# データ再配置によるストレージ装置の負荷分散システム

3L−2

桂島 航　石川 潤　菊地 芳秀
NECインターネットシステム研究所

## 1. はじめに

近年，ストレージ装置の最大容量は数十テラバイトの単位にまで達しており，従来オペレータが手動で行ってきたデータの再配置による負荷分散のような作業が，非常に困難になってきている．本研究は，大規模ストレージ装置において，データを再配置することで装置全体での平均レスポンスタイムを改善するように負荷を分散するシステムの一構成法を提案する．このようなシステムを用いることで，ストレージ装置にホットスポットを自動的に解消するなどの機能を与えることができる．

データの再配置により磁気ディスク間の負荷分散を図るシステムは，近年いくつかの方式[1][2]が提案されている．しかしながら，これらの方法は，磁気ディスク間での性能差，I/O 負荷の時間的な変化について十分な考慮がなされていない．本研究では，これらの二つの問題点を解決する負荷分散システムを提案する．すなわち、磁気ディスク間での性能差に関しては，各磁気ディスクを M/M/1 待ち行列としてモデリングすることで，そのサービス率によって性能差を考慮する．また、I/O 負荷の時間的な変化には，アクセスパターンに周期性が有ることを前提として，アクセスパターンの一周期を複数の区間に分割し，区間別に評価関数を計算し、その加重平均をもとにデータを再配置することで対応している．

## 2. 原理

データ再配置によるストレージ装置の負荷分散の基本的原理は，アクセスの統計値から現状を分析し，装置全体での平均レスポンスタイムが改善されるような負荷状態へ個々の磁気ディスクへのアクセス量を調整することにある(図1)．本研究では，装置全体での平均レスポンスタイムを待ち行列モデルから推定し，その推定値が最も改善されるような論理ボリュームの再配置(交換)を行う方式を提案する．

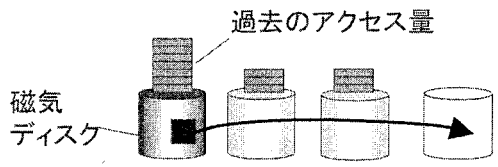


図1 データ再配置によるストレージ装置の負荷分散

### 2.1 待ち行列モデルに基づく装置全体での平均レスポンスタイムの推定

本研究では，$m$ 個の磁気ディスクによる並列 I/O 処理を M/M/1 待ち行列の並列モデルとして仮定する．そして公式から得られる各磁気ディスクでの平均レスポンスタイムを到着率で重み付けし，その総和を単位時間当たりの総 I/O 要求数 $c$ で割ることで装置全体での平均レスポンスタイム $f$ を(2.1)式のように推定する．

$$f = \frac{1}{c}\cdot\sum_{j=1}^{n}\frac{1}{\mu_j(1-\rho_j)}\cdot\lambda_j = \frac{1}{c}\cdot\sum_{j=1}^{n}\frac{\rho_j}{1-\rho_j} \qquad (2.1)$$

ただし、$\rho_j$ は磁気ディスク $j$ の利用率，$\lambda_j$ は I/O 要求の到着率，$\mu_j$ はサービス率である．

論理ボリュームの再配置に伴う各磁気ディスクの利用率の変化は，以下のように推定する．磁気ディスク A に属する論理ボリューム t1 と磁気ディスク B に属する論理ボリューム t2 を交換する場合，それぞれ交換後の利用率 $\hat{\rho}_A$，$\hat{\rho}_B$ は $c$ が一定であるという拘束条件の下で，(2.2)，(2.3)式のように推定する．

$$\hat{\rho}_A = \rho_A - \rho_{A\_t1} + \frac{\mu_B}{\mu_A}\rho_{B\_t2} \qquad (2.2)$$

$$\hat{\rho}_B = \rho_B - \rho_{B\_t2} + \frac{\mu_A}{\mu_B}\rho_{A\_t1} \qquad (2.3)$$

ここで $\rho_{A\_t1}$ は論理ボリューム t1 にアクセスが行われている率を示している．この $\hat{\rho}_A$，$\hat{\rho}_B$ を用いて，再配置後の装置全体での平均レスポンスタイム $\hat{f}$ は(2.4)式の様に推定することができる．

Wataru Katsurashima, Jun Ishikawa, Yoshihide Kikuchi
Internet Systems Research Laboratories, NEC Corporation

$$\hat{f} = \frac{1}{c} \cdot \left( \sum_{j=1, j \neq A,B}^{m} \frac{\rho_j}{1-\rho_j} + \sum_{j=A,B} \frac{\hat{\rho}_j}{1-\hat{\rho}_j} \right) \quad (2.4)$$

### 2.2 複数区間の統計情報による目的関数の構成

ストレージ装置に対するアクセスパターンは，ある程度の周期性を持つ場合が多いことが知られている．従来，データ再配置による負荷分散では，この一周期に渡る利用率の分散を負荷分散の目的関数として用いることが多く，その周期は一日や一週間という期間が多かった．しかし，その期間全体に渡って測定した利用率をそのまま(2.4)式に用いた場合，I/O 負荷が大きく変化するアクセスパターンに対しては適切な再配置を行うことが難しかった．そこで本研究では，このアクセスパターンの一周期を複数の区間に分割し，区間別の統計情報を基にその区間における平均レスポンスタイムを計算，それらを到着率で重み付けしたものを目的関数として用いる．目的関数 $g$ を(2.5)式に示す．基本的には(2.4)式を時間軸方向に分割した形となっている．ここで $d$ は単位時間当たりの総 I/O 要求数，添字 $i$ は区間 $i$ であることを示す．

$$g = \frac{1}{d} \cdot \left( \sum_{i=1}^{l} \sum_{j=1, j \neq A,B}^{m} \frac{\rho_{ij}}{1-\rho_{ij}} + \sum_{i=1}^{l} \sum_{j=A,B} \frac{\hat{\rho}_{ij}}{1-\hat{\rho}_{ij}} \right) \quad (2.5)$$

再配置案の選択方法は以下による．まず，容量等の制約条件から再配置が可能な論理ボリューム交換リストを作成し，その中で目的関数 $g$ が最小となるような案を探索する．探索された案を基に，その次に最も目的関数 $g$ が小さくなる案を再び探索する．この探索を数回繰り返すことで，複数回の論理ボリューム再配置に相当する案を得ることができる．繰り返し上限回数は，論理ボリューム交換コスト，再配置可能期間の長さ，また目的関数の改善度から決める．

## 3. 評価

本章では，提案するアルゴリズムを以下のように二つのケースに分けてシミュレーションで評価する．

1) 磁気ディスク間の性能が異なる場合
2) I/O 負荷が大きく変化する場合

比較対象には，従来の方法の代表格として，期間全体に渡って集計した磁気ディスク利用率の分散を目的関数に設定する方法(利用率を平準化することが狙いとなる)を取り上げた．磁気ディスク間の性能が異なる場合のシミュレーション結果を図2に示す．磁気ディスクの性能比は，[Disk1 : Disk2 : Disk3] = [1 : 5 : 10]である．この例ではレスポンスタイムが約26%改善された．

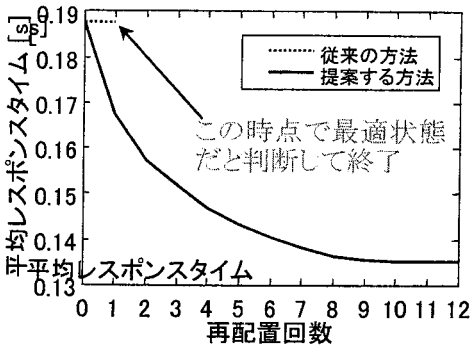


図2 磁気ディスク間の性能が異なる場合

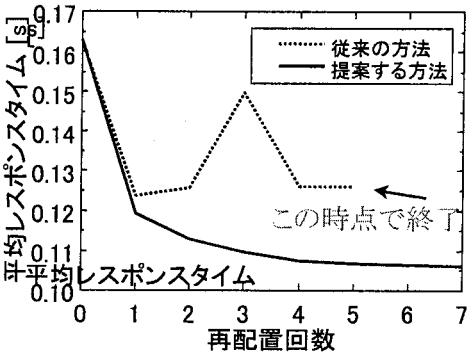


図3 I/O 負荷が大きく変化する場合

I/O 負荷が大きく変化する場合のシミュレーション結果を図3に示す．解析条件は，アクセスパターンの周期は24時間で，提案する方法では1時間単位でI/O 処理効率を計算した．この例では，従来の方法に比べて，約18%の改善がみられた．

## 4. おわりに

並列 M/M/1 待ち行列モデルと複数区間の統計情報による目的関数を用いた，データ再配置によるストレージ装置の負荷分散システムの一構成法を提案した．従来提案されている方法よりも，性能が異なる磁気ディスク間で用いるとき，アクセスパターンが平坦ではないときに有効である．今後の課題としては，再配置コストを見積もった動的な負荷分散，ネットワーク上に存在する複数ストレージ装置間での負荷分散等が挙げられる．

参考文献

[1] 米国特許 USP6061761
[2] 公開特許公報 特開 2001-67187