リソース利用状況を考慮した仮想計算機予備イメージ管理手法

松尾英治† 金木佑介† 金田典久† 鶴薫† 飯塚剛†
†三菱電機株式会社 情報技術総合研究所

1. はじめに

近年，物理計算機(以下サーバ)の障害に伴う仮想計算機(以下 VM)の停止期間を短縮する構成として，VM と冗長構成制御の連携が注目されている[1]．運用系サーバと待機系サーバの両系ダウンを防ぐためには，運用系の迅速な復旧が重要となる．

これに対し，サーバ停止後に，その上で動作していたVMを予備VMとして他のサーバにて再構築し，冗長構成を復旧する機構が提案されている[2]．この機構を，共有ディスクを用いずに実現する場合，VMの動作に必要なVMの予備イメージを，障害に先立ち予め他のサーバ内に保存し，予備VMの構築時に，サーバ上に動作可能なイメージとして復元する必要がある(図1)．

本稿では，この非共有ディスク環境における冗長構成の維持において，所要ディスク領域を削減するために VM イメージ間で差分を取りつつ，保存する差分データを予備 VM 構築時にリソース(CPU，メモリ，ディスク I/O，ネットワーク I/O 等)が不足しない範囲で同一のサーバに集約する VM イメージの管理手法について提案する，これにより，復旧時におけるサーバ間での転送量を削減し，短時間での冗長構成維持を実現する．

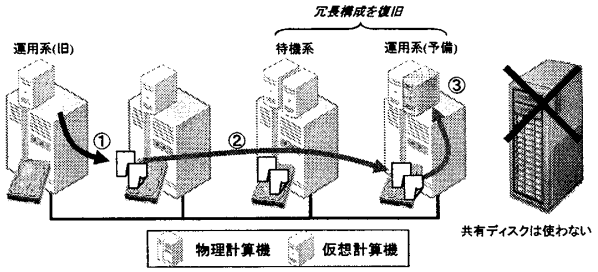
2. 課題

非共有ディスク環境における冗長構成の維持において，次の2点が課題となる．

(1) VM イメージの保存時に差分処理を用いた場合，複数の VM が同一の差分データに依存することとなる．この時，復旧時に別々のサーバにおいて同一のデータが必要になる状況が生じる場合がある．

(2) VM 起動に必要なリソースが差分データを保存しているサーバにて不足する場合，他のサーバにデータを転送し，復旧時の VM 起動をその転送先のサーバに委ねる必要がある．

これら2つの課題は，共に，予備 VM の構築時にサーバ間でのデータ転送を生じさせ，冗長構成の復旧おいて遅延を生じさせる原因となる．



①障害に先立ち予めVMイメージを他のサーバにデータとして保存
②障害に伴い、保存したデータを基に他のサーバの上にVMイメージを構築
③VMイメージを基に予備VMを起動し冗長構成を復旧 (③の処理は外部が行う)

図1 冗長構成復旧の流れ

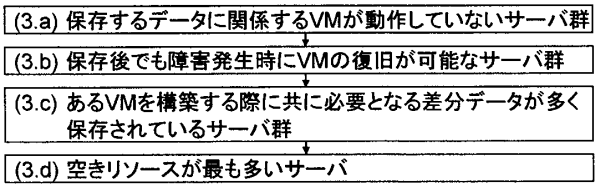


図2 保存先サーバ選択の流れ

3. 提案手法

3.1. 予備イメージ保存時の処理の流れ

予備イメージの保存では，VM 起動の保証と、復旧時のデータ転送量削減を目的に，各データの保存先サーバを適切に選択し保存する．保存時の処理は主に次の5ステップからなる．

(1) 既に保存済みの VM イメージと差分を取る．

(2) 保存先の選択に必要な情報を取得する．

(3) 適切な保存先サーバを選択する．適切なサーバが存在しない場合，データの二重化や分割を行い，再度保存先のサーバを選択する．

(4) 選択したサーバに差分データを転送する．

(5) 管理情報を更新する．

ステップ(2)において取得する情報は，保存する差分データ量と基のデータ量，保存対象となる VM が動作するのに必要なリソース量である．また，保存先として利用している全サーバの空きリソース量も取得対象となる．これら情報に関しては，VM 設定ファイル，現在のリソース利用状況を基とする．

ステップ(3)では，図2に示す4ステップの順に保存先のサーバを絞り込み，選択する．

ステップ(3.a)では，復旧に必要なデータが停止したサーバ上に存在しないように，保存するデータに関係する VM が動作しているサーバとは

"Spare Virtual Machine Image Management Considering Resource Availability"
†Hideharu MATSUO, Yusuke KANEKI, Norihisa KANEDA, Kaoru TSURU, Tsuyoshi IIZUKA
Information Technology R&D Center, Mitsubishi Electric Corporation. (†)

別のサーバを保存先の候補とする．もし，候補となるサーバが存在しない場合，同じデータを 2 つのサーバに保存するものとして，全てのサーバを候補とする．

ステップ(3.b)では，データを保存した後でも，他のサーバの停止に伴う VM 復旧が可能なサーバを(3.a)の候補から選択する．VM 復旧が可能かは，障害により停止する VM が必要とするリソースの合計を基に判定する．ディスク領域が不足する場合，保存するデータを複数のサーバに分けて保存する．(3.a)の候補全てに分けた場合でも領域が不足する場合は，保存が不可能と見なす．

3.2. 予備イメージ構築時の処理の流れ

構築時の処理は次の 2 ステップからなる．

(1) 構築先のサーバにて差分データを基に，保存した元データを復元する．

(2) 復元後，再度の障害に備え，復元に用いた差分データを他のサーバに移動させる．

ステップ(2)における移動は，予備イメージ保存時の(2)～(5)のステップを自サーバ以外のサーバを対象に行う．

4. 効果

本手法の効果として，本手手法と，少なくとも復旧時に VM が起動する範囲で空き容量が均一になるよう保存した従来例との復旧時におけるデータ転送量を比較する．

Si～Sl 4 台のサーバ(各容量 270GB)を対象とし，VM A～J が動作しているものとする(図 3)．保存するデータは，図 4 の差分ツリーを構成するものとし，サーバへは，図の左上から下へ，a0，b0，a1，a2，b1，b2･･･の順に登録する．なお，マスタを 40GB，差分データを各 5GB，VM イメージを 40GB とする．

従来手法と提案手法との保存状態を図 3 に示す．提案手法では，保存ステップ(3.a～b)によって b0 が分割され，Si，Sj 2 台で保存するものとして配置が行わる．また，保存ステップ(3.a～c)によってリソースが潤沢に存在する Sl に a0～a14 の集約が行わる．

復旧時の例として，この状態で，Si，または，Sk が停止した場合の転送量を表 1 に示す．Sj について，Si と類似するので省略する．Si の停止に伴い Sl にて VM A～C を復旧する際、従来では合計 15GB の転送が生じるが，提案手法では転送が生じない．また，Sk の停止に伴い Si にて VM G を，Sj にて VM H，I を復旧する際，従来では合計 75GB の転送が生じるが，提案手法では合計 40GB の転送となる．

結果，従来手法と比較し，復旧時のデータ転送量の削減が可能となる．

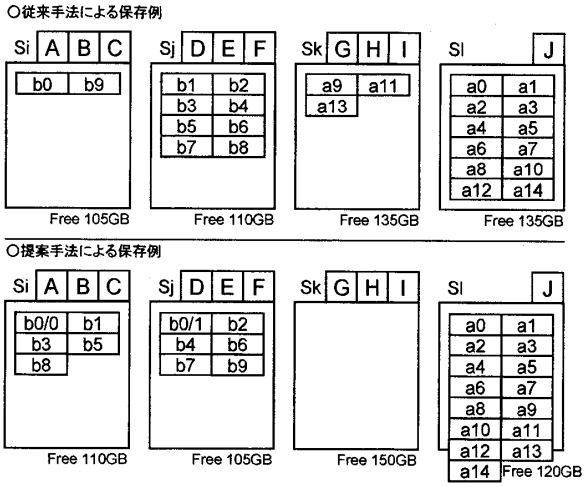


図 3 予備 VM 保存例(上:従来手法 下:提案手法)

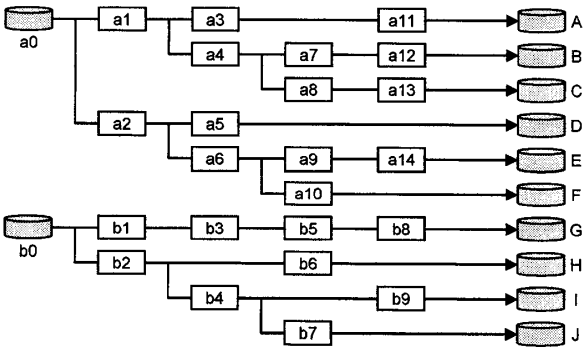


図 4 VM A～J に関する差分ツリー例

表 1 サーバ Si，Sk が停止した際の転送量比較

手法	Siの停止		Skの停止	
	転送データ	転送量	転送データ	転送量
従来	a9,a11,a13	15GB	b0,b1,b3,b5,b8,b9	65GB
提案	-	0GB	b0/0,b0/1	40GB

5. おわりに

非共有ディスクを用いたサーバ環境において，リソース使用量を考慮しつつ予備となる VM イメージを管理する手法を提案した．本手法を用いることで，非共有ディスク環境における，他サーバでの短時間での VM 復旧を実現する．

今後の課題としては，実装と評価などが挙げられる．

参考文献

[1] Pelleg, D. Schulz, C. Spainhower, LF. Ta-Shma, P. Tomek, LA.: Using virtualization for high availability and disaster recovery, IBM journal, Vol.53, No.4, paper.8 (2009).

[2] NEC: 高可用性クラスタリング CLUSTERPRO, NEC(オンライン), 入手先 <http://www.nec.co.jp/pfsoft/clusterpro/index.html> (参照 2010-01-06)