Vol.2013-HPC-139 No.9
2013/5/29

Gfarm/Pwrake による NICT サイエンスクラウドの並列分散処理技法

村田健史[†1]　渡邉英伸[†1]　山本和憲[†1]　久保田康文[†1]
建部修見[†2]　田中昌宏[†2]　深沢圭一郎[†3]　木村映善[†4]
鵜川健太郎[†5]　村永 和哉[†5]　鈴木 豊[†5]　磯田総子[†6]

NICT サイエンスクラウドは，情報通信研究機構が科学研究目的のために構築したクラウドシステムであり，特にビッグデータサイエンスを主対象の一つとしている．10TB を超えるデータ処理（例えば大規模可視化）や 100TB を超えるデータ検索（たとえば特徴検出）などの大規模データ処理は，これからのデータ指向型科学研究では重要である．近年の CPU 処理速度の向上に伴いこれらのデータ処理は処理時間とデータ読み書き（I/O）時間が同規模となることがある．そのため，レガシーな HPC 型数値計算環境ではなく I/O の高速化がなされているクラウド環境が有効となる．本稿では，NICT サイエンスクラウドにおいて分散ストレージシステム（Gfarm）と Gfarm のためのワークフローシステム（Pwrake）を用いた並列分散処理実験結果について報告する．

A Parallel Processing Technique on the NICT Science Cloud via Gfarm/Pwrake

KEN T. MURATA[†1]　HIDENOBU WATANABE[†1]
KAZUNORI YAMAMOTO[†1]　YASUBUMI KUBOTA[†1]
OSAMU TATEBE[†2]　MASAHIRO TANAKA[†2]
KEIICHIRO FUKAZAWA[†3]　EIZEN KIMURA[†4]
KENTARO UKAWA[†5]　KAZUYA MURANAGA[†5]　YUTAKA SUZUKI[†5]
FUSAKO ISODA[†6]

For data intensive science on cloud systems, we need development of techniques for DIC (Data-Intensive Computing) as well as HTC (High-Through-put Computing), MTC (Many-Task Computing), and HPC (High-Performance Computing). The DIC is a new concept of large-scale data processing paying attentions to data distribution, data-parallel execution, and harnessing data locality by scheduling of computations close to the data. As the data file size is getting larger, I/O time to read and/or write data is not negligible compared with data processing time. We herein develop a DIC technique on a science cloud using Gfarm/Pwrak. The Gfarm/Pwrake has been developed as an integrated system of both distributed file system and parallel data processing system. With identifying file system nodes (FSN) and processing client node (CN) and giving higher priority to process files on the local disk than on remote disks, we succeeded in progress of total performance in processing large-scale data files.

1. はじめに

科学研究の分野には，3 つの研究手法があると言われてきた．第 1 の手法は理論研究手法，第 2 の手法は観測や実験による研究である[3]. 19 世紀までに始まったこれらの伝統的な研究手法に加えて，20 世紀に計算機シミュレーション技法が登場した（第 3 の手法）．21 世紀に入り，これら 3 つの研究手法に加えて，第 4 の手法としてデータ指向型研究手法（The Fourth Paradigm: Data-Intensive Science）が 提言されている[1]．インフォマティクスは，データ指向型科学において データ(特に大規模データや複雑で多種多様なデータ）を解析する技術を示す．

インフォマティクスが研究手法として提言されてきた背景には，科学研究で扱うデータのほとんどがデジタル化された（すなわち,コンピュータ上で処理することができる）ことと，データサイズや種類が大規模化・多様化していることが挙げられる． 科学データは量・種類とも増え続け，多くの研究者は「一生かけても解析できない程の量と種類のデータ」に埋もれつつある．いわゆる，科学研究分野における BigData 問題である．インフォマティクスへの期待の一つは，コンピュータのデータ処理能力を存分に活用して， これらの BigData 問題を解決することである．

実験的アプローチが様々な実験装置や観測装置を用い，数値シミュレーションがスーパーコンピュータを活用するのと同様に，データ指向型科学研究やインフォマティクス技術のためにも基盤となるインフラストラクチャが必要で

†1 情報通信研究機構
National Institute of Information and Communications Technology
†2 筑波大学計算科学研究センター
Center for Computational Sciences, University of Tsukuba
†3 九州大学情報基盤研究開発センター
Research Institute for Information Technology, Kyushu University
†4 愛媛大学 大学院医学系研究科
Department of Medical Informatics Ehime University Graduate School of Medicine
†5 株式会社セック
Systems Engineering Consultants Co., LTD
†6 株式会社サイエンス・サービス
Science Service Inc.

ある．筆者らは，サイエンスクラウドがそのインフラストラクチャであると提唱している[4]．サイエンスクラウドは，2008 年ごろにイリノイ大学などにおいて提案された概念であり，機能や有効性の議論が様々に行われている．

データ指向型科学研究において，10TB を超えるデータ処理（例えば大規模可視化）や 100TB を超えるデータ検索（たとえば特徴検出）などの大規模データ処理は，主要な技術課題の一つである．近年の CPU 処理速度の向上に伴いこれらのデータ処理は処理時間とデータ読み書き（I/O）時間が同規模となることがある．そのため，レガシーな HPC 型数値計算環境ではなく I/O の高速化がなされているクラウド環境が有効となる．本稿では，NICT サイエンスクラウドにおいて分散ストレージシステム（Gfarm）と Gfarm のためのワークフローシステム（Pwrake）を用いた並列分散処理実験結果について報告する．

2. NICT サイエンスクラウド

NICT サイエンスクラウド（以下，サイエンスクラウドと記述することもある）は，情報通信研究機構（NICT）が構築した科学研究専用のクラウドシステムである[4]． サイエンスクラウドは，高速ネットワークバックボーンである JGN-X 上に分散型クラウドシステムとして構築されている． 大規模分散ストレージ，並列分散処理環境，スパコン，大規模可視化環境等のリソースから構成される統合型のデータ指向科学研究の基盤として設計されている（図 1）．

このような大規模・広域サイエンスクラウドにおいてデータ処理を行う場合に高いスケーラビリティーを達成することは容易ではない．負荷分散はもちろんのこと，ディスク I/O 分散，ネットワークスループット分散，データ領域分散など，多くの要素からなる計算効率の最適化を行わねばならないからである．特に，クラウドのようなヘテロ環境下での計算負荷の効率的な配分は，スーパーコンピュータに代表される HPC 系計算環境の負荷分散とは異なるクラウド技術が必要である．本研究では，広域分散型のヘテロ計算環境である NICT サイエンスクラウドにおいて高いスケーラビリティーを達成するための基本実験として，ディスク I/O 時間が計算時間と比較して無視できない場合の並列分散処理性能の基本的な実験を行う．具体的には，スーパーコンピュータによる時系列数値シミュレーションデータ（1TB 超）のデータファイルを 6 台のクラスタ計算機環境で可視化処理する．分散ファイルシステム Gfarm（バージョン 2.5.8）と Gfarm のためのワークフローシステム Pwrake[2]を用いることで高速で高い並列化効率のデータ処理を目指す．

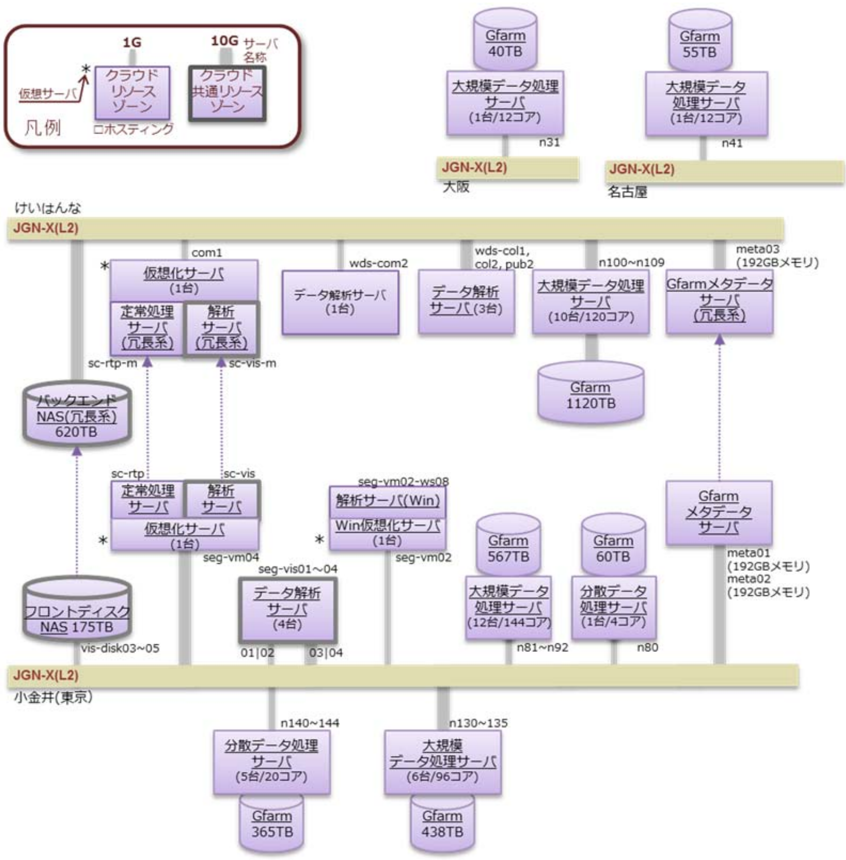

図 1　NICT サイエンスクラウド構成図（一部計画を含む）

Figure 1　Construction of the NICT Science Cloud.

3. 実験

3.1 実験環境

図2に,本実験の計算機環境を示す.本実験は図1のNICTサイエンスクラウドとは独立の閉じたネットワーク系内に構築した.6台のノードはGfarmのファイルシステムノード(FSN)とクライアントノード(CN)を兼ねている.(以下では,単にノードと示す.)すべてのノードとGfarmメタデータサーバは DELL 社製 PowerConnect 6224 により10GbEで接続されている.表1は,図2の実験システムの計算機スペックである.

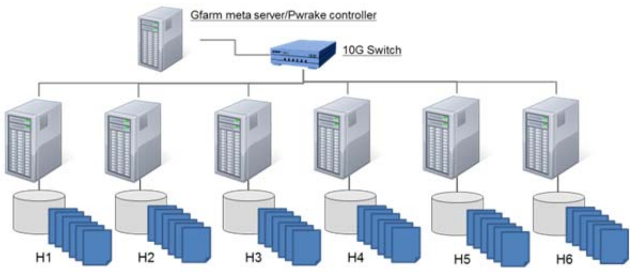

図 2 本実験システム

Figure 2 The computer system for the experiment.

表 1 実験環境の計算機スペック

Table 1 Spec. of computers for the present experiments.

	Spec.
CPU number/node	8
CPU	Intel Xeon X5550@2.67GHz
Main Memory	144GB
OS	openSUSE 11.1 (x86_64)
HDD	SATA 3 x4 (RAID5)
HDD (read)	371 MB/sec
HDD (write)	137MB/sec
NIC	10GbE

本実験のデータ処理の対象となるデータを表2および図3に示す.データは地球磁気圏を対象としたグローバルMHDシミュレーションにより生成された数値データファイルである.数値シミュレーションは時系列に計算されるため,データファイルは時系列に出力される.本実験では,数値シミュレーション終了後に,出力された782の数値データファイルを6台のノード(FSN)が管理するGfarm分散ストレージ上に保存し,同じ6台のノード(CN)により可視化処理を行う.可視化処理には,NICTが開発したバーチャルオーロラツール(3次元可視化ツール)を用いた.なお,本実験の可視化処理は時間ステップ間の相関はないため,実験においてはシミュレーションデータの時刻(すなわちデータファイルの番号)を無視して可視化を行うことができる.

Gfarmは分散環境においてデータファイルを管理する分散ファイルシステムであり,各ファイルの複製を作成することによりファイルの冗長性を高めることができる.同時に,本研究のような大規模データの分散処理においては,データファイルを処理するCNがFSNを兼ねている場合には,CN自身がFSNとして管理するデータファイルを読み込むことでI/O処理の高速化が期待できる.本研究では,実験のため,782のすべてのデータファイルをすべてのFSNに配置した.すなわち,最もコストが高いが高速化が期待できるデータ配置を行った.

可視化処理(データ処理)はGfarmのためのワークフローツールであるPwrakeを用いた[2].Pwrakeは,Gfarmが管理するデータファイル処理において,並列I/O処理の効率化のために最適なデータファイルとCN(データ処理ノード)の組み合わせによりデータファイル処理を行う.

表 2 実験対象データファイル

Table 2 Data files for the present experiments.

	Spec.
Number of data files	782
File size	2.2GB/file
Total file size	1.72TB

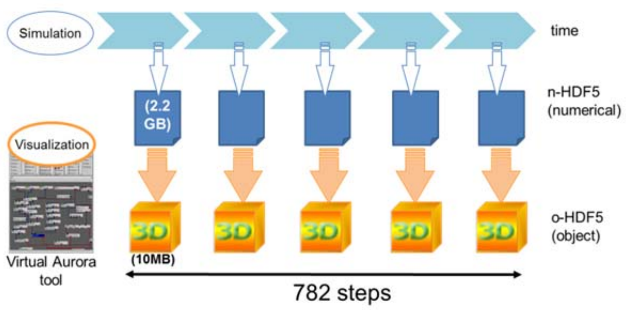

図 3 実験対象データファイル

Figure 3 Data files for the present experiments.

3.2 実験結果

図4は,782ファイルのデータ処理(可視化処理)を各ノードに割り当てて処理を行った結果である.図4の上図は各ノードにおいて6コア(すなわち6プロセス)で処理を行った場合であり,下図は1コアの場合である.横軸はシミュレーションのステップ数を表しており,これは処理を行ったデータファイルの番号に一致する.

図4の縦軸は,各データファイルに対する全処理時間(I/O時間と可視化時間の合計)を示している.棒グラフの下側(赤色)が可視化時間を表し,上側(白色)がI/O時間を示している.

図4より,本実験のデータセットの場合にはどのデータファイルについても可視化処理時間はほぼ一定(約40秒)であることが分かる.これは,782ステップの中では数値

シミュレーションの変化が大きくないため，可視化処理時間も変化が小さいためである．

一方，同図において I/O 時間はばらつきが大きい．このバラつきは，同じ環境で各 CN のデータ処理（可視化）を行うコア数を 1 とした場合（図 4 下図）には見られなかった．すなわち，一台の CN において複数のコア（プロセス）が並列にデータを読み込む場合には，I/O のコンフリクトが発生し，I/O 時間のばらつきが発生すると考えられる．

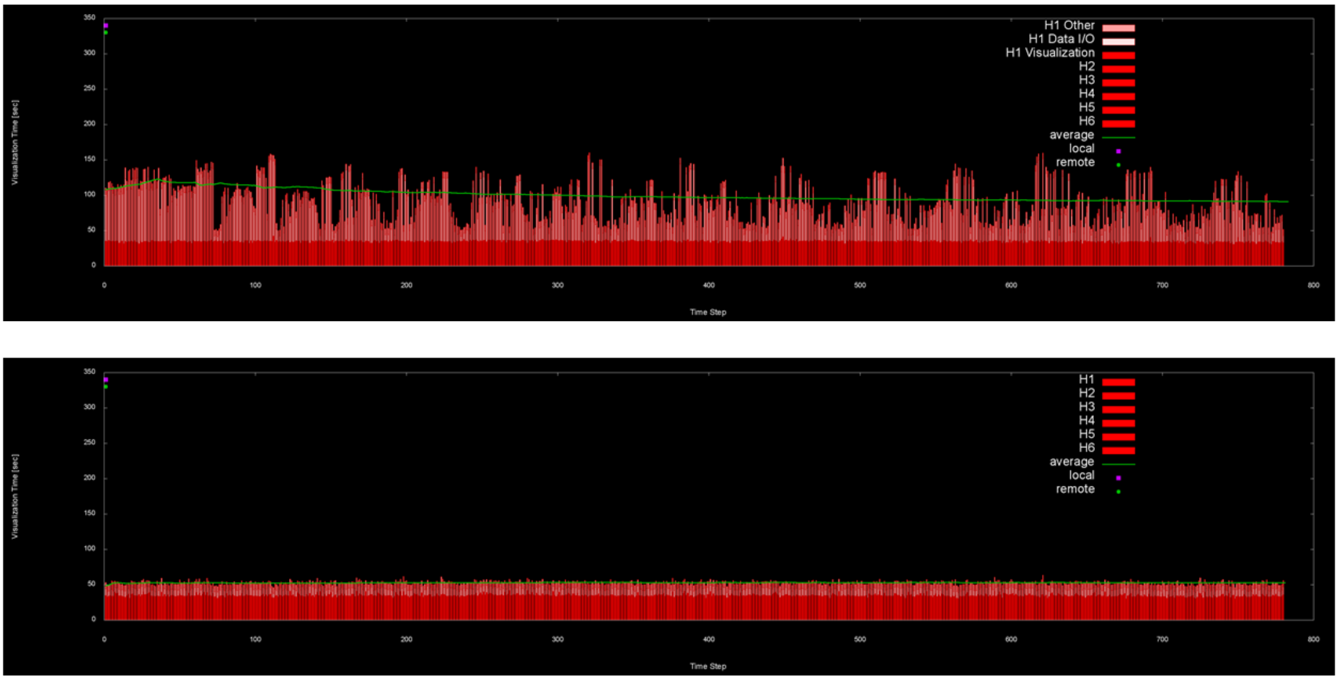

図 4　上図：各 CN（データ処理ノード）6 コアを用いた全処理時間（ステップ毎の処理時間と I/O 時間）．下図：上図で各 CN の処理コア数を 1 とした場合の比較実験結果．（ともに赤はデータ処理時間・白は I/O 時間）

Figure 4　Upper: Processing time and I/O time at each step with 6 cores on each node. Lower: Same result in case with 1 core (process) on each node. Red part: data processing time. White part: data I/O time.

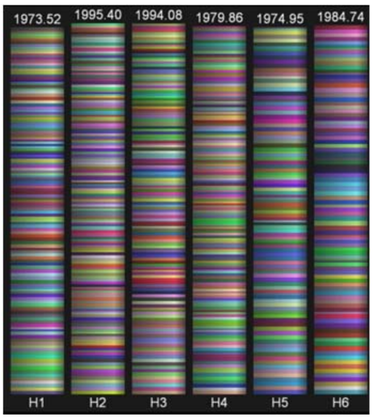

図 5　各 CN（データ処理ノード）6 コアを用いたロードバランス（ノード毎の処理および I/O 時間）：それぞれのブロックの高さがデータファイルごとの処理時間と I/O 時間の和を示す

Figure 5　Load balance between nodes: Total time (data processing time and I/O time) for each data file.

表 3　各ノードの処理結果（図 5）

Table 3　Data processing results on each node.

Node	Core (process)	Step (file) number	Average time (sec.)	Total processing time (sec.)
H1	6	140	84.58	1973.52
H2	6	151	79.29	1995.40
H3	6	142	84.26	1994.08
H4	6	155	76.64	1979.86
H5	6	100	118.50	1974.95
H6	6	95	125.35	1984.74

3.3 考察

本実験では，データ処理（可視化）については，すべてのノード（コア）が常時可視化処理を行っており，また，データファイル間での依存性がない．したがって，各 CN（コア）は FIFO 的に順次データの可視化処理を行っており，データ処理についてはほぼ 100%の並列化効率が得られている．しかし，図 5 および表 3 によると各ノードの全データ処理時間（または処理データファイル数）にはばら

つきがあり，完全な負荷分散が達成できていない．可視化時間はほぼ一定であることから，高い負荷分散によるデータ処理の高速化を達成するためにはデータファイルの読み込み時間（I/O 時間）の高速化が求められる．

I/O 高速化にはいくつか方法がある．まず，データ処理を行う CN が FSN となり，自らが管理するデータファイルを優先的に処理することで I/O 時間を短縮できる[5]．本実験では，前述のとおり，すべての FSN と CN が一致しており，すべての FSN が対象となる全データファイルを有するため，高コストであるが I/O 分散としては理想的な状態である．もう一つは，同じ FSN 内での I/O の分散である．本研究では，一つの CN（すなわち FSN）で 6 コア（6 プロセス）の処理を並列に行った．そのために，1 コア/ノードの場合と比較して I/O 時間が増加した．Gfarm/Pwrake 処理において処理コア数と I/O 時間の関係は明らかになっておらず，最適化を行うためには今後の研究が必要である．

本実験は 6 台のクラスタにより行ったが，Gfarm のワークフローシステムである Pwrake を活用することで，本研究結果をヘテロ環境下での分散処理に拡張することができる．図 6 に本実験と同様の実験をヘテロ環境において行った実験結果を示す．(紙面の都合上，実験の詳細は別稿で述べる．）図 5 のフォーマットは，図 4 と同じである．実験を行った計算機環境は図 2 のシステム図の 6 台に加えて，さらに 14 台のノード（FSN 兼 CN）を追加している（図 7）．14 台の追加ノードは低スペック（8 台）および中スペック（6 台）から構成される．（表 1 のノードは高スペックとなる．）

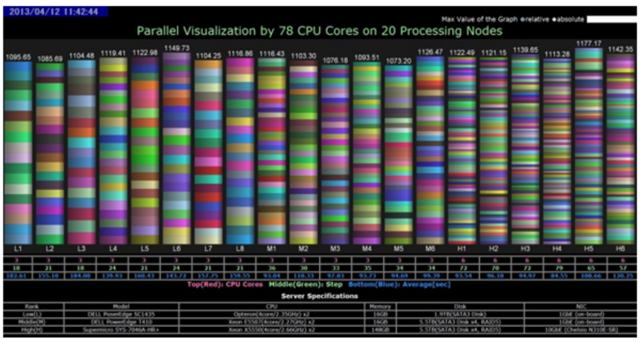

図 6　比較実験結果①ヘテロ環境でのロードバランス（ノード毎の処理および I/O 時間）

Figure 6　Reference Experiment: Load balance between nodes.

図 6 の結果については本稿では詳細を省略するが，この比較実験では高い負荷バランスが達成できていることが分かる．また，データファイルごとの処理時間（図中のブロックの高さ）が各スペックのクラスタにより異なっており，高いスペックの CN ほど多くのデータファイルを処理したことが分かる．このような Gfarm/Pwrake が有する計算機の処理能力に合わせた処理を割り当てるタスク機能をヘテロ計算機環境で活用することで，ディスク I/O の高速化と負荷分散の高効率化を同時に達成できることが示唆された．

なお，図 4 より，本実験のデータセットではデータ処理時間（赤）と I/O 時間（白）の時間は同程度であることが分かる．近年，CPU 高速化に伴う数値計算の大規模化とデータ処理の高速化により，ポスト処理においてはデータ処理時間の相対的な短縮が実現している．すなわち，今後，サイエンスクラウドにおける分散データ処理を考える場合には，データ処理の並列化だけではなく，データ I/O 時間の短縮または並列化が重要であること分かる．

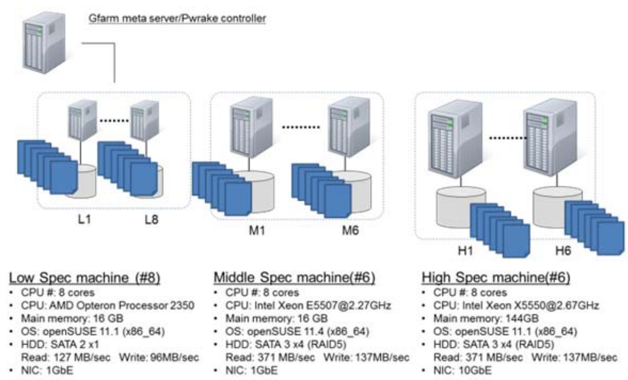

図 7　比較実験システム

Figure 7　The computer system for the reference experiment.

4. おわりに

データ指向型科学は，これまでに発展してきた実験科学，理論科学，計算科学に対して，科学的発見のための第 4 の柱と言われている[1]. IT ビジネス分野の中心となりつつあるビッグデータという概念は，科学研究分野でも適用され始めている．実験科学を支えるインフラは実験装置や観測装置であり，計算科学を行う基盤となるのがスーパーコンピュータである．それらに対して，データ指向型科学を支えるインフラとして提案するのが，科学研究専用のクラウド（サイエンスクラウド）であると提唱している[4].

データ指向型科学では，複雑さと量の両面において指数関数的に増大している科学データセット処理が重要である．大規模なデータセットを処理し，視覚化し，解析・解釈するために，高度な情報処理環境へのニーズが高まっている．この 10 年で，データ保存のためのデータストレージは大規模化し，データ処理のための中央処理装置（CPU）処理速度も高速化している．しかし，それらの基盤環境だけではデータ指向型科学研究を推進することはできない．

データ指向型科学という概念が提唱され，我が国では京コンピュータの利用も始まり，TB または PB 以上の大規模科学データが研究対象となっている．一方，これらのような特別なプロジェクトではなく，大学や研究機関の小〜中規模研究プロジェクトにおいてもデータの大規模化は始ま

っている．10TB を超えるデータ処理（例えば大規模可視化）や 100TB を超えるデータ検索（たとえば特徴検出）などは，プロジェクト規模によらずこれからのデータ指向型科学研究では重要である．

これらのデータ処理は, 処理時間とデータ読み書き（I/O）時間が同規模となることがあるため, これまでの HPC 型数値計算環境ではなく I/O の高速化がなされているクラウド環境が有効となる．本稿では，NICT サイエンスクラウドにおいて分散ストレージシステム（Gfarm）[5]と Gfarm のためのワークフローシステム（Pwrake）[2]を用いた並列分散処理実験を行った．その結果，TB スケールの大規模・大量のデータファイルをクラウド環境下で並列処理する場合には，(1)データファイルの配置，(2)I/O の分散化を考慮した最適化が必要であることが分かった．

謝辞　本論文の研究は情報通信研究機構の NICT サイエンスクラウドを用いて行われました．

参考文献

1) EditEd by Tony Hey, STewarT TanSley, and KriSTin Tolle, The Fourth Paradigm: Data-Intensive Scientific Discovery, ISBN 978-0-9825442-0-4, 2009.
2) 田中昌宏, 建部修見, 並列分散ワークフローシステム Pwrake による大規模データ処理，宇宙航空研究開発機構研究開発報告（JAXA Research and development report） JAXA-RR-11-007, pp.67-76, 2012-03-30.
http://office.microsoft.com/ja-jp/word-help/CH010097020.aspx
3) 松本　紘著，宇宙開拓とコンピュータ，共立出版，情報フロンティアシリーズ，情報処理学会編，1996.
4) Murata, K., T, Watari, S., Nagatsuma, T., Kunitake, M., Watanabe, H., Yamamoto, K., Kubota, Y., Kato, H., Tsugawa, T., Ukawa, K., Muranaga, K., Kimura, E., Tatebe, O., Fukazawa, K. and Murayama, Y., A Science Cloud for Data Intensive Sciences, Data Science Journal, Vol. 12, pp. WDS139-WDS146 (2013) .
5) Gfarm File System, ISBN-10: 6133490381, ISBN-13: 978-6133490383.