータの変換パターンが十分に予測できていれば、より良好な圧縮率が期待できたかもしれない可能性がある。

２．研究の目的

しかしながら、従来からの圧縮器の動作である局所のエントロピーの変化から圧縮器へ未来に入力されるデータの傾向を予測して、圧縮器の符号化動作を動的に変更できれば、最適な圧縮率を得られるのではないか？という疑問がわく。具体的には、圧縮器での圧縮率やテーブルヒット率といった制御情報を未来の入力データの圧縮制御に利用する。すなわち、圧縮率に起因する制御条件を動的に変更し、データ全体のエントロピーを予測しながら最良の圧縮率を得るハード化可能で、無限遠のデータストリームを扱うことのできる圧縮方式を開発することを本研究の目的とした。

３．研究の方法

（1）研究課題

本研究は以下の 3 つの小課題に分けて実施した。
課題１：データストリームの局所エントロピーの数値化技法の開発
無限遠のデータストリームの局所エントロピーを圧縮の制御情報から数値化する技法を確立する。実験的な結果に数学的証明を与え、その制御情報でデータの複雑さを表現する技法を開発する。
課題２：データストリームの大域的エントロピーを予測の制御方式の開発
課題 1 での局所エントロピーを予測しながら圧縮率を最適化する基本原理をソフトウェア、ハードウェアで実装するための具体的な構成について、その技法を開発する。
課題３：ハードウェアへの実装
課題 2 の手法を適用できるハードウェア向けのアーキテクチャを開発し、FPGA で試作する。

（2）研究体制

本研究は以下の 3 つのチームに分け、協働して進めた。
・アーキテクチャ開発チーム
山際と和田は計算機システムの専門家であり、ハード向けアルゴリズムとシステムの開発、実験用ソフトウェア開発、研究全体の総括を担当した。
・圧縮アルゴリズム開発チーム
坂本はデータ圧縮理論の専門家であり、アーキテクチャ開発チームと連携して理論形成を担当した。
・圧縮率制御方式開発チーム
河原は機械学習分野の専門家であり、圧縮率の制御部分における制御方式について機械学習からアプローチを担当し、情報理論・アーキテクチャ開発両チームと連携した。

（3）研究計画

本研究は課題 1、2 のそれぞれを 1 年目と 2 年目に実施し、原理の解明とハード実装を繰り返しながら進行した。3 年目は手法のパラメタと圧縮性能の評価を実装と共に進めた。最終年度は実装に注力し、実社会への応用を探求した。

４．研究成果

本研究課題での研究成果は無限遠のデータストリームを扱うことができ、そのデータ単位を 1 ビットにまで圧縮可能な新方式 ASE Coding（文献①）を実現したことがあげられる。この圧縮方式はロスレスであり、圧縮されたデータは、元のデータに完全に戻すことができる。さらに、ハ

ードウェア実装を意識したアルゴリズムを開発することで、FPGA や LSI といったデジタル回路にコンパクトに実装することができる。
ASE Coding を開発するにあたり、従来技術である LCA-DLT（文献②）でのテーブル利用率と圧縮率との挙動を観察した。圧縮率はすなわち、データの複雑さであるエントロピーを表しているため、テーブルの利用率がデータの複雑さに追従し、それをデータストリームの局所で判定できることに気づき、大域的なデータの複雑さを局所のデータの挙動から判定する方法を発見した。その手法は、圧縮する際にオリジナルのデータから圧縮データへと変換するシンボル変換テーブルの占有数をkとすると、圧縮対象となるオリジナルデータに一致するテーブルインデクスの必要ビット数はエントロピー計算$e = ceil(\log_2 k)$によって計算することができる。

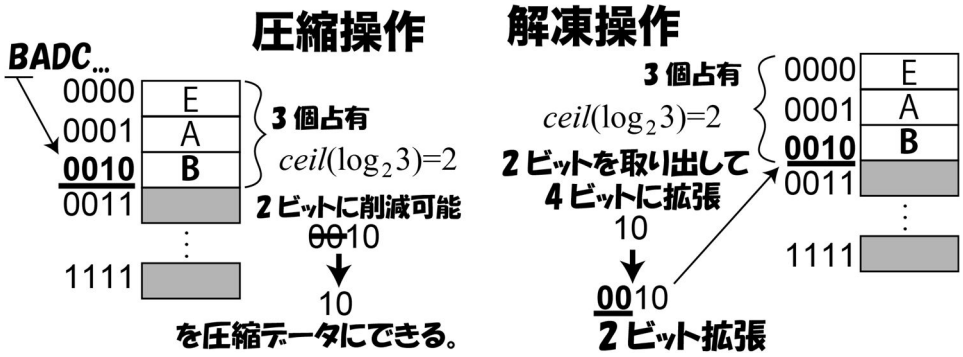

図 1　ASE Coding の圧縮・解凍の仕組み

エントロピー計算により、図 1、図 2 に示す ASE Coding の仕組みを開発した。図 1 では、圧縮器（左図）にデータストリームが入力され、ASCII コードの B が圧縮されるパターンを示している。この例ではテーブルにすでに B が登録されており、データが圧縮されるパターンである。B をテーブル検索し、その一致するエントリのインデックス 0010 に変換でき、圧縮する。さらに、このとき、テーブルの使用数は 3 であり、上述のエントロピー計算から 2 ビットにインデックスを削減できることがわかる。そこで、インデックスの上位桁を削減し、10 の 2 ビットに削減し、これを圧縮データとする。このように、テーブルの使用率が少ないときにはエントロピーが低くなるため、少ないビット数に圧縮データを変換でき、最少で 1 ビットにまで削減可能である。一方、解凍側（右図）では、圧縮側でミスの場合はオリジナルのデータが出力されるため、圧縮器の変換テーブルの内容と同一のエントリを保持している。圧縮データが解凍側に到着するとテーブルの使用数を調べ、ここでは 3 であり、エントロピー計算をすると 2 と求まるため、圧縮データストリームのうち 2 ビットを取り出す。この例では 10 であり、このビット列テーブルのインデクス幅に 0 を追加して拡張すると 0010 となり、B がヒットする。以上から、圧縮器によって削減されたデータは解凍側で完全に戻すことができる。

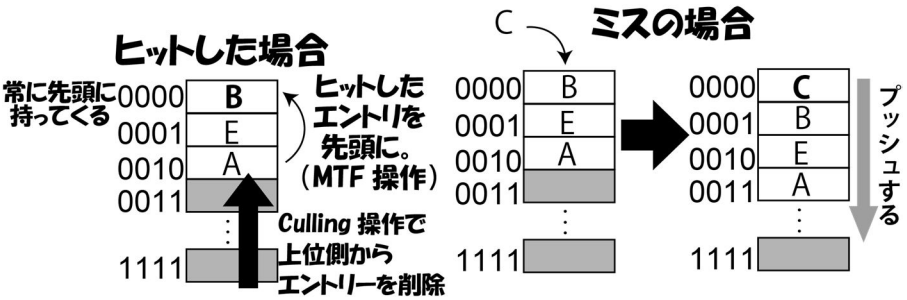

図 2　ASE Coding でのテーブル管理の方法

ASE Coding の圧縮器・解凍器で管理している変換テーブルは、圧縮時に検索する入力データが変換テーブルにヒット/ミスする場合にテーブルエントリの操作がそれぞれ特徴的な操作を行っている。図 2 ではその変換テーブルの操作方法の例を示している。ヒットする場合（左図）には、ヒットしたエントリをテーブルの先頭に移動し、元のエントリをそれぞれ 1 つ下げた位置に移動する Move to Front（MTF）操作を行う。ここでの例は、上述の圧縮操作の際に B が 0010 のエントリにヒットするが、その際に行われる MTF 操作を示している。また、ミスの場合（右図）には、テーブルの先頭にデータをプッシュし、既存のエントリを 1 つ下に移動する。例えば、上述の圧縮操作の後、C が圧縮器に入力された場合、C をテーブルの先頭に配置し、それ以外を 1 つずつしたのエントリに移動する。これらの操作は解凍側においても同様に行われ、圧縮器のテー

ブル内容と同一のエントリ構成を維持している。ミスの場合にテーブルエントリが全て登録されてしまうと、つねに圧縮データのビット数がテーブルのインデックスのビット数になってしまうため、Entropy Culling と呼ぶ、エントリのヒット回数に従って、テーブルの後ろのエントリから使用されているエントリを無効化していく操作を同時に実施する。以上の圧縮操作のうち、データビット数、テーブルのエントリ数、Culling のヒット回数、といったパラメタを環境に合わせ適宜決定し、圧縮器・解凍器の間で同一のパラメタセットを用いることでロスレス圧縮を実現する。
上記の圧縮・解凍の操作はハードウェアで実装可能であり、さらに、メモリを用いないで実装が可能であるため、ラッチと組み合わせ回路での高速動作が可能なロスレス圧縮ハードウェアを実現できる。我々による FPGA への実装例では、従来の圧縮方式である LCA-DLT に比べ、1／10 のハードウェアリソース量で同程度の圧縮性能を実現できることを確認した。
以上の ASE Coding に関し、詳細なパラメタなどの分析を行う事で、パラメタの自動判定技法の開発（文献③）を行った。そして、従来からのロスレス圧縮技法には埋め込むことができなかった例外発生イベントを埋め込むことにより、圧縮器から解凍器へと例外の発生を伝達する手法（文献④）も開発した。また、ASE Coding と従来から音声圧縮などに使われている ADPCM と呼ばれる量子化手法を組合わせ、ロッシー圧縮とロスレス圧縮を組み合わせて高精細な画像伝送が可能であることを示した（文献⑤）。さらに、上述の例外を利用することでデータストリームをチャンクに分割することにより、ASE Coding の並列実行に向けた予備的実験と圧縮率の考察（文献⑥、⑦）を実施した。上記の実験結果や知見に関しては、ジャーナル論文と国際会議での発表を行った。

＜引用文献＞

① Shinichi Yamagiwa, Eisaku Hayakawa, Koichi Marumo. Stream-Based Lossless Data Compression Applying Adaptive Entropy Coding for Hardware-Based Implementation, Algorithms, 159, (2020-06-30), DOI:10.3390/a13070159
② Koichi Marumo, Shinichi Yamagiwa, Ryuta Morita, Hiroshi Sakamoto. Lazy Management for Frequency Table on Hardware-Based Stream Lossless Data Compression, Information, 63, (2016-10-31), DOI:10.3390/info7040063
③ Shinichi Yamagiwa, Suzukaze Kuwabara. Autonomous Parameter Adjustment Method for Lossless Data Compression on Adaptive Stream-Based Entropy Coding, IEEE Access, 186890--186903, (2020-10-08), DOI:10.1109/access.2020.3029705
④ Shinichi Yamagiwa, Koichi Marumo, Suzukaze Kuwabara. Exception Handling Method Based on Event from Look-Up Table Applying Stream-Based Lossless Data Compression, Electronics, 240, (2021-01-21), DOI:10.3390/electronics10030240
⑤ Shinichi Yamagiwa, Yuma Ichinomiya. Stream-Based Visually Lossless Data Compression Applying Variable Bit-Length ADPCM Encoding, Sensors, 4602, (2021-07-05), DOI:10.3390/s21134602
⑥ Taiki Kato, Shinichi Yamagiwa, Koichi Wada. Toward Parallelization Technique for Stream-based Lossless Data Compression, In proceedings of 2023 IEEE International Conference on Big Data, 2667-2672, (2023-12-15), DOI:10.1109/BigData59044.2023.10386184
⑦ Taiki Kato, Shinichi Yamagiwa, Koichi Marumo. Performance Enhancement of Stream-Based Decompression Process by Notifying Compression Buffer Size, 2023 IEEE Symposium on Computers and Communications (ISCC), 491-494, (2023-08-28), DOI:10.1109/iscc58397.2023.10218261

５．主な発表論文等

〔雑誌論文〕　計6件（うち査読付論文　6件／うち国際共著　0件／うちオープンアクセス　4件）

１．著者名 Yamagiwa Shinichi、Ichinomiya Yuma	４．巻 21
２．論文標題 Stream-Based Visually Lossless Data Compression Applying Variable Bit-Length ADPCM Encoding	５．発行年 2021年
３．雑誌名 Sensors	６．最初と最後の頁 4602～4602
掲載論文のDOI（デジタルオブジェクト識別子） 10.3390/s21134602	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Yamagiwa Shinichi、Marumo Koichi、Kuwabara Suzukaze	４．巻 10
２．論文標題 Exception Handling Method Based on Event from Look-Up Table Applying Stream-Based Lossless Data Compression	５．発行年 2021年
３．雑誌名 Electronics	６．最初と最後の頁 240～240
掲載論文のDOI（デジタルオブジェクト識別子） 10.3390/electronics10030240	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Yamagiwa Shinichi、Kuwabara Suzukaze	４．巻 8
２．論文標題 Autonomous Parameter Adjustment Method for Lossless Data Compression on Adaptive Stream-Based Entropy Coding	５．発行年 2020年
３．雑誌名 IEEE Access	６．最初と最後の頁 186890～186903
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/access.2020.3029705	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Yamagiwa Shinichi、Hayakawa Eisaku、Marumo Koichi	４．巻 13
２．論文標題 Stream-Based Lossless Data Compression Applying Adaptive Entropy Coding for Hardware-Based Implementation	５．発行年 2020年
３．雑誌名 Algorithms	６．最初と最後の頁 159～159
掲載論文のDOI（デジタルオブジェクト識別子） 10.3390/a13070159	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

1．著者名 Kato Taiki、Yamagiwa Shinichi、Wada Koichi	4．巻 1
2．論文標題 Toward Parallelization Technique for Stream-based Lossless Data Compression	5．発行年 2023年
3．雑誌名 Proceedings of IEEE International Conference on Big Data	6．最初と最後の頁 2667～2672
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/BigData59044.2023.10386184	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 Kato Taiki、Yamagiwa Shinichi、Marumo Koichi	4．巻 1
2．論文標題 Performance Enhancement of Stream-Based Decompression Process by Notifying Compression Buffer Size	5．発行年 2023年
3．雑誌名 Proceedings of IEEE Symposium on Computers and Communications (ISCC)	6．最初と最後の頁 491～494
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/iscc58397.2023.10218261	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0件

〔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
研究分担者	河原　吉伸 (Kawahara Yoshinobu) (00514796)	大阪大学・大学院情報科学研究科・教授 (14401)	
研究分担者	和田　耕一 (Wada Koichi) (30175145)	筑波大学・システム情報系・名誉教授 (12102)	
研究分担者	坂本　比呂志 (Sakamoto Hiroshi) (50315123)	九州工業大学・大学院情報工学研究院・教授 (17104)	

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