大規模 CFD 解析におけるポスト処理効率化のためのデータステージング技術に関する研究

報告書番号：R19JACA42
利用分野：JSS2 大学共同利用
URL：https://www.jss.jaxa.jp/ar/j2019/11419/

責任者

高橋慧智，奈良先端科学技術大学院大学 先端科学技術研究科 情報科学領域

問い合せ先

高橋慧智(keichi@is.naist.jp)

メンバ

高橋 慧智

事業概要

CFD シミュレーションの大規模化にともない，従来のように全てのシミュレーション結果を並列ファイルシステムに保存し，シミュレーション終了後にポスト処理を実行することはストレージ容量や IO 帯域幅の制約により困難になると予想される．そのため，シミュレーション実行中にポスト処理プログラムへ計算結果をリアルタイムに転送する，データステージング技術が注目されている．本研究では，JSS2 に代表される HPC 環境におけるデータステージング技術の実現可能性を検討し，データステージングミドルウェア，および，HPC 環境に求められる要件を分析する．

JAXA スーパーコンピュータを使用する理由と利点

JSS2 はアーキテクチャの異なる計算用主システム（SORA-MA）とプリポスト処理用副システム（SORA-PP）から構成され，2 システム間で通信が可能であるため．

今年度の成果

今年度は，昨年度に概念実証した SORA-MA および SORA-PP を跨ぐデータステージングの実用性向上に取組んだ．昨年度は，米オークリッジ国立研究所が開発を推進しているデータステージング用ミドルウェア ADIOS2 を JSS2 に移植し，MA/PP 間のデータステージングを実現した．この際，MA の IO ノード上で通信ブリッジを動作させ，MA から PP への通信を中継させることによってデータステージングを実現した．

しかし，数千プロセスに及ぶ大規模な CFD 解析に対してデータステージングを適用した際，次の 2 点の問題がある：(1) ステージング通信のスループット：MA/PP の両サブシステムにおいて，ステージング通信のスループットが相互結合網の帯域幅の 10%未満に留まる (2) 通信ブリッジのメモリ消費：MA の BIO/GIO ノード上で動作させる通信ブリッジのメモリ消費量が，送信データ量の 3～4 倍

に及ぶ．これら 2 点の問題は，大規模な CFD 解析に対するデータステージングの適用の障壁となるため，今年度ではこれらの問題の解決に取組んだ．

（1）については，ADIOS2 の通信エンジン SST が TCP/IP 通信を使用することに起因すると推測した．事実，通信ベンチマークを用いて TCP/IP と RDMA 通信の性能を比較したところ，TCP/IP 通信の方が RDMA 通信に比べスループットが低かった．(2) については，ヒーププロファイラ Massif を使用し通信ブリッジのメモリ確保・開放を解析したところ，SST エンジンが複数の重複する通信バッファを確保すること，また，通信バッファが動的拡張される際のメモリ再確保によりメモリ消費が増大していることが明らかになった．

これらの問題は，ADIOS2 の開発中の新たな通信エンジン SSC により解決できると期待できる．SSC はバックエンドに MPI を使用するため，RDMA 通信を利用する．また，SST よりも使用する通信バッファの数が少ないため，メモリ消費量が少ない．以上の分析を踏まえ，SSC エンジンを MA/PP の両システムに移植するとともに，正常に動作することを確認した．今後は SSC エンジンの性能評価を実施し，問題（1）および（2）を解決できたか確認する．

成果の公表

-査読なし論文

堤誠司，藤田直行，伊藤浩之，大日向大地，井上敬介，松村洋祐，高橋慧智，Greg Eisenhauer, Norbert Podhorszki, Scott Klasky, "In Situ/In Transit アプローチを 用いた大規模数値解析におけるポスト処理効率化"，第 33 回数値流体力学シンポジウム．

-口頭発表

Seiji Tsutsumi, Naoyuki Fujita, Hiroyuki Ito, Daichi Obinata, Keisuke Inoue, Yosuke Matsumura, Keichi Takahashi, Greg Eisenhauer, Norbert Podhorszki, Scott Klasky, "In Situ and In Transit Visualization for Numerical Simulations in HPC", In Situ Infrastructures for Enabling Extreme-scale Analysis and Visualization (ISAV 2019) .

JSS2 利用状況

計算情報

プロセス並列手法	MPI
スレッド並列手法	非該当
プロセス並列数	1 - 128
1 ケースあたりの経過時間	5 分

● **利用量**

総資源に占める利用割合※1 (%)：0.00

内訳

計算資源		
計算システム名	コア時間(コア・h)	資源の利用割合※2 (%)
SORA-MA	11,189.95	0.00
SORA-PP	275.58	0.00
SORA-LM	0.00	0.00
SORA-TPP	0.00	0.00

ファイルシステム資源		
ファイルシステム名	ストレージ割当量(GiB)	資源の利用割合※2 (%)
/home	9.54	0.01
/data	95.37	0.00
/ltmp	1,953.13	0.17

アーカイバ資源		
アーカイバシステム名	利用量(TiB)	資源の利用割合※2 (%)
J-SPACE	0.00	0.00

※1 総資源に占める利用割合：3 つの資源(計算,ファイルシステム,アーカイバ)の利用割合の加重平均
※2 資源の利用割合：対象資源一年間の総利用量に対する利用割合