データサイズの相関係数を用いた未使用VMの特定と圧縮によるバックアップフォルダのストレージの削減

川端 ももの[1]　増田 和範[1]　串田 高幸[1]

概要：東京工科大学クラウド・分散システム研究室では，サーバ内の VM のバックアップデータを NAS で管理している．サーバ内の全 VM をフルバックアップすると 1 日あたり 3TB のストレージが必要になる．このバックアップデータは 5 日間保持されるため，合計で 15TB のストレージが必要になる．バックアップデータの圧縮を行うことでフォルダのデータサイズを縮小させることができる．しかし，バックアップデータから VM を復元するためには解凍する必要がある．そのため，作業を行う VM のフォルダを圧縮するべきではない．よって，作業をしていない VM のフォルダのみを圧縮する必要がある．作業をしていない VM を判断する方法として，バックアップデータのファイルのサイズを分析することが挙げられる．課題はバックアップファイルのサイズのみを考慮した場合，圧縮するフォルダと圧縮しないフォルダの判断が出来ないことである．この課題の解決方法として，圧縮対象のバックアップデータ量と作業していない VM のデータ量の変化相関係数を算出し，圧縮するフォルダを自動選別する手法を提案する．評価実験では 20 台の VM で相関係数を算出し，選別の際の相関係数の閾値を 0.5 から 0.1 ずつ増加させた際の正解率を比較した．評価実験では，VM を使用している人にアンケートを行い，圧縮するフォルダの正解率を出して評価を行った．評価実験を行った結果，一番高い正解率は 85.0%であり，その際の相関係数の閾値は 0.5 と 0.6 である．

1. はじめに

背景

サーバーにあるデータは，機器の故障や人為的ミス，災害によりデータが消失または破損し，使用できなくなることがある [1]．その際，バックアップは消失したデータの復元や障害から回復することができるため，バックアップはデータの保護に必須である [2][3].

バックアップ方法の 1 つとしてフルバックアップが挙げられる．フルバックアップはバックアップ毎にディレクトリを含む全てのデータをバックアップする形式のことである [4]．1 回のリストアで全てのデータを復元することができるため差分バックアップや増分バックアップと比較してリストアする時間が早い [5]．しかし，毎回全てのデータをバックアップするため，データサイズが大きい．また，バックアップの世代管理を行うことで，数日分のバックアップを残しておくことが可能である．世代管理は，数日前のデータを復元することができるため重要である[*1]．圧縮を行うことで記録されている情報を失わずに，データサイズを小さくすることができる．バックアップの圧縮を行うことで，ストレージの削減やデータセキュリティの強化，転送時間の削減の利点がある [6]．しかし，バックアップファイルを圧縮した場合，リストアを実施する前に解凍する時間が必要である．

仮想環境技術の 1 つにハイパーバイザー型がある．ハイパーバイザー型は，ゲスト OS を直接サーバへインストールし VM（Virtual Machine）を稼働させる方式のことである[*2]．ハイパーバイザーが直接ハードウェアを制御することでホスト OS を使用せず VM の実行を行うことができる．VMWare ESXi は完全仮想化に基づくハイパーバイザーである．完全仮想化はハードウェア全体のシミュレーションが含まれており，変更を加えることなくゲスト OS をインストールして実行する事が可能である [7].

課題

課題として，データサイズのみでは，作業しているフォルダと作業していないフォルダの見分けがつかないことがあげられる．圧縮はデータサイズが小さくなるが再度使用する際に解凍時間がかかり，作業が遅くなる．そのため，圧縮するフォルダと圧縮しないフォルダを分けることが必

[1] 東京工科大学コンピュータサイエンス学部
〒 192–0982 東京都八王子市片倉町 1404-1

[*1] https://business.ntt-east.co.jp/service/coworkstorage/column/sedaikanri/

[*2] https://thinkit.co.jp/story/2012/10/17/3722

要である．現在作業しているフォルダはすぐ使用する可能性が高いため圧縮せず，作業していないフォルダを圧縮することでストレージを削減する．

図 1 はバックアップサーバのフォルダ構成についてである．バックアップサーバには作業している VM と作業していない VM のフォルダが混在している．そのため，利用者が見てもどれが作業している VM のフォルダなのか判断できない．

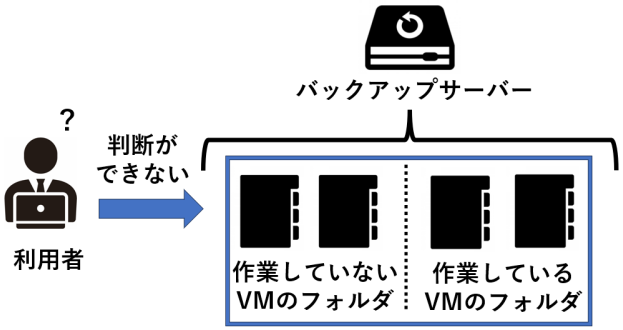

図 1 バックアップサーバのフォルダ構成

課題の基礎実験

作業していない VM フォルダの 5 日間のデータサイズの変化を測定した．VM1 と VM2 と VM3 の VM を使用した．3 つの VM は，作業をしていない VM とする．対象期間において，フォルダ作成や SSH を行わずにデータサイズを測定した．データサイズは du コマンドで vmdk ファイルのデータサイズを取得している. 測定期間は 2023 年 12 月 15 日から 12 月 19 日の 5 日間である．

作業していない VM フォルダのデータサイズを表 1 に表す．前日とのデータサイズの差を表 2 に表す．表 2 の VM3 のデータサイズの差が 544 と増えていることから，データサイズは使用していない VM でも変更されることが分かった．また，16 日-15 日は 544,17 日-16 日の差は 1920 と日にちごとにデータの変化サイズが違うことも分かる．VM2 の 16 日-15 日の差は 1924,VM3 の 16 日-15 日の差は-11504 と VM 別にデータの変化サイズを見てもそれぞれ異なっている．さらに，データサイズは増加するだけではなく，VM3 の 15 日と 16 日の差が-11504 でデータサイズが減少することもある．これは，tmpfs ファイルのデータサイズが減少しているからである．

このように，データの変化サイズには法則性が見られない．そのため，データサイズで見た際に，どれが作業しているフォルダか分からず，圧縮するフォルダと圧縮しないフォルダの判断が出来ない．

各章の概要

2 章は関連研究について述べる．3 章はデータサイズの相関係数を用いた圧縮するフォルダの選別方法の提案について述べる．4 章は提案ソフトウェアの実装について述べる．5 章は正解率とストレージに関する評価実験について述べる．6 章は議論について述べる．7 章はまとめについて述べる．

表 1 作業していない VM フォルダのデータサイズ (Byte)

	VM1	VM2	VM3
2023/12/15	3631392	3593536	3589054
2023/12/16	3631936	35955456	3587552
2023/12/17	3636832	3596032	3350400
2023/12/18	4761536	4911168	4871168
2023/12/19	5133312	4919744	5049152

表 2 前日とのデータサイズの差 (Byte)

	VM1	VM2	VM3
16 日-15 日	544	1920	-11504
17 日-16 日	4896	576	-237152
18 日-17 日	1124704	1315136	1520768
19 日-18 日	371776	8576	177984

2. 関連研究

浮動小数点データ圧縮のための圧縮方法を提案する研究がある [8]．この研究では，CPU サイクルを使用しデータサイズを削減することで，I/O 帯域幅を超えると CPU がアイドル状態になってしまうことを解決する．この研究は，圧縮率を上げることで，転送データサイズを削減しているが，再度使用する際にリストアする必要があるため，時間がかかる．

ファイル名を概念のリストに区切ることで，ファイルをクラスタリングする研究がある [9]．この研究では，ファイル名にある英単語や，アプリケーションドメインの用語を略語している．略語を用いて，ファイルのクラスタリングを行う．その結果，90%が正しくクラスタリングされた．しかし，この分別方法では，そのファイルが現在使われているかの判断をすることはできない．

ファイルを 3 種類にクラスタリングし，使用率によって圧縮レベルを変更することでストレージを削減する研究がある [10]．この研究では，1 か月あたりのファイルの使用率によって圧縮率を変更している．使用率が高いものは圧縮率が低く，解凍時間が短い．その結果，既存の方法より効率的にストレージを削減できた．しかし，ファイルの使用率で分別しているため期間内に 1 回だけ大幅に変更されたようなファイルは圧縮率が高くなってしまう．そのため，作業したが圧縮されてしまい解凍に時間がかかることがある．

3. 提案

提案の基礎実験

作業しているフォルダと作業していないフォルダのデータサイズの変化がそれぞれどのくらい類似しているのかについて実験を行った．作業していない VM は課題の基礎実験の 3 つの VM である．作業している VM は VM4 と VM5 と VM6 である．作業している VM も課題の基礎実験と同様，5 日間データサイズの測定を行った．

作業していない VM のデータサイズの増減のグラフを図 2 に表す．作業している VM のデータサイズの変化を図 3 に表す．図 2 において VM1 のデータサイズが少量増加している日は VM2 と VM3 のデータサイズも少量増加している．さらに VM1 のデータサイズが少量減少している日は VM2 と VM3 のデータサイズも少量減少している．このように作業していない VM のデータサイズが同じように変化していることが分かる．図 3 において，2022/12/19 のように VM4 のデータサイズが少サイズ増加している日でも，VM5 のデータサイズは大幅に増減しているように，データサイズの増減は VM ごとに異なったためグラフの形にばらつきがある．そのため，データサイズの変化はそれぞれの VM で違う．

次に，相関係数を算出した．相関係数を式 1 に表す．相関係数は，数値が 1 と-1 に近づくほど相関があると言える．1 は正の相関で片方のデータサイズが増えるともう片方のデータサイズも増える相関であり，-1 は片方のデータサイズが増えるともう片方のデータサイズが減る相関である．作業していない VM は VM1 と VM2 の相関，VM1 と VM3 の相関を算出した．作業している VM は VM4 と VM5 の相関，VM4 と VM6 の相関を算出した．

作業していない VM の相関はそれぞれ 0.98 と 0.99 であり，どちらも強い正の相関であった．作業している VM の相関はそれぞれ 0.18 と 0.54 であり相関関係なしと正の相関であった．このことから，作業していない VM は強い正の相関になるが，作業している VM は相関にばらつきがあることが分かった．

$$\rho = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \qquad (1)$$

ρ　ピアソンの相関係数
n　バックアップフォルダの数
x_i と y_i　各日付のデータサイズ
$\bar{x}$　x_i の平均データサイズ
$\bar{y}$　y_i の平均データサイズ

提案方式

圧縮するファイルを自動判断するソフトウェアを提案する．提案方式を図 4 に表す．このソフトウェアは 3 つの機能がある．1 つ目は，NAS サーバーに接続し，バックアップフォルダ内の vmdk ファイルのデータサイズをバックアップフォルダ内の全ての VM で取得する機能である．

2 つ目は，相関係数を求め，圧縮するフォルダの判別を行う機能である．データサイズはバックアップフォルダに保存されている日数分取得する．作業していない VM を 1 台用意しデータサイズを測定する．作業してない VM の vmdk ファイルとサーバーの VM の vmdk ファイルをそれぞれ 1 つずつ相関係数を計測する．相関係数をもとに圧縮するフォルダの選別を行う．

3 つ目は，相関係数が一定値以上のフォルダを圧縮する機能である．フォルダを圧縮後，元フォルダは削除される．

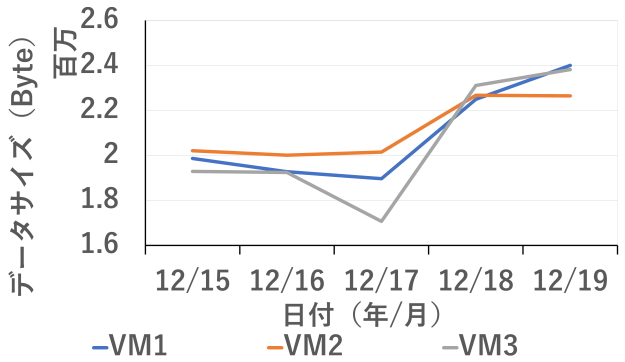

図 2　作業していない VM のデータサイズの増減

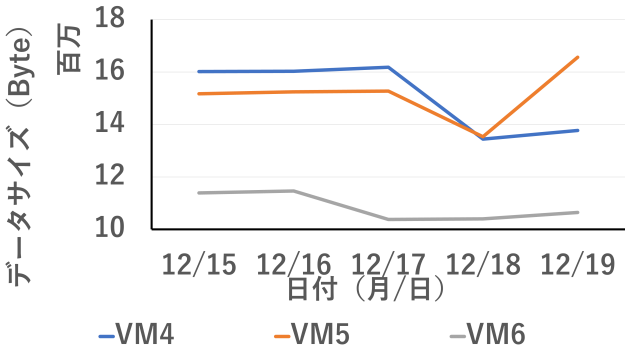

図 3　作業している VM のデータサイズの増減

ユースケース・シナリオ

東京工科大学のクラウド・分散システム研究室の lotus サーバで VM をつくり，実験を行っている状況を想定する．提案ソフトウェア導入前のユースケースシナリオを図 5 に表す．研究生達は VM を作り実験を行っていた．実験は，テスト環境で行い，ファイルの圧縮やログの観察，wordpress のデータ移行，監視などに使われていた．この研究室では外部の web サイトやアプリケーションは作成しておらず，VM で作業する際は作業したい VM に SSH 接続を行っていた．lotus サーバでは VM2,VM4,VM5 のような使用している VM と VM1 と VM3 のような使用さ

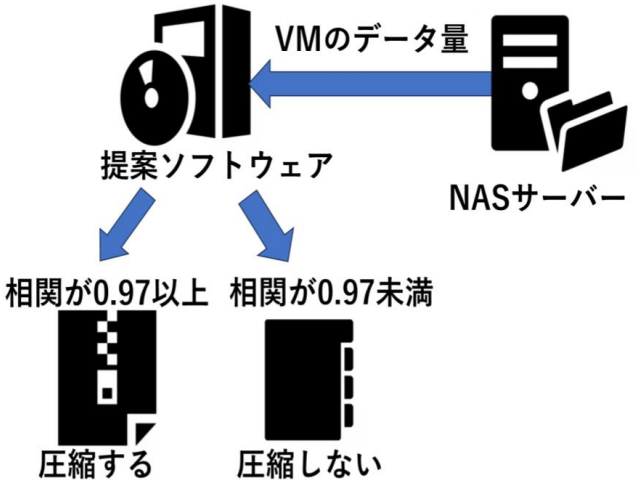

図 4 提案方式

れていない VM が混在していた．これらの VM のバックアップデータは 1 つ 1 つフォルダになっており NAS で管理している．仮想環境のデータ破損やウイルス感染の際にデータを復元するために NAS で毎日 1 回，18 時にフルバックアップを行っている．フルバックアップは 5 日分行われており，1 回のバックアップに 3TB 使用されている．そのため，バックアップのみで 15TB のストレージを使用し，NAS サーバのストレージの 46%を使用していた．バックアップフォルダを圧縮しようとしたが，解凍時間がかかるため，すぐ使用する可能性があるフォルダを圧縮してしまうと作業する時間が遅くなり実験に支障が出る．そのため，現在使用しないフォルダを圧縮しようとしたが，作業しているフォルダとしていないフォルダの判断がつかず，圧縮するフォルダが決められない．

そこで本稿の提案方式を用いる．提案ソフトウェア導入後のユースケースを図 6 に表す．データサイズの相関係数により，作業している VM のフォルダと作業していない VM のフォルダを自動選別する．そのため作業していない VM フォルダのみを圧縮でき，バックアップフォルダのストレージを削減しつつ，すぐ使う可能性のあるフォルダを解凍する時間もかからず作業を行うことができる．

4. 実装

Linux で Python を使用し作成する．提案ソフトウェアを作成するための VM を立てる．次に，作業していない VM として VM α を作成する．VM α は SSH 接続を行わない VM である．

提案ソフトウェアの流れを図 7 に表す．NAS サーバーに接続するために pexpect を使用し SSH 接続を自動で行う機能を作成した．SSH 後 NAS サーバーでバックアップフォルダのディスク上のサイズをバックアップフォルダ日分取得し，リスト化する．取得した値を用いて，VM α のデータサイズと各 VM のデータサイズで相関係数を算出する．算出した相関係数によって圧縮するフォルダと圧縮しないフォルダを分ける．相関係数が一定値以上の VM は圧

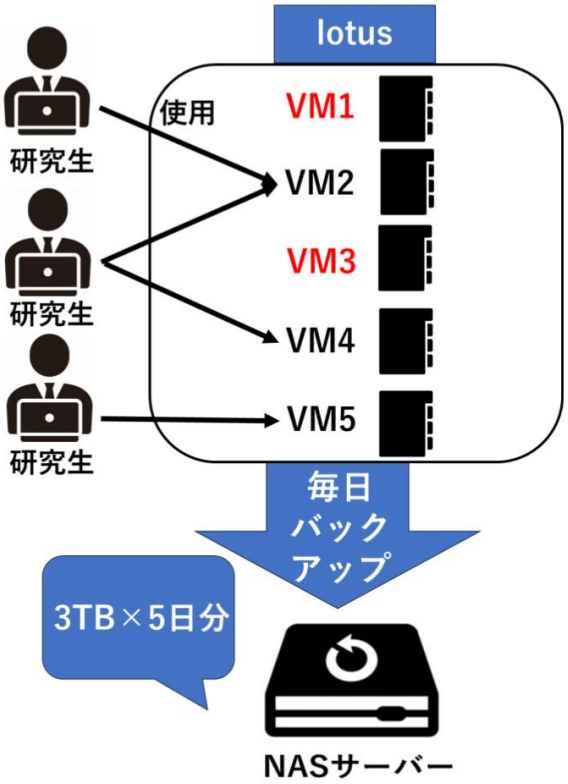

図 5 提案ソフトウェア導入前のユースケースシナリオ

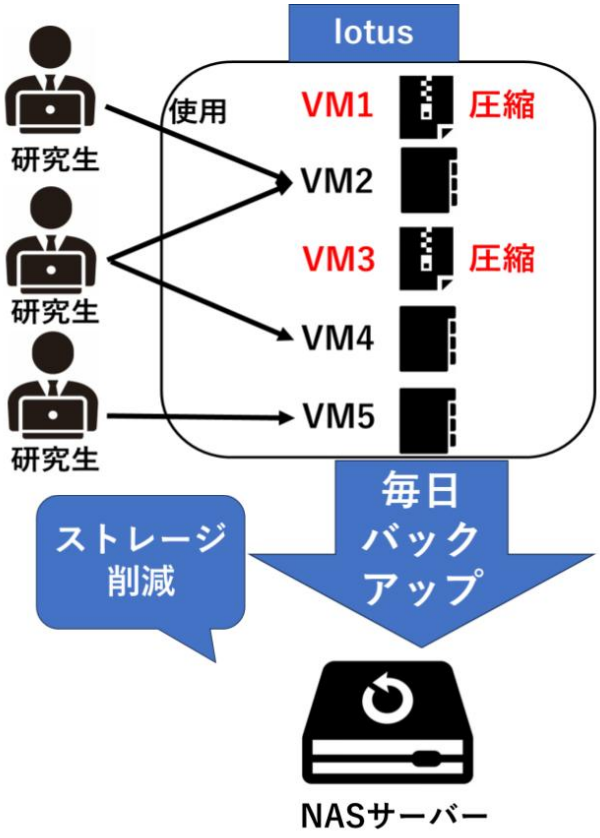

図 6 提案ソフトウェア導入後のユースケースシナリオ

縮を行う．一定値未満は圧縮を行わない．シャットダウンしている VM はデータサイズが変化しないため，常にデータサイズの変化が 0 になる．そのため，全ての日付でデータサイズが同じ VM は作業していない VM とし，圧縮を行う．

圧縮したいフォルダを tar コマンドと gzip コマンドでアーカイブ化し，圧縮する．圧縮後，元のフォルダは削除する．

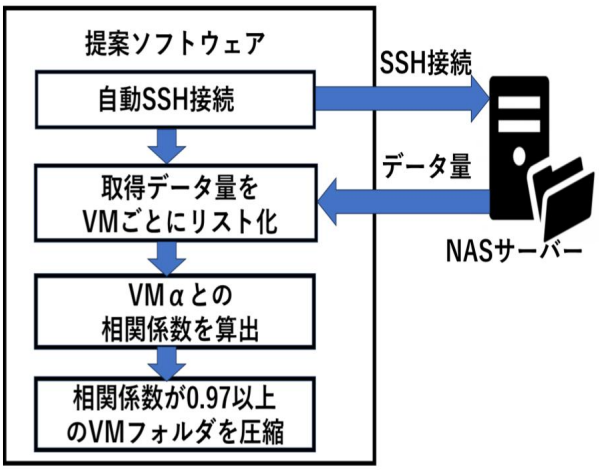

図 7　実装の流れ

5.　評価実験

提案ソフトウェアを使用しているサーバーで VM を使っている人に VM で作業したかアンケートを取り，作業していないフォルダのみを圧縮できたかを正解率用いて評価する．相関係数の閾値を 0.5 から 0.1 ずつ増加させた際の正解率を比較した．Kubernetes クラスターの一部である VM はデータサイズの変化が異なるため対象から除外する．SSH 接続のみを行い，VM 内で変更を行っていない場合は，作業していない VM とする．

実験環境

提案ソフトウェアを使用する VM のサーバーは ESXi を使用し，サーバー名は jasmine である．バックアップフォルダは NAS で管理されており，バックアップフォルダ内には 41 台分のバックアップフォルダが保存されている．jasmine のバックアップフォルダ 6 日分保存されておりトータルで 3TB である．

41 台の内 21 台は SSH していない間も VM が稼働する場合，または Kubernetes クラスター用途して稼働しているため，データ量の変化が異なる．そのため，21 台を省いた 20 台で正解率を測定する．

バックアップフォルダの日付は 2023 年 12 月 19 日，2023 年 12 月 20 日，2023 年 12 月 21 日，2023 年 12 月 22 日，2024 年 1 月 9 日の 5 日間である．日付が飛んでいるのは，12 月 22 日から 3 日間の大学内での計画停電の影響で，バックアップを行っている VM がシャットダウンされてしまいバックアップフォルダが取れていなかったのが原因である．また，12 月 22 日はバックアップが取れていたフォルダが半分であった．そのため，フォルダが無い日のデータサイズは VM がシャットダウンしている状態と考え，前日と同じデータサイズとした．

実験結果と分析

それぞれの正解率を図 8 に表す．図 8 から圧縮する相関係数を下げるほど正解率が上がっている．このことから，作業していない VM の相関係数が 0.9 以上より 0.6 や 0.5 の方が多いことが分かる．

基礎実験と作業していない VM のデータサイズの相関係数が違う原因として，停電でバックアップフォルダのデータサイズが正しくにバックアップできていない可能性が挙げられる．さらに，2023 年 12 月 22 日から 2024 年 1 月 9 日にバックアップ飛んでいるため，2023 年 12 月 23 日から，2024 年 1 月 8 日に新しく作成された VM のデータサイズが取得できなかった．そのため，2023 年 12 月 22 日に VM が無く，2024 年 1 月 9 日に VM がある場合，2024 年 1 月 9 日に新しく作成された VM として，2 データサイズが 1 日分のみになり，4 日分はデータサイズを 0 と仮定した．そのため，正しいデータサイズが取得出来ず，本当は作業していないが作業している VM として，判断されてしまうことも正解率が低い要因の 1 つだと考えられる．

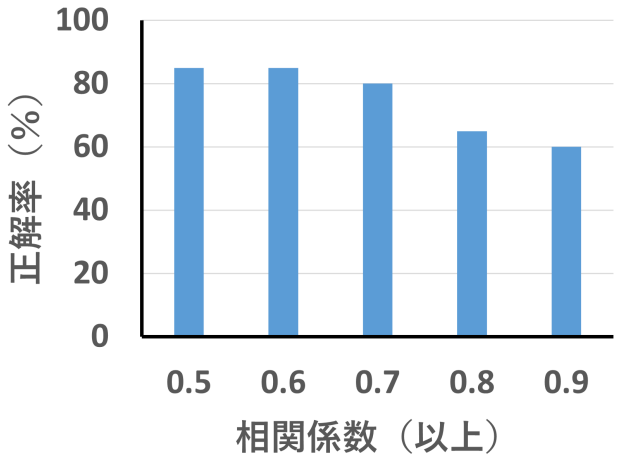

図 8　正解率

6.　議論

本提案ソフトウェアはフルバックアップのみ対応している．増分バックアップや差分バックアップは，数日ごとにフルバックアップを行い，前日やフルバックアップ時との差分を取っていく形式であるため，データサイズが大幅に変更される．そのため，相関係数がフルバックアップ時と

異なる．また，増分バックアップや差分バックアップは全てのデータを残すには，数日分のバックアップデータが必要である．本提案ソフトウェアでは，バックアップを取っている日付分より古いバックアップフォルダは削除しているが，増分バックアップや差分バックアップはデータの一部が消去されてしまう場合がある．

本稿では，停電や人為的ミスで，VM のバックアップフォルダが無い日があった場合，相関係数の測定ができない．本稿の提案ソフトウェアでは，フォルダが無い場合はシャットダウンしていると仮定し，前日と同じデータサイズを入れているが作業していない相関係数が，低いため値の見直しが必要である．また，提案ソフトウェアではその日のバックアップフォルダ自体が無くその期間に新しい VM が作成されている場合，VM が作成されたと認識されない．例えば，3 日に新規 VM を作成したが，5 日からのバックアップフォルダしかない場合，この VM は 5 日に新規作成されたと認識される．そのため，3 日と 4 日のデータサイズは，VM が存在しないとし，0 となる．これらは，データサイズが 0 の日付の前後の日が 0 ではない場合，またはバックアップフォルダに存在するフォルダの日付が飛んでいた場合，バックアップが正常に行われなかったとし，データ量が 0 だった日付は相関係数の際に含めないことで解決することができる．

本稿では，相関係数の計測対象期間をバックアップフォルダにある日数分とした．本稿で想定した課題はフルバックアップデータがストレージを圧迫することである．そのため，バックアップフォルダに使用できるストレージの閾値を設定し，閾値までストレージを使用した時点で，提案ソフトウェアを使用する．それ以降は，提案ソフトウェアを使用し，フォルダを圧縮した次の日から，現在までの日数を計測対象期間とする．

7. おわりに

本稿は，バックアップフォルダのサイズが大きいため圧縮を行いたいが，圧縮するフォルダと圧縮しないフォルダを判断することが難しいことを課題に挙げた．判断が難しい原因は，フォルダのデータサイズは日々変化しており，増減に法則性もないことである．しかし，作業していないフォルダのデータサイズの増減には，強い相関関係があり，相関係数は 0.97 以上である．作業しているフォルダには相関係数がばらばらで，相関関係がほぼ無いことが基礎実験から分かった．

そこで，本稿では，作業しない VM を一つ作成し，その VM と NAS サーバーのバックアップフォルダの dmvk ファイルのデータサイズの相関係数を算出し，相関係数が一定値以上のフォルダを圧縮する方法を提案した．NAS サーバーへの SSH 接続を行い，それぞれの vmdk のデータサイズをリストで取得し，作業していない VM と相関係数を算出し，相関係数が一定値以上のフォルダを圧縮するソフトウェアを作成した．

評価実験では，相関係数が 0.5,0.6 以上のときが 85.0% と一番正解率が高かった．

参考文献

[1] Nakagawa, A., Furukawa, H., Konishi, R., Kudo, D., Matsumura, T., Sato, D., Abe, Y., Washio, T., Arafune, T., Yamanouchi, S., Kushimoto, S. and Tominaga, T.: The Great East Japan Earthquake: Lessons Learned at Tohoku University Hospital During the First 72 Hours, *IEEE Pulse*, Vol. 4, No. 3, pp. 20–27 (online), DOI: 10.1109/MPUL.2013.2250851 (2013).

[2] Chervenak, A., Vellanki, V. and Kurmas, Z.: Protecting file systems: A survey of backup techniques, *Joint NASA and IEEE Mass Storage Conference*, Vol. 99, Ft. Lauderdale, FL (1998).

[3] Kumar, P. M. A., Pugazhendhi, E. and Nayak, R. K.: Cloud Storage Performance Improvement Using Deduplication and Compression Techniques, *2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT)*, pp. 443–449 (online), DOI: 10.1109/ICSSIT53264.2022.9716524 (2022).

[4] Hilmi, M. R., Sudarma, M. and Linawati: Virtual Backup Server optimization on Data Centers using Neural Network, *2018 International Conference on Smart Green Technology in Electrical and Information Systems (ICSGTEIS)*, pp. 162–167 (online), DOI: 10.1109/ICSGTEIS.2018.8709101 (2018).

[5] Xia, R., Yin, X., Alonso Lopez, J., Machida, F. and Trivedi, K. S.: Performance and Availability Modeling of ITSystems with Data Backup and Restore, *IEEE Transactions on Dependable and Secure Computing*, Vol. 11, No. 4, pp. 375–389 (online), DOI: 10.1109/TDSC.2013.50 (2014).

[6] Murugesan, M. and Ravichandran, T.: Evaluate database compression performance and parallel backup, *International Journal of Database Management Systems*, Vol. 5, No. 4, p. 17 (2013).

[7] Đorđević, B., Timčenko, V., Kraljević, N. and Davidović, N.: File system performance comparison in full hardware virtualization with ESXi and Xen hypervisors, *2019 18th International Symposium INFOTEH-JAHORINA (INFOTEH)*, pp. 1–5 (online), DOI: 10.1109/INFOTEH.2019.8717664 (2019).

[8] Lindstrom, P. and Isenburg, M.: Fast and Efficient Compression of Floating-Point Data, *IEEE Transactions on Visualization and Computer Graphics*, Vol. 12, No. 5, pp. 1245–1250 (online), DOI: 10.1109/TVCG.2006.143 (2006).

[9] Anquetil, N. and Lethbridge, T.: Extracting concepts from file names; a new file clustering criterion, *Proceedings of the 20th International Conference on Software Engineering*, pp. 84–93 (online), DOI: 10.1109/ICSE.1998.671105 (1998).

[10] Yahyaoui, H. and Moalla, S.: CloudFC: Files Clustering for Storage Space Optimization in Clouds, *2016 IEEE International Conference on Cloud Computing Technology and Science (CloudCom)*, pp. 193–197 (online), DOI: 10.1109/CloudCom.2016.0042 (2016).

