様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成２５年５月３１日現在

機関番号：８２６２６
研究種目：若手研究（A）
研究期間：2011 ～ 2012
課題番号：２３６８０００４
研究課題名（和文）　大規模ＨＰＣクラスタにおける高性能共有ストレージの性能保証に関する研究
研究課題名（英文）　QoS Guarantees on High Performance Shared Storage Systems for Large-scale HPC Clusters
研究代表者
谷村　勇輔（TANIMURA YUSUKE）
産業技術総合研究所・情報技術研究部門・主任研究員
研究者番号：８０４１５７１０

研究成果の概要（和文）：
予約に基づいて，データ入出力に関わるストレージシステムの各コンポーネントを制御し，データアクセスの性能を保証するストレージ（Papio ストレージ）において，データの書き込み性能を制御する SSD 向けの I/O スケジューリング手法を開発した．その上で，HPC の並列プログラム，データステージングの実行において Papio ストレージにアクセスするためのソフトウェアを開発し，複数のデータアクセスが競合した場合にも，Papio ストレージが個々のアプリケーションに対して必要な I/O 性能を提供できることを評価実験により示し，その有効性を明らかにした．

研究成果の概要（英文）：
We developed an I/O scheduling method which controls write performance to our object-based storage device (OSD) backed by the Solid State Drive. The OSD is used in the Papio storage system, which provides I/O performance guarantees by controlling all storage components based on advanced reservations. Subsequently, we developed software which allows application users to access the Papio storage system from their parallel programs or in their data-staging scenarios. Through our evaluation, we showed that our approach can provide requested I/O performance to each application, under the situations where many applications access the same shared storage system of the HPC cluster.

交付決定額

（金額単位：円）

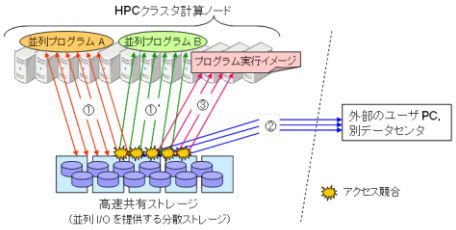
	直接経費	間接経費	合　計
２０１１年度	17,900,000	5,370,000	23,270,000
２０１２年度	4,200,000	1,260,000	5,460,000
年度			
年度			
年度			
総　計	22,100,000	6,630,000	28,730,000

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：ハイパフォーマンスコンピューティング、並列分散ストレージ

１．研究開始当初の背景

グリッドやクラウド技術を用いて，スパコンなどの高性能計算機を複数のユーザ，かつネットワーク越しに遠隔のユーザと共有利用することは科学技術の発展のために重要であり，年々，共有利用化が進んでいる．そ

の一方で，高性能計算機（以下では，HPC（High Performance Computing）クラスタと記す）に付随する，データ入出力や実行プログラムの中間状態を保存するための高速共有ストレージには，並列アプリケーション同士（図１-①および①’）に加えて，外部とのデータ転送（図１-②），内部のデータステージング（図１-③）の各種アクセスの競合が多発し，個々のアクセスにおいて十分な性能が得られない問題が生じている．



① HPC 並列プログラムのデータ入出力・中間データの保存
② 遠隔地のユーザPC，データセンタとの間のデータステージング
③ アプリケーション実行イメージ（OSの仮想化イメージやアプリケーションプログラム，関連ライブラリ等）の展開・保存

図１ HPC クラスタの共有利用において生じるストレージアクセスの競合事例

現状，こうしたストレージアクセスの競合に起因する I/O の性能低下・ボトルネック問題を防ぐ手段はなく，コストをかけて，ピーク性能を大幅に上回るストレージを予め用意しておくか，アプリケーション毎にストレージシステムを分離する等の対応しかできない．また，そのような場合でもデータ配置とアクセスに偏りが生じれば性能低下が起こりうる．そして，近年の HPC クラスタの規模，アプリケーションが扱うデータ量の増大を考慮すると，本問題はますます深刻になっているといえる．

２．研究の目的

本研究課題では，HPC クラスタにおいて同時間帯に実行される並列プログラム，データ転送プログラムの双方が明示的に I/O 性能を予約するコンセプトを採用し，予約に基づいて，各アプリケーションに対し，要求された I/O 性能を提供できる高速共有ストレージの実現を目指す．これにより，１節で述べた「データアクセスの競合による性能低下の問題」を解決し，各種 HPC アプリケーションの安定的，かつ効率的な実行を可能にし，HPC クラスタの共有利用の促進に貢献する．

３．研究の方法

これまでの研究において，研究代表者は，事前予約に基づいてストレージ資源（ディスク等のストレージデバイス，ネットワークやバス等の I/O パス，サーバやクライアント上のバッファスペース等）を適切に割り当てて I/O 制御を行うコンセプトとその基本アーキテクチャを明らかにし，それを実装した Papio ストレージソフトウェアを開発している．本研究課題では，上記の提案コンセプトを HPC アプリケーションに適用し，その有効性を明らかにすることを目指した．

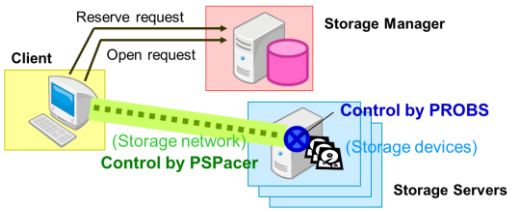


図２ Papio ストレージソフトウェアの概要

まず，図２に示すような Papio の内部で用いられている I/O 性能制御の一部である，PROBS のディスク I/O スケジューリングを拡張し，Solid State Drive（SSD）向けに書き込み性能を制御できる手法の開発に取り組んだ．H23 年度の前半に手法の提案・プロトタイプ実装を完了し，単体での評価，Papio に組み込んでの評価を行った．

次に，HPC クラスタで実行される並列プログラムとデータ転送プログラムから Papio ストレージシステムにアクセスするためのソフトウェアを開発し，実際的なアプリケーションからの利用において性能予約のコンセプトの有用性を検証した．並列プログラム向けには，MPI-IO のバックエンドストレージに Papio を利用できるようにするための Papio-ADIO モジュールを開発した．データ転送プログラムに関しては，クラウドなどでよく用いられる Amazon S3（Simple Storage Service）の互換インタフェースを Papio ストレージのフロントエンドに実装するとともに，Papio の予約機能を明示的に利用するための Papio 向けの S3 クライアントプログラムを開発した．これらの開発は H23 年度の後半に開始し，H24 年度の 12 月末までに主要な部分を終えた．また，H24 年度は開発と並行して評価検証を行った．

なお，Papio-ADIO の開発においては，Edinburgh 大学で開発され，データの局所性を活用することで Two-Phase I/O の高速化が図られた Dynamic-CoMPI を利用することとし，Edinburgh 大学と共同で研究を進めた．

このように本研究課題は３つのテーマから構成されており，それぞれにおいて手法の提案と実装，評価検証を行い，その成果の研究発表を行うという形で進めた．

４．研究成果

３節で述べたように，本研究課題は Papio

内部の I/O 制御機構と，Papio を利用するための上位ソフトウェアの研究からなる．後者に関しては，さらに並列プログラムからの利用とデータステージングにおける利用のための研究の２つに分類される．以下では，それぞれの研究毎に，成果として，その研究開発内容と得られた効果について記し，最後にまとめを述べる．

(1) I/O スケジューリング機構の Write 性能制御向けの拡張に関する研究

Papio のストレージサーバでは，SSD をバックエンドに用いた Object-based Storage Device（PROBS と呼ぶ）を利用している．その PROBS の I/O スケジューリングに関して，SSD の特性を考慮した Write 性能制御の提案・実装を行った．基本設計としては，I/O 処理キューをメタデータ用の高優先度キューとデータ用のデータキューに分離した．そして，メタデータの更新を規則化させるとともに，アクセス毎に，要求性能に応じたデータキューを動的に作成し，重みづけラウンドロビン方式によって，要求性能を満たすように各 I/O を処理していく仕組みとした．

評価実験により，Write の同時アクセスが発生するアクセス競合の状況において，提供可能な最大性能の 15~20% のオーバーヘッドを見積もることで性能の違反率を 10% 以下に抑制できる I/O 制御が行えること，そして，各 PROBS において同時アクセス数を 6 程度まで増やせることを確認した．

本開発により，従来 Papio がサポートしていたデータの読み込みにおける性能保証に加えて，データの書き込みに関しても同様の性能制御を可能にすることができた．なお，開発した PROBS は，後述する Papio の Web ページにて LGPL v2.1 のライセンスのもと，公開を行った．

(2)MPI-IO を通して Papio を利用するための上位ソフトウェア（Papio-ADIO）の研究

HPC の並列プログラムの多くは MPI（Message Passing Interface）を利用して記述されており，Dynamic-CoMPI は MPI 実装の１つである．図３に示すように，Dynamic-CoMPI では，MPI-IO の代表的な実装である ROMIO をベースにデータの局所性を活用した Two-Phase I/O（LA-Two-Phase I/O）を実装している．ROMIO の ADIO では，バックエンドに複数のストレージを利用することが可能であり，本研究ではその枠組を利用して，Papio ストレージにアクセスするための Papio-ADIO を実装した．そして，N 個のプロセスが１つのファイルに対して並列に I/O を行うアクセスパターン（N-to-1）において，予約に基づいた性能制御を行う仕組みを提案し，実装した．

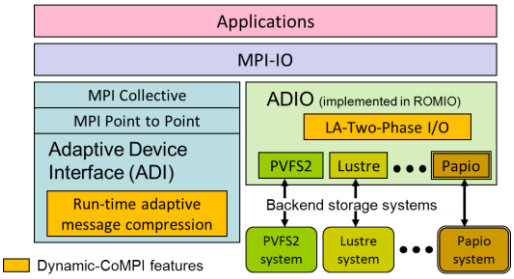


図３ Dynamic-CoMPI と Papio-ADIO の構造

評価実験では，ベンチマーク・アプリケーションを用いて，各プロセスが連続領域にアクセスする場合と不連続な領域にアクセスする場合を比較し，Papio-ADIO の性能制御の効果を検証した．また，バックエンドに PVFS2（または OrangeFS）と Lustre を用いた場合に，アクセス競合がどの程度の性能低下をもたらすのかを明らかにし，Papio-ADIO のアプローチの有効性を示した．図４は BISP3D アプリケーションを用いた評価実験結果の一例であり，PVFS2 や Lustre ではアクセス競合により性能低下が見られるが，Papio を利用した場合には，予約により性能を維持できているのが確認できる．

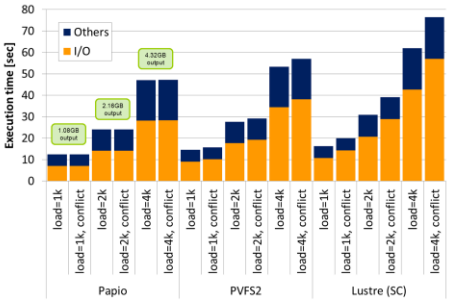


図４ BISP3D を用いた評価実験の結果

(3)S3 を通して Papio を利用するための上位ソフトウェア（PapioS3）の研究

HPC クラスタの外部と Papio ストレージ間のデータ転送や，Papio ストレージ内の各データを HPC クラスタの各ノードに展開するためのソフトウェアとして，クラウドで標準的に用いられている S3 の互換インタフェースを備えたデータ転送クライアント・サーバ（PapioS3）の研究開発を行った．PapioS3 では，S3 の互換インタフェースだけでなく，Papio の性能予約向けに拡張したインタフェースを提供し，ユーザによる明示的な性能予約を可能にした．拡張された API は，PUT Bucket，PUT Object，GET Object，Initiate Multipart Upload である．このうち，PUT Bucket は Papio のディスクスペース（容量）の予約に対応するために拡張を行い，ユーザが確保したいスペースに関して，容量等のパラメータを指定可能にした．それ以外は

Papioストレージに保存したObjectにアクセスするための操作に関する拡張であり，個々の操作要求において，予約時にPapioストレージから取得した予約 ID を設定できるようにした．

図 4 は開発した PapioS3 の概要である．PapioS3 サーバは，RADOS 向けの S3 フロントエンドである RADOS Gateway（RGW）がベースになっており，下位レイヤにおいて，Papio へアクセスするための拡張機能を有する．PapioS3 クライアントは JetS3t がベースになっており，マルチパート・データ転送において，並列 I/O を効率よく実行できるような改良が施されている．

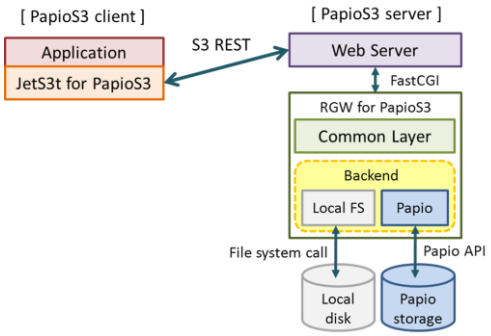


図４ PapioS3 の概要

評価実験では PapioS3 を用いたデータ転送（アップロード，およびダウンロード）の基本性能を調査した後，複数の PapioS3 クライアントが同時に PapioS3 サーバにアクセスした時の性能制御の効果を検証した．図５は同時アップロードの性能試験の結果であり，A～E までのクライアントがそれぞれ異なる性能要求（Request）を行った場合に，実際に提供できた性能（Achieved）を示しており，性能保証が達成できていることが確認できる．ダウンロードに関しても同様の実験を行い，PapioS3 の効果を確認できている．

図５ 同時アップロードの性能

(4) まとめ

これらの成果により，HPC クラスタで動作する並列プログラム，およびデータステージングの実行において，予約を用いた性能保証がアクセス競合による性能低下の問題解決に有用であることを示した．すなわち，２節で述べた研究目的を概ね達成できたと考える．なお，個々の成果については，HPC に関する国内外の会議や研究会において発表を行った．

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５件）

① Yusuke Tanimura, Rosa Filgueira, Isao Kojima, Malcolm Atkinson, Reservation-based I/O Performance Guarantee for MPI-IO Applications using Shared Storage Systems, SC '12 Companion (Proceedings of the 2012 companion on High Performance Computing Networking, Storage and Analysis), 査読有, 2012, 1-2

② 谷村勇輔、柳田誠也、予約に基づくストレージ QoS 実現のための S3 REST API の拡張実装、情報処理学会研究報告、査読無、2012-HPC-136、2012、1－6

③ Yusuke Tanimura, Rosa Filgueira, Isao Kojima, Malcolm Atkinson, I/O Performance Isolation on A Shared Storage System for MPI-IO Applications, SACSIS2012 論文集, 査読無, 2012, 24-25

④ 谷村勇輔、鯉江英隆、工藤知宏、小島功、田中良夫、ユーザによる明示的な予約に基づき I/O 性能を保証する分散ストレージシステム、情報処理学会論文誌コンピューティングシステム(ACS38)、査読有、5 巻 3 号、2012、42－56

⑤ 谷村勇輔、鯉江英隆、工藤知宏、小島功、SSD を用いたオブジェクトベース・ストレージデバイスの I/O 性能制御、査読無、2011-HPC-130 巻、2011、1-8

〔学会発表〕（計８件）

① 谷村勇輔、Extension of S3 REST API for Providing QoS Support in Cloud Storage、11th USENIX Conference on File and Storage Technologies、2013 年 2 月 13 日、Fairmont San Jose (California、USA)

② 谷村勇輔、Reservation-based I/O Performance Guarantee for MPI-IO Applications using Shared Storage Systems、SC12 (The International Conference for High Performance Computing Networking, Storage and Analysis)、2012 年 11 月 13 日、Salt Palace Convention Center (Utah、USA)

③ 谷村勇輔、予約に基づくストレージ QoS 実現のための S3 REST API の拡張実装、第 136 回ハイパフォーマンスコンピュー

ティング研究発表会、2012 年 10 月 4 日、沖縄産業振興センター（沖縄県）

④ 谷村勇輔、I/O Performance Isolation on A Shared Storage System for MPI-IO Applications、先進的計算基盤システムシンポジウム（SACSIS 2012）、2012 年 5 月 16 日、神戸国際会議場（兵庫県）

⑤ 谷村勇輔、Storage QoS for HPC Environments、Research Exhibit of AIST in SC11、2011 年 11 月 13－16 日、Washington State Convension Center（Washington、USA）

⑥ 谷村勇輔、Storage Research for Data-Intensive Applications、DIR Seminar of the University of Edinburgh、2011 年 8 月 19 日、Informatics Forum、The University of Edinburgh（Edinburgh、UK）

⑦ 谷村勇輔、SSD を用いたオブジェクトベース・ストレージデバイスの I/O 性能制御、2011 年並列／分散／協調処理に関する『鹿児島』サマー・ワークショップ（SWoPP 鹿児島 2011）、2011 年 7 月 28 日、かごしま県民交流センター（鹿児島県）

⑧ 谷村勇輔、Towards Performance-assured Cloud Storage、Workshop on Science Agency Uses of Clouds and Grids、2011 年 7 月 18 日、Salt Lake City Marriott Downtown（Utah、USA）

〔その他〕
Papio に関する Web ページ（本研究成果の開発物、関連ソフトウェアについて記載）
http://papio.apgrid.org/

６．研究組織
(1)研究代表者
谷村 勇輔（TANIMURA YUSUKE）
産業技術総合研究所・情報技術研究部門・主任研究員
研究者番号：80415710