社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2007－DPS－132 (20)
2007－GN－ 65 (20)
2007－EIP－ 37 (20)
2007／9／21

並列可変長アクセス方式による書込みスループット向上についての検証

奥村恭弘† 境美樹† 伊織生美† 新井克也†

概要
APレイヤで1ファイルに対する書込み要求が連続的に発行された場合、各要求に従った書込みデータ転送は不連続となり、データ転送効率が低下する。
本論文では、上記点に着目し、APレイヤの書込み方式を改善することで、データ転送効率を向上させる方式を提案、検証する。

Write Performance Evaluation of the Parallel Variable Access method

YASUHIRO OKUMURA† MIKI SAKAI†
KIYOSHI IORI† KATSUYA ARAI†

Abstract
When a number of write demands against a file occur simultaneously at AP layer, the transfer of data becomes discontinuous. The discontinuity deteriorates the data transfer efficiency.
In this paper, we propose a novel method called Parallel Variable Access to improve the data transfer efficiency. The performance evaluation confirms the effectiveness of the proposed method.

1. はじめに *

一般的に、システムは、耐障害性を実現するため、処理内で実施するデータ更新処理の度、取得する更新内容の差分データ(ジャーナル)や故障原因追跡、監査等を目的とした、システムに対する操作の度、取得する操作内容データ(ログ)を取得する。

上記ジャーナルやログを、高速かつ確実に二次記憶装置に追記型記録する(書込む)ことが要求される領域に対して、単位時間当たりの総書込みデータ実効転送量(トータル書込みスループット性能)を向上する方式を提案、検証する。

二次記憶装置に対してデータ書込みを実施する場合、要求毎に要求データ及び応答データの転送が実施される。また、応答データ転送は、要求データ転送完了後、応答待ち時間後に実施される。

従って、書込み要求が連続的に発生すると、各要求に従ったデータ転送は不連続に実施されることとなる。

一般的にジャーナルやログ取得では、図1に示すように1ファイルに対して書込み依頼が連続的に発生する。

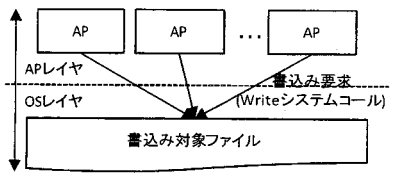

図1 書込み要求

本研究では、AP(OS上で動作するプロセスまたはスレッド)レイヤで1ファイルに対して書込み要求を連続的に発行した場合、前述の通り、データ転送が不連続となり、データ転送効率が低下する点に着目し、APレイヤの書込み方式を改善することで、データ転送効率を向上させる方式を提案する。

本論文ではまず、2章において提案方式及び実現方式を説明し、3章において提案方式の性能検証結果及び評価について述べる。最後に、4章にて本論文をまとめる。

2. 並列可変長書込み方式(PVA[a]方式)

2.1 前提

本研究では、高速かつ確実なジャーナル、ログ書込みを実現するため、RAIDコントローラにライトバック設定のバッテリバックアップ式キャッシュの搭載、かつ二次記憶装置(キャッシュ)のライトスルー設定を前提とする。(図2)

前者により、高速に書込み要求元に書込み応答を返却することが可能となると同時に、後者により書込み要求データが不意の電源断等により二次記憶装置(プラッタ)に書き込まれる前に消失してしまうことを防止することが可能となる。また、RAIDコントローラにより、RAIDを構成することが可能となり、二次記憶装置の信頼性の向上が可能となる。

一般的に二次記憶装置への書込みに要する時間は、主記憶装置とRAIDコントローラキャッシュ/RAIDコントローラキャッシュと二次記憶装置(キャッシュ)/二次記憶装置(キャッシュ)と二次記憶装置(プラッタ)間の書込み要求・応答データ転送時間と二次記憶装置(プラッタ)へのデータ書込み時間から構成される。

本研究では、前述の通り、RAIDコントローラにライトバック設定のバッテリバックアップ式キャッシュを搭載することにより、二次記憶装置への書込みに要する時間は、主記憶装置とRAIDコントローラキャッシュ間のデータ転送時間(バス転送時間)とすることが可能となる。また、RAIDコントローラは複数要求を受付可能とする。

本研究では上記前提の下、バス転送におけるデータ転送効率の向上方式を提案する。

*† 日本電信電話(株)
NTT

a) PVA : Parallel Variable Access

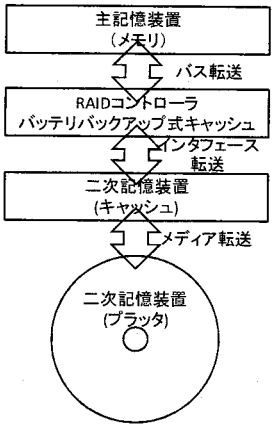

図 2　システム構成

2.2　提案方式

1 書込み要求時、バスを通して書込み要求データ及び書込み応答データの転送が実施される。また、書込み応答データ転送は、書込み要求データ転送完了後、RAID コントローラ内部処理時間(応答待ち時間)後に実施される。

応答待ち時間中は、CPU は使用されていないため、他の書込み要求データ(もしくは書込み応答データ)の転送が可能である。

ただし、AP レイヤで 1 ファイルに対して書込み要求が連続的に発生した場合は、ファイルの整合性を保持するため、各書込み依頼は同期直列に処理されなければならない。そのため、要求データ、応答データは、図 3 上図のように同期直列にバス転送(CPU 処理)されることとなる。

提案方式では、1 ファイルに対して連続的に発生した複数の書込み要求を非同期並列に処理することで、図 3 下図に示すように、応答待ち時間中に他の書込み要求データ転送(もしくは書込み応答データ転送)を実施することにより、バス転送におけるデータ転送効率を向上させ、結果トータル書込みスループット性能の向上を図る。

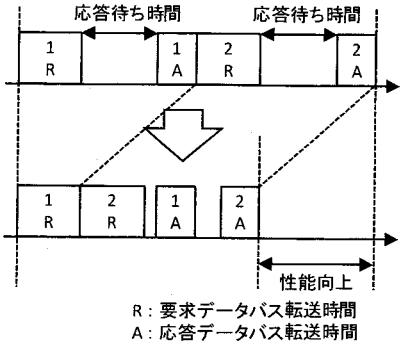

図 3　データ転送効率向上

2.3　実現方式

AP レイヤで連続的に発行された 1 ファイルに対する各書込み要求を非同期かつ並列に処理することで、応答待ち時間中に他の書込み要求データ転送(もしくは書込み応答データ転送)を実施することが可能となる。

PVA では、下記実現方式(1)－(3)により、1 ファイルに対して連続的に発生した書込み要求を非同期かつ並列に処理することを可能とする。

また、下記実現方式(4)により、書込み処理を効率的に実施することを可能とする。

(1)　書込み対象ファイル

[問題・課題]

EXT3 等の汎用ファイルシステム上のファイルに対する書込みは、ファイル内のデータ整合性を保持するため、複数非同期並列に実施することはできない。

[PVA 方式]

図 4 に示すように書込み対象となるファイル(論理ファイル)を、EXT3 等の汎用ファイルシステム上の複数ファイル(物理ファイル)から構成する。この際の物理ファイルの数を並列数とする。

上記により、1 論理ファイルに対して連続的に発生した書込み要求を各物理ファイルへの単一の書込み要求とすることが可能となる。

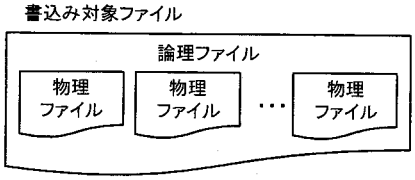

図 4　論理ファイル

(2) **書込み処理主体**

[問題・課題]

論理ファイルを複数物理ファイルから構成し、書込み先ファイルを複数にしても、論理ファイルに対する書込み処理を実施する主体が 1 つしか存在しない場合、1 論理ファイルに対して連続的に発生した書込み要求は、唯一存在する書込み処理主体により同期直列に処理される。

[PVA 方式]

図 5 に示すように書込み処理主体を各物理ファイルにひも付けて配置する。上記により、複数の書込み要求処理を非同期並列に処理することが可能となる。

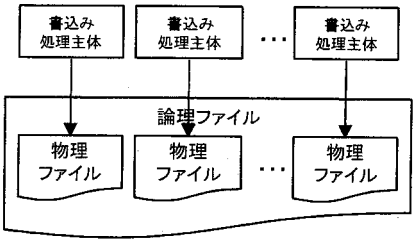

図 5　書込み処理主体配置方法

(3) **書込み要求応答**

[問題・課題]

書込み処理主体を物理ファイルにひも付けて複数配置しても、書込み要求処理に加えて、書込み応答処理も非同期並列に処理できなければ、複数の書込み要求を非同期並列に処理することはできない。

[PVA 方式]

書込み要求データを分割することなく、書込み処理主体へ処理を依頼することにより、各書込み処理主体は他の書込み処理主体の処理完了を待つことなく、書込み応答を返却することが可能となる。また、書込み要求データの分割オーバヘッドも発生しない。

上記に加えて、図 6 に示すように書込み処理主体から直接書込み要求元へ書込み応答を返却するにより、複数の書込み処理を非同期並列に処理することを可能とする。

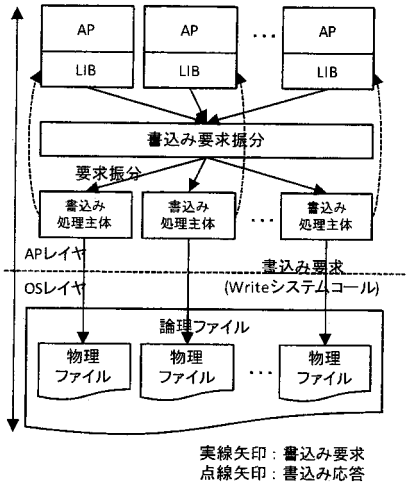

図 6　実現方式

(4) **書込み要求の振分**

[問題・課題]

書込みデータサイズの差により、各書込み処理主体の処理時間に差が発生する。

書込み要求を書込み処理主体にラウンドロビン等により振分けると、書込み処理を実施していない書込み処理主体が存在していても、書込み要求を振分けできない状況が発生する。

[PVA 方式]

図 7 に示すように書込み処理を実施していない書込み処理主体のリスト(依頼待ち書込み処理主体リスト)を管理し、書込み処理を実施していない書込み処理主体へ書込み要求を振分ける。

具体的には、依頼待ち書込み処理主体リストからリスト取得後、書込み要求をリストに従った書込み処理主体へ振分ける。また、書込み処理主体はライブラリ(LIB)へ書込み応答返却後、依頼待ち書込み処理主体リストへ取得していたリストを再登録する。

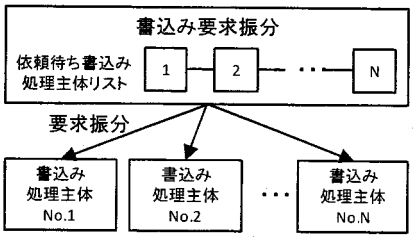

図 7　依頼待ち書込み処理主体管理

3. 検証・評価

3.1 検証環境・構成・条件

本検証では、通常方式とPVA方式双方のトータル書込みスループット性能を測定する。

通常方式は、複数のアプリケーションプログラム(AP)が、1書込み先ファイルに対して、書込み先ファイルに対する整合性保持を目的とした排他制御機能を提供するライブラリ(LIB)経由で書込み要求を同時に発行する。(図 8)

一方、PVA方式では、複数のAPが、1書込み先論理ファイルに対して、前述した実現方式を実装したPVAライブラリ(PVA LIB)経由で書込み要求を同時に発行する。(図 9)

検証環境・条件については、表 1、表 2 及び表 3 に示す。2.1で述べたように、高速かつ確実なデータ書込みを実現するため、二次記憶装置キャッシュはライトスルー設定、RAID コントローラキャッシュはライトバック設定とする。

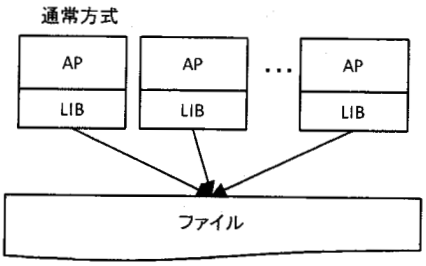

図 8 検証構成（通常方式）

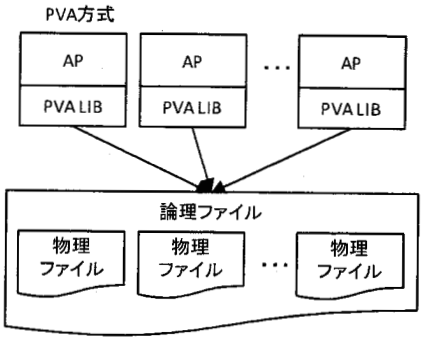

図 9 検証構成（PVA方式）

表 1 検証環境(ハードウェア)

ハードウェア名	仕様
IBM System X 3550	CPU：DualCoreIntel 1.6GHz
	メモリ：2GB
	バス：PCI Express x4 転送速度：送受信、計 2GB/sec
	二次記憶装置：36.4GB × 2
	二次記憶装置キャッシュ：16MB ※ ライトスルー設定
	RAID コントローラキャッシュ： 256MB ※ ライトバック設定

表 2 検証環境(ソフトウェア)

ソフトウェア名	備考
RedHatEnterpriseLinux ES	Version4 Update4

表 3 検証条件

項目	条件
同時書込み要求AP数	20
1AP当たり書込み要求数	500 回
総書込み要求データサイズ	10 MB

3.2 検証結果及び評価

3.2.1 検証 1 結果

1書込み要求のデータサイズを1KB(ジャーナル、ログデータを想定)と固定し、並列数を変化させた場合のトータル書込みスループット性能を検証・評価した。

表 4 に通常方式及びPVA方式それぞれのトータル書込みスループット性能値と両方式のトータル書込みスループット性能比を示す。

図 10 に通常方式及びPVA方式のトータル書込みスループット性能の遷移をグラフとして示す。

スループットはMB/s単位、性能比は%単位表示とする。

表 4 検証 1 結果

通常方式	PVA方式		性能比
スループット	並列数	スループット	
10.78	1	8.57	79
	2	11.33	105
	4	14.19	132
	8	14.47	134
	16	15.11	140

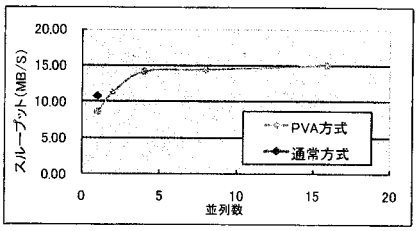

図 10　検証 1 結果

3.2.1 検証 1　評価

(1)　PVA 方式オーバヘッド

表 4 に示すように通常方式と PVA 方式(並列数：1)を比較することにより、PVA 方式のオーバヘッド約 20%であることが分かる。

PVA 方式では、書込み先ファイルに対して実際に書込み処理を実施するのは書込み処理主体であるため、AP から書込み要求データを書込み処理主体へ受け渡す必要がある。また、通常方式では実施不要である書込み要求の振分け処理も実施している。

約 20%のオーバヘッドは上記 2 点から発生するものと考える。

(2)　PVA 方式性能向上

表 4 に示すように通常方式と比較して PVA 方式(並列数：2 以上)では、最大約 40%のトータル書込みスループット性能優位性を確認した。

図 10 から PVA 方式のトータル書込みスループット性能は並列数：4 付近から、収束していることが分かる。今回の検証環境条件下では、並列数:4 以上では、総書込み要求データ転送量が RAID コントローラの要求受付許容量を超え、2.2 で述べた応答待ち時間中に、他の書込み要求データ転送を実施できなくなってしまうことが限界性能に達した一因と考えられる。

3.2.1 検証 2　結果

検証 1 の結果・評価より、並列数を 4 と固定し、1 書込み要求のデータサイズを変化させた場合のトータル書込みスループット性能を検証・評価した。

表 5 に通常方式及び PVA 方式それぞれのトータル書込みスループット性能値と両方式のトータル書込みスループット性能比を示す。

図 11 に通常方式及び PVA 方式のトータル書込みスループット性能の遷移をグラフとして示す。

データサイズは KB 単位、スループットは MB/s 単位、性能比は%単位表示とする。

表 5　検証 2 結果

	通常方式	PVA 方式	性能比
データサイズ	スループット	スループット	
1	11	14	132
4	20	54	276
10	38	127	336
50	71	361	511
100	84	412	491

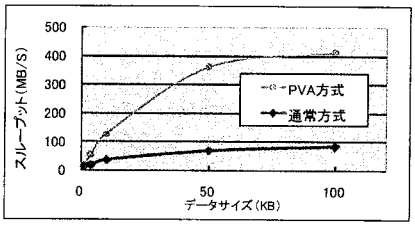

図 11　検証 2 結果

3.2.1 検証 2　結果

(1)　PVA 方式性能向上

表 5 及び図 11 に示すように通常方式と比較して PVA 方式では、最大約 4 倍のトータル書込みスループット性能優位性を確認した。

通常方式については約 80MB/sec、PVA 方式については約 400MB/sec で限界性能に達しているが、両限界性能共に、バス転送速度（理論値）2GB/sec に達していない。要因としては、書込み要求を発行する AP 内部処理、RAID コントローラ内部処理にボトルネックが発生していると考えられる。

4.　おわりに

本論文では、AP レイヤで 1 ファイルに対して書込み要求が連続的に発行された場合のトータル書込みスループット向上方式を提案し、その性能優位性を検証、確認した。

2.2 より、PVA 方式の性能向上率は応答待ち時間及びデータ転送時間に大きく依存する。今後、上記 2 条件と PVA 方式の性能向上率の相関、更に PVA 方式の性能の依存する環境条件の有無を明らかにすることが課題となる。

参考文献

1)　菅谷誠一: SCSI-2 詳細解説、CQ 出版社(1994)