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　３ 年　 ６ 月　 ４ 日現在

機関番号：８２４０１
研究種目：若手研究(B)
研究期間：2017～2020
課題番号：１７Ｋ１２６９３

研究課題名（和文）超エクサシステムのコデザインのための疑似プロファイルに基づく性能予測手法の研究

研究課題名（英文）Scalable communication performance prediction using pseudo trace files for beyond-exascale system co-design

研究代表者
辻　美和子（Tusji, Miwako）

国立研究開発法人理化学研究所・計算科学研究センター・研究員

研究者番号：８０４６６４６６

交付決定額（研究期間全体）：（直接経費）　　3,200,000円

研究成果の概要（和文）：未知のシステム上での通信制能推定手法としては，既存のシステムでアプリを実行し，各ノードから得られた通信ログを用いてネットワークシミュレータに駆動させる手法が広く用いられている．しかし，将来システムが既存システムよりも大規模な場合，この通信ログをそのまま用いることはできない．本研究では，実システムから得られた通信ログを大規模システム向けに水増しし，これらのPseudoファイルをネットワークシミュレータの入力として用いる手法を提案した．また，ソースコードの解析によりPseudoファイルを生成する手法を開発した．これにより，既存システムよりも大規模なシステムでの通信性能推定が，平易に可能となった．

研究成果の学術的意義や社会的意義
アプリケーションを効率的に実行するためには，アプリケーションをシステム向けに最適化するだけではなく，設計段階からアプリケーションが性能を発揮できるシステムを設計することが重要である．システムとアプリケーションの双方向的な設計アプローチはコデザインと呼ばれ，富岳をはじめ多くのシステムの設計課程で採用されている．この過程では，複数のアプリを複数のシステムの候補で評価する必要があるため，平易かつ正確な性能推定ツールが求められる．本手法では，とくに通信性能推定に焦点を当て，既存の手法では難しかった大規模なシステムでの通信性能を平易に行えるようにした．

研究成果の概要（英文）：Trace-driven network simulators, which use MPI event trace files recorded in existing systems, have been widely used because of its simplicity to estimate communication performance of future systems. However, if a future system is larger than a current system, it is difficult to adopt the trace files obtained from the current system directly. In order to address the scaling problem in the trace driven network simulators, we have proposed to create "Pseudo" trace files from the real files and use them as the inputs of simulators. We have also developed a method to create pseudo files automatically by analyzing source codes of applications. These enable us to obtain a first-order approximation of the communication performance of large system easily.

研究分野：高性能計算

キーワード：性能推定　通信性能　コデザイン

様　式　C－１９、F－１９－１、Z－１９、CK－１９（共通）

１．研究開始当初の背景

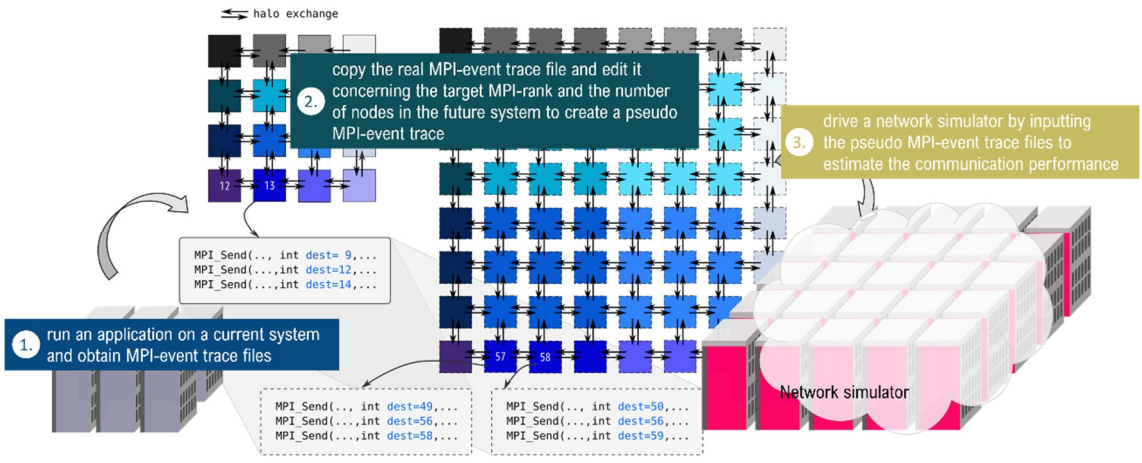
アプリケーションを効率的に実行できる大規模システムの設計においては、アプリケーションとシステムのコデザインが不可欠である。コデザインは、システム設計における双方向的アプローチであり、アプリケーションがシステムに向けて最適化するだけでなく、ハードウエアやシステムソフトウエアをアプリケーションがより性能が発揮できるように設計する。コデザインにおけるシステム設計においては、さまざまなシステムの候補の上で、アプリケーション性能を推定し、より良いシステムを模索することが重要である。実際には存在しないシステム上でのアプリケーションの推定を行うためには、シミュレータをはじめとするさまざまなツールが用いられる。
本研究では、通信性能推定に焦点を当てる。通信性能は、通信バンド幅、レイテンシ、およびネットワークトポロジに加えて、システムの大きさに大きく影響される。通信性能推定において広く用いられる手法のひとつに、既存のシステムでアプリケーションを実行し、各ノードで取得された通信ログを、パラメータをさまざまに変化させたネットワークシミュレータに入力して性能推定を行うトレースドリブンシミュレーションがある。トレースドリブンシミュレーションは、実機実行とシミュレータ実行を機械的に行うだけでよく、アプリケーションに大きく手を加えるなど必要はなく、平易である点にアドバンテージがある。しかしながら、実実行でノードごとに得られたログを利用するため、既存のシステムよりもはるかに大きな将来システムの通信性能推定が難しいという問題があった。本研究では、実行ログを用いて平易かつスケーラブルな通信性能推定手法を開発した。

２．研究の目的

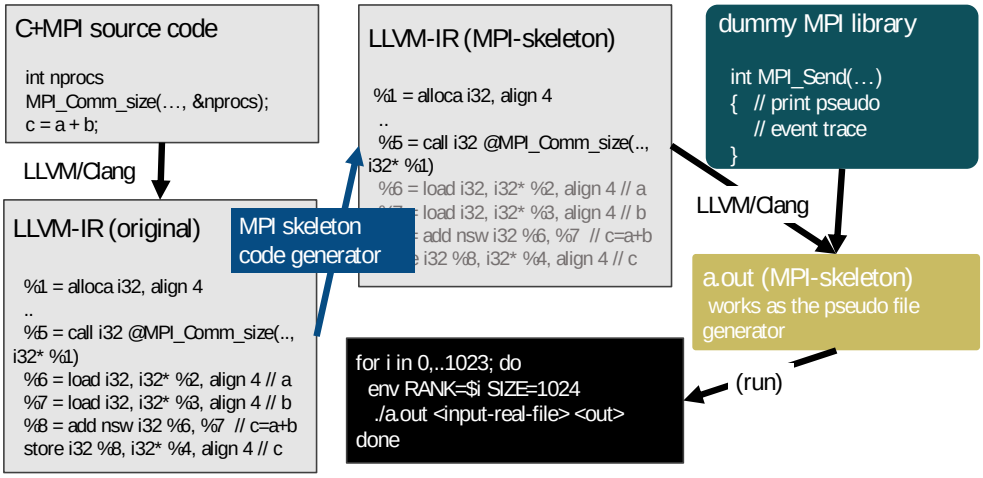
前述のように、既存の実システムでの通信ログを入力として用いてネットワークシミュレータを実行するトレースドリブンシミュレーションを既存システムよりもはるかに大規模な将来システムに適用する場合、将来システムに対応するだけの通信ログを得ることができないことから、困難であった。
本研究の目的は、トレースドリブンシミュレータの平易さを残しつつ、小規模な現在システムからより大規模システムの推定に利用可能なものに拡張することである。このために、数百?数千程度の比較的小規模な実行から得られた通信ログを、大規模システムで想定するノード数に合わせて「水増し」生成し、これを用いて数千?数万ノードからなる将来システムの性能推定を行う。アプリケーションの通信手法としては、Massage Passing Interface (MPI) がもっとも広く用いられ将来的にも有用であろうことから、MPI を想定し、この手法を SCAMP (SCAlable Mpi Profiler) と名づけた。また、水増しされた擬似的な通信ログファイルを Pseudo トレースファイルと呼ぶ

３．研究の方法

本研究で提案する通信性能推定手法は手順を以下に述べるとともに下図に示す：
(1) 既存システムでアプリケーションを実行し、各ノードから通信ログを得る
(2) 得られた通信ログを目標システムのサイズに合わせて「水増し」する
(3) 水増しした通信ログをトレースドリブン通信シミュレータの入力として将来システムの性能を推定する



もっとも重要な過程は (2) である。本研究では通信手法としてもっとも広く用いられている MPI を想定した。よって、各ノードから得られる通信ログは、各 MPI 関数呼び出しごとに、呼び出された関数およびその入力引数である。入力引数は、通信相手、通信量などの情報を含む。例えば、 rank=0 から rank=(ノード数)-1 への通信について、1000 ノードからなる実システムの通信ログが rank=0 から 999 への通信であっても、10000 ノードを想定する通信性能推定用の Pseudo トレースファイルは、rank=0 から 9999 への通信であると書き換える必要がある。また、ストロングスケーリングでの通信性能を推定する場合、ノードあたりの通信バッファサイズを小さくするなどの変更も必要である。
これをユーザに大きな負担をかけることなく実現するために、ソースコードの解析による自動 Pseudo トレースファイル生成手法を実装した（下図）。まず、オリジナルのソースコードを LLVM/Clang コンパイラで LLVM-IR と呼ばれる中間ファイルに変換する。この中間ファイル対して MPI 関数への入力パラメータを逆解析することで、通信と通信先ランクの計算など関連する命令のみを残した中間ファイルを生成し、これを LLVM/Clang コンパイラでコンパイルし、Pseudo トレースファイル生成のための通信ダミー関数ライブラリとリンクする。得られたバイナリに、環境変数として想定ノードサイズと MPI ランクを与えることで、任意のサイズのシステムの任意のランクについて Pseudo トレースファイルを生成することを可能とした。

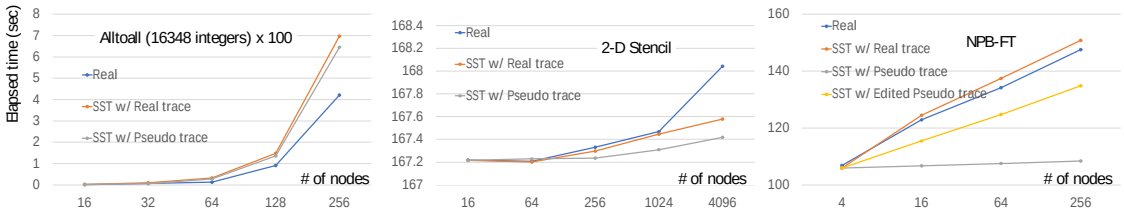


通信シミュレータは SST/macro と呼ばれる既存のシミュレータを用いた。将来システムのパラメータ -- トポロジやバンド幅など --- と Pseudo トレースファイルを SST/macro に入力として与え、実行することで推定結果を得ることができる。
本研究では主に通信性能に焦点をあてたが、演算部については、将来システムを想定した CPU シミュレータと連携して修正を加えることが望ましいと考えられる。CPU シミュレータは、その速度の限界から 1 ノードもしくは 1 スレッドのシミュレーションのみ行うのが通例であり、通信を含む大規模アプリケーションはこのためにしばしば手動で演算部を切り出す必要があった。著者らの別の研究で開発した MPI 通信を、別途取得しておいた通信バッファの読み込みに置き換えるダミー関数の使用により、ソースコードにほとんど手を加えることなく、通信を省略することができる。ただし、通信バッファの読み込みに IO を伴うため、推定される性能には実際の性能と差が生じる可能性がある。本研究では、実実行の結果と CPU シミュレータの結果を用いて本ツールの性能を検証し、要求メモリ量とキャッシュサイズが近い場合などにおいて、性能の差が大きく、一方で、キャッシュにまったく乗らない場合などは、適切な推定が行えることを明らかにした。

４．研究成果

SCAMP 法について、手法の有効性を検証するために、実際の大規模とごく少数のノードから得られた実トレースファイルからの Pseudo トレースファイルによる大規模実行時の性能推定の結果を比較した。ここでは、Pseudo トレースファイルは手動で生成した。ベンチマークアプリケーションの NPB-EP、NPB-FT、および実アプリケーションから抽出された格子 QCD 計算を行うミニアプリケーション CCS-QCD において結果を検証した。その結果、SCAMP 法ではとくに通信が単純な NPB-EP においてはかなり正確な推定が得られ、NPB-FT および CCS-QCD においては実際よりも高速であると推定する傾向が見られるものの、大規模システムの通信性能について知見を与えることが可能であった。推定結果を次ページの図に示す.
さらに、ソースコードの解析による通信性能推定について、NPB-FT、ステンシル計算アプリケーションを用いて実験を行った。実験結果は、上記と同様に実際よりも高速であると推定する傾向が見られるものの、大規模システムの通信性能について知見を与えることが可能であった。また、Pseudo トレースファイルの生成のコストや、シミュレーションの時間について検証した。シミュレーションは 1 スレッドのみを用いて行われたにもかかわらず、実際にアプリケーションを実行するよりも高速であった。また、Pseudo トレースファイルの生成についても１つ１つのフ

ァイルにかかる時間は僅かであり、各ファイルは独立に生成できるため、複数のコアを用いれば、大規模システムを想定して大量の Pseudo トレースファイルを生成する場合でも、リーズナブルな時間ですべての Pseudo トレースファイルが生成できることを確認した。



５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計3件（うち招待講演　0件／うち国際学会　1件）

１．発表者名 辻 美和子, 佐藤 三久
２．発表標題 将来システムのコデザインのための CPU シミュレータによ る MPI リプレイ環境および性能推定手法の検討
３．学会等名 第173回HPC研究会
４．発表年 2020年

１．発表者名 Miwako Tsuji, Taisuke Boku, Mitsuhisa Sato
２．発表標題 Scalable communication performance prediction using auto- generated pseudo mpi event trace
３．学会等名 Proceedings of the International Conference on High Performance Computing in Asia-Pacific Region, HPC Asia 2019（国際学会）
４．発表年 2019年

１．発表者名 辻 美和子, 李 珍泌, 朴 泰祐, 佐藤 三久
２．発表標題 疑似MPIトレースプロファイルを用いた通信性能推定手法SCAMPのための疑似トレースファイル自動作成手法の検討
３．学会等名 第161回HPC研究発表会
４．発表年 2017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