自動生成された疑似 MPI トレースファイル と CPU シミュレータを用いたアプリケーション性能予測

Scalable communication performance prediction using auto generated pseudo MPI event trace and CPU simulator

辻 美和子
理化学研究所

１．研究目的

大規模システムの導入後の速やかな成果創出のためには，システム設計段階からシステムをアプリケーションの性能を最大限に発揮するよう調整し，一方で，アプリケーションをシステムに向けて最適化するコデザインが重要とされる．コデザインの過程においては，実際には存在しない設計段階のシステムにおけるアプリケーションの性能を推定するための性能推定ツールが重要である．本研究ではとくに通信性能推定に焦点をあてる．

通信性能を推定するために，アプリケーションを実システム上で実行して，通信トレースファイルを取得し，このトレースを入力としたネットワークシミュレータを利用する方法がある．この手法はアプリケーションの内部を細かく調査することなく機械的に実行できるため平易であるが，将来システムが現在利用可能なシステムよりも大規模な場合，得られた通信ログをそのまま用いることができないとうい問題点があった．

われわれは，この問題を解決するために，少数の実トレースを将来システムの規模に合わせて「水増し」して疑似トレースファイルを生成し，大量の疑似トレースファイルを用いてネットワークシミュレーションを行う手法である SCAMP を提案してきた．

昨年度まではトーラスネットワークである京コンピュータを中心に用いて提案手法を検証してきた．本年の目的は FatTree ネットワークである OFP を用いて手法の汎用性を検証することである．さらに，将来システムにおけるアプリケーションの実行性能を包括的に推定するためにネットワークシミュレータと CPU シミュレータを併用した性能予測手法について検討する．

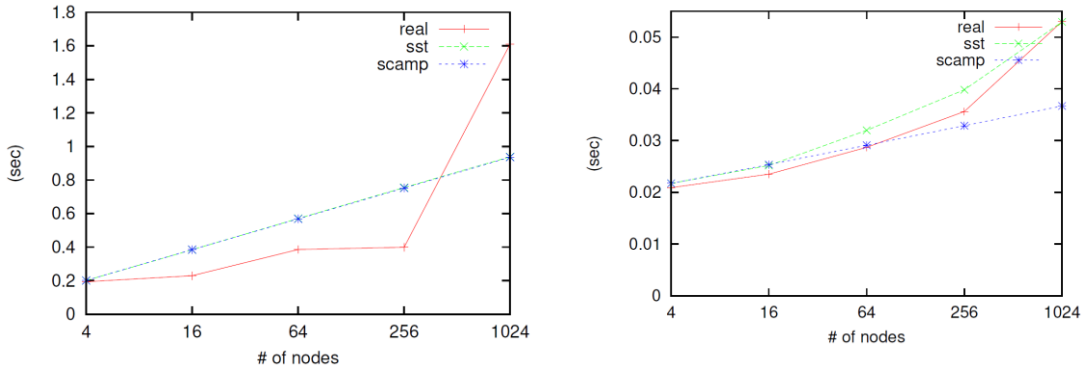
２．研究成果の内容

１）全体全通信（MPI_Allreduce）および 1 対 1 通信（pingpong）を複数回繰り返すシンプルなコードを用いて通信性能推定ツール sst/macro のパラメータ調整を行った．この過程で，

・OFP で同種の通信関数を複数呼び出す場合，それぞれの通信関数で最初の 1 回が他に比べて長い時間を要する

・MPI_Finalize に，比較的長い時間を要する

ことがわかった．sst/macro ではこのようなイレギュラなケースをシミュレーションすることは不可能であり，また短時間で終了するコードについては上記の影響が大きく，性能推定の精度を検証することがむずかしくなる．そこで，実トレースファイルから最初の通信を削除した中間トレースファイルを生成し，中間トレースファイルによりパラメータの合わせこみを行うとともに，疑似トレースファイルを生成した．



上図の左は実ファイルによる実行時間・推定実行時間である．右は中間ファイルによる実行時間・推定実行時間である．推定に用いるパラメータ（レイテンシ）はそれぞれの場合で合わせこみを行った．右から，少数のトレースファイルから疑似トレースファイルを生成する提案手法（scamp）は，推定規模と同数のトレースファイルを用いる既存手法(sst)と比較して，より楽天的な見積もりをする傾向があり，またその度合いはトーラスネットワーク（京コンピュータ）を対象とした実験よりも大きくなった．なお京コンピュータに対する推定結果は Tsuji らの HPCAsia2019 発表を参照されたい．これは，実ファイルには通信そのものの時間に加えて，集合通信の時間に影響を与えるランクごとの演算時間のばらつきも記録されているのに対して，少数のファイルから生成する疑似ファイルでは，演算時間のばらつきが再現できていないためである．OFP では京コンピュータと比較してより演算時間のばらつきが大きく，推定結果に大きな差がでたと考えられる．

また，CPU シミュレータと通信シミュレータを併用してアプリケーションの全体性能推定を行うために，実際に通信を行うことができない CPU シミュレータ用に通信部をトレース取得時に同時に取得した通信バッファのログと置き換えることを考えた．CPU シミュレータ実行時は，通信呼び出しはデスク等に保存されたログの読み込みに置き換えられる．このような置き換えを行った場合の CPU シミュレータの精度について検証した．扱うデータがキャッシュのサイズに近い場合，読み込みによるキャッシュの汚れなどから，シミュレーション結果が不正確になることがわかった．逆に各プロセスのデータが十分に大きい場合，通信をデータ読み込みに置き換えても，演算性能推定結果には大きな影響はない．

３．学際共同利用が果たした役割と意義

将来システムのコデザインのための通信性能推定手法として提案された SCAMP について，汎用性を検証し，課題をフィードバックすることができた.

４．今後の展望

CPU シミュレータと通信シミュレータの融合については年度内に実装することができなかったため今後の課題としたい.

５．成果発表

(1) 学術論文

(2) 学会発表

(3) その他

・Jeffly Vetter, Miwako Tsuji, S. Moore and Mitsuhisa Sato,"Exascale-codesign and performance modeling tools",DOE/MEXT Workshop 2019,2019.05.29-30,Gleacher Center University of Chicago Chicago IL USA,2019,

・辻 美和子 and 佐藤 三久,"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",情報処理学会研究報告,2020-HPC-173,On Line,-,-,情報処理学会,2020,

使用計算機	使用計算機に○	配分リソース※	
		当初配分	追加配分
Cygnus			
Oakforest-PACS	○	9500	
※配分リソースについてはノード時間積をご記入ください。			