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　４ 年　 ５ 月 ２８ 日現在

機関番号：１２６０１
研究種目：基盤研究(A)（一般）
研究期間：2016～2020
課題番号：１６Ｈ０１７１５

研究課題名（和文）大規模なad-hocデータに対する高速処理基盤

研究課題名（英文）High Performance Data Processing System for Ad-hoc Data

研究代表者
田浦　健次朗（Taura, Kenjiro）

東京大学・大学院情報理工学系研究科・教授

研究者番号：９０２８２７１４

交付決定額（研究期間全体）：（直接経費）　　29,900,000円

研究成果の概要（和文）：並列化，SIMD化を用いた高性能テキスト処理を達成するため，正規表現または文脈自由文法に対する並列化・ベクトル化された字句解析器・構文解析器を自動生成するアプローチに沿って研究を行った．字句解析器を用いない(スキャナレス)構文解析に対してSIMD命令を用いて高速化を行うアプローチ，字句解析器の並列化と，局所的に構文解析可能な(したがって並列に処理しやすい)文法に対する構文解析器の並列化を行うアプローチを追求した．

研究成果の学術的意義や社会的意義
データ活用はSociety 5.0の要諦である．多くの利用可能なデータはテキスト形式で保存されている(XML，JSONなど標準的な形式のものもあれば，決まった形式のないものもある)．文字列に対するデータ処理の一番はじめの段階が字句解析または構文解析と呼ばれる，一種のパターンマッチング処理である．本研究はそれらを容易に，かつ高速に処理することを目指したもので，社会で利用可能なビッグデータの増大に対して有用な貢献を果たしうる研究である．

研究成果の概要（英文）：Toward the goal of high performance text processing using parallelization and vectorization, we studied lexer (or parser) generators that generate parallelized/vectorized lexers (or parsers) from regular expressions or context free grammars. We investigate an approach that vectorizes scannerless parser and an approach that parallelizes both lexers and locally parsable (thus relatively simple-to-parallelize) parsers.

研究分野：並列処理

キーワード：大規模データ処理　ad-hocデータ処理　字句解析　構文解析　データ抽出

様　式　C－１９、F－１９－１、Z－１９（共通）

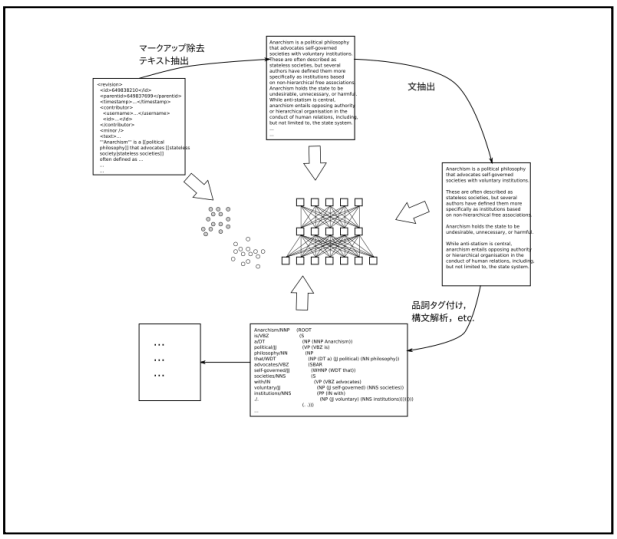
１．研究開始当初の背景

大規模データ解析処理では，ad-hoc データと呼ばれる，標準的なデータ形式(XML, json など)に従わず，したがってそのデータ解析のためのツール(パーザ)が存在しない非構造データが頻繁に登場する(ログデータやプログラムの出力など)．また，標準的な形式で書かれたデータであったとしても，フィールドのデータ自身が任意のテキストでありそれを解析しなくてはならなかったり，標準的な形式を逸脱した，実質的には ad-hoc データと呼ぶべきデータも存在する．さらに，さらに，標準的な形式に整合したデータであっても解析したい部分はデータのごく一部であって,そのデータ形式の汎用的なパーザを通さずに，解析したい部分だけを文法として記述することで高速化が達成できる可能性もある．

そのようなデータの解析は通常，正規表現などのライブラリを用いてスクリプト言語などを用いて行われる．正規表現では済まないデータ(括弧の対応を取る必要がある形式)の場合，文脈自由文法でパターンを記述する必要がある．そのような文法記述から，その文法に沿って文字列を解析するプログラムを生成するプログラムは，パーサ生成器と呼ばれ，プログラミング言語の構文解析器を作るのに広く使われている．ad-hoc なデータ処理を専門にした，ドメイン特化型言語(PADS)も存在していた．しかし，正規表現ライブラリ，パーサ生成器，PADS，いずれのも処理は逐次処理であり，マルチコア，SIMD 命令，GPU，分散メモリ並列計算機を用いた高速化はできていなかった．パターンや文法に対するマッチングは文字列を先頭から順に読み込んで状態遷移を繰り返す処理が基本であり，SIMD 命令の利用も並列化も容易ではない．

２．研究の目的

そこで本研究はそのような状況を打破して，ad-hoc データに対して，宣言的・簡潔なデータ形式 (パターン・文法) の記述を行うだけで，並列化やベクトル化を施した高性能なデータ処理系を生成し，大規模データ処理の生産性と性能を同時に，飛躍的に高める基盤技術を確立することを目的とした．またそのようなデータ処理を複雑に組みあわせてつくられるデータ処理ワークフローを実行する際のスケジューリング最適化問題を，並列度，メモリ消費量，入出力コストのトレードオフという観点から研究することを目的とした．自然言語処理を主な応用として，生データから機械学習に至るまでの実タスクに対して評価を行うことを目指した



３．研究の方法

テキスト解析のアルゴリズムは大別すると，正規言語を解析するためのアルゴリズム(字句解析またはスキャナ)と文脈自由文法を解析するためのアルゴリズム(構文解析またはパーザ)に別れる．また，構文解析を行う際に，最初に文字列をスキャナに入力して，それによって認識(生成)された字句の列を構文解析が処理するのが通常であるが，スキャナレスと呼ばれる，構文解析器に直接文字を入力する方式も存在する．

並列化，高速化の方法：したがって並列化，高速化の方法も幾つかの戦略，部分問題が考えられ，それぞれの方式を模索することとした

(1) 字句解析の並列化
(2) スキャナ(字句解析)ありの構文解析の並列化
(3) スキャナレス構文解析の並列化

評価の方法：高速なテキスト処理に関する研究は，しばしば XML や json 形式のデータを対象として，それらの文法に特化した高性能パーザを実装している．我々の研究は特定データ形式に特化したものではなく，あくまで任意の文法に対して高速なパーザを生成するが，その上で性能比較対象としては，それらの XML や json に対するパーザとの比較を行っている．

４．研究成果

[1] LL(*) 文法に基づくスキャナレス構文解析器の提案

2017 年-2018 年にかけて，文法の表現力の高さと高速実装の両立をめざして，LL(*)文法のスキャナレス化というアプローチで研究を行った．LL(*)は，Parr らによって提唱された，任意の先読みを許す(そのため表現力が高い)文法クラスであり，それを元にした ANTLR 3 というパーサジェネレータが実装されている．スキャナレス化は，ほとんどの構文解析処理で必要なスキャナを取り除き，文字レベルの高速な処理と構文解析を一体化でき，高速な実装が可能なのではないかという着想に基づき本研究を行った．

LL(k)文法は，ある非終端記号を左辺に持つ生成規則が複数ある場合，どの生成規則を適用すべきかを k 個以内の字句を先読みすることで決定できる文法のクラスである．LL(*)はその個数に制限がなく，変わりに，字句列の正規表現によって区別する．通常の(スキャナありの)構文解析器において，個々の字句は文字の正規表現で表されるため，先読みに必要なのは結局，文字の正規表現に他ならない．

文字レベルの最適化として，SIMD 命令（SSE4.2（V)PCMPISTRI 命令）を用いたリテラル文字列のマッチングを実装した．本処理系を用いて JSON 形式の記述を行い，180MB 程度の JSON ファイルを処理し，性能評価を行った．RapidJson 1.1.0，PEGTL 2.3.4 との比較において，提案した処理系が他の処理系を上回る性能(341MB/秒)を示した．

本研究は情報処理学会 CS 領域賞を受賞した。

[2] バックトラックのない prescan による並列字句解析

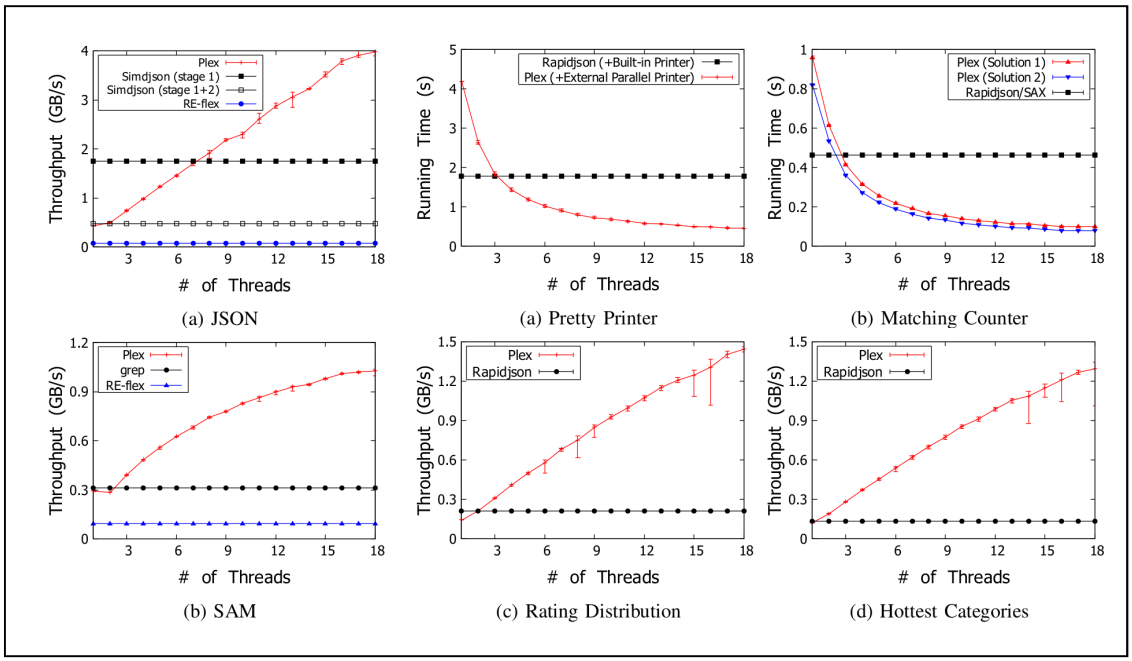
字句解析の並列化にあたっての基本的な方針は入力を複数の部分(チャンク)に区切りってそれぞれのチャンクをスレッドで処理することであり，その際の基本的な困難は，(最初のチャンク以外)チャンクにおける初期状態(逐次的に処理した場合のその文字における状態)がわからないことである．

そのため実際の字句解析の前に prescan と呼ばれるフェーズを設けて，各チャンクの先頭における状態を求めるためだけの，前処理的な(実際には字句を生成しない)字句解析を行うが，この prescan においても結局，チャンクの初期状態がわからないという同じ問題が生ずる．

また，字句解析は通常，字句を構成する正規表現を字句の種類の数だけ定義することで行われるが，ほとんどの場合，どこまでを一つの字句と認識するかについて多くの可能性(曖昧性)がある．そのため通常は longest prefix (maximal munch) という方針で，なるべく長い文字列をひとつの字句として認識しようとするが，この方針の実現には一般にはバックトラックを要する．並列処理をする際はこのバックトラックによってチャンクをまたがったバックトラックが行われる可能性もあり，並列処理は非常に複雑になる．

そこで本研究では prescanning をバックトラックなしで行う方法，それを用いて並列に字句解析を行う方法を提案した．さらに，prescanning ではチャンク先頭における初期状態が不明のため，すべての可能な初期状態から状態を遷移させる必要があるがそれを SIMD 命令を用いて行う方法を実装した．性能評価では，高速な JSON 専用パーザとして知られる simdjson と比較して，1 コア実行の性能は劣るものの，マルチコアを用いることでそれを大幅に凌駕する性能が得られることを確認した(下図)．

本研究は並列分散処理に関するトップクラス会議の一つである IPDPS 2021 に採択された。



[3] 局所的に構文解析可能な文脈自由文法の並列化

構文解析，字句解析いずれにあたっても，並列化にあたっての基本的な方針は，入力を複数の部分(チャンク)に区切りってそれぞれのチャンクをスレッドで処理することであり，その際の基本的な困難は，(最初のチャンク以外)チャンクの先頭の状態がわからないことである．

これに対する方針には様々なものが考えられるが，構文解析においては，どのような場合に，その前後の入力を見なくても，一部の入力に対する構文木を決定できるかというのがもっとも基本的なといとなる．例えば通常の算術式の文法を考えると，「... 2 * 3 ...」という入力においては，... の部分がなんであっても，(2 * 3) 部分が一つの(部分)木を作ることは直感的に頷けるであろう．それに対し「... 2 + 3 ...」の場合そうはいえない．実際例えば，それは「... 1 * 2 + 3 * 4 ...」という入力の一部だったかも知れない．前後の入力がなんであっても部分木への還元が可能である性質を，「局所的に構文解析可能(locally parsable)」といい，並列構文解析のために有用な性質である．

演算子順位文法と呼ばれる文法に対して，局所的に構文解析可能であるケースを判定し，それに対する並列構文解析を行う既存研究として，PAPAGENO が存在した．PAPAGENO は，スキャナの存在を前提とした処理系で，字句解析器の並列化に際しては文法ごとに人間が作成した方法で入力の分割(字句の終わりの検出)を行っている．本研究では[2]で提案した手法で字句解析も完全に自動並列化し，PAPAGENO 同様の手法で構文解析を並列化している．また，詳細は省略するが PAPAGENO ではほとんど台数効果が得られない文法が頻繁に生ずることを見出しており，それに対する解決策も提案し，現在論文を執筆中である．

５．主な発表論文等

〔雑誌論文〕　計15件（うち査読付論文　14件／うち国際共著　2件／うちオープンアクセス　0件）

１．著者名 井原 央翔, 佐藤 重幸, 田浦 健次朗	４．巻 -
２．論文標題 LL(*) 文法に基づくスキャナレス構文解析器の提案	５．発行年 2018年
３．雑誌名 Cross-disciplinary workshop on computing Systems, Infrastructures, and programming (xSIG) 2018	６．最初と最後の頁 -
掲載論文のDOI（デジタルオブジェクト識別子） なし	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 リュウ ケイコウ, 井原 央翔, 田浦 健次朗	４．巻 12
２．論文標題 OPG を利用したアドホックな並列データ処理系	５．発行年 2018年
３．雑誌名 情報処理学会論文誌プログラミング（PRO）	６．最初と最後の頁 1-8
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
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 井原 央翔, 佐藤 重幸, 田浦 健次朗	４．巻 -
２．論文標題 LL(*) 文法に基づくスキャナレス構文解析器の提案	５．発行年 2018年
３．雑誌名 xSIG 2018 workshop	６．最初と最後の頁 -
掲載論文のDOI（デジタルオブジェクト識別子） なし	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 Yuchen Qiao, Kazuma Hashimoto, Akiko Eriguchi, Haixia Wang, Dongsheng Wang, Yoshimasa Tsuruoka, and Kenjiro Taura.	４．巻 -
２．論文標題 Cache friendly parallelization of neural encoder-decoder models without padding on multi-core architecture.	５．発行年 2017年
３．雑誌名 The 6th International Workshop on Parallel and Distributed Computing for Large Scale Machine Learning and Big Data Analytics	６．最初と最後の頁 437-440
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/IPDPSW.2017.165	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 Kazuma Hashimoto and Yoshimasa Tsuruoka	4．巻 -
2．論文標題 Neural Machine Translation with Source-Side Latent Graph Parsing	5．発行年 2017年
3．雑誌名 Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)	6．最初と最後の頁 125-135
掲載論文のDOI（デジタルオブジェクト識別子） 10.18653/v1/D17-1012	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 Akiko Eriguchi, Yoshimasa Tsuruoka, and Kyunghyun Cho	4．巻 -
2．論文標題 Learning to Parse and Translate Improves Neural Machine Translation	5．発行年 2017年
3．雑誌名 Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL2017)	6．最初と最後の頁 72-78
掲載論文のDOI（デジタルオブジェクト識別子） 10.18653/v1/P17-2012	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 Wataru Endo and Kenjiro Taura	4．巻 -
2．論文標題 Parallelized software offloading of low-level communication with user-level threads	5．発行年 2018年
3．雑誌名 Proceedings of the International Conference on High Performance Computing in Asia-Pacific Region	6．最初と最後の頁 289-298
掲載論文のDOI（デジタルオブジェクト識別子） 10.1145/3149457.3149475	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 中澤 弘樹, 田浦 健次朗	4．巻 -
2．論文標題 低レイテンシ SSD をメモリ拡張として利用したときの性能評価	5．発行年 2018年
3．雑誌名 xSIG 2018 workshop	6．最初と最後の頁 -
掲載論文のDOI（デジタルオブジェクト識別子） なし	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 Masahiro Tanaka, Kenjiro Taura, and Kentaro Torisawa	4．巻 -
2．論文標題 Low latency and resource-aware program composition for large-scale data analysis	5．発行年 2016年
3．雑誌名 16th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid)	6．最初と最後の頁 325-330
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/CCGrid.2016.88	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 Shintaro Iwasaki, Kenjiro Taura	4．巻 -
2．論文標題 A static cut-off for task parallel programs	5．発行年 2016年
3．雑誌名 Proceedings of the 2016 International Conference on Parallel Architectures and Compilation	6．最初と最後の頁 139-150
掲載論文のDOI（デジタルオブジェクト識別子） 10.1145/2967938.2967968	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 Masaru Ito, Hiroshi Inoue, and Kenjiro Taura	4．巻 -
2．論文標題 Fragmented BWT: An Extended BWT for Full-Text Indexing	5．発行年 2016年
3．雑誌名 International Symposium on String Processing and Information Retrieval	6．最初と最後の頁 97-109
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/978-3-319-46049-9	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 Shintaro Iwasaki, Kenjiro Taura	4．巻 -
2．論文標題 Autotuning of a Cut-Off for Task Parallel Programs	5．発行年 2016年
3．雑誌名 IEEE 10th International Symposium on Embedded Multicore/Many-core Systems-on-Chip (MCSoC)	6．最初と最後の頁 353-360
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/MCSoC.2016.51	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 Yuchen Qiao, Kenjiro Taura, Kazuma Hashimoto, Yoshimasa Tsuruoka and Akkiko Eriguchi	4．巻 -
2．論文標題 Cache Friendly Parallel Encoder-Decoder Model without Padding on Mulit-core Architecture	5．発行年 2017年
3．雑誌名 Proceedings of The 6th International Workshop on Parallel and Distributed Computing for Large Scale Machine Learning and Big Data Analytics	6．最初と最後の頁 -
掲載論文のDOI（デジタルオブジェクト識別子） なし	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 Kazuma Hashimoto, Akiko Eriguchi, and Yoshimasa Tsuruoka	4．巻 -
2．論文標題 Domain Adaptation and Attention-Based Unknown Word Replacement in Chinese-to-Japanese Neural Machine Translation	5．発行年 2016年
3．雑誌名 the 3rd Workshop on Asian Translation (WAT2016)	6．最初と最後の頁 75-83
掲載論文のDOI（デジタルオブジェクト識別子） なし	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 Akiko Eriguchi, Kazuma Hashimoto, and Yoshimasa Tsuruoka	4．巻 -
2．論文標題 Character-based Decoding in Tree-to-Sequence Attention-based Neural Machine Translation	5．発行年 2016年
3．雑誌名 the 3rd Workshop on Asian Translation (WAT2016)	6．最初と最後の頁 175-183
掲載論文のDOI（デジタルオブジェクト識別子） なし	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2件（うち招待講演　1件／うち国際学会　1件）

1．発表者名 リュウ ケイコウ
2．発表標題 OPGを使ったアドホックな大規模な文字列データ解析のための並列処理系
3．学会等名 情報処理学会論文誌プログラミング（PRO）
4．発表年 2019年

１．発表者名 Kenjiro Taura
２．発表標題 A Quest for Unified, Global View Parallel Programming Models for Our Future
３．学会等名 A Quest for Unified, Global View Parallel Programming Models for Our Future（招待講演）（国際学会）
４．発表年 2016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