基幹系システム向け仮想化技術「Virtage」の開発　その 3

對馬雄次(†)、森木俊臣(†)、服部直也(†)、荻本貴宏(‡)
(株)日立製作所　中央研究所(†)、(株)日立製作所　エンタープライズサーバ事業部(‡)

1. はじめに

IT システムへの性能や信頼性向上要求の高まりで多数のサーバが利用され、運用管理コスト増加が問題となっている。本問題に対し、仮想化技術を用いたサーバ統合でのサーバ台数削減が注目を集めている。

日立では、独自の仮想化技術として Virtage を開発した。Virtage は Intel 社の IPF/x86 に対応しており、物理サーバと仮想サーバの運用環境が同一であるという機器透過性を特長としている。

本論文では、Virtage が実現している機器透過性を実現するハードウェア支援機構について述べる。

2. 仮想化時の IO 支援機構の必要性

Virtage ではサーバ(以下、物理サーバ)が有する CPU やメモリ、IO デバイスを複数の区画に分ける。各区画では割当てられた資源を用いて仮想サーバを構築し、OS やデバイスドライバを無修正で動作させる。

仮想サーバ上で動作する OS(以下、ゲスト OS)は仮想サーバと物理サーバを区別せず扱う。このため、ゲスト OS はメモリを含む全資源を占有していると仮定して動作し、ゲスト OS はアドレス(以下、ゲスト物理アドレス)が 0 番地開始と認識している。しかし、実際のアドレス(以下、ホスト物理アドレス)を複数のゲスト物理アドレスで分割利用するため、必ずしも 0 番地から開始しているとは限らない。ゲスト物理アドレス・ホスト物理アドレスおよびアプリケーションが利用する仮想アドレスの関係を図 1 に示す。

本状況に対し、Virtage では CPU からのメモリアクセスについては、CPU が参照するページテーブルを Virtage が用意して、ホスト物理アドレスになるようにしている。

しかし、IO デバイスからのメモリアクセスについてはページテーブルが利用されない直接メモリアクセス(以下、DMA)となる。また、Virtage ではゲスト OS およびデバイスドライバの修正を行わないため、ゲスト OS が認識するゲスト物理アドレスを用いてデバイスドライバに動作指示を行う。このため、IO デバイスはゲスト物理アドレスに基づいて DMA を行う。

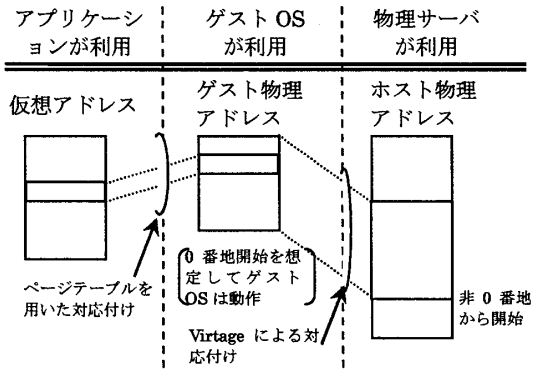

図 1：アドレス空間の関係

先に述べた通りゲスト物理アドレスとホスト物理アドレスは異なるため、ゲスト物理アドレスによって DMA が実施されると誤動作の原因となる。

誤動作防止のため、DMA を直接ゲスト OS が利用するホスト物理アドレスに変換するハードウェアによる IO 支援機構が必要不可欠となる。

3. Virtage における IO 支援機構

3.1 Virtage における対応方針

DMA は IO デバイスが接続される PCI バス上のトランザクション(以下、Tx)である。したがって、Tx を受けて処理するチップセットで実現する方針とした(図 2)。

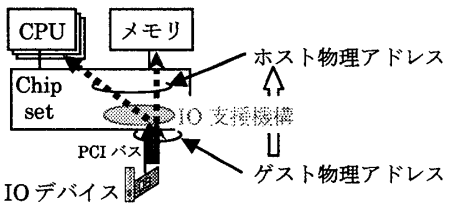

図 2：IO 支援機構の搭載方針

Development of Server Virtualization Feature "Virtage" (3)
† Yuji Tsushima, Toshiomi Moriki, Naoya Hattori,　Hitachi,Ltd. Central Research Laboratory
‡ Takahiro Ogimoto,　Hitachi,Ltd. Enterprise Server Department

3.2 IO 支援機構の提供機能

チップセットに搭載される IO 支援機構が提供する機能と概要(図 3)を以下記す。

(1) アドレス変換/検査機能

IO デバイスからの Tx を受け付けて、当該IOデバイスに対応する仮想サーバのホスト物理アドレスに変換する機能である。本機能の実現のため、Virtage では仮想サーバ毎にゲスト物理アドレスとホスト物理アドレスの差分(オフセット)をチップセットがレジスタに記憶している。本差分を DMA アドレスに加算してアドレス変換を実施している。また、DMA アクセス先が仮想サーバ用のホスト物理アドレスか否かの検査も同時に実施している。

(2) 割り込み先検査機能

IO デバイスからは(1)で示した DMA 以外に割り込み Tx もある。割り込み Tx については、割り込み先の CPU が当該 IO デバイスに対応した仮想サーバに割り付けられた CPU か否かの検査を実施する。

(3) 不正アクセス処理機能

DMA と割り込み Tx の検査により不正アクセスを検出すると、PCI バスでのタイムアウトや、仮想サーバの誤動作といった副作用がないように、Virtage が用意するメモリ領域へのアクセスに変換し、Tx 自体を正常終了させる。

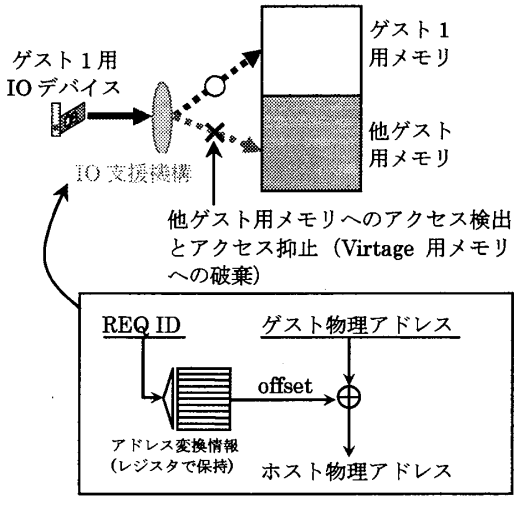

図 3：IO 支援機構のアドレス変換機能

4. IO 支援機構の効果

IO 支援機構を用いた場合の効果を表 1 に示す。本効果は、1Gbps の NIC と FC HBA 利用時の物理サーバの性能(スループット)を 1 とした場合の仮想サーバの性能を示す。物理サーバに比べて 15%程度のオーバヘッドであり高い性能を発揮している。

表 1：IO 支援機構利用時の相対性能

	1Gbps NIC	FC HBA*
物理サーバ	1(基準)	1(基準)
仮想サーバ	0.96	0.95

6. 関連研究との比較

仮想化技術の普及により IO 支援機構に関する研究開発も活発である。Intel 社と AMD 社からは、本研究と同様の IO 支援機構の対応を表明している[1][2]。これらは、IO デバイスからの DMA をハードウェアで変換している。しかし、両機構共に OS が用意するページテーブルのようなデータ構造をメモリ上に配置し、ハードウェアがこれを参照して動作する。このため、IO 側にも TLB に類似する構成を有するため、IO 側 TLB ミスによりメモリアクセスが発生し性能変動を引き起こすと推定される。

7. おわりに

本研究では、日立仮想化機構 Virtage における IO 支援機構について述べた。本機構では仮想サーバ実現に必要となる DMA に関するアドレス変換・検査機能や割り込み先の検査機能を有している。これらによって、デバイスドライバを含む動作環境が物理サーバと仮想サーバで同一である機器透過性を実現している。また、アドレス変換等の処理を全てレジスタで行うことによって性能変動がないのも特長となっている。

参考文献

[1] Intel Technology Journal, Aug. 10, 2006
[2] AMD64 Architecture Programmer's Manual, Vol 2: System Programming

* ブロックサイズ＝256Byte 時