遠隔ライブマイグレーションに対応した仮想計算機ストレージ再配置機構

広渕崇宏† 小川宏高† 中田秀基†
伊藤智† 関口智嗣†

1. はじめに

計算機センタやデータセンタにおいて仮想化技術の導入が進みつつある. 物理的な計算機資源を抽象化し論理的に分割・共有可能になり, 資源利用効率を向上させ運用コストを低減できる. 特に注目を集める仮想化技術として仮想マシンモニタ (VMM) が備えるライブマイグレーション機能が挙げられる. 仮想計算機 (VM) を一切停止することなく異なる物理ノード上に移動できる.

我々は計算資源運用コストの削減や電力消費の効率化を実現するため, 拠点横断的な仮想化資源運用技術の確立に取り組んできた[5]. 特に, ライブマイグレーション機能を利用して遠隔拠点にまたがって VM を動的に再配置できれば, 拠点横断的な負荷バランスや省電力化が可能になり, また施設メンテナンスも容易になると考えている.

しかし現状では VM のライブマイグレーション機能の実用化は単一拠点内での利用にとどまっている. ライブマイグレーションに際しては, VM の継続的な I/O を可能にするため, 移動先・移動元双方のホストからアクセスできる共有ストレージが必要とされる. しかし, 遠隔拠点間にまたがって共有ストレージを構築すると LAN 環境よりもはるかに大きいネットワーク遅延を原因として十分な I/O 性能を得ることができない. また VM が拠点外のストレージサーバに常に依存するため運用の柔軟性にも乏しい.

そこで我々は, VM ストレージの動的かつ透過的な再配置機構を提案する. 提案機構はブロックレベルのストレージ I/O プロトコルに対するターゲットサーバおよびプロキシサーバとして振る舞い, VM の動作に支障を与えることなく透過的に仮想ディスクを遠隔拠点間で移動する. VM の I/O にともなうオンデマンドなデータコピー, およびバックグラウンドでのデータコピーによって, VM のライブマイグレーションと連動してストレージを再配置する. WAN 環境下での VM 再配置に取り組んだ既存研究 (先行発表[3] 参照) と比較しても, VM を一切停止させることなく迅速な実行ホストの変更を可能にし, VMM に非依存な機構となる点に優位性がある.

† 産業技術総合研究所 / National Institute of Advanced Industrial Science and Technology (AIST)

2. VM ストレージ移動手法

提案手法はブロックレベルのストレージ I/O プロトコル (iSCSI[2] や NBD[1] プロトコル) のストレージサーバとして振舞いながらも, VMM による VM 再配置と連動して遠隔拠点間でストレージデータのコピーを行う. VM はホスト OS 上のブロックデバイス (/dev/nbd0 等) を介して仮想ディスクを読み書きする.

2.1 オンデマンドコピー

VMM はライブマイグレーションを開始すると, VM の実行メモリイメージを移動先ホストにコピーし始める. すべてのメモリイメージがコピーできた時点で移動元ホストでの VM の実行を停止し, 移動先ホストにて実行を再開する. この時点以降, VM のすべての I/O は移動先ホストを介して行われる. このとき, 提案機構のプロキシサーバは, 移動先拠点において VM の I/O リクエストに応じて, 移動元拠点のターゲットサーバからストレージ I/O プロトコルによってデータを取得していく. さらに移動先拠点にデータを保存 (キャッシュ) していく.

図 1 において, 読み込み及び書き込みリクエストそれぞれにおける処理を示す. 未キャッシュのブロックに対する読み込みリクエストは, 移動元拠点ストレージサーバからデータを取得し, キャッシュ済みとマークする. 書き込みリクエストは, 移動先拠点にデータを保存し, 対象ブロックが未キャッシュだったならキャッ

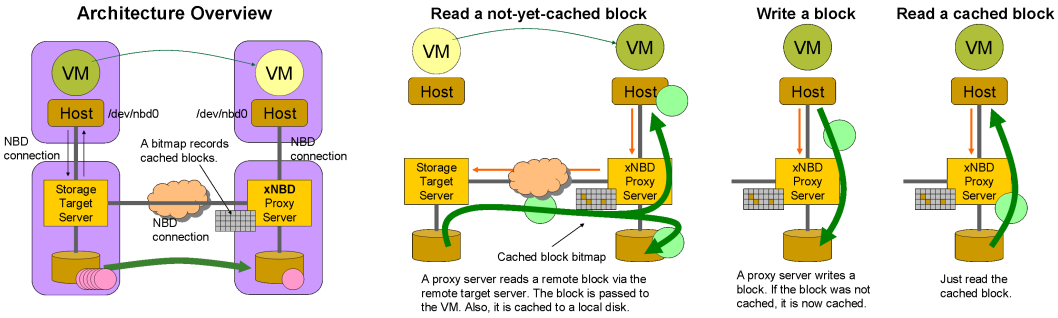

図 1 提案機構とその基本動作概要

シュ済みとマークする．キャッシュ済みのブロックに対する読み込みリクエストは，単に移動先拠点からデータを読み込めばよい．一度キャッシュされた領域については移動元への遠隔読み込みは発生しない．VM によって一度アクセスされた領域から徐々に再配置が完了していく．

2.2 バックグラウンドコピー

上述した VM の I/O に応じたオンデマンドなデータ再配置に加えて，未キャッシュなブロックはバックグラウンドでもコピーされる．最終的に仮想ディスクのすべての領域がキャッシュされると，仮想ディスクの再配置は完了し，プロキシサーバから移動元拠点ターゲットサーバへのストレージ I/O プロトコル接続は終了する．バックグラウンドコピーは，VM が特定のブロックに初めてアクセスする前にあらかじめブロックをコピーしてキャッシュしておくことで，移動元への遠隔読み込みを防ぐ側面もある．

バックグラウンドコピーにおいては，仮想ディスクの先頭から未キャッシュブロックを順にコピーするという単純な動作だけではなく，利用頻度の高いブロックから先にコピーすることも可能である．実際，我々は仮想ディスク上に作られた ext3 ファイルシステムを解析することによって，利用頻度が高い領域を判別し優先的にコピーする仕組みを実装している．ext3 はディスク領域を複数のグループに分割して，関連性の高いブロックを同じグループ内に保存することで，キャッシュ効率を高めディスクヘッド移動を減らすよう試みる．同じディレクトリに属するファイルは同じグループに，同じファイルに属するブロックは同じグループに属することが多い．また使用中のディスクブロックを管理するために，inode ブロックビットマップおよびデータブロックビットマップとしてディスク上にビットマップを保存している．

そこで，ターゲットサーバで I/O リクエスト領域の統計を取っておくことで，直近にアクセス頻度の高いグループからバックグラウンドコピーに取り掛かれるようにしている．さらにターゲットサーバにおいて残された ext3 イメージのビットマップブロックを解析して，データが保存されているブロックから優先的にコピーを開始できる．

3. おわりに

VM の遠隔ライブマイグレーションに対応した，VM ストレージの透過的再配置機構を提案した．その評価実験においては WAN 環境における基本的な有効性を確認している[4]．今後は提案機構を我々が開発するマルチサイトクラウド構築ツールキット[5] に適用していく．

謝辞 本研究は科研費 (20700038) および CREST の助成を受けたものである．

参考文献

1) Breuer, P. T., Lopez, A. M. and Ares, A. G.: The network block device (1999).
2) Satran, J., Meth, K., Sapuntzakis, C., Chadalapaka, M. and Zeidner, E.: Internet Small Computer Systems Interface (iSCSI), RFC 3720 (2004).
3) 広渕崇宏, 小川宏高, 中田秀基, 伊藤智, 関口智嗣: 仮想クラスタ遠隔ライブマイグレーションにおけるストレージアクセス最適化機構, 情報処理学会研究報告 (2008-HPC-116), 情報処理学会, pp. 19–24 (2008).
4) 広渕崇宏, 小川宏高, 中田秀基, 伊藤智, 関口智嗣: 仮想クラスタ遠隔ライブマイグレーションにむけた仮想計算機ストレージの透過的再配置機構の評価, 情報処理学会研究報告 (2008-HPC-117), 情報処理学会, pp. 7–12 (2008).
5) 広渕崇宏, 中田秀基, 横井威, 江原忠士, 谷村勇輔, 小川宏高, 関口智嗣: 複数サイトにまたがる仮想クラスタの構築手法, 第 6 回先進的計算基盤システムシンポジウム SACSIS 2008, pp. 333–340 (2008).