筑波大学 University of Tsukuba　国立大学法人 九州工業大学

平成 27 年 8 月 21 日

報道関係者各位

国立大学法人　筑波大学
国立大学法人　九州工業大学

ビッグデータ時代に対応した新しいロスレスデータ圧縮[注1)]技術を開発
～コンパクトにハードウェア実装可能な高速ストリームデータ圧縮・復号化技術～

研究成果のポイント

1. データストリームを一時的に溜めることなく連続的に圧縮・復号化可能な、高速データ圧縮技術を開発しました。
2. この技術は、リアルタイムに、理論的にはデータを 1/10 のサイズにまで縮小でき、従って、銅線や光ケーブルなどの伝送媒体を変更することなく、ネットワークなどの通信速度を最大 10 倍高速化できます。
3. この技術により、ビッグデータを扱う機器のデータ伝送路の通信性能、および、ストレージのデータ保存容量を格段に飛躍させることができ、ライフログ[注2)]などの次世代アプリケーションをコンパクトに実装できるようになります。

国立大学法人筑波大学 システム情報系　山際伸一准教授は、国立大学法人九州工業大学　坂本比呂志教授と共同で、ビッグデータ時代を見据えた新しいロスレスデータ圧縮技術LCA-DLT(Lowest Common Ancestor-Dynamic Lookup Table)を開発しました。

本研究では専用ハードウェア(LSIチップ)によるデータ圧縮技術として、データの出現傾向を自動認識する、新しいヒストグラム[注3)]管理技術を開発しました。さらに、圧縮されたデータに、圧縮の規則を割り当てた変換表を復元する情報を埋め込むことによって、次々と圧縮されたデータが復号側に送られていき、それを受け取ったところから、順次復号化が可能な技術を確立しました。これにより、従来は圧縮データと別々に復号側に送られていたデータ変換規則を送る必要がなく、流れるデータを連続して圧縮・復号できます。

この技術はハードウェアとの親和性が高く、最大50%の圧縮が可能なモジュールを多段接続することができ、4段接続で、最大10%のサイズにまでデータ圧縮が可能です。このように、ハードウェア量によって圧縮率を自由に調整できるため資源コストが選べるという特徴を有し、さらに、ZIP形式などソフトウェアによるデータ圧縮よりも少ない電力で高速処理ができる、といったメリットがあります。

本研究成果はすでに特許出願済みで、9月4日にハワイで開催されるVLDB(Very Large Data Base) 2015のワークショップで発表予定です。また、8月27～28日に開催のイノベーションジャパンにデモを出展します。

＊ 本研究は、学術振興会による科学研究費補助金基盤研究B(15H02674「データストリーム伝送路におけるロスレス圧縮技術の開発」H27～H30、26280088「圧縮情報処理によるストリームデータからの知識発見」H26～H29)、および、科学技術振興機構さきがけ(H22～H25)、CREST(H26～)によってサポートされています。

研究の背景

ビッグデータ時代の到来に伴い、センサーや、画像といったデータの量は爆発的に増大しており、それを伝送する手段は日々、複雑化しています。データ伝送路により多くのデータを流すには、GHz単位での高速化、または、通信用のケーブルなどデータ伝送路の並列化といった手段がありますが、これだけではコストの永続的な増大が避けられず、技術限界がすぐそこまで来ているのが現実です。

そこで、データ伝送路に流れるデータ量を削減するために、データ圧縮技術が導入されています。これは、データ列に現れる頻出パターンを解析し、そのパターンの出現頻度に応じて代わりの短い符号を割り当ててパターンを符号化することでデータ量を圧縮するもので、ZIP圧縮などソフトウェアによって行うものと、専用のハードウェア(LSIチップ)によって行うものがあります。前者は、ソフトウェアを動かすためにプロセッサパワーを必要とします。後者は、圧縮専用のハードウェアで高速処理が可能です。ただし、実用化されている圧縮技術の多くはソフトウェア処理を基本にしているためプロセッサの搭載が必須となり、消費電力やコストの増大を避けられません。

従来の圧縮技術では、データの中に現れる頻出パターンを集めて、少ないデータに変換するための規則を割り当てる変換表を準備します。この方法だと、その変換表が復号化の際に、どのくらいの大きさになるのかは未知であり、元のデータより大きくなることもありました。さらに、復号するためには、圧縮されたデータとは別に、その変換表をまとめて復号化側に渡さなければならず、データの切れ目ができてしまう上に、復号化する際の一時記憶領域(メモリ)を十分な大きさで取っておく必要がありました。しかしこのメモリ領域は圧縮状態に依存するため大きさが予測できず、動的にメモリ容量を増減できないハードウェアでの実装には、致命的な欠点となっていました。

研究内容と成果

そこで本研究では、ハードウェアによる新しいデータ圧縮技術LCA-DLTを開発しました。これは、データの出現傾向を自動認識するヒストグラム管理技術で、圧縮されたあとのデータに上述の変換表に匹敵する出現頻度を埋め込むことで、変換表を復号側で再構成しながら元のデータに復元できます。従って、変換表の送信が不要で、圧縮処理の切れ目がなく、圧縮側から送られるデータが次々と復号化でき、まさにデータストリームがそのまま圧縮・復号化されていきます(図1)。また、復号化のためのメモリが必要なくなり、一定のメモリ量で実装できます。さらに、変換表に用いるメモリ量の上限を設定することによってデータの出現頻度を動的に再構成する新技術も搭載し、限られた少ないメモリ量で圧縮・復号ができるハードウェアを実現できるようになりました。

本技術では、最大50%の圧縮が可能なモジュールを多段接続することができ、4段接続で、圧縮前のデータが最大 10%のサイズにまで圧縮が可能です(図 2)。つまり、ハードウェア量とメモリ量を圧縮率との間で自由に調整可能な圧縮ハードウェアを開発できます。また、ソフトウェアによるデータ圧縮のようなプロセッサ＋メモリの構成が不要で、消費電力の低減と高速化を通して、システムの低価格化を可能にします。本研究チームでは、圧縮時に用いるパラメーターの組合せを鍵とする独自の暗号化手法も開発しており、それと組み合わせることで、安全で高速なデータ通信が実現できます。

今後の展開

本技術により、既存のインフラでデータ伝送量や伝送速度を大幅に向上させることが可能となり、コスト削減にも貢献します。また、HDD の物理容量を、見かけ上、数倍に拡張したデータセンター向けストレージサーバや、超高音質音源の無線伝送などの用途への応用展開が期待されます。

参考図

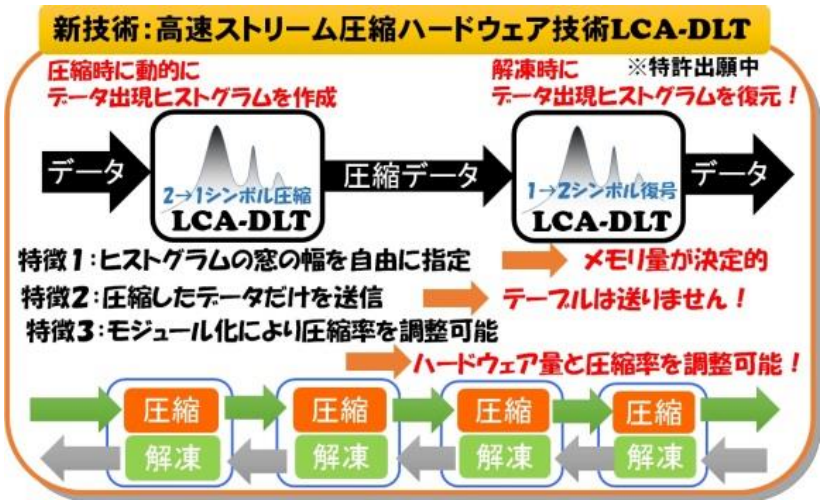

図1 新圧縮技術の特徴。データが圧縮器に入力されると次々に圧縮されたデータが出力され、復号化側に伝搬し、復号化側では圧縮データを 1 つでも受け取ると、圧縮側で作られた変換テーブルが復元され、復号されていきます。さらに、1 段で 50%圧縮（2→1 シンボル圧縮）可能なモジュールを多段接続することで、ハードウェア量と圧縮率を選択できます。

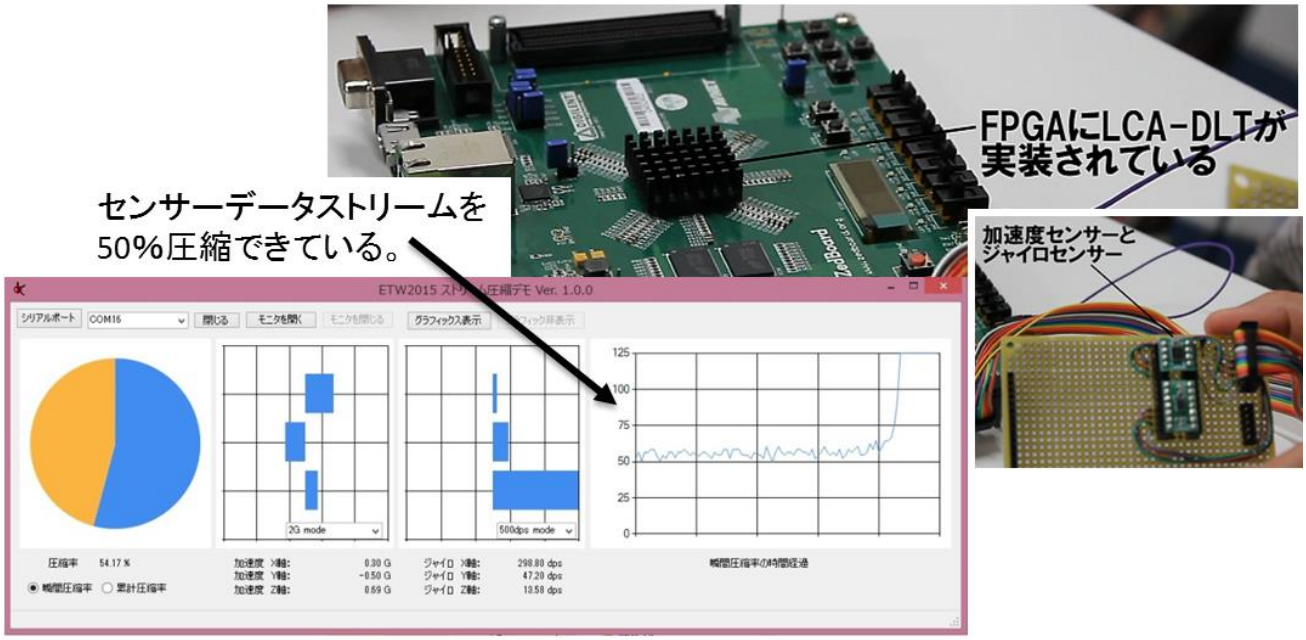

図2 センサーデータをリアルタイム圧縮するデモ。FPGA（集積回路の一種、field-programmable gate array）に 2 段の LCA-DLT モジュールを搭載し、加速度センサーとジャイロセンサーのデータを 50%程度にまでリアルタイム圧縮しています。

用語解説

注1） ロスレスデータ圧縮：圧縮されたデータが復号化によって元に戻る圧縮方法。可逆圧縮ともいう。JPEG などの画像圧縮は非可逆圧縮と呼ばれ、圧縮前のデータに戻すことができない。

注2） ライフログ：ウェアラブルデバイスを用いて日々の身体の活動や行動、心拍や脈拍といったすべての生体データから映像、画像、といった五感で捉えた情報をビッグデータとして蓄え、健康管理や生活のアドバイスを行う次世代アプリケーション。莫大なデータ量を扱うことが問題となっている。

注3）　ヒストグラム：データの種類毎に出現傾向の統計をまとめた表

掲載論文

【題　名】 Stream-based Lossless Data Compression Hardware using Adaptive Frequency Table Management
（適応的にシンボル変換テーブルを管理するロスレス・ストリームデータ圧縮ハードウェア）

【著者名】 Shinichi Yamagiwa, Koichi Marumo and Hiroshi Sakamoto

【掲載誌】 In Proceedings of Very Large Data Base (VLDB2015) / the sixth workshop on big data benchmarks, Performance optimization, and Emerging Hardware (BPOE-6)

問合わせ先

山際伸一（やまぎわ　しんいち）
筑波大学 システム情報系　准教授

坂本比呂志（さかもと　ひろし）
九州工業大学　大学院情報工学研究院　教授