準仮想化ページフォルトによる ポストコピー型ライブマイグレーションの性能向上手法

広渕崇宏[†] 伊藤智[†]

1. はじめに

ポストコピー型の仮想マシン再配置機構（ライブマイグレーション）は、仮想マシンの実行ホストを他のホストに素早く切り替えることを可能にする。データセンタの運用効率を向上させる上で有用な技術であり、特に資源消費量に応じて動的に仮想マシンの配置を調整するシステムにおいて、今日一般的に普及しているプレコピー型のライブマイグレーションを利用するよりも、高いレベルの性能保証や消費電力削減[1),2)]が可能になる。

我々は、先行研究[3)]において、仮想マシンモニタ Qemu/KVM[4)]に対するポストコピー型ライブマイグレーションのプロトタイプを実装した。そして現在、Qemu/KVM のメインラインに統合できる、実用化レベルの品質を持ったポストコピー型ライブマイグレーション機構 Yabusame[5)~7)]を開発している。

本稿では、Yabusame において実装した、準仮想化ページフォルトによるポストコピー型ライブマイグレーションの性能向上手法について紹介する。

2. ポストコピー型ライブマイグレーションの課題

ポストコピー型ライブマイグレーションは、基本的には実行ホストを切り替えた後にオンデマンドにメモリを転送する機構である。数百ミリ秒での実行ホストの切り替え、およびプレコピー型よりも短いマイグレーション完了時間を提供できる。その動作は、

(1) 移動元ホストで仮想マシンを停止。
(2) 仮想マシンの CPU レジスタやデバイスの状態を移動先ホストに転送。
(3) 移動先ホストで仮想マシンを開始。
(4) A. 仮想マシンが新たなメモリページにアクセスするたび、オンデマンドに移動元ホストからメモリページを転送。B. オンデマンドな転送と並行して、バックグラウンドで残りのメモリページを移動先に転送。
(5) 全てのメモリページが転送できたら、移動元ホストに残しておいたメモリページを破棄。

となる。段階（4）において、仮想マシンが新たなメモリページにアクセスするとページフォルトが発生し、仮想マシンから仮想マシンモニタに実行が移る。仮想マシンモニタは対象メモリページの内容を用意し、仮想マシンの実行を再開する。すでに対象メモリページが移動先ホストに転送済みであった場合、ごく短時間仮想マシンを停止するのみである。本稿ではこれをマイナーフォルトと呼ぶ。一方で、対象メモリページが移動先ホストに転送されていなかった場合には、ネットワークを介して移動元ホストからメモリページを取得するため、比較的長時間（ギガビットイーサネットで構成された LAN 環境でれば百マイクロ秒程度）仮想マシンを一時停止する必要がある。本稿ではこれをメジャーフォルトと呼ぶ。仮想マシン全体の実行が一時的に停止してしまうため、メジャーフォルトが頻出するとアプリケーションの性能が低下してしまう。☆

3. 提案手法

そこで、ポストコピー型ライブマイグレーションにおける仮想マシンのページフォルトを準仮想化することで、実行ホスト切り替え直後の性能低下を緩和する手法を提案する。メジャーフォルトが発生した際に、仮想マシン全体の実行を対象メモリページが準備できるまで停止するのではなく、メジャーフォルトの原因となったプロセス（注: ゲスト OS 上のプロセス）のみをその間停止する。他のプロセスの実行は引き続き継続する。複数のプロセス（もしくはネイティブスレッド）からなるアプリケーション（例えば Apache ウェブサーバ等）において、実行ホスト切り替え直後の性能低下が抑制できることが期待される。

KVM が備える Asynchronous Page Fault (APF) 機能を応用して実装した。APF は、仮想マシンがアクセスしたメモリページがホスト OS 上でスワップアウ

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

☆ 厳密には、仮想マシンが複数の仮想 CPU を持つ場合は、ページフォルトをおこした仮想 CPU が一時的に停止する。しかし、議論の簡略化のため、本稿では仮想マシンが単一の仮想 CPU を持つ場合のみを考える。

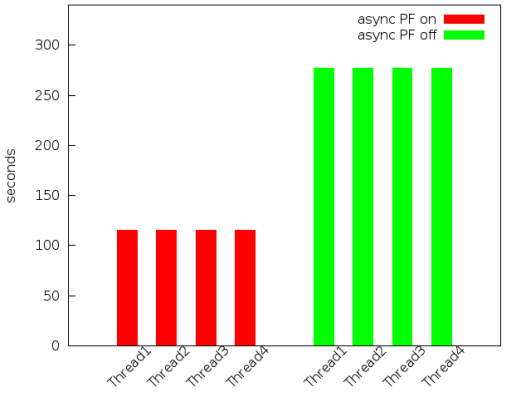

図 1 実験結果（左: 提案手法、右：既存手法）
Fig. 1 Results of Experiments (Left: Proposed Mechanism, Right: Existing Mechanism)

トされていた際に、当該メモリページにアクセスしたプロセスのみの実行を遅延させる仕組みである。APF に対応したゲスト OS をそのまま Yabusame においても利用できる。

4. 評価実験

提案手法の初期的な評価実験を 2011 年 12 月版の Yabusame を用いて行った。ポストコピー型ライブマイグレーションのページフォルトを準仮想化した効果を詳しく調べるため、オンデマンド転送のみを有効にし、バックグラウンド転送は無効にしている。仮想マシンに対してメモリを 8GB 割り当て、ホスト A からホスト B にポストコピー型ライブマイグレーションを行う。ただし、オンデマンド転送のみを有効にしているため、マイグレーションは完了しない。仮想マシンが新たなメモリページにアクセスするたび、メジャーフォルトが発生し、ホスト A からホスト B にメモリページの内容が転送される。

ホスト A およびホスト B はギガビットイーサネットで接続されている。両ホスト間のネットワーク遅延は `ping` コマンドで百マイクロ秒程度であった。マイグレーション実行中に、仮想マシン上で、メモリページにアクセスするマイクロベンチマークを動かした。ベンチマークプログラムは、あらかじめ 4 つのスレッドを起動し、各スレッドにそれぞれ 1GBytes のメモリ領域を割り当てる。その後、マイグレーションの開始と同時に、各スレッドは各自のメモリ領域を先頭から 1Byte ずつアクセスし、メモリ領域の最後に到達するまでの時間を各スレッドそれぞれで出力する。

提案機構が有効な場合および無効な場合を比較した。**図 1** に実験結果を示す。提案機構を有効にした場合は、約 120 秒程度で各スレッドのメモリアクセスが完了している。一方、無効にした場合には、約 275 秒程度を要している。提案機構によって、マイクロベンチマークにおいてメモリアクセスに要する時間を半分以下にまで短縮できた。メジャーフォルトを発生させたスレッドのみ一時的に停止し、他のスレッドの実行を継続できた効果が現れている。

本実験結果より提案機構が基本的な部分では正しく動作していることが確認できた。今後は実際的なアプリケーションにおいてどの程度性能向上が得られるのか評価していく。また提案機構の実装において改善点があればさらに改良を進めていく。

謝辞 本研究は、科研費（23700048）および CREST（情報システムの超低消費電力化を目指した技術革新と統合化技術）の助成を受けた。Yabusame の開発は、経済産業省およびエヌ・ティ・ティ・コミュニケーションズ株式会社の支援を受けた。Yabusame の現行バージョンは産業技術総合研究所の依頼にもとづき VA Linux Systems Japan 株式会社が中心となって開発している。本稿執筆にあたり VA Linux Systems Japan 株式会社山幡為佐久氏から助言をいただいた。

参 考 文 献

1) Takahiro Hirofuchi, Hidemoto Nakada, Satoshi Itoh, and Satoshi Sekiguchi. Reactive cloud: Consolidating virtual machines with postcopy live migration. *IPSJ Transactions on Advanced Computing Systems*, Vol. ACS37, pp. 86–98, Mar 2012.
2) Takahiro Hirofuchi, Hidemoto Nakada, Satoshi Itoh, and Satoshi Sekiguchi. Making vm consolidation more energy-efficient by postcopy live migration. In *The Second International Conference on Cloud Computing, GRIDs, and Virtualization*, pp. 195–204. IARIA, Sep 2011.
3) 広渕崇宏, 中田秀基, 伊藤智, 関口智嗣. 既存 VMM への適用が容易でゲスト透過なポストコピー型仮想マシン再配置機構. 情報処理学会論文誌:コンピューティングシステム, Vol. ACS31, pp. 248–262, Sep 2010.
4) Avi Kivity, Yaniv Kamay, Dor Laor, and Anthony Liguori. kvm: the Linux virtual machine monitor. In *Proceedings of the Linux Symposium*, pp. 225–230. The Linux Symposium, 2007.
5) Postcopy Live Migration for Qemu/KVM (Yabusame). `http://grivon.apgrid.org/quick-kvm-migration`.
6) Takahiro Hirofuchi and Isaku Yamahata. Yabusame: Postcopy live migration for qemu/kvm. KVM Forum 2011, Aug 2011.
7) Takahiro Hirofuchi and Isaku Yamahata. Yabusame: Postcopy live migration for qemu/kvm. Linux Plumbers Conference, Sep 2011.