様式 C-19

科学研究費補助金研究成果報告書

平成 21 年 4 月 1 日現在

研究種目：若手研究（B）
研究期間：2006～2008
課題番号：18700074
研究課題名（和文）　可搬性のあるデータ形式をサポートする大規模並列入出力システムに関する研究
研究課題名（英文）　Study of a Parallel I/O System with a Portable Data Format for Data-Intensive Applications

研究代表者
辻田　祐一（TSUJITA YUICHI）
近畿大学・工学部・講師
研究者番号：70360435

研究成果の概要：
可搬性のあるデータ形式をサポートする大規模並列入出力を実現するため、異機種計算機間通信ライブラリ Stampi をベースに、可搬性のあるデータ形式をサポートする並列入出力インタフェースである Parallel netCDF（以下、PnetCDF）を計算機間でも透過的に利用可能にする研究開発を行った。その結果、計算機間で並列入出力を利用する際に生じる性能低下を低く抑える改善を行い、大規模なデータを可搬性のあるデータ形式で効率的に入出力を行うことが可能になった。

交付額

（金額単位：円）

	直接経費	間接経費	合　計
2006 年度	2,800,000	0	2,800,000
2007 年度	500,000	0	500,000
2008 年度	500,000	150,000	650,000
年度			
年度			
総　計	3,800,000	150,000	3,950,000

研究分野：総合領域
科研費の分科・細目：情報学、計算機システム・ネットワーク
キーワード：ネットワークコンピューティング、並列入出力、PC クラスタ

１．研究開始当初の背景

計算機シミュレーションにおいて、並列処理は処理時間の短縮だけでなく、計算規模の大規模化を可能にし、且つ、より精度の高い計算を支援する重要な基盤技術となってきている。その中で、並列処理を実現する為の共通のインタフェース仕様を定めた Message Passing Interface（以下、MPI）では、複数のプロセス間の通信を容易に実現する機能を提供しており、広く用いられている。計算アプリケーションの規模が大きくなるにつれ、扱うデータ量も増大し、高速にデータの入出力を行う並列入出力インタフェースである MPI-I/O が MPI の中で定められている。しかしながら、このインタフェースにおいては、データ形式の決定はユーザに委ねられており、データの可搬性に乏しい状況が続いていた。そこで、データの可搬性を高めるインタフェース仕様がいくつか提案され、その一つとして、netCDF がある。netCDF は逐次的な入出力のみのサポートのため、このデータ形式を保ちつつ、並列入出力を実現した PnetCDF が提案された。PnetCDF では、並列

入出力のために MPI-I/O インタフェースを下層レイヤに配置している。計算機内の並列入出力は可能だが、出力されたデータをさらに別の計算機間で可視化を行う場合など、計算機間の並列入出力は利用できない状況であった。

２．研究の目的

近年、可搬性を保つデータ形式を策定する試みが特に計算アプリケーション分野でなされており、データを相互利用出来ることには対応しているが、異なる計算機間で動的に入出力を行うことは出来ない。そこで、可搬性を保つデータ形式で、異なる計算機間でも自由に、且つ透過的に入出力操作が可能になる実装を行う。さらに PC クラスタの並列ファイルシステムを用いた大規模並列入出力機能を実現する。実装したシステムの性能評価を行い、問題点を洗い出し、実装方法の見直しや新しい機能の追加など、更なる機能向上を狙う。

３．研究の方法

申請者らは、異なる計算機間でも動的に MPI-I/O インタフェースによる並列入出力が可能になる Stampi-I/O を開発している。Stampi-I/O は、

(1) 異なる計算機間でも MPI インタフェースによるプロセス間通信を利用可能
(2) MPI-2 仕様に準拠した MPI-I/O 機能を計算機間で利用可能
(3) MPI 通信並びに MPI-I/O 機能において、いくつかの派生データ型による操作をサポート

などの特徴がある。

一方、本研究の拡張実装を行う対象である PnetCDF ライブラリは、

(1) 計算アプリケーション向けに可搬性を持った入出力インタフェースをサポート
(2) 下層レイヤの並列入出力部分に MPI-I/O インタフェースを含む MPI ライブラリを利用している。以上の特徴を利用
(3) 計算アプリケーションに特徴的な飛び飛びのデータアクセスパターンは、MPI-2 仕様における派生データ型による MPI-I/O 機能で並列入出力を実現

などの特徴を持つ。

以上の特徴を踏まえて、PnetCDF の下層レイヤ部分の MPI 関数を Stampi の MPI 関数に置き換えることにより、PnetCDF が提供する可搬性を保つデータ形式をサポートする入出力インタフェースでも計算機間並列入出力機能を利用することを可能にする。また、計算機間の並列入出力という、データ通信上不利な構成において、なるべく性能の低下を招かないような効率的な入出力機能の実現を目指す。

４．研究成果

1 年目の成果としては、まず PnetCDF の基礎的な入出力インタフェースである逐次型入出力関数に関して、計算機間の入出力機能の設計・実装を手掛けた。さらに集団型入出力の実装も行ったが、この集団型入出力では、各プロセスの担当するデータ領域はそれぞれ連続的なものとなっており、入出力パターンは限定されていた。また、この実装の性能評価を通して、計算機間のネットワーク通信性能に大きく影響することを確認した。同じネットワーク上で計算機同士がつながっている場合だけでなく、通信遅延の大きい WAN などで接続されることもあり得る。よって、この部分の改善策を次年度に試みることにした。

2 年目は、1 年目の成果を踏まえて、ネットワーク遅延の影響を詳細に調査した。一般に、標準的なネットワークでは、通信遅延が大きくなるにつれて、発揮できる通信性能は低下する。そこで、（１）ネットワーク通信に用いるソケットバッファの容量を増やすことや（２）ネットワーク遅延を配慮した効率的なデータ転送を実現する通信基盤（PSPacer）の導入を行い、性能向上を試みた。その結果、通信遅延がミリ秒オーダーになる場合、（１）と（２）をバランス良く組み合わせることで、性能向上が図れることを確認した。図 1 では、通信遅延が 50 ミリ秒の時に、ソケットバッファの大きさを変えながら、PSPacer 無し／有り（図中の PSP）でのリモート書き込みに要する時間を示している。

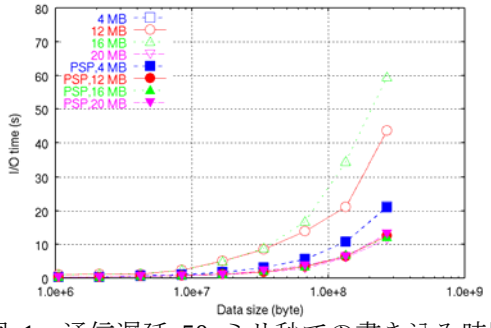

図 1：通信遅延 50 ミリ秒での書き込み時間（MPI プロセス数=4）

PSPacer がある方が、少ないソケットバッファの大きさでも時間を短縮できることが分かる。ソケットバッファを多く用意するには、それだけ計算資源を多く利用するため、出来

るだけ少ない大きさで時間を短縮できる PSPacer 有りの手法が有用であることが分かる。

3 年目は、これまでの成果を踏まえ、より実アプリケーションに適用できるための実装と評価を行った。具体的には、1 年目に行った実装では、非連続的なデータパターンには未対応であったため、Stampi-I/O による派生データ型を用いた計算機間並列入出力機能を利用した PnetCDF インタフェースを用いた入出力の設計・実装を行った。性能評価として C 言語によるプログラムで 3 次元配列を用意し、左側の添え字から順に x、y、z 軸に対応させてデータを定義した。データの分割数は MPI プロセス数と同じとし、3 つの軸それぞれでデータを分割し、入出力性能を計測した。図 2 に集団型書き込み操作における分割軸による性能評価の結果を示す。

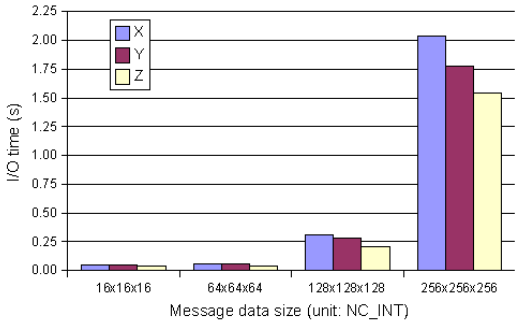

図 2：分割軸に対する計算機間集団型書き込み操作の処理時間（MPI プロセス数=4）

z 軸で分割する場合、各プロセスでアクセスするデータアクセス領域が連続になるため、最も処理時間が短くなる。一方、x 軸での分割では、各プロセスにおいて、最も複雑な飛び飛びのデータアクセス領域となり、処理時間が長くなっている。分割する軸による処理時間の増加の原因を調査するために、PnetCDF 関数内で使用している集団型 MPI-I/O 関数に要する時間を計測した。図 3 に集団型 MPI-I/O 関数による書き込み操作に要した時間を示す。

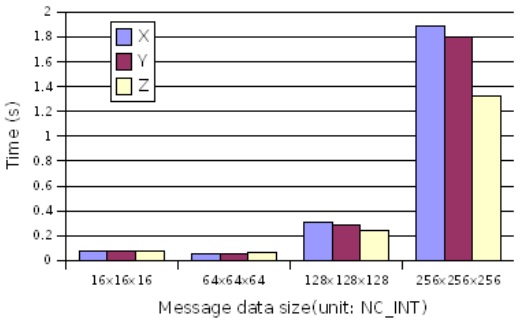

図 3：集団型 MPI-I/O 関数（書き込み）を計算機間で実行するのに要した時間（MPI プロセス数=4）

この結果、MPI-I/O レイヤでも、分割軸に依存した処理時間の差が確認できた。そこで、同じ MPI-I/O 関数について、ローカル入出力に要する時間を計測した。図 4 に集団型 MPI-I/O 関数による書き込み操作に要した時間を示す。

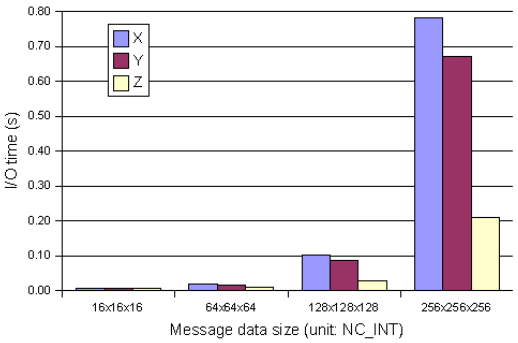

図 4：集団型 MPI-I/O 関数（書き込み）によるローカル入出力に要した時間（MPI プロセス数=4）

この図から、ローカル入出力においても、分割する軸に依存して、計算機間入出力で生じているのと同程度の処理時間の差が生じていることが分かった。

以上の結果、計算機間の入出力において、分割する軸に依存して実行時間が増加する原因として、ローカル入出力における派生データ型による集団型入出力によるものであることが分かった。つまり、データ分割方法により、非連続なデータアクセスパターンは異なるが、本研究での実装では、計算機間のデータ通信には、このデータパターンの影響は殆ど無く、自計算機及びリモート計算機内部でのデータ交換に要する処理に大きく起因するため、本実装に起因する大幅な性能低下は無いことが確認された。

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 6 件）

① Yuichi Tsujita, “Implementing A Parallel NetCDF Interface for Seamless Remote I/O Using Multi-dimensional Data,” High Performance Computing for Computational Science – VECPAR 2008, Revised Selected Papers, Lecture Notes in Computer Science, Vol. 5336, pp. 69-82, 2008, 査読有り

② Yuichi Tsujita, “Feasibility Study of Effective Remote I/O Using a Parallel NetCDF Interface in a Long-Latency Network,” Proceedings of the International Multi-Conference of Engineers and Computer Scientists 2008, Vol. 1, pp. 242-247, 2008, 査読有り

③ Yuichi Tsujita, “An Empirical Study of Optimization in Seamless Remote MPI-I/O for Long Latency Network,” Lecture Notes in Computer Science, Vol. 4757, pp. 389-390, 2007, 査読有り
④ Yuichi Tsujita, “Supporting Seamless Remote I/O Using a Parallel NetCDF Interface,” Distributed and Parallel Systems - From Cluster to Grid Computing, 81-90, 2007, 査読有り
⑤ Yuichi Tsujita, “Implementation of a Parallel NetCDF Interface for Seamless Collective Remote I/O,” Proceedings of IADIS International Conference on Applied Computing 2007, 732-736, 2007, 査読有り
⑥ Yuichi Tsujita, “Effective Seamless Remote MPI-I/O with Derived Data Types Using PVFS2,”, Lecture Notes in Computer Science, Vol. 4192, pp. 230-237, 2006, 査読有り

〔学会発表〕（計８件）
① Yuichi Tsujita, “Implementing A Parallel NetCDF Interface for Seamless Remote I/O Using Multi-Dimensional Data,” VECPAR2008, Toulouse, France, June 25-27, 2008
② Yuichi Tsujita, “Feasibility Study of Effective Remote I/O Using a Parallel NetCDF Interface in a Long-Latency Network,” The International Multi-Conference of Engineers and Computer Scientists 2008, Hong-King, China, March 19-21, 2008
③ Yuichi Tsujita, “An Empirical Study of Optimization in Seamless Remote MPI-I/O for Long Latency Network,” 14th European PVM/MPI User’s Group Meeting, Paris, France, September 30-October 3, 2007
④ Yuichi Tsujita, “Implementation of a Parallel NetCDF Interface for Seamless Collective Remote I/O,” IADIS International Conference on Applied Computing 2007, Salamanca, Spain, February 18-20, 2007
⑤ 辻田　祐一, “Parallel netCDFインタフェースによる計算機間集団型入出力機能の実装,” 情報処理学会　第108回HPC研究会, 京都大学, 2006年10月6日
⑥ Yuichi Tsujita, “Supporting Seamless Remote I/O Using A Parallel NetCDF Interface,” 6th Austrian-Hungarian Workshop on Distributed and Parallel Systems (DAPSYS2006), Innsbruck, Austria, September 21-23, 2006
⑦ Yuichi Tsujita, “Implementation of a Parallel NetCDF Interface for Seamless Collective Remote I/O,” 13th European PVM/MPI User’s Group Meeting, *Late and Breaking Results*, Bonn, Germany, September 17-20, 2006
⑧ Yuichi Tsujita, “Effective Seamless Remote MPI-I/O with Derived Data Types Using PVFS2,” 13th European PVM/MPI User’s Group Meeting, Bonn, Germany, September 17-20, 2006

６．研究組織
(1)研究代表者
辻田　祐一（TSUJITA　YUICHI）
近畿大学・工学部・講師
研究者番号：70360435

(2)研究分担者
なし　　（　　）

研究者番号：

(3)連携研究者
なし　　（　　）

研究者番号：