情報処理学会論文誌 Vol.54 No.3 1131–1140 (Mar. 2013)

仮想環境向け自動データ適正配置方式の提案

坂下 幸徳[1,a)] 三神 京子[1] 金子 聡[1] 敷田 幹文[2]

受付日 2012年6月29日, 採録日 2012年12月7日

概要：近年，サーバやストレージ装置がデータセンタへ集約され，データセンタが大規模化している．さらに，設備コストを削減するため，仮想技術の適用が進み，データセンタのシステム構成は複雑化している．このような中，性能やメディアコストの最適化を狙い，仮想サーバ上の VM を停止させずにデータを移動するデータ移動技術が複数登場している．しかし，これらのデータ移動技術は，仮想サーバやストレージ装置など，それぞれの装置で個別最適となるようデータ配置を行っており，データセンタ全体視点での，性能やメディアコストを考慮したデータ配置ができない．これを行うためには，システム全体の構成情報，各種メディアの性能やコストといった高度な知識を持つ管理者に頼らざるをえなかった．しかし，データセンタの大規模化，複雑化が年々進んでいる中，管理者負担が増加し運用ができなくなっている．そこで，本論文では，システム全体の接続関係や各種リソースの性能，メディアコストの情報を使い，性能とメディアコストを適正化する自動データ適正配置方式を提案する．本提案により，管理者負荷を削減しつつ，2012 年現在のメディア性能・コストの場合において，すべてのデータを SSD に配置したときと比較し，ユーザの要求性能を満たしたまま 47%以上のメディアコストを削減するデータ配置方式を実現した．

キーワード：ストレージ，仮想サーバ，データ移動，自動化

Proposal of Automatic Optimization of Data Location for Virtual Environment

Yukinori Sakashita[1,a)] Kyoko Mikami[1] Satoshi Kaneko[1] Mikifumi Shikida[2]

Received: June 29, 2012, Accepted: December 7, 2012

Abstract: In recent years, servers and storages have been consolidated to a data center. As a result, they become large-scale. Moreover, the servers and storages have been virtualized for CAPEX reduction. Therefore, the system configuration of the data center comes to be more complex. In this situation, the data mobility technology is to move data without stopping the VM for performance and media cost optimization. Nonetheless, the technology focuses only on data location such as the best location in each server or storage. There is no considering in the data location for performance and media cost of whole data center. Thus, the operation of data migration relies on an administrator with advanced knowledge, such as configuration information for the whole data center, performance and media cost. However, the data center has become large-scale and complicated every year, the administrators won't be able to operate because the load is increased. In this paper, we propose a method of automatic optimization of data location for virtual environment. This method uses the configuration information, the performance data of resource and the media cost data of the whole data center. This proposal also reduces the administrator's load as well as the media cost. Furthermore, in the case of the current media cost in 2012, our method can economize the media cost of more than 47%.

Keywords: storage, virtual server, data mobility, automation

[1] 株式会社日立製作所横浜研究所
Yokohama Research Laboratory, Hitachi Ltd., Yokohama, Kanagawa 244–8555, Japan
[2] 北陸先端科学技術大学院大学情報社会基盤研究センター
Research Center for Advanced Computing Infrastructure, Japan Advanced Institute of Science and Technology, Nomi, Ishikawa 923–1292, Japan
[a)] yukinori.sakashita.hk@hitachi.com

1. はじめに

近年，デジタルデータの量が爆発的に増加し，2020 年には世界のデジタルデータの総容量が，2012 年現在の約 10 倍となる 73ZBytes に到達する見込みである．さらに，クラウドの登場で IT インフラの「所有」から「利用」への流れが進み，これまで分散され設置されていたサーバやストレージ装置のデータセンタへの集約が進んでいる．このような中，大規模化するデータセンタでは，設備コストを削減するために，仮想技術を使いサーバやストレージ装置のコンソリデーションが進められている．しかし，一方でサーバやデータセンタの大規模化と仮想環境によるシステムの複雑化により，管理コストは増加傾向にある．さらに，これを管理する管理者の数は横ばいの傾向であり，大規模化，複雑化するデータセンタを変わらぬ管理者の数で管理しなければならない．

このような中，仮想環境を柔軟に構成変更を行うことで，性能やメディアコストを適正化するために，仮想サーバ（Hypervisor）上の VM（Virtual Machine）へ割り当てられたデータを，VM を停止することなくデータを自動移動するデータ移動技術が登場している．このデータ移動技術は仮想サーバやストレージ装置など異なるレイヤの実装がある．しかし，これらのデータ移動技術は，各々のレイヤの視点でデータ移動を行うため，性能やメディアコストが個別最適となり，データセンタ全体視点での性能やメディアコストを考慮したデータ移動はできない．そのため，データセンタ全体視点での性能やメディアコストを考慮しようとした場合，高度な知識を持った管理者の経験に頼らざるをえず，管理者の負担となっていた．データセンタの大規模化，複雑化が進んでいる状況を考慮すると，管理者の負担がますます増加し，その結果，設備コストを下げるために管理コストが増加してしまう．

そこで，本論文では，管理者が有していたデータセンタ全体の構成情報や各種ストレージリソースの性能，メディアコストの情報を使い，データセンタ全体での性能とメディアコストを適正に配置する自動データ適正配置方式を提案する．これにより，データセンタ全体視点でユーザの要求する性能を保ったままメディアコストを削減し，さらに，管理者の知識を使ったデータの自動配置を行うことで管理者の負担を削減する．

以下，2 章では従来のデータ移動技術と問題点について述べ，3 章では提案方式について述べる．次に 4 章は，提案方式の試作システムを使った測定結果を述べ，5 章で測定結果に基づいた提案方式の有効性を議論する．

2. 従来のデータ移動技術と問題点

本章では，従来のデータ移動技術と特徴を紹介した後，問題点について述べる．

2.1 従来技術

近年，サーバやストレージ装置の仮想化が進み，CPU やメモリ，ストレージ（Volume）など様々なリソースが仮想化されている．さらに，仮想化されたリソースは，利用していない VM のリソースを性能不足の VM に割り当てることで，コストを抑えつつ要求性能を満たすことが求められている．そのため，性能の観点によるネットワークや CPU，メモリの割当てだけでなく，性能とコストの両観点からデータを格納するメディアの割当てが重要視されている．このような中，VM に割り当てられた Volume 上の仮想ディスク（VMDK や VHD）を無停止で移動する技術が仮想サーバ，ストレージ装置と異なるレイヤの装置向けに研究 [1], [2] や製品化が進められている．**図 1** (a) に仮想サーバ，(b) にストレージ装置のデータ移動技術の詳細を示し，その特徴を**表 1** にまとめる．

仮想サーバのデータ移動技術は，仮想サーバが受信した I/O の統計情報をもとに，VM に割り当てられた Volume 上の仮想ディスクを，別 Volume へ移動させる技術である．この技術により，仮想サーバが備え持つ複数内蔵 Volume 間やストレージ装置間をまたがったデータ移動ができる．代表的な製品として，VMware の Storage Distributed Resource Scheduler [3] がある．

ストレージ装置のデータ移動技術は，ストレージ仮想化の一部として提供されている．まず，ストレージ仮想化技術の代表的なものに Thin Provisioning [4] と呼ばれる容量の仮想化技術がある．容量の仮想化技術は，ストレージ装置への書き込みがあった際に，書き込みデータ分の領域（以降，ページと呼ぶ）をメディア上に確保することで，事前に大量のメディアを用意する必要がなくなる技術である．さらに，ストレージ仮想化は，容量の仮想化を進化させ，複数メディアの階層も仮想化している．階層の仮想化

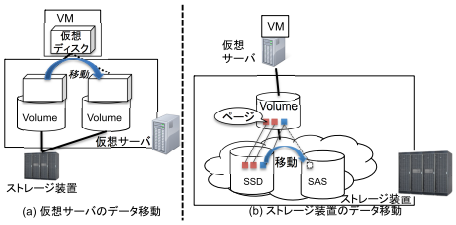

図 1 データ移動技術
Fig. 1 Data migration technology.

表 1 データ移動技術の特徴
Table 1 Peculiar features of the data migration.

特徴点 ＼ 提供装置	仮想サーバ（Hypervisor）	ストレージ装置
移動範囲	装置内のメディア間の移動，装置外のメディア間の移動	装置内のメディア間の移動
VM への負荷	あり	なし
移動判断材料	仮想サーバの I/O 性能	ストレージ装置の I/O 性能
移動データの単位	仮想ディスク単位	ページ
自動移動契機	I/O 性能劣化時	一定時間ごと

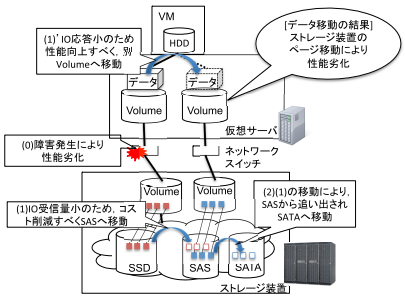

図 2 データ移動技術の併用時に発生する矛盾の一例
Fig. 2 Inconsistency of multi layer data migration.

では，SSD や SAS，SATA のハードディスクなど異なる種別のメディアを 1 つの Volume にまとめサーバへ割り当て，ストレージ装置が受信した I/O の統計情報をもとに，性能やメディアコストの異なる複数メディア（SSD，SAS など）間をページ単位でデータ移動させ，階層の仮想化を実現している．このような，階層の仮想化のデータ移動技術により，仮想サーバや VM へ負荷をかけずデータ移動ができる．代表的な製品として，Hitachi Virtual Storage Platform の Dynamic Tiering [5] がある．

このように，仮想サーバ，ストレージ装置のデータ移動技術には，それぞれ特徴がある．

2.2 問題点

2.1 節の技術を使い，管理者はユーザの要求性能を満たすべくデータ移動を行っている．しかし，従来技術では，異なるレイヤのデータ移動技術の矛盾動作，性能とメディアコストの個別最適，管理者への負担，の 3 つの点で問題がある．

(1) 異なるレイヤのデータ移動技術の矛盾

仮想サーバとストレージ装置のデータ移動技術を併用すると，異なるレイヤで相反するデータ移動を行うことがある．相反するデータ移動の例を**図 2** で紹介する．本例では，障害発生でネットワークの性能劣化が発生すると，仮想サーバは性能向上をさせるために別 Volume へデータ移動しようとするが，ストレージ装置はメディアコスト削減のために低速なメディアへデータ移動してしまい，よりいっそう性能が低下する．理想の動作としては，仮想サーバで性能向上をさせるために別 Volume へデータ移動をした際，ストレージ装置側は移動は行わないか，もしくは性能向上させるように移動させる，といったシステム全体として両方のデータ移動技術が協調し性能向上を図ることである．しかし，2012 年現在では，このような矛盾したデータ移動を行わないように，ベンダ自らいずれか一方のデータ移動技術のみ利用することを推奨している [6]．

(2) 性能とメディアコストの個別最適

仮想サーバ，ストレージ装置のデータ移動技術は，仮想サーバは I/O 性能によってデータの移動先を決定し，ストレージ装置は，I/O 性能とメディアコストを意識したメディア種別の情報を使い決定している．そのため，仮想サーバのデータ移動では，ストレージ装置のメディアの種別が取得できないため，メディアコストを意識した移動ができず，性能が最適になるようにデータ移動する．一方，ストレージ装置は，仮想サーバが備えるメディアコストやネットワークの性能を考慮せず，ストレージ装置内のみ性能とメディアコストが最適になるようデータ移動する．そのため，仮想サーバ，ストレージ装置がそれぞれの視点での個別最適となり，データセンタ視点での全体最適な性能とメディアコストになっていない．

(3) 管理者への高負荷

データ移動技術は登場してきたものの，データセンタ全体を考慮した適切なデータ配置を実現するためには，(1)，(2) の問題により，自動のデータ移動技術は利用できない．そのため，データセンタ全体のストレージの性能・メディアコスト，仮想サーバやストレージ装置の構成情報といった知識を持った高度な管理者の知識や経験に頼り，データの配置を検討しデータ移動を行っていた．これは，管理者の負荷となっていた．さらに，2012 年現在において，銀行や通信キャリアなど大規模なデータセンタを所有するエンタープライズ企業の中には，すでに 80 台を超えるストレージ装置を備えるデータセンタが登場し始めており，調査会社の報告によると 2015 年には，ストレージ装置が 100 台以上，VM は 50,000 台以上となる大規模なデータセンタの登場が予測されている [7]．このような大規模データセンタでは，管理対象が増加することで管理者の負担も増加している．

このデータ配置に関する管理者の負担を定式化する．管理者の 1 人あたりの 1 カ月あたりの作業時間を Wt，1 サービスの初期構築にかかる時間を It，1 カ月あたりのデータ配置の見直し回数を N，1 サービスあたりのデータ配置計画時間を Pt，提供するサービス数を S，管理者数を A，とすると管理者の負荷はサービスを開始する始めの月の作業時間を式 (1)，その翌月以降の月々の作業時間を式 (2) のように表すことができる．

$$W0 = It \times S/A + Wt \quad (1)$$

$$Wt = NPt \times S/A \quad (2)$$

式 (1)，式 (2) を使い，2015 年の大規模環境における管理者の負担を試算する．VM 10 台で 1 つのサービスを提供，1 サービスの初期構築にかかる時間を 60 分，1 カ月あたりのデータの見直し回数を 4 回，1 サービスあたりのデータ配置の計画時間を 10 分，高度な管理者 30 人が分担して管理していると仮定する．この場合，式 (1)，式 (2) に当てはめて管理者の作業時間を算出すると，サービスを開始する始めの月は 16,667 分かかり，その翌月からは 6,667 分かかることになる．また，管理者が 1 日あたり平均 8 時間の労

働，月 20 日間勤務したとすると，1 カ月あたりの作業時間は 9,600 分である．つまり，サービスを開始する始めの月の作業負荷は 174%，その翌月からは 70%となる．サービスを開始する月に関しては，規定の業務時間だけでは対応できず，翌月からもデータ配置の作業だけで，管理者が行う作業の 70%がデータ配置の作業になる．これは，他の管理業務も考えると，現実的ではない．

また，仮想環境の進展によりシステムが複雑化し各リソースの依存関係も考慮しなければならないことを考慮すると，1 サービスあたりのデータ配置計画時間が長くなるだけでなく，複数管理者による分業が難しくなり，さらに負荷は増大する．

3. 自動データ適正配置方式の提案

本章では，2 章の従来技術の問題点を解決すべく，自動データ適正配置方式を提案する．まず初めに，方針，システム構成を述べた後，本提案方式の特徴であるデータ適正配置アルゴリズムについて述べる．

3.1 方針

従来，管理者がデータセンタ全体の性能とメディアコストを考慮しデータ移動を行う際，以下のステップで運用するのが一般的であった．

(1) 監視/分析
仮想サーバ，ストレージ装置各々の管理ソフトウェアを使い性能，構成情報を監視し，性能比較を行うことで，性能ボトルネックの部位を特定．

(2) 計画
性能ボトルネックの部位情報をもとに，移動するデータおよび移動先を管理者の知識や経験により決定．

(3) 実行
計画に従い，仮想サーバもしくはストレージ装置のデータ移動技術によりデータ移動を実行．

そこで，本提案では，この運用ステップをベースとし，2.2 節の問題点を解決すべく以下の方針で解決する．

[方針 1] 管理ソフトウェアによる仮想サーバ，ストレージ装置の性能・構成情報の一元管理

[方針 2] 管理者の知識や経験に依存していた計画ステップをアルゴリズム化

[方針 3] 方針 2 と方針 3 により，監視/分析，計画，実行の各ステップを連携させ自動化

次節より，本方針に従い設計したシステム構成およびデータ適正配置アルゴリズムについて説明する．

3.2 システム構成

提案方式のシステム構成を図 3 に示す．本システムでは，まず初めに，仮想サーバ，ストレージ装置のインフラ機器より性能・構成情報を収集する．次に，収集した情報

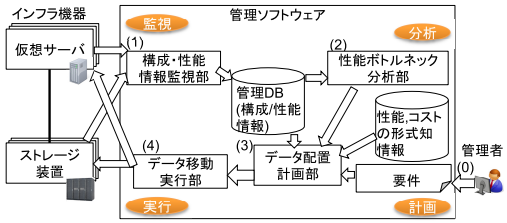

図 3 システム構成
Fig. 3 System architecture for this proposal.

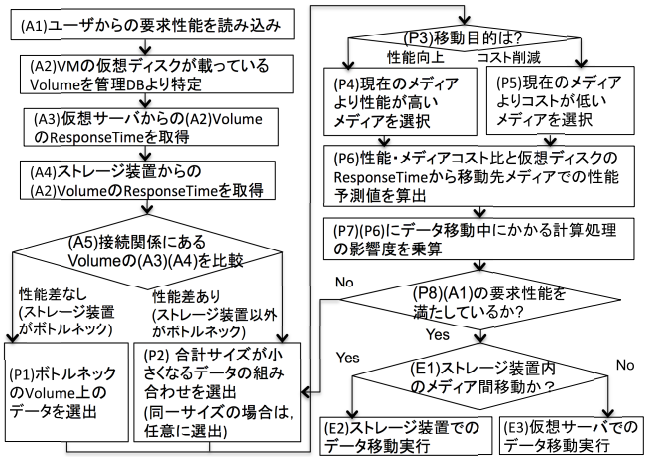

図 4 データ適正配置アルゴリズムのフローチャート
Fig. 4 Flow of automatic optimization of data location.

から性能ボトルネック分析部，データ配置計画部でデータ配置の計画を立て，最後に仮想サーバまたはストレージ装置に対しデータ移動を指示する．これをユーザの性能要件を満たすまで繰り返し，データの適正配置を行う．

なお，本システムでは，データ配置計画を行うにあたり，ユーザからの性能要件を事前設定し，データ配置計画を立てる．この性能要件については，管理者とユーザの間で取り交わされるサービスレベルの保証契約である SLA（Service Level Agreement）を想定している．

3.3 データ適正配置アルゴリズム

3.2 節で示したシステムで動作するデータ配置を行うためのアルゴリズムであるデータ適正配置アルゴリズムについて述べる．データ適正配置アルゴリズムのフローを図 4 に示す．

データ適正配置アルゴリズムは，3.1 節に示した従来管理者が行っていた監視/分析，計画，実行の 3 つのフェーズで実現する．監視/分析フェーズは，図 4 (A1)–(A5) で示されるフェーズ，計画フェーズは，図 4 (P1)–(P8) のフェーズ，実行フェーズは，図 4 (E1)–(E3) のフェーズ，である．

監視/分析フェーズでは，仮想サーバとストレージ装置から取得した構成情報から VM に割り当てられた仮想ディスクが載っている Volume を特定する．その後，仮想サーバとストレージ装置から取得した Volume の性能を比較し，ボトルネック箇所を特定する．

計画フェーズでは，まず初めにボトルネックの箇所に応じ移動対象となるデータ（仮想ディスク）の決定を行う．次に，ユーザからの要求性能から移動目的が性能向上なのか，コスト削減なのかを判断し，目的に応じて移動先のメディアを決定する．この移動対象データの候補決定ステップ（P1，P2）と，移動先の決定ステップ（P4-P8）では，従来管理者が有しているメディアの性能やコストに関する知識を形式知化し算出する．

最後の実行フェーズでは，計画フェーズで決定した移動対象データと移動先から，ストレージ装置のデータ移動を使いデータ移動を行うか，または仮想サーバのデータ移動技術を使いデータ移動を行うかを決定し，データ移動を実行する．

次項より，このデータ適正配置アルゴリズムの計画フェーズの移動対象データの候補決定と移動先の決定について詳細に述べる．

3.3.1 移動対象データの候補決定

図 4 (P1，P2) に示す移動対象データの候補決定では，単純に性能ボトルネックとなっているデータを移動すればよいとは限らない．性能ボトルネックとなっているデータを移動させることで移動中にさらなる性能低下が発生する場合には，他のデータを移動させ，性能改善を行うケースもある．そこで，性能ボトルネックの部位により整理する．

性能ボトルネックがストレージ装置にある場合，性能ボトルネックになっているデータを，I/O 性能の高いメディアに移動させる必要がある．この場合，表 1 に示したデータ移動技術の特徴より，VM への負荷がないストレージ装置でのデータ移動を行うのがよい．そのため，本ケースの場合，データ適正配置アルゴリズムでは，性能ボトルネックになっているデータを移動対象データの候補とする．

しかし，性能ボトルネックが仮想サーバもしくはネットワークにある場合，仮想サーバのデータ移動技術により移動を行うことになる．これは，表 1 の特徴に示すように仮想サーバへ負荷がかかり，仮想サーバが提供する VM 上で実行している計算処理が低下する．そのため，仮想サーバへの負荷が少なくなるように VM への影響が小さいストレージリソースを選出することが重要である．VM への影響が小さくなるために考慮すべき重要なポイントは移動時間の短さである．そのため，本ケースの場合，データ適正配置アルゴリズムの移動対象データの候補決定では，移動時間を短くするため，移動するデータのサイズが最小となるデータの集合を求める．

移動対象データの候補は，図 4(A2) の情報から性能ボトルネックの VM が載っている Volume 上に存在する VM 群の中から選出する．各々の VM に割り当てられた仮想ディスクのサイズ S とすると，式 (3) に示すように移動対象の候補となるデータの合計サイズ D が算出できる．この D が最小となる移動対象データの組合せを求め，データ適正配置アルゴリズムにおける図 4 (P2) に示す移動対象データの候補とする．

$$D = \sum_{k=1}^{n} S_k \tag{3}$$

3.3.2 移動先の決定

次に，図 4 (P3–P8) の移動先の決定について述べる．

移動先の決定では，移動後の性能・メディアコストのバランスと，移動中にかかる VM 上の計算処理への影響，について考慮する必要がある．移動後の性能・メディアコストのバランスについては，ユーザの性能要件を満たしつつ最もメディアコストの低価格なメディアの選別が重要である．

そこで，仮想サーバ，ストレージ装置が備える各種メディアにデータを配置したときの VM 上の計算処理性能とメディアコストの関係を示した性能・メディアコスト比，およびデータ移動中の VM 上の計算処理への影響度，を形式知化し利用することで，データの移動先を決定する．

データの移動先決定のステップである図 4(P3–P8) を以下に説明する．

(P3)–(P5) 1 段階性能が高いメディアもしくはコストが安いメディアを移動先として選出する．

(P6)現時点の VM の仮想ディスクの性能（ResponseTime）に，性能・メディアコスト比（後述の表 3）の性能値を乗算し，移動後の性能予測値を算出．

(P7)データの移動中にかかる VM 上の計算処理への影響（後述の表 4）を加味し (P6) の性能予測値を補正する．

(P8)(P7) の性能予測値が，ユーザの性能要件（ResponseTime）を満たさない場合，再度データ移動候補の選出 (P2) に戻り，次候補を選出しデータ移動先決定ステップを繰り返す．

これにより，ユーザの性能要件を満たしつつ，低価格なストレージを移動先として選出する．

4. 測定

測定では，3.3 節で述べた性能・メディアコスト比とデータ移動中の VM 上の計算処理への影響を形式知化すべく事前測定を行う．その後，形式知を適用した試作システムを用い，効果を測定する．

4.1 測定環境

測定環境を**表 2** と**図 5** に示す．測定では，複数 VM が協調して処理を実施する分散処理環境を用い測定を行った．

4.2 データ適正配置アルゴリズム向け事前測定

データ適正配置アルゴリズムで利用する管理者が従来有していたサーバ/ストレージ装置横断での各種メディアの性能・メディアコストの情報とデータ移動による VM 上の

表 2 測定環境
Table 2 Measurement environment.

物理サーバ	Dell OptiPlex （CPU: Intel Corei7 3.4 GHz, Memory16G）x2
仮想サーバ	VMware ESXi5
ストレージ装置	Hitachi Virtual Storage Platform
接続形態	物理サーバ/ストレージ装置間 （FC 直接続，8Gb/秒）
ミドルウェア	分散処理環境 Hadoop 1.0.2
VM	CentOS 6 x 10 台（1 VM あたり仮想ディスク 1 個）
測定プログラム	Hadoop1.0.2 付属の terasort（2 G のデータをソート）
試作プログラム	Windows7, Java6 で開発

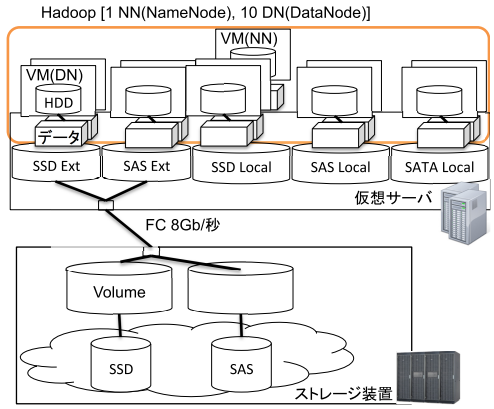

図 5 測定システム
Fig. 5 Measurement system.

表 3 性能・メディアコスト比
Table 3 Relation performance and media cost.

	SSD(Local)	SAS(Local)	SATA(Local)	SSD(Ext)	SAS(Ext)
性能比	1.0	2.9	5.2	1.0	2.9
コスト比	49.5	17.3	1.0	99.0	34.6

計算処理へ与える影響，を形式知化すべく測定を実施した．

4.2.1 性能・メディアコスト比の測定

本測定では，3.3.2 項の移動先の決定で利用する性能・メディアコスト比を形式知化すべく測定を行った．本測定では，4.1 節の環境で，単一種類メディアで構成された Volume 上に全 10 個の DataNode の VM のデータを配置した後，測定プログラムを実施し，各種メディアにデータを配置した場合における VM 上の計算処理性能を測定した．さらに，各種メディアのコストに関しては，2011 年 SNIA Data Protection and Capacity Optimization（DPCO）Committee [8], [9] で公開されている GBytes あたりのコストをもとに比率を算出した．

測定結果を**表 3** に示す．メディア種別の Local は，仮想サーバが備える内蔵のメディアを示し，Ext は，ストレージ装置のメディアをそれぞれ示す．性能は，実行時間の最も短かった SSD(Local) を基準（1.0）とし，メディアごとの実行時間を比率で示し，コストは，$/GBytes の最も安価であった SAS(Local) を基準（1.0）とし，それぞれ比率で示す．表 3 を図 4 (P6) で利用する形式知として提案方式の試作システムに適用する．

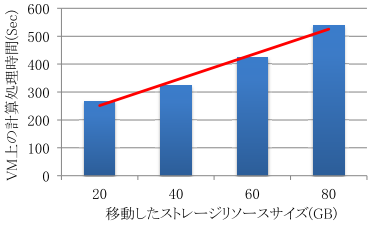

図 6 移動データのサイズが VM に与える影響
Fig. 6 Influence which size of the data gives to VM.

4.2.2 データ移動中の VM 上の計算処理への影響度

本測定では，移動データのサイズ，異種メディア間のデータ移動，がそれぞれ VM 上の計算処理に与える影響について測定した．

(1) 移動データのサイズが与える影響

最初の測定では，移動するデータのサイズが VM 上の計算処理に与える影響の傾向を測定すべく 20，40，60，80 Gbytes のデータを用意し測定を行った．測定では，VM 上で測定プログラムを実行中に，20–80 Gbytes のデータ 1 個を SSD(Local) から SAS(Local) へ移動させ，測定プログラムの実行時間について測定した．測定結果を**図 6** に示す．本結果より，移動するデータのサイズと VM 上の計算処理の時間の関係は，単調増加であることが判明した．

(2) 異種メディア間のデータ移動が与える影響

次の測定では，20GBytes のデータを異種メディア間を 1 個移動させ，VM 上での実行時間を計測した．結果を，**表 4** に示す．図 6 の 20 Gbytes のデータを SSD(Local) から SAS(Local) への移動した場合の VM 上の実行時間を基準（1.0）とし，各メディア間の移動にかかる計算処理への影響を示す．

その結果，VM 上の実行時間には，各種メディアの性能が大きく影響し，最も性能の低い SATA(Local) から SAS(Ext) への移動が，実行時間に最も影響があることが判明した．この表 4 をデータ適正配置アルゴリズム図 4 (P7) で利用する形式知として提案方式の試作システムに適用する．

さらに，この表 4 と図 6 の測定結果を使い，ユーザの性能要件を満たすか否かの判定を行う．図 6 の測定結果より，データのサイズが VM 上の計算処理に与える影響は単調増加であり，1 Gbytes あたり 4.6 秒の傾きで増加であることを加味し，表 4 の異種メディア間のデータ移動にかかる影響を M，式 (3) により求めた移動対象データの候補のデータサイズ D，より移動中の ResponseTime を求める．さらに，現時点の VM の仮想ディスクの性能（ResponseTime）に表 3 の性能値を乗算し算出した移動後の ResponseTime の予測値 $T0$ を加算することで，式 (4) の右辺で移動中の負荷も加味した ResponseTime の予測値を求める．さらに，ユーザの性能要件の ResponseTime を U，とすると，式 (4) のようになる．

$$U \geq 4.6DM + T0 \tag{4}$$

表 4 データ移動にかかる計算処理への影響度
Table 4 Influence to computing which depends on data migration.

移動元＼移動先	SSD(Local)	SAS(Local)	SATA(Local)	SSD(Ext)	SAS(Ext)
SSD(Local)	-	1.0	3.0	0.8	1.1
SAS(Local)	1.1	-	4.8	1.1	1.6
SATA(Local)	3.2	4.7	-	3.3	4.8
SSD(Ext)	0.7	1.1	3.4	-	0.0
SAS(Ext)	1.1	1.6	5.3	0.0	-

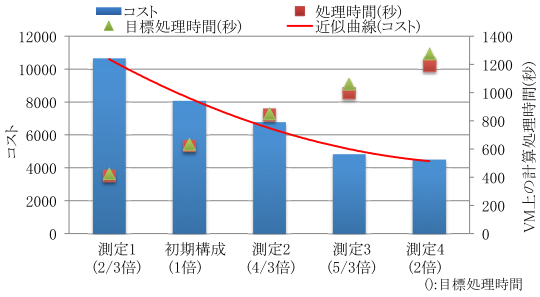

図 7 試作システムによる効果
Fig. 7 Performance and media cost for the proposal system.

表 5 提案方式による移動後のデータ配置
Table 5 Data location after a data migration.

配置先＼測定ケース	測定 1	初期構成	測定 2	測定 3	測定 4
SSD(Local)	2	2	2	1	1
SAS(Local)	2	2	2	3	4
SATA(Local)	2	2	2	2	2
SSD(Ext)	4	2	1	0	0
SAS(Ext)	0	2	3	4	3

この式 (4) の条件を満たしつつ，右辺と左辺の差分が最も小さくなる移動データの集合が，最終決定する移動対象のデータとなる．また，右辺の値が同一となる移動対象のデータの集合が複数存在する場合には，いずれの移動対象のデータでも効果は同じと考え，データ適正配置アルゴリズムでは任意の集合 1 つを選択する．

提案方式の試作システムでは，式 (4) の判定をデータ適正配置アルゴリズム図 4 (P8) のユーザの性能要件を満たすか否かの判定に適用した．

4.3 提案方式の試作システムによる測定

次の測定では，3.2 節に示す試作システムに，4.2 節の測定結果を管理者の形式知として用いたデータ適正配置アルゴリズムを適用した試作システムを使い測定を行った．

測定方法は，10 個の DataNode の VM が各メディアに 2 個ずつ均等に配置されている場合を初期構成とし，ユーザの性能要件を初期構成の性能より 2/3 倍，4/3 倍，5/3 倍，2 倍の ResponseTime を指定し，VM 上での測定プログラムの実行時間を計測した．

結果を**図 7** に示し，それぞれの測定ケースのデータ配置結果を**表 5** を示す．表 5 の数値は各メディアに搭載されている VM の数を示す．たとえば，ユーザの性能要求が初期構成より 2/3 倍の ResponseTime の向上を目指す測定 1 の場合，表 5 の初期構成と比較し，SAS(Ext) のデータ 2 個が SSD(Ext) へ移動したことが分かる．また，図 7 のメディアコストに関しては，表 5 の結果から表 3 のコスト比を用い算出した結果である．

この図 7 と表 5 の結果より，本提案方式により，ユーザの性能要件を満たしつつメディアコストの低減に成功しているのが分かる．また，データの移動傾向としては，VM 上の計算処理に影響のないストレージ装置によるデータ移動が優先的に実施されている．そのため，ストレージ装置のメディアの価格差と，仮想サーバのメディアの価格差より，メディアコスト低減は 2 次関数の削減傾向となった．

5. 議論

本章では，2.2 節に述べた 3 つの問題点，異なるレイヤのデータ移動技術の矛盾，性能とメディアコストの個別最適，管理者への高負荷，について提案方式の有効性を議論する．

5.1 データ移動技術の使い分け

本提案方式では，3.1 節と 3.2 節に示すように，まず仮想サーバとストレージ装置より，性能・構成情報を収集し一元管理行い，本情報をもとに性能ボトルネックの部位を特定し，ボトルネック部位に応じデータ移動技術を使い分けるデータ適正配置アルゴリズムを実現した．

これにより，4.3 節の表 5 測定 3，4 のようにメディアコストの削減を目指す場合には，仮想サーバ，ストレージ装置ともメディアコストを削減するといったように同一のポリシでデータ配置を実現できた．つまり，従来技術では，異なるレイヤで別々の性能・構成情報，別々のアルゴリズムでデータ配置を行っていたため発生していたデータ移動技術の矛盾の問題を解決した．

続いて，データ移動技術を使い分けの有用性について議論する．

大規模データセンタでは，仮想サーバやストレージ装置を使い，複数サービスの IT インフラをコンソリデーション

し設備コストを削減しようという動きが活発化している．しかし，大規模データセンタのようにユーザ要件の異なるサービスが複数存在する環境においては，コンソリデーションが困難であった．たとえば，あるサービスはデータ分析用途のようにコスト削減よりも性能を重要視し，I/O性能を劣化させずより高速なメディアへのデータ移動が求められたり，別のサービスではVDI（Virtual Desktop Infrastructure）用途のようにユーザの利用頻度の少ない時間帯であれば，性能が多少劣化してもコスト削減を優先させるデータ移動が求められたり，と様々なユーザ要件が混在している．このような場合，表1に示すようにデータ移動技術の特徴から，前者のユーザ要件であればストレージ装置が有するデータ移動技術の利用がよく，後者であれば仮想サーバが有するデータ移動技術を利用するのがよい．しかし，従来のデータ移動技術では，2.2節 (1) に示すように異なるレイヤのデータ移動技術を混在させることができなかった．そのため，これらのユーザ要件を満たすためには，それぞれの用途別に仮想サーバやストレージ装置を用意し，仮想サーバもしくはストレージ装置のどちらか一方のデータ移動のみ有効化することで対応せざるをえなかった．

これに対し，本提案方式では，3.3節に示すデータ適正配置アルゴリズムにより，異なるレイヤのデータ移動技術の使い分けを実現した．これにより，仮想サーバとストレージ装置の両方のデータ移動技術を有効化しても矛盾したデータ移動を行うことなく一貫したポリシでデータ移動が可能となった．本提案方式を大規模データセンタに適用する場合，ユーザ要件の異なるサービスが複数存在する場合であっても，それらのサービスがデータを共有していない構成であれば，同一の仮想サーバやストレージ装置で異なるユーザ要件を満たすことが可能となる．その結果，よりいっそうのコンソリデーションを促進させることができ，設備コスト削減が見込めるようになった．

このように，提案方式は，異なるレイヤのデータ移動技術を使い分け矛盾なくデータ移動を実現したことで，大規模データセンタにおいても設備コストの削減において有用であるといえる．

5.2 性能とメディアコストのバランス

本提案方式では，4.3節に示すようにユーザの性能要件を満たしつつメディアコストの削減を実現した．以下，メディアコスト，性能面から有用性を議論する．

まず，メディアコスト面について議論する．本測定環境において，本提案方式を使わず性能のみを考慮し，最短時間で計算処理がおわるようにすべてのデータをSSD(Ext)に置くという単純なデータ配置と仮定した場合，メディアコストは表3より，19,800となる．これに対し，ユーザの性能要件が4.3節の測定1（2/3倍）のケースでは，すべてのデータをSSD(Ext)に置いた場合と比較し，ユーザ要件を満たしつつ47%のメディアコスト削減，測定4のケースでは77%ものメディアコスト削減となる．

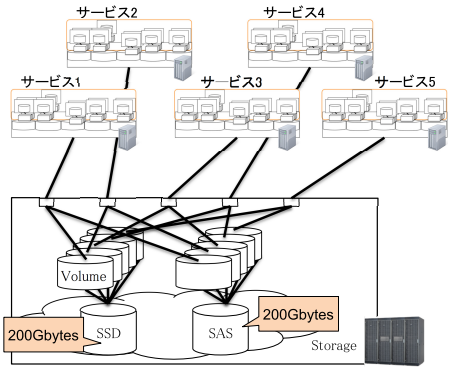

図 8 5つのサービスの並列実行環境
Fig. 8 Environment of 5 parallel services.

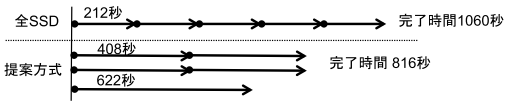

図 9 5つのサービスを並列実行したケースの実行時間
Fig. 9 Time in executing 5 parallel services.

この結果より，提案方式がメディアコスト削減に有効な方式であるといえる．

次に，性能面について議論する．データセンタでは，ストレージのリソースは容量制限があり，さらに複数の計算処理が同時に実行されるのが一般的である．そこで，メディアごとのストレージ容量は200 Gbytes，4.1節のHadoop環境を1サービスとし5つのサービスが並列実行される環境を仮定する．この仮定する環境を**図 8** に示す．

この環境で，すべてのデータをSSD(Ext)におき処理するという単純なデータ配置ポリシで運用した場合，最大ストレージ容量から，1度に1つのサービス（10 VM）分のデータしかSSDに配置できず，SSD(Ext)が空くまで他のサービスが待ち状態となり，5つのサービスをシーケンシャルに実行することになる．これに対し，提案方式では，表5から，測定1と同一構成のプロセスを2つ，初期構成と同一構成のプロセスを1つ並列に実行できる．その結果，**図 9** に示すようにすべてのデータをSSD(Ext)に置いた場合の実行時間212秒を，シーケンシャル実行する場合の完了時間1,060秒にくらべ，提案方式では，測定1と同一構成の実行時間408秒，測定2と同一構成の実行時間622秒が並列実行されるため，完了時間は816秒となる．つまり，すべてのデータをSSD(Ext)に置いた場合より提案方式の方がトータルでの性能が向上し，23%もの時間が短縮できる．なお，この性能向上は，並列実行される計算処理が増えれば増えるだけ性能改善効果は大きくなる．つまり，複数のサービスを提供する大規模データセンタにおいては，よりいっそうの効果が見込める．

このように，本提案方式は，仮想サーバやストレージ装置の個別最適の場合と比べ，仮想サーバやストレージ装置といったトータルなシステムで判断しデータセンタ全体最適となるようにデータ配置を決定することから，メディアコスト面・性能面のいずれの観点においても有効である．

5.3 管理者への負荷

本提案方式では，これまで管理者が有していた各種メディアの性能・コスト，仮想サーバやストレージ装置の構成情報といった知識を使い実施していたデータの配置計画に関し，管理者の思考パターンを 3.3 節のデータ適正配置アルゴリズムで実現した，また，本アルゴリズムの実現にあたり，従来の管理者の知識に該当する情報については，性能・メディアコスト比（表 3），移動にかかる計算処理への影響度（表 4），ユーザの性能要件を満たすか否かの判定条件（式 (4)）を測定により求め形式知化した．これにより，管理者がユーザの性能要件をサービスの初期構築時に設定するだけで，データセンタ全体の性能・メディアコストを考慮したデータの配置計画を行うデータ適正配置の自動化を実現した．

本提案方式の効果について，2.2 節で示した式 (1)，式 (2) にあてはめて算出する．本提案方式では，データ配置の計画をアルゴリズム化し自動化を実現したことで，管理者による 1 サービスあたりのデータ配置計画時間 Pt を 0 とすることができる．これは，式 (2) の Wt が 0 となること意味し，式 (5) の作業時間のみとなる．

$$W0 = ItS/A \tag{5}$$

これを 2015 年に予測されるストレージ装置が 100 台以上，VM50,000 台以上となる大規模なデータセンタでの作業を，2.2 節の仮定をベースとし，1 サービスの初期構築にかかる時間のみ提案方式のユーザの性能要件を設定する時間 5 分を加算した 65 分として算出すると，従来まではサービスを開始する月の作業負荷は 174% だったのに対し，113%と低減し，翌月からの作業負荷は 70% が 0%となる．つまり本提案方式により，サービスを開始する月の作業負荷は 61%，その翌月からの作業負荷は 70%削減の効果がある．

さらに，サービスを開始する月の作業負荷に関しては，依然 100%を超えているものの，サービスを開始する月のみのため，一時期のみの増員や開始するサービス数を 2 月に分割するなどの対策が可能である．また，1 サービスの初期構築にかかる時間を短縮させる技術として，近年，VM のクローニング技術や一括デプロイ技術が進展しているため，仮定の 65 分よりも短縮され管理者の負荷が削減されると予測する．

このように，本提案方式は，管理者の負荷の削減に有効であり，特に年々大規模化しているデータセンタにおいて必要性は年々高まってくると考える．

6. おわりに

従来のデータ移動技術では，異なるレイヤのデータ移動技術の矛盾，性能とメディアコストの個別最適，管理者への高負荷，という 3 つの問題をかかえていた．そこで，本論文では，従来管理者の知識と経験に頼っていたデータ配置の計画をアルゴリズムし，仮想サーバとストレージ装置のデータ移動技術を使い分けを行うことで，ユーザの性能要件を満たしつつデータセンタ全体でのメディアコストが最も削減可能なメディアへ配置する自動データ適正配置方式を提案し，問題を解決した．

本提案方式を適用した試作プログラムを開発し測定した結果，2012 年時点のメディアの性能・メディアコストの場合において，ユーザの応答性能を満たしつつ，すべてのデータを SSD に配置したときに比べ 47%のメディアコストを削減できることを実証した．また，管理者の負荷に関しては，2015 年の大規模データセンタにおける管理者の負荷を 70%削減が見込める試算である．

本論文では，2012 年現在のメディアの性能・コストを使い効果を実証したが，データを保存するメディアの性能・コストは，新メディアの登場によっても変動する．しかし，将来，新メディアが登場することで，2012 年現在の各種メディアの性能とメディアコストの差が大きくなると，本提案方式の効果はよりいっそう大きくなる．さらに，近年ビックデータの分析が注目されており，I/O 性能がビックデータの高速な分析を実現するために重要な要素となっている状況を考えると，データの保存場所であるメディアの性能・コストに着目した本提案方式の重要度はますます大きくなる．このように，年々大規模化するデータセンタにおいて，管理者の負担を削減しつつ，性能とメディアコストを適正化したデータの配置を行う本提案方式は有効であり，本提案方式がなければ，仮想サーバやストレージ装置が提供しているデータ移動技術を使いこなした運用管理は実現できない．

今後は，普及が予見される複数データセンタをまたがったデータの移動についても，本提案方式の有効性を検証するのが課題である．

参考文献

[1] Kikuchi, S. and Matsumoto, Y.: Performance Modeling of Concurrent Live Migration Operations in Cloud Computing System using PRISM Probabilistic Model Checker, *2011 IEEE 4th International Conference on Cloud Computing*, pp.49–56 (2011).
[2] 江丸裕教，高井昌彰：仮想ボリュームクラスタリング法による動的階層制御ストレージの性能管理，情報処理学会論文誌，Vol.52, No.7, pp.2234–2244 (2011).
[3] VMware: Storage Distributed Resource Scheduler (SDRS)，available from ⟨http://www.vmware.com/jp/

products/datacenter-virtualization/vsphere/vsphere-storage-drs/overview.html⟩
[4] （株）日立製作所：Dynamic Provisioning，入手先⟨http://www.hitachi.co.jp/products/it/storage-solutions/products/software/allsofts/index.html?hard=Virtual%20Storage%20Platform&soft=Dynamic%20Provisioning⟩
[5] （株）日立製作所：Hitachi Dynamic Tiering，入手先⟨http://www.hitachi.co.jp/products/it/storage-solutions/products/software/function.html#02⟩
[6] Holler, A. and Lohani, M.: VMware Storage Distributed Resource Scheduler, *Vmworld 2011*, pp.1–54 (2011).
[7] IDC: Worldwide Enterprise Storage System 2009-2013 Forecast Update (2009).
[8] Freeman, L.: What's Old is New Again Storage Tiering. SNW Spring2012, available from ⟨https://www.eiseverywhere.com/ehome/SNWS2012/56399/⟩
[9] SNIA: SNIA Data Protection And Capacity Optimization Product Selection Guide, available from ⟨http://www.snia.org/forums/dpco/⟩

坂下 幸徳 （正会員）

2003 年北陸先端科学技術大学院大学情報科学研究科情報システム学専行修士課程修了．同年株式会社日立製作所システム開発研究所（現，横浜研究所）入所．現在に至る．IT システム運用管理の研究開発に従事．SNIA 日本支部技術委員会委員長．

三神 京子

2006 年津田塾大学大学院理学研究科修士課程修了．同年株式会社日立製作所システム開発研究所（現，横浜研究所）入所．現在に至る．ストレージシステムおよびシステム運用管理の研究に従事．

金子 聡 （正会員）

2008 年電気通信大学大学院人間コミュニケーション学研究科修士課程修了．同年株式会社日立製作所システム開発研究所（現，横浜研究所）入所．現在に至る．ストレージシステムおよびシステム運用管理の研究に従事．

敷田 幹文 （正会員）

1965 年生．1995 年東京工業大学大学院理工学研究科情報工学専攻博士後期課程修了．博士（工学）．同年北陸先端科学技術大学院大学情報科学センター助手．2012 年同教授．大規模分散システム，グループウェアに関する研究に従事．ACM，電子情報通信学会，日本ソフトウェア科学会各会員．