B-014

ストレージ階層管理連携サーバフラッシュキャッシュ機能の提案

Proposal of Server Flash Cache Cooperating with Storage Tier Management

高田 有時† 林 真一†
Aritoki Takada Shinichi Hayashi

1．はじめに

近年，コンシューマ製品におけるフラッシュデバイスの普及に伴い，エンタープライズ・高信頼情報システムにおいてもフラッシュデバイス導入が本格化している．そのようなシステムにおいて，フラッシュデバイスの用途の一つとして，フラッシュデバイスに比べて I/O 性能の低い記憶領域に対する補助記憶領域として用いることで，I/O 性能向上を図るものがある．

それらの用途を実現した機能の中に，外付けストレージ装置内部においてデータへのアクセス頻度に応じ割り当てる記憶デバイスの種類，すなわち記憶デバイスの階層を自動的に制御するストレージ階層管理機能[1]や，サーバ装置内部において，内蔵 HDD や外付けストレージのデータを一時的に保持するサーバフラッシュキャッシュ機能[2]がある．いずれの機能も，全データ量に対して少ない容量のフラッシュデバイス，すなわち低コストにて，全データをフラッシュデバイスに格納する場合に準じた性能向上効果が得ることを目的とするものである．

2．本研究の課題と目的

サーバ装置と外付けストレージ装置からなるシステムにおいて，前述のストレージ階層管理機能とサーバフラッシュキャッシュ機能を組み合わせて利用する場合がある．このとき，各々の機能による相乗的な性能向上効果が得られることが望ましい．

しかし，両者を単に組み合わせて利用した場合，一方のみを 利用する場合に対して性能向上効果が限定的となる場合がある．これは，両者ともアクセス頻度が高いデータをフラッシュデバイス等の高性能デバイス上に配置することで上位からみた I/O 性能を向上させる機能であり，サーバフラッシュキャッシュ機能とストレージ階層管理機能の両者が同一のデータを各々高性能デバイス上に配置し，これにより高性能デバイスの容量が消費されることによる．

従って，両者を組み合わせて利用する場合，性能向上効果の観点から，同一のデータを二重に高性能デバイス上に保持しないよう工夫することが望ましい．

そこで本研究では，ストレージ階層管理機能と連携することにより，同一データの二重配置を回避する「ストレージ階層管理連携サーバフラッシュキャッシュ機能」を提案する．

3. ストレージ階層管理連携方式の提案

本研究が想定するシステム構成，及びこれにおける従来のサーバフラッシュキャッシュ機能(以降，「従来方式」と表記)，及び，ストレージ階層管理連携サーバフラッシュキャッシュ機能(以降，「提案方式」と表記)の概要を図 1 に示す．

ストレージ装置内の階層管理機能はデータをフラッシュデバイス等の高性能デバイスからなる高階層，あるいは HDD 等からなる低階層に配置し，これらを仮想ボリュームとしてサーバ装置から透過的にアクセス可能とする．サーバ装置はアプリケーション・ソフトウェアからの I/O 要求に対して， サーバフラッシュキャッシュ機能を介してストレージ装置の仮想ボリュームに I/O 要求を発行する．従来方式ではサーバ装置及びキャッシュ機能はデータが格納されている階層を識別できず，従って階層により処理内容を変更することもできないため，高階層にあるデータもサーバフラッシュ内にキャッシュする．

これに対して提案方式では以下の処理を行う．まず，ストレージ装置から階層情報，すなわち，仮想ボリューム上のアドレス範囲とそこに割り当てられている階層の種類の対応関係を表す情報を取得する．これにより，任意のアドレス範囲のデータについて，格納されている階層が判別可能となる．

次に，ある I/O 要求をキャッシュ機能が受け付けた際に，前述の階層情報を用いて，I/O 要求が対象とするデータが存在する階層を判断する．低階層にあると判断した場合はキャッシュ処理の対象とするが，高階層にあると判断した場合はキャッシュ処理の対象外とする．

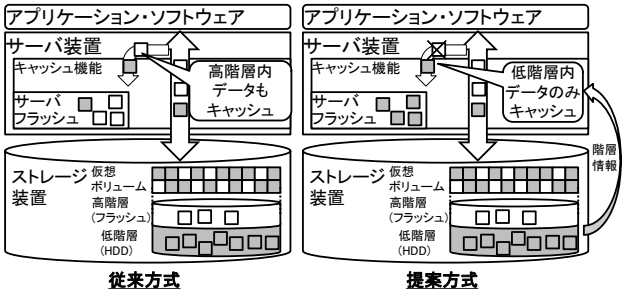

図 1 ストレージ階層管理連携方式

4. 性能評価

4.1 性能評価モデル

本研究では，提案方式の性能向上効果について従来方式，および，サーバフラッシュキャッシュ機能を用いない場合と机上評価を行う．評価にあたって，下記のようなシステムを仮定する．

- アプリケーション・ソフトウェア(以降，「アプリケーション」と表記)がボリュームに対し，ランダム Read を行う．ただし，例えば Zipf 分布[3]のような偏りのある頻度分布に従いアクセスするものとする．分布は一定の時間内では不変であるものとする．
- アプリケーションは上記の頻度分布に従いつつ，サーバフラッシュ，ストレージ内高階層，低階層のいずれかがボトルネックとなるまで I/O 要求を発行する．

†(株)日立製作所 横浜研究所

・ サーバフラッシュキャッシュはデータの入換を行わず，アクセス頻度の高いデータから順にキャッシュするものとする．

このようなシステムの I/O 性能，すなわち，アプリケーションが単位時間内で発行可能な I/O 要求数を以下の方法で見積もる．まず，キャッシュ機能及び階層管理機能の制御により発生する，アプリケーションが発行する I/O 要求量に対するサーバフラッシュ，ストレージ装置内高階層，低階層の各々への I/O 要求量の比率をそれぞれP_f，P_h，P_lとする．また，サーバフラッシュ，ストレージ装置内高階層，低階層の最大スループット(IOPS)を各々M_f，M_h，M_lとする．このとき，アプリケーションが単位時間内で発行可能な I/O 要求数T_{max}は式(1)の通りとなる．

$$T_{max} = \min\left(M_f/P_f, M_h/P_h, M_l/P_l\right) \quad (1)$$

このモデルについて，サーバフラッシュキャッシュが存在しないシステム，階層管理連携を行わないサーバフラッシュキャッシュ機能を備えたシステム，階層管理連携を行うサーバフラッシュキャッシュ機能を備えたシステムについて個別にP_f，P_h，P_lを算出することにより，各システムの性能を求める．

尚，以降にてアプリケーションが発行する I/O 要求のうち，アプリケーションがアクセス可能な仮想ボリュームのうち I/O 頻度の上位から$\mathrm{x}(0 \le x \le 1)$の割合の部分への I/O 頻度を$\mathrm{P}(x)$，サーバフラッシュ，ストレージ内高階層，低階層の容量を各々C_f，C_h，C_lとし，これを用いてシステム性能の算出式を求める．

4.1.1 キャッシュなしの場合の I/O 量比率

サーバフラッシュキャッシュを適用しないシステムの場合，アプリケーションが発行する I/O 要求のうち I/O 頻度上位$C_h/(C_h + C_l)$の部分への I/O 要求が高階層への I/O 要求となり，残りの部分が低階層への I/O 要求になると考えられる．すなわち，高階層，低階層への I/O 量の比率は各々式(2)，式(3)の通りとなる．

$$P_\mathrm{h} = \int_0^{C_h/(C_h+C_l)} \mathrm{P}(x)dx \quad (2)$$

$$P_\mathrm{l} = \int_{C_h/(C_h+C_l)}^{1} \mathrm{P}(x)dx \quad (3)$$

4.1.2 従来方式での I/O 量比率

従来方式のサーバフラッシュキャッシュを適用したシステムの場合，アプリケーションが発行する I/O 要求のうち I/O 頻度上位$C_\mathrm{f}/(C_h + C_l)$の部分への I/O 要求がサーバフラッシュへの I/O 要求，上位$C_\mathrm{h}/(C_h + C_l)$の部分からサーバフラッシュへの I/O 要求となった部分を除いた部分が高階層への I/O 要求となり，残りの部分が低階層への I/O 要求になると考えられる．従って，サーバフラッシュキャッシュ，高階層，低階層への I/O 量の比率は各々式(4)～式(6)の通りとなる．

$$P_f = \int_0^{C_f/(C_h+C_l)} \mathrm{P}(x)dx \quad (4)$$

$$P_\mathrm{h} = \int_{C_f/(C_h+C_l)}^{C_h/(C_h+C_l)} \mathrm{P}(x)dx \quad (5)$$

$$P_\mathrm{l} = \int_{C_h/(C_h+C_l)}^{1} \mathrm{P}(x)dx \quad (6)$$

4.1.3 提案方式での I/O 量比率

提案方式のサーバフラッシュキャッシュを適用したシステムの場合，アプリケーションが発行する I/O 要求のうち I/O 頻度上位$C_\mathrm{h}/(C_h + C_l)$の部分への I/O 要求が高階層への I/O 要求となり，$(C_h + C_f)/(C_h + C_l)$の部分から高階層への I/O 要求となった部分を除いた部分がサーバフラッシュへの I/O 要求となり，残りの部分が低階層への I/O 要求になると考えられる．従って，サーバフラッシュキャッシュ，高階層，低階層への I/O 量の比率は各々式(7)～式(9)の通りとなる．

$$P_f = \int_{C_h/(C_h+C_l)}^{(C_h+C_f)/(C_h+C_l)} \mathrm{P}(x)dx \quad (7)$$

$$P_\mathrm{h} = \int_0^{C_h/(C_h+C_l)} \mathrm{P}(x)dx \quad (8)$$

$$P_\mathrm{l} = \int_{(C_h+C_f)/(C_h+C_l)}^{1} \mathrm{P}(x)dx \quad (9)$$

4.2 評価結果

4.1 節にて述べたモデルに対して，表 1 に述べるシステム構成における容量及び最大スループットの値(市販品カタログ値から算出)を適用し，性能の評価を行った．また，I/O 分布としては Zipf 分布に偏りを表す分布指数 s = 1.1を適用したものを用いた．

表 1 システム構成

#	構成要素	デバイス	容量 [GB]	I/O 性能 [KIOPS]
1	サーバフラッシュ	PCIe 接続フラッシュデバイス	1,200	580
2	ストレージ高階層	SAS 接続 SSD (3 台)	2,400	435
3	ストレージ低階層	SAS 接続 HDD (10 台)	9,000	4.6

評価の結果，キャッシュなしの場合及び従来方式に対して 38%の性能向上の見込みを得た(図 2)．これは，サーバフラッシュキャッシュの適用対象を低階層内データに限定することで， 低階層ボトルネックの緩和によりシステム全体の I/O 性能が向上したためと考えられる．

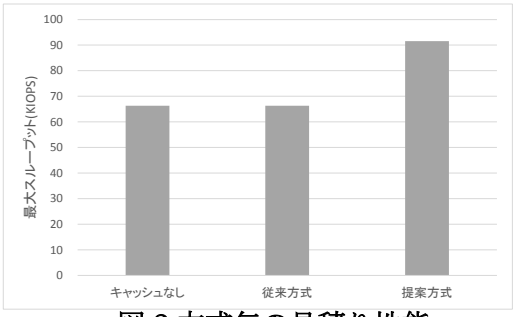

図 2 方式毎の見積り性能

5. まとめ

本報告では，フラッシュデバイスをサーバ装置内およびストレージ装置内にて併用する場合に性能向上効果を高めるためのストレージ階層管理連携サーバフラッシュキャッシュ方式を提案した．また，分布に偏りのあるランダム Read 要求を処理する場合の性能について I/O 分布モデルによる机上評価を行い，想定したシステム構成に於いて従来キャッシュ方式に対して 38%の性能向上の見込みを得た．

今後は，より多様な I/O パターンについて机上及び実測による効果の確認，及び装置構成の最適化の検討を行う予定である．

参考文献

[1] 圷 弘明 他，“ストレージ自動階層配置機能におけるデータ再配置の最適化”, 情報処理学会論文誌 54(4), 1592-1608, 2013
[2] E. Van Hensbergen et al, “Dynamic policy disk caching for storage networking”, IBM Research Report (RC24123), 2006
[3] Zipf’s law, http://en.wikipedia.org/wiki/Zipf%27s_law