中断可能なポストコピーライブ仮想マシン移送

小川遥加[†]　山田浩史[†]

ライブ仮想マシン (VM) 移送は，VM を稼働させたまま別の物理マシンへ移動するテクニックである[2]．クラウド環境やデータセンタなどの仮想化されたプラットフォームにおいては，ライブ VM 移送を用いることで，VM の再配置による物理マシンのメンテナンスや負荷分散が容易化することが知られている[4),5),9),10)]．実際に，Google のデータセンタにおいてはライブ VM 移送を用いた資源管理がなされている．ライブ VM 移送の有用さから，ライブ VM 移送の機構に関する研究がこれまでに多くなされてきた．

ライブ VM 移送の効率的な実現方式としてポストコピー方式が提案されている[1),6),7)]．ライブ VM 移送の実現方式はプレコピー方式[2)] が主流であり，多くの VM モニタ (VMM) でサポートされている．プレコピー方式では，VM のメモリ内容をすべて移送先へ転送する．その後，その間に更新のあったメモリページを転送し，この工程を繰り返す．これを繰り返した回数，もしくは更新のあったページ数が閾値より小さくなったら，VM を停止する．そして，更新のあったページおよび仮想 CPU の状態を移送元へ転送して，移送先の VM を再開する．これによりライブ VM 移送は実現可能であるが，稼動しているワークロードが頻繁に多くのページを更新する場合，総移送除間の長期化やメモリページ転送量の増加が問題であった．ポストコピー方式では，最初に仮想 CPU の状態を移送先に送り，移送先で VM を稼動させる．転送されていないメモリページは順次転送し，VM が未転送のページにアクセスしたら，そのページフォールトを VMM が補足し，該当ページを優先的に移送元からフェッチする．プレコピー方式と比較すると，ポストコピー方式はメモリページの転送はその VM が有するメモリサイズに収まり，それに伴って移送の長期化を避けることができる．また，早期に移送先で VM を稼働させるため，即座に負荷分散効果を得ることができる．

ポストコピー方式は強力であるものの，ネットワーク障害などによって移送が中断した場合に VM が破壊されてしまうという問題点がある．プレコピー方式では，移送元に動作に必要なすべての状態が保持されているため，移送が中断した場合においても VM は移送元で稼動しつづけることができる．一方ポストコピー方式では，移送開始直後に実行を移送先に移すため，稼動に必要な資源が移送元と移送先とで分断されてしまう．具体的には，仮想 CPU などの実行状態は移送先に，また転送が完了していないメモリページは移送元に点在することになる．そのため，ポストコピー方式で移送を実行している最中に移送をキャンセルすると，VM が移送元や移送先で再開することができず，そのまま停止してしまい，VM のメモリ内容などの稼動情報に復元できなくなってしまう．

本研究では，ポストコピー方式の弱点を解決するべく，中断可能なポストコピーライブ VM 移送を提案する．提案方式では，ポストコピー方式で移送するものの，移送途中に実行を中断しても VM の稼動を可能にする．これにより，ポストコピー方式が有する迅速な負荷分散効果や移送時間の短縮を達成し，かつプレコピー方式と同様に移送を中断しても VM を稼動可能にする．

ポストコピー方式を中断しても VM を稼動し続けられるようにするために，提案方式では移送元と I/O View の一貫性を維持しながらスナップショットを作成する．Remus[3)] や MicroCheckpointing[8)] といった VM の複製を高速に作成する方式を応用し，移送先から移送元へ VM の稼働に必要なオブジェクトを送信する．具体的には，移送先に VM の実行状態を移した直後から，VM の実行状態を移送元へと定期的に転送する．この際，スナップショット復帰後のディスクやネットワークといった I/O の一貫性を考慮し，スナップショットを取得するまでに生じた出力はバッファリングしておき，取得後にデバイスへと解放する．入力に関しては，再度読み込みを行なう．ネットワークに関しては TCP などのプロトコルによる再送機能に頼る．これにより，スナップショットから復帰するときには，厳密には実行が多少戻るものの，I/O の一貫性は保持されているため，外部からは VM が連続稼動しているように見える．

現在，Qemu 2.6.0 を基盤として提案方式の初期プロトタイプの実装を行なっている．今後は，実装を完了させて，メモリのみを更新する，ファイルをシーケンシャルに読み込む，ファイルを転送するなどの単純なワークロードを稼動させるところから始まり，SPEC CPU や Memcached，Apache や MySQL な

† 東京農工大学
Tokyo University of Agriculture and Technology

どの Real-world なアプリケーションを用いた実験を実施する予定である．

参 考 文 献

1) Y. Abe, R. Geambasu, K. Joshi, and M. Satyanarayanan. Urgent Virtual Machine Eviction with Enlightened Post-Copy. In *Proceedings of the12th ACM SIGPLAN/SIGOPS Int'l Conf. on Virtual Execution Environments (VEE '16)*, pages 51–64, 2016.
2) C. Clark, K. Fraser, S. Hand, J. G. Hansen, E. Jul, C. Limpach, I. Pratt, and A. Warfield. Live Migration of Virtual Machines. In *Proceedings of the 2nd USENIX Symposium on Networked Systems Design and Implementation (NSDI '05)*, pages 273–286, May 2005.
3) B. Cully, G. Lefebvre, D. Meyer, M. Feeley, N. Hutchinson, and A. Warfield. Remus: High Availability via Asynchronous Virtual Machine Replication. In *Proc. of the 5th USENIX Symp. on Networked Systems Design and implementation (NSDI '08)*, pages 161–174, Apr. 2008.
4) S. Govindan, D. Wang, A. Sivasubramaniam, and B. Urgaonkar. Leveraging Stored Energy for Handling Power Emergencies in Aggressively Provisioned Datacenters. In *Proc. of the 17th ACM International Conference on Architectural Support for Programming Languages and Operating Systems(ASPLOS '12)*, pages 75–86, Mar. 2012.
5) F. Hermenier, X. Lorca, J.-M. Menaud, G. Muller, and J. Lawall. Entropy: a Consolidation Manager for Clusters. In *Proc. of the 2009 ACM International Conference on Virtual Execution Environments (VEE '09)*, pages 41–50, Mar. 2009.
6) M.R. Hines and K. Gopalan. Post-Copy Based Live Virtual Macine Migration Using Adaptie Pre-Paging and Dynamic Self-Ballooning. In *Proceedings of the 2009 ACM International Conference on Virtual Execution Environments (VEE '09)*, pages 51–60, Mar. 2009.
7) P. Lu, A. Barbalace, and B. Ravindran. HSG-LM: Hybrid-Copy Speculative Guest OS Live Migration without Hypervisor. In *Proc. of the 6th International Systems and Storage Conference (SYSTOR '13)*, pages 2:1–2:11, 2013.
8) QEMU Features/MicroCheckpointing.
9) A. Verma, P. Ahuja, and A. Neogi. pMapper: Power and Migration Cost Aware Application Placement in Virtualized Systems. In *Proc. of the 9th ACM/IFIP/USENIX International Conference on Middleware (Middleware '08)*, pages 243–264, Dec. 2008.
10) T. Wood, A. Venkataramani, and M. Yousif. Black-box and Gray-box Strategies for Virtual Machine Migration. In *Proceedings of the 4th USENIX Symposium on Networked Systems Design and Implementation (NSDI '07)*, pages 229–242, Apr. 2007.