仮想サーバ環境のためのリモートスワップ性能評価

永井 洋太郎[†] 岡本 慶大[†] 奥田 剛[†]
河合 栄治[†††] 砂原 秀樹[†,††]

1. はじめに

昨今のデータセンターにおいては，複数台の実計算機に仮想計算機モニタ (VMM) を導入し，その上に多数のサーバを仮想計算機 (VM) として集約し，マネージメントすることが多い．CPU，メモリ，ストレージ，ネットワークなどのリソース共有による総合的なコスト削減が可能になることが，VM 技術を用いたサーバ集約のメリットである．

仮想環境のリソースマネジメントの手法として，ライブマイグレーションがある．ライブマイグレーションを用いると，ある実計算機においてリソースが不足した場合に当該 VM を動的に他の実計算機に対して，移動させることができる．また，系全体としてリソースが余っている場合には，一部の実計算機を解放することで電源断を行うといった利用も可能となる．

しかし，既存技術では，ある VM が利用できるリソースはその時点で当該 VM をホストしている実計算機上のリソースに限られる．例えば，ある実計算機上でメモリリソースだけが余剰であるという状況では，このリソースは有効に利用できない．そこで我々は，実計算機を越えたメモリリソース共有を可能とする仮想化サーバ機構を提案する．

本提案では，動的に余剰なメモリリソースを他の実計算機上の VM にスワップデバイスという形で提供する．この機構を用いることで，利用する術のないメモリリソースを他の実計算機に一時的に受け渡すことが可能となる．本機構を用いることで，VM をホストする実計算機群全体のリソース利用率向上を図る．

2. 関連研究

ネットワークを介するリモートスワップデバイスを利用した研究として，DLM[1] や Teramem[2] 等が存在する．これらは，High Performance Computing・クラスタなどの用途において，メモリを大量に必要とする計算を行うために，遠隔の計算機のメモリをリモートスワップデバイスとして使用することを目的としている．

どちらも，OS のスワップ機構とは独立したスワップ機構を提供し，ディスクではなくリモートメモリをスワップデバイスとして使用するという特徴がある．これは，計算性能を改善したい特定のアプリケーションでのみリモートスワップを使用させたいという目的であることと，ディスクへのスワップを前提として設計されている OS のスワップ機構の最適化がリモートスワップにおいては性能的に不利になる[2] などが理由である．

3. 提案システム

我々は，各 VM のメモリ使用状況を監視して，状況に応じてメモリリソースを実計算機間で貸し借りすることを可能にするシステムを提案する．

提案システムでは，ある実計算機にホストされる VM において余剰のメモリが生じた場合に，その領域を他の実計算機にリモートスワップデバイスとして貸し出せる．また，ある VM にスラッシングが発生した場合に，スワップデバイスをディスクからリモートスワップデバイスに切り替えることで，スラッシングを起こした VM の救出も行える．

本システムを利用することで，利用されることがなかったメモリリソースが他計算機に貸し出されることで系全体のリソース利用率が向上し，計算性能の向上につながると考えている．

3.1 リモートスワップ機構

2 章で述べたように，既存研究では，リモートスワップは，OS のスワップ機構と独立して実装されることが多い．本研究では，ゲスト OS 全体でリモートスワップデバイスを使用させるため，OS のスワップ機構を利用する．

今回は，一方の計算機のメモリ上に ramfs を用いてファイルシステムを構築し，それを iSCSI でエクスポートすることで，他方の計算機から SCSI デバイス

† 奈良先端科学技術大学院大学 情報科学研究科
†† 慶應義塾大学 メディアデザイン研究科
††† 情報通信研究機構

としてアクセス可能とした．このデバイスをスワップ領域としてフォーマットし，OS に使用させた．次章で，この方法でのリモートスワップを実施した場合の性能評価を行う．

4. 性能評価

3.1 章で述べたようなリモートスワップを使用した際の性能に関して評価を行った．

4.1 測定環境

測定環境は表 1 の通りである．．以下で行った実験では，メモリサーバに 4GB の ramfs を構築した．その上に，4GB のイメージファイルを作成し，iSCSI target として，エクスポートする．メモリクライアントは，iSCSI initiator を動作させて，メモリサーバ上のメモリを SCSI デバイスとしてマウントする．メモリクライアントは，実メモリを 1GB 搭載している。スワップデバイスとして，ローカルディスク (4GB)，リモートメモリ (4GB) を切り替えて使用し，測定を行った．

メモリサーバとクライアントは 1000BASE-T での接続時には，L2 スイッチ 1 台を挟んで接続されている．10GBASE-T での接続時には，ケーブルで直結している．

4.2 data size に関する性能評価

data size(仮想メモリ空間に確保を行うメモリ量) に関する性能評価を行った結果が，図 1 である．この実験では blocksize を 1,024byte に固定して，data size を変化させた場合の性能特性を調査した．data size は 512MB から 4GB まで変化させた．縦軸は，local swap に対するスループット比である．本実験環境では，実メモリ搭載量は 1GB であるので，実際にベンチマークを行うプロセスが使用できる実メモリは 900MB 程度であり，それを超える data size を指定した場合にスラッシングが起こり，性能が劣化する．

sequential read においては，リモートメモリを使用するとローカルディスクを使用した場合の 2.5〜3.2 倍の性能向上がみられた．また，random read においては sequential read に比較して極端な性能劣化がみられるが，リモートメモリを使用するとローカルディスクを使用した場合の 22〜27 倍の性能向上がみられた．

表 1 実験環境

	メモリサーバ	メモリクライアント
CPU	Intel Xeon X3350	Intel Core2 Quad Q9450
Memory	8GB	1GB
NIC(1G)	Intel Gigabit Ether(PCIe)	Intel Gigabit Ether (PCI)
NIC(10G)	Intel 10GbE-T(PCIe)	Intel 10GbE-T(PCIe)
OS	openSUSE 11.1(Native)	openSUSE 11.1(Native)

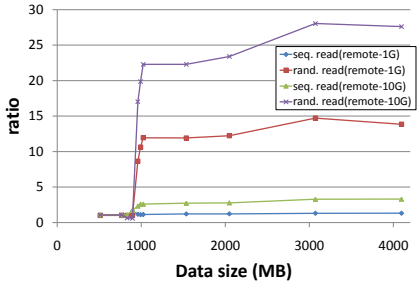

図 1 datasize を変化させた場合の throughput（blocksize=1KB）

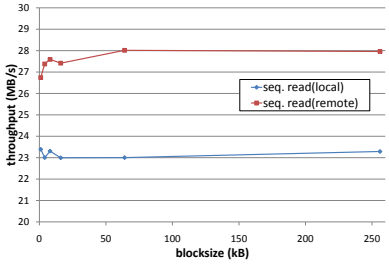

図 2 blocksize を変化させた場合の throughput（datasize=1GB）

4.3 blocksize に関する性能評価

data size を 1GB に固定して，blocksize に関する性能評価を行った結果が，図 2 である．64kB 以上の blocksize を指定した場合には，性能は安定しており，1.1〜1.2 倍の性能向上がみられた．

5. まとめと今後の課題

既存技術の組み合わせによるリモートスワップを利用することにより，最大で 27 倍の性能向上を確認できた．今後，ゲスト OS を監視することでメモリ使用状況を判断し，動的なリモートメモリ貸し出し・借り入れを可能とするツールを作成する．さらに本システムを，複数台の実計算機が存在する環境で動作させ，システムのスケーラビリティと系全体の性能向上に関して調査を行う予定である．

参考文献

1) 緑川博子, 黒川原佳, 姫野龍太郎: 遠隔メモリを利用する分散大容量メモリシステム DLM の設計と 10GbEthernet における初期性能評価, 情報処理学会論文誌コンピューティングシステム, Vol.1, No.3 (2008).
2) 山本和典, 石川裕: テラスケールコンピューティングのための遠隔スワップシステム Teramem, *SACSIS 2009* (2009).