C-006

All Flash Array 向けデータ圧縮・重複排除機能の性能向上方式
Performance Improvement of Data Compression and Deduplication for All Flash Array

出口 彰† 阿部 高大‡ 吉原 朋宏† 吉井 義裕†
Akira Deguchi Takahiro Abe Tomohiro Yoshihara Yoshihiro Yoshii

1. はじめに

近年，フラッシュドライブのみを搭載するストレージである AFA(All Flash Array)の利用が進んでいる[1]．AFA は HDD や，HDD とフラッシュドライブが混載するストレージに比べ，安定した高性能が特徴である．しかし，フラッシュドライブは HDD に比べ高価であり，AFA には圧縮や重複排除といった方法で，格納データ容量を削減し，記録媒体コストを低減することが求められる．

圧縮機能は，ランレングス，辞書式圧縮，ハフマン符号化等の可逆圧縮アルゴリズムを利用して入力データよりも小サイズの圧縮データを生成し格納する機能である．重複排除機能はライトされたデータと同一のデータが既にストレージ内に存在するかを，各データのハッシュを利用して検索し，同一データが検出された場合には，新たに書き込まれたデータを物理的に保存せず，検出された同一データへのポインタで管理する機能である．これらの機能を使用すると，ライト処理に圧縮・重複排除処理が追加となり，ライトのスループット性能や，応答時間が悪化する．

圧縮・重複排除処理をライト要求に同期して実行する方式と非同期に実行する方式を組み合わせる従来技術がある．本稿では，従来技術では応答時間が処理方式により異なり安定しないことを示し，これを回避する方法を提案する．

2. 圧縮・重複排除を伴うライト処理方式

圧縮・重複排除機能の処理方式には大きく二つの方式が存在する．一つ目は単位時間当たりの処理 IO 数であるスループット性能(単位は IOPS(Input/Output per Second))に優れている同期方式，二つ目はライト完了をサーバに報告するまでの時間である応答時間に優れている非同期方式である．図 1 に同期方式，非同期方式の処理内容を示す．

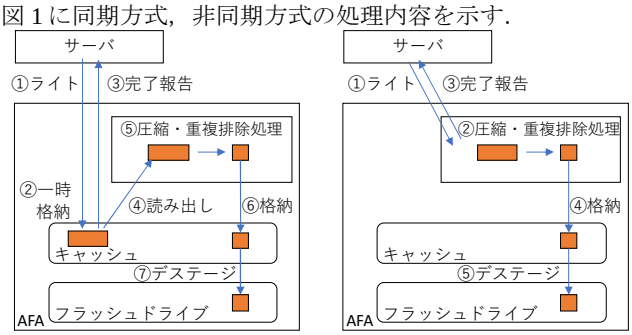

図 1　非同期方式(左)と同期方式(右)

非同期方式は，圧縮・重複排除処理前のデータを一時的にキャッシュに格納しライトの完了報告を返す．その後，一時格納したデータを再び読み出し，圧縮・重複排除処理を実行する．圧縮・重複排除を未使用時と同等の応答時間を実現できるが，キャッシュへの一時格納とキャッシュから読み出す処理が必要となる(図の②④)．このため，処理量が多くなり同期方式よりもスループット性能が低い．

同期方式は，ライトの完了報告を返す前に，圧縮・重複排除処理を実行する(図の②)．このため，ライトの応答時間に圧縮・重複排除の時間が加算され，非同期方式よりも応答時間が長くなる．非同期方式の様なキャッシュへの一時格納やキャッシュからの読み出し処理は不要であり，スループット性能は非同期方式よりも高い．

関連研究として，重複排除機能を対象に IOPS の値に基づき，同期方式と非同期方式切り替える方式が提案されている[2]．IOPS は，圧縮・重複排除の適用・非適用や，リードとライトの比率，シーケンシャルとランダムの比率などにより大きく変動する．このため，本稿ではプロセッサ稼働率を基に切り替える方式を従来方式とする．

3. 従来技術の課題と解決方式

本章では，安定した応答時間の目標値を設定し，従来技術である単一閾値方式は応答時間が安定しないことを示す．最後に，解決策として複数閾値方式を提案する．

3.1 応答時間安定の目標設定

応答時間の安定とは，同一負荷における平均応答時間からの変動量が小さいことと定義する．本研究では，同期方式と非同期方式を切り替えない場合と同等の安定性をめざす．方式を切り替えない場合，同一負荷における平均応答時間からの変動は 50%以下であった．このため，方式切り替えを行う場合も同等の 50%以下を目標とする．

3.2 単一閾値方式とその課題

単一閾値方式について述べる．図 2 は，同期方式と非同期方式の応答時間とスループット性能の関係を示している．図 2 は X 軸がスループット性能，Y 軸が応答時間である．スループットが高くなると，プロセッサ空き待ちが発生し，応答時間が伸びる．同期方式はスループット性能，非同期方式は応答時間が優位であり，両方式は図 1 の様な関係となる．単一閾値方式は，同期方式と非同期方式が交差するプロセッサ稼働率で，方式を切り替えることで，非同期方式の応答時間と同期方式のスループット性能を得る．

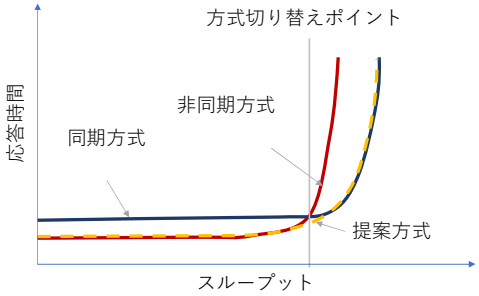

図 2　同期・非同期方式の切り替え概念図

† (株)日立製作所デジタルテクノロジーイノベーションセンタ, Center for Technology Innovation, Hitachi, Ltd.
‡ (株)日立製作所サービス&プラットフォームビジネスユニット, Service & Platform Business Unit, Hitachi, Ltd.

次に，単一閾値方式の問題点について説明する．単一閾値方式は，閾値付近の稼働率となる負荷を継続した場合に，同期方式と非同期方式の切り替えが頻発し応答時間が不安定となる．方式切り替え頻発の原因について，図 3 に示す方式と閾値の組み合わせに基づく状態遷移表を用いて説明する．非同期方式で動作中に閾値を上回った場合(状態 1 から状態 2 へ遷移)，処理方式が同期方式へ切り替わる(状態 4 へ遷移)．前述した通り，同期方式は非同期方式に比べて処理量が少ない．処理量が少なくなると，プロセッサ稼働率が下がり，再び閾値を下回り(状態 3 へ遷移)，非同期方式へ戻ってしまう(状態 1 へ遷移)． この切り替えの頻発により，同期方式と非同期方式の応答時間が混在し，応答時間が不安定となる．

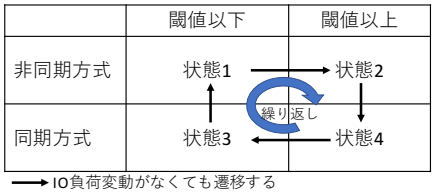

図 3　方式切り替えの頻発

3.3 複数閾値方式の提案

二つの閾値を用いた切り替えの頻発の回避方式を提案する．二つの閾値として閾値 1 と閾値 2 を設ける．閾値 1 は同期方式へ切り替えるための閾値であり，プロセッサ稼働率が，閾値 1 を上回るとスループットが有利な同期方式へ切り替える．閾値 2 は非同期方式へ切り替えるための閾値であり，閾値 2 を下回ると応答時間が有利な非同期方式へ切り替える．これらの閾値は，閾値 2 ＜ 閾値 1，かつ，閾値 2 +（非同期方式時のプロセッサ稼働率－同期方式時のプロセッサ稼働率）＜閾値 1 を満たすように定義する．

図 4 の状態遷移表を用いて課題の解決を説明する．プロセッサ稼働率が閾値 1 を上回り非同期方式から同期方式へ切り替わる(状態 2 から状態 4 へ遷移)．同期方式となりプロセッサ稼働率が低下するが，閾値 1 と閾値 2 の間に方式切り替えによるプロセッサ稼働率の変動分以上の差を設けているため，状態 6 で状態遷移が止まり，非同期方式の状態 1 へ戻ってしまうことがない．状態 6 から状態 3，状態 5 から状態 2 への遷移は IO 負荷の変動がない限り発生せず，同一負荷を継続中の大きな応答時間の変動を回避できる．

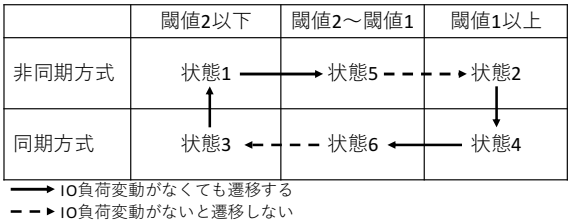

図 4　複数閾値方式による切り替え頻発の回避

4. 検証

表 1 に示す環境を構築し単一閾値方式と複数閾値方式の検証を行った．本研究では，ストレージのプロセッサ稼働率に基づき方式を切り替えるため，ストレージのプロセッサ以外がネックとならない様にサーバとストレージを構築した．ライト性能検証のため，IO 生成ツールである vdbench を用いてライト 100%の負荷をストレージに発行した．単一閾値方式の閾値は 65%，提案方式である複数閾値方式の閾値 1 は 65%，閾値 2 は 35%とした．

表 1　評価環境

IO 負荷 (vdbench パラメタ)	圧縮率	2:1
	重複排除率	2:1
	ライト比率	100%
	データ長	8KB
	アクセス範囲	6TB
ストレージ構成	ボリューム数	240
	ボリュームサイズ	27GB
	SSD	48 台
単一閾値方式	閾値	65%
複数閾値方式	閾値 1	65%
	閾値 2	35%

図 5 に示す検証結果を示す．検証では最大 IOPS の 10%,25%,45%,60%,70%,80%,90%の負荷の応答時間を測定した．各負荷に対して，ライト処理の応答時間を 15 回測定した．1 回当たり 5 秒間の負荷を発行し応答時間は 15 秒間の平均応答時間である．図 5 は，15 回の平均応答時間に対する変動量を平均応答時間に対する比率で示している．

単一閾値方式は閾値である 65%を含む 45%～80%の間で応答時間が大きく変動してしまっていることが分かる．同期方式で動作する最大スループットの 45%の IOPS で負荷をかけた場合，処理量が多い非同期方式が動作するとプロセッサ稼働率が 65%を上回る場合があり，45%から 60%の負荷においても方式の切り替えが発生してしまっている．平均応答時間からの変動は最大で 150%以上の増加となり目標の 50%以内を大きく上回る結果となった．

提案した複数閾値方式は単一閾値方式と異なり安定した応答時間を実現できている．平均応答時間からの変動は最大でも 36%となり，目標の 50%を達成した．従来方式の単一閾値方式からも大幅に改善する結果となった．

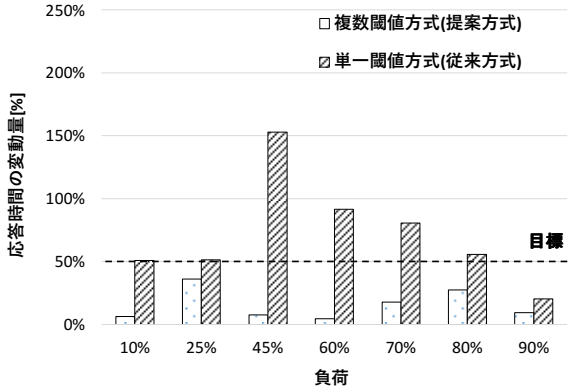

図 5　切り替え頻発回避による応答時間の安定

5. まとめ

本稿では，複数閾値を用いた同期方式と非同期方式の切り替えを提案し，本方式により応答時間を安定させることができることを検証した．

参考文献

[1] IDC, “Worldwide and U.S. Enterprise Storage Systems Forecast Update, 2017-2021”, (2017).
[2] 加藤, 大辻, 鈴木, 佐藤, 吉田, “インメモリー重複除去における書き込み高速化”, ComSys2016 (2016).