

2 段階仮想化構成のマルチコア性能改善と評価

服部 直也[†] 今田 貴之[†] 森木 俊臣[†]
日立製作所中央研究所[†]

1. はじめに

今日の IT システムでは物理サーバに Virtual Machine Manager (VMM) をインストールして、複数の仮想サーバ (VM : Virtual Machine) を運用する構成が広く用いられている。物理サーバの性能は年々改善しているため、既存のシステム群をより高性能な物理サーバで統合するニーズが繰り返し発生する。本ニーズに応える選択肢の1つとして、我々は日立論理分割機構 Virtage の上で VMM を稼働させる 2 段階仮想化構成を提案している[2][3]。

2 段階仮想化構成の性能を分析したところ、VM にマルチコアを割り当てるとコア間の同期処理が増えて 1 コアより性能が下がるケースがあった。そこで本研究では分析と改善に取り組んだ。

2. サーバ仮想化の仕組み

サーバ仮想化とは、物理サーバに搭載された CPU コアなどを用いて仮想的なサーバ (VM) を作り出し、複数の OS/AP を稼働させる技術である。サーバ仮想化は VMM を用いて実現する (図 1)。

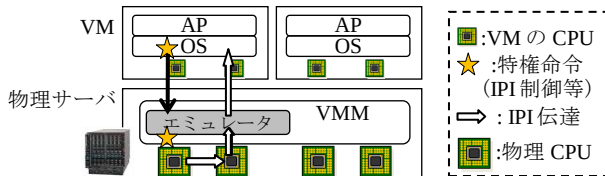


図 1 : サーバ仮想化の基本構成

一般に OS は CPU コア間割り込み (IPI : Inter Processor Interrupt) を送受信するなどの目的で特権命令を使用する。しかし、特権命令はシステム内で最高権限を持つプログラムしか実行できないため、図 1 の OS は特権命令を実行できない。そこで VMM はエミュレータを用いて OS の特権命令を代理実行する。エミュレータは直接実行に比べて低速なので、サーバ仮想化を用いると性能は低下する。サーバ仮想化のために CPU ベンダは、エミュレータの実行を支援する

Multi-Core Performance Improvement and
Evaluation of Nested Virtualization Server
Naoya Hattori[†], Takayuki Imada[†] and Toshiomi Moriki[†]
[†]Hitachi, Ltd., Central Research Laboratory

VT (Virtualization Technology) や頻度を減らす APICv (Advanced Programmable Interrupt Controller Virtualization) などの機能を実装し、性能低下を軽減している[1]。

3. 2 段階仮想化構成の概要

2 段階仮想化構成では物理サーバを複数の論理的な区画 (LPAR : Logical Partition) に分割し、LPAR 上で VMM を稼働させる (図 2)。LPAR を制御するハイパバイザには日立論理分割機構 Virtage などが該当する。各 LPAR では VMM が動作し、VM を構築する。各 VM では OS/AP が動作する。その際、OS の特権命令で VMM のエミュレータが実行され、VMM の特権命令で Virtage のエミュレータが実行される。2 段階仮想化構成では、エミュレータを実行する契機が多いため、性能が低下しやすい。

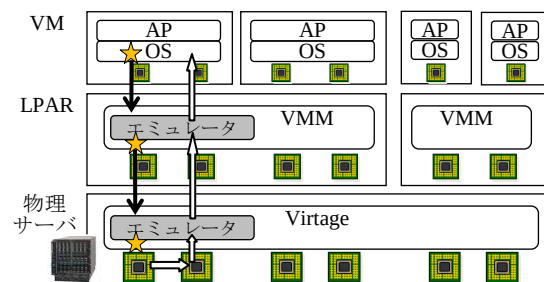


図 2 : 2 段階仮想化構成の例

4. 2 段階仮想化構成のマルチコア性能分析

2012 年製の物理サーバを用いて 2 段階仮想化構成のプロトタイプで OLTP (On Line Transaction Processing) を実行した時の相対処理時間を図 3 に示す。横軸は VM に 1 コアを割り当てた場合の処理時間を 1.0 としているが、VM に 2 コアを割り当てると予想に反して性能が劣化している。原因を調べると、2 段階仮想化構成では同期処理が 0.77 と非常に長くなっている。

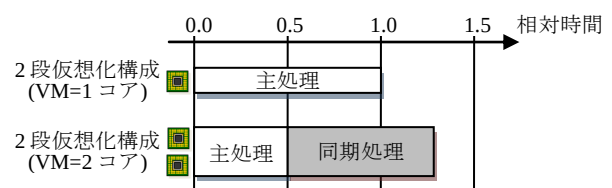


図 3 : 2 段階仮想化構成での OLTP 処理時間

同期処理の時間が長い原因を更に分析したところ、IPI 送受信時に実行される VMM のエミュレータが問題と判明した。OS は IPI を 1 回送受信するために、割り込み情報を操作する特権命令 3 つと IPI を送信する特権命令 1 つを実行する(図 4)。各特権命令は VMM のエミュレータによって処理されるが、VMM のエミュレータは VT 制御などの特権命令を多数含むので、更に Virtage のエミュレータを何度も呼び出す。そのため 2 段仮想化構成では VMM のエミュレータ実行に長時間を要していた。

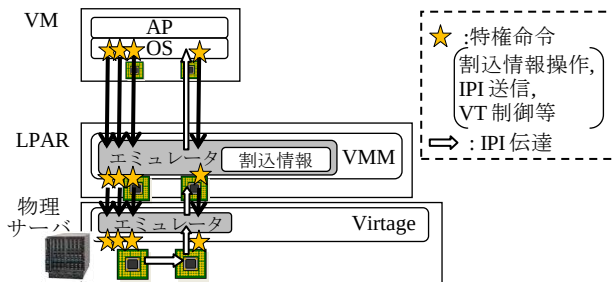


図 4: IPI 送受信時の特権命令処理

5. 仮想 APICv 方式の提案

2 段仮想化に限らず同期処理に伴うエミュレータ実行は仮想化一般で問題になっていた。そこで CPU ベンダは 2013 年から APICv と呼ばれる機能を搭載し始めた。APICv は、割り込み情報を取得／更新する特権命令に対して、エミュレータ相当の処理を CPU コアが直接実行する機能である。本機能を用いると VMM のエミュレータ実行が減るためマルチコア性能が改善する。

以上を踏まえて我々は 2 段仮想化構成向けに、LPAR 上で APICv 機能を提供する仮想 APICv 方式を提案する。本方式では、VMM が APICv を有効化する際に、Virtage のエミュレータが CPU コアの APICv を有効化する(図 5)。以降は OS が IPI を送受信する際に、CPU コアが割り込み情報を操作する特権命令を直接実行する。そのため VMM と Virtage のエミュレータ実行が減って、同期処理時間が短縮される。

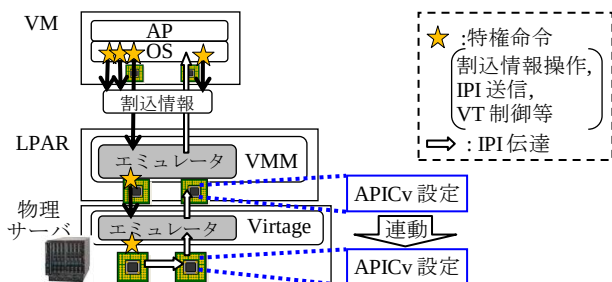


図 5: 仮想 APICv 方式

6. 性能評価

本研究では仮想 APICv 方式のプロトタイプを実装し、表 1 の構成で同期処理時間を測定した。

同期処理時間は仮想 APICv の効果で 67%減少した(図 6)。OLTP 処理時間の推定値も 2 コアが 1 コアより短縮されており(図 7)、本方式でマルチコア性能が改善する見通しが得られた。

表 1: 性能評価に用いた 2 段仮想化構成

VM	SW	AP	セマフォによる同期処理ベンチマーク
		OS	Fedora 20(64bit)
	HW	CPU	4Core
LPAR	SW	VMM	gemu-kvm-1.6.2-12.fc20.x86_64
		CPU	8Core
	HW	Memory	8GB
物理	SW	Hypervisor	仮想 APICv 方式のプロトタイプ
		Server	BS2000 BladeSymphony E55R4
		CPU	Intel® Xeon® CPU E5-2697 v2 (2socket : 計 12Core) (HyperThread/ターボ/省電力は無効)
		Memory	16GB (1333 MHz)
		NIC	Intel 82576
		HBA	Emulex LPe12002 (PCI)
		Storage	Hitachi AMS2100 (40GB LU x1)

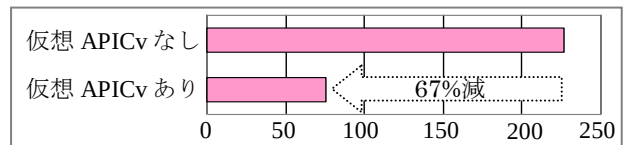


図 6: 仮想 APICv による同期処理時間[ns]の変化

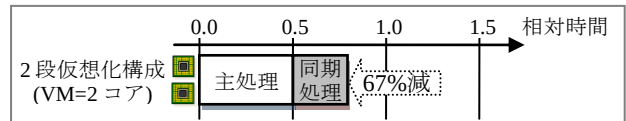


図 7: 仮想 APICv 適用時の推定 OLTP 処理時間

7. まとめ

本研究では、2 段仮想化構成に於けるマルチコアの同期処理時間を短縮する仮想 APICv 方式を提案し、プロトタイプを用いて同期処理時間の 67%減少を確認した。本方式により 2 段仮想化構成のマルチコア性能が改善する見込みである。

今後は実機評価を充実させると共に、より新しい CPU を活用する性能改善に取り組むたい。

参考文献

- [1] Intel®64 and IA-32 Architectures Software Developer's Manual
- [2] 水野和彦他, 2 段仮想化構成のユースケース提案と評価, 電子情報通信学会総合大会講演論文集 2012 年 情報・システム(1), 91, 2012-03-06
- [3] 服部直也他, 2 段仮想化構成のメモリ性能改善と評価, 電子情報通信学会総合大会講演論文集 2012 年 情報・システム(1), 90, 2012-03-06