複数ホストで動作する分割メモリVMのチェックポイント・リストア

村田 時人[1]　光来 健一[1]

1. はじめに

近年，IaaS 型クラウドでは大容量メモリを持つ仮想マシン（VM）が提供されるようになっている．例えば，Amazon EC2 では 24TB のメモリを持つ VM が提供されており，インメモリ・データベースやビッグデータの解析などに利用されている．このような VM のマイグレーションを容易にするために，メインホストと複数のサブホストにメモリを分割して転送する分割マイグレーション [1] が提案されている．しかし，マイグレーション後の分割メモリ VM は複数のホストにまたがって動作するため，ホストやネットワークの障害の影響を受ける可能性が高くなる．障害対策として VM の状態を保存・復元するチェックポイント・リストアが用いられているが，従来のチェックポイント手法を分割メモリ VM に適用すると，ホスト間でメモリデータのやりとりを行うリモートページングが大量に発生するためオーバヘッドが大きい．また，従来のリストア手法では複数ホストに分割された状態で VM を復元することができない．

本研究では，複数ホストにまたがって動作する分割メモリ VM の柔軟で効率のよいチェックポイント・リストアを可能とするシステム D-CRES を提案する．

2. D-CRES

D-CRES のチェックポイントは，図 1 のように各ホストで並列に VM の状態の保存を行う．これにより，メインホストとサブホスト間でリモートページングを発生させないようにし，分割メモリ VM の高速なチェックポイントを実現する．チェックポイントを行う際は，メインホストではそのホスト上に存在する VM のメモリと仮想 CPU や仮想デバイスなどの VM 本体の状態だけを保存する．また，

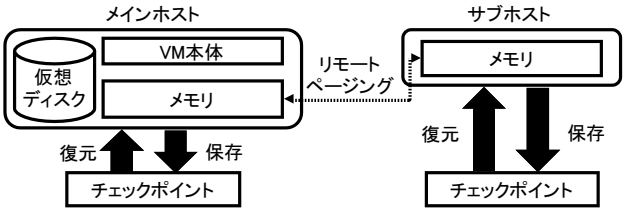

図 1　D-CRES のチェックポイント・リストア

VM の仮想ディスクの状態も通常の VM と同様にメインホストにおいて保存する．一方，サブホストではそのホスト上に存在する VM のメモリだけを保存する．メインホストとすべてのサブホストで VM の状態を保存し終えるとチェックポイント処理が完了する．

D-CRES はライブチェックポイントにより VM をほとんど停止させずに VM の状態を取得することができる．従来のライブチェックポイントは VM を動かしたままメモリを保存し，保存中に更新されたメモリについて追加で保存する．しかし，分割メモリ VM に適用するとチェックポイント中に VM が発生させたリモートページングにより，保存したメモリの状態に不整合が生じる場合もある．D-CRES ではファイルのオフセットがメモリのアドレスに 1 対 1 に対応するスパースファイルを用いることで，リモートページングによるホスト間でのメモリの移動を考慮し，それぞれのホストで過不足なくメモリの保存を行う．

一方，D-CRES のリストアは最新のチェックポイントを用いて複数のホストそれぞれで並列に VM の状態の復元を行う．これにより，チェックポイント時と同様に複数ホストに分割された VM を復元することができる．リストアを行う際には，まず，十分な空きメモリを持つメインホストとサブホストを探す．そして，チェックポイントを用いて，メインホストでは VM のメモリと VM 本体の状態，仮想ディスクの状態を復元し，サブホストでは VM のメモリだけを復元する．メインホストとサブホストで状態の復

[1] 九州工業大学
Kyushu Institute of Technology

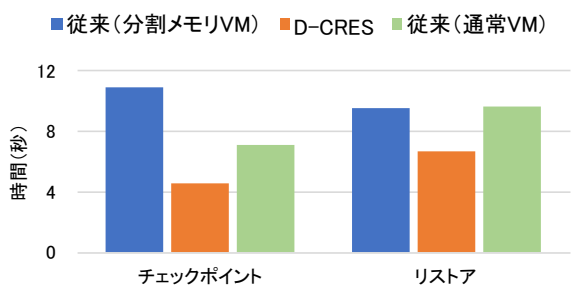

図 2　チェックポイント・リストアの実行時間

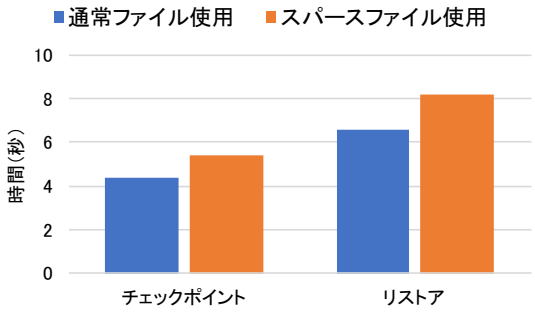

図 3　異なるチェックポイント・フォーマットを用いたチェックポイント・リストア時間

元が完了すると，ホスト間でリモートページングのためのネットワーク接続を確立して分割メモリ VM の実行を再開する．

3. 実験

D-CRES を用いて分割メモリ VM のチェックポイント・リストアにかかる時間を測定した．比較のために，分割メモリ VM に対して従来手法を用いた場合の時間も測定した．また，1 台のホストで動作する通常の VM に対して従来手法を適用した場合の時間も測定した．この実験では 1GB のメモリをもつ VM を用い，2 つのホストに 512MB ずつにメモリを分割した．測定結果を図 2 に示す．

D-CRES のチェックポイントは従来手法を分割メモリ VM に適用する場合よりも 58%高速であることが分かった．通常 VM のチェックポイントを取得する場合と比べても 30%高速であった．一方，D-CRES のリストアについては，従来手法で VM を復元する場合よりも 36%高速に復元できることが分かった．従来手法の 2 倍高速にならないのは，メモリの保存・復元は並列に行えるが，VM 本体の保存・復元はメインホストのみで行われるためである．

D-CRES において，異なるチェックポイント・フォーマットを用いた場合にチェックポイント・リストアにかかる時間を測定した．一つは従来手法と同様に通常のファイルを用いた場合であり，もう一つはライブチェックポイントのためにスパースファイルを用いた場合である．それぞれのチェックポイント・リストアの測定結果を図 3 に示す．

スパースファイルを用いた場合は通常のファイルを用いた場合と比べてチェックポイント時間が 19%，リストア時間が 20%長くなった．これはスパースファイルを用いた場合には多くのシークが発生したためと考えられる．しかし，通常ファイルを用いた場合には，復元時に繰り返し同じデータを復元する可能性があり，その場合にはリストア時間が長くなると考えられる．

4. まとめ

本研究では，複数ホストにまたがって動作する分割メモリ VM の柔軟で効率のよいチェックポイント・リストアを可能とするシステム D-CRES を提案した．今後の課題は，リモートページングに対応したライブチェックポイントの実装を完成させることである．その際に，スパースファイルを用いるとチェックポイント・リストアの性能が低下することが分かったため，他の方法も検討する．チェックポイントをさらに高速化するには，メモリの差分チェックポイントを実現する必要もある．分割メモリ VM ではリモートページングによってメモリの配置が変わるため，差分の検出に工夫が必要である．また，現在のリストアの実装ではチェックポイント取得時のホスト構成でしか復元できないため、チェックポイント時とは異なるホスト構成で分割メモリ VM を復元できるようにすることも検討している．

参考文献

[1] M. Suetake, T. Kashiwagi, H. Kizu, and K. Kourai: S-memV: Split Migration of Large-Memory Virtual Machines in IaaS Clouds, In Proc. Int. Conf. Cloud Computing, pp.285-293, 2018.