ホワイトペーパー

HITACHI
Inspire the Next

BladeSymphony

サーバ仮想化機構「Virtage（バタージュ）」ハードウェア透過性

2008年9月発行

株式会社　日立製作所

目　次

１．はじめに

企業の基幹系システムをサーバ仮想化環境で構築するためには、物理サーバと同様な運用を可能とするハードウェア透過性が必要です。本解説では、日立のブレードサーバ BladeSymphony 向け サーバ仮想化機構 Virtage(バタージュ)における、ハードウェア透過性の実現方法とその利用シーンについて説明します。

２．ハードウェア透過性

２．１ ハードウェア透過性とは

ハードウェア透過性とは、論理サーバ上の OS(以下、ゲスト OS)が、ハードウェアが仮想化のためにエミュレートされていることを意識することなく、すなわちゲスト OS や標準提供されるドライバが無修正で、物理サーバ上と同様に動作できる性質のことを言います。これを実現するためには、論理サーバが動作しているプロセッサやメモリ資源だけでなく、物理デバイスを直接アクセスできる必要があります。ゲスト OS と物理デバイスの間に介在し、このハードウェア透過性を実現しているのが、論理サーバから物理デバイスへの直接アクセスを支援する I/O 仮想化支援機構と呼ばれるハードウェアとハイパバイザの制御です。

２．２ ハードウェア透過性の条件

このようなハードウェア透過性を実現するためには、仮想化制御方式において、以下の二つの特徴的な条件を満たす必要があり、Virtage(バタージュ)ではこれを実現しています。

条件1：ゲストOS(仮想計算機上のOS)が発行する全てのI/Oコマンドとそのレスポンスが、物理サーバの場合のそれらと等しいこと。

条件2：ゲストOSが利用するディスクのフォーマットが、物理サーバの場合と等しいこと。

２．３ ハードウェア透過性実現の目的

Virtage におけるハードウェア透過性実現の目的について、それぞれの性質のメリットから説明します。

ゲスト OS がディスクに対して行う処理が単純な読み書きだけであれば、必ずしも条件1を満たす必要はありません。むしろ、他社仮想化ソフトウェアでは、条件1も条件2も満たしていないことが普通となっています。しかし、 クラスタ制御ソフトなどが発行する制御系コマンドアクセスに対しても、正しい応答を返すことができるためには、条件1を満たす必要があります。逆に言えば、この条件を満たすことにより、論理サーバ間で連携したクラスタシステムを、特別な制約なしに構築することが可能になります。

また、条件2を満たすことで、得られるメリットが二つあります。一つはバックアップ環境を構築する際に、ファイルシステムを直接アクセスできるため、仮想化を利用しないバックアップサーバからのアクセスが可能になる点です。これにより LAN を介さずに SAN 経由で、ゲスト OS が使用しているディスクの差分バックアップをファイル単位でとる、いわゆる LAN フリーバックアップ構成の採

用が可能になります。

もう一つのメリットは、ゲスト OS が確実にディスク装置に書き込むべきデータを確実に書き込むことができるという点です。一般に、他社仮想化ソフトウェアでは、ディスク装置上の記録形式として、仮想ファイルシステムを採用することが多く、この場合には、メモリ上にディスクキャッシュを用意することが普通です。このように、仮想化ソフトウェアがディスクキャッシュを持っていると、ゲスト OS が raw write(物理ディスクに確実に記録するための動作)を実行する場合でも、ゲスト OS に書き込み完了の応答を返した後、一定時間キャッシュ上にのみデータが保持されてしまう恐れがあります。このような動作はトランザクションモニタやデータベースソフトウェアなどの信頼性保持に必要なジャーナルファイル用途には利用できないことを意味します。仮想化ソフトウェアによっては、raw write 対応に必要なファイルについては、個別に対応する機能を用意していますが、システム構成が複雑になります。Virtage では、条件2を満たすことで、高いシステム的信頼性を満たしています。

3. 仮想化制御アーキテクチャの比較

ここでは I/O 仮想化方式について、代表的な二つの方式を比較します。ひとつは、Virtage が採用しているパススルー方式であり、もうひとつは、他社仮想化ソフトウェアによく採用されているハイパバイザエミュレーション方式です。ハードウェア透過性実現には、これらのうちのパススルー方式が必要なことを説明します。

(1) パススルー方式

パススルー方式とは、ゲスト OS が I/O 動作を起動・実行する際にハイパバイザの介入を必要としない方式のことです。I/O 動作を起動する際には、ゲスト OS は、I/O デバイス(例:物理 FC カード)に、データ転送領域のメモリアドレスを設定する必要があります。I/O デバイスは、設定されたメモリアドレスを用いて、I/O 機器とメモリ間のデータ転送動作を実行します。この I/O デバイスの動作のことを、DMA(Direct Memory Access)転送と言います。

ゲスト OS が、I/O デバイスに設定する DMA アドレスは、仮想化されたメモリ上でのアドレスなので、このまま I/O デバイスがデータ転送を実行すると、誤ったメモリ上にデータが転送されてしまいます。これをハイパバイザの介入無しに防止するためには、ハードウェアの支援が必要になります。Virtage では、I/O 仮想化支援機構と呼ぶハードウェア機能を提供することにより仮想化されたメモリアドレスの課題を解決し、パススルー方式を実現しました。

(2) ハイパバイザエミュレーション方式

ハイパバイザエミュレーション方式とは、ゲスト OS の DMA 転送のアドレス変換を、I/O 起動時にハイパバイザがトラップしてエミュレーションすることによって行う仮想化方式です。この方式ではゲスト OS に対するエミュレータと物理デバイスに対するドライバ、及び DMA 転送を正しく設定するためのアドレス変換制御が必要となります。この方式を採用してハードウェア透過性を実現しようとすると、I/O デバイス 1 種に対して、ゲスト OS に対するエミュレータを 1 種とハイパバイザ内のドライバ 1 種をセットにして開発しなければならず、開発量が多くなるという問題があります。この問題

を避けるため、ゲスト OS に対しては、仮想化固有の標準インタフェースのみをエミュレートして提供し、ハイパバイザ内のドライバを I/O デバイス毎に用意する方式が採用されることが多く、他社仮想化ソフトウェアのほとんどが、このような方式により問題点を回避しています。したがって、ゲスト OS に見せる I/O インタフェースと実際の I/O デバイスのインタフェースが異なることとなり、ハードウェア透過性が実現できなくなります。

図 1 に各方式の概念図を、表1 に機能上の特徴比較を示します。

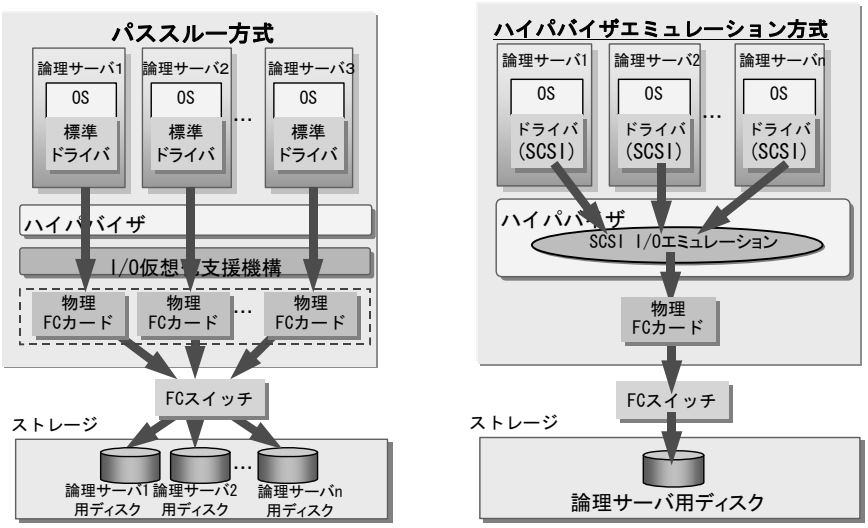


図 1　I/O 仮想化方式の概念図

表 1　I/O 仮想化方式の特徴比較

	パススルー方式		ハイパバイザ エミュレーション方式	
I/O 性能	○	H/W による直接実行	△	ハイパバイザ・エミュレーション
ﾊｰﾄﾞｳｪｱ透過性(1)	○	クラスタソフトなど対応可	×	制御系アクセスへの応答が困難
ﾊｰﾄﾞｳｪｱ透過性(2)	○	LAN フリーバックアップ/ DB ソフトなど対応可	×	仮想ファイルシステム対応要
ディスク仮想化	×	仮想ディスクをサポートしない	○	VM の移動性良い

このように、企業の基幹系システム向けに用いられる BladeSymphony BS1000 に搭載する仮想化環境では、確実な構成を組むことを重視するため、Virtage では I/O デバイスを直接アクセスするパススルー方式を採用こととし、ハードウェア機構として独自の I/O 仮想化支援機構を開発し、パススルー方式を実現しました。

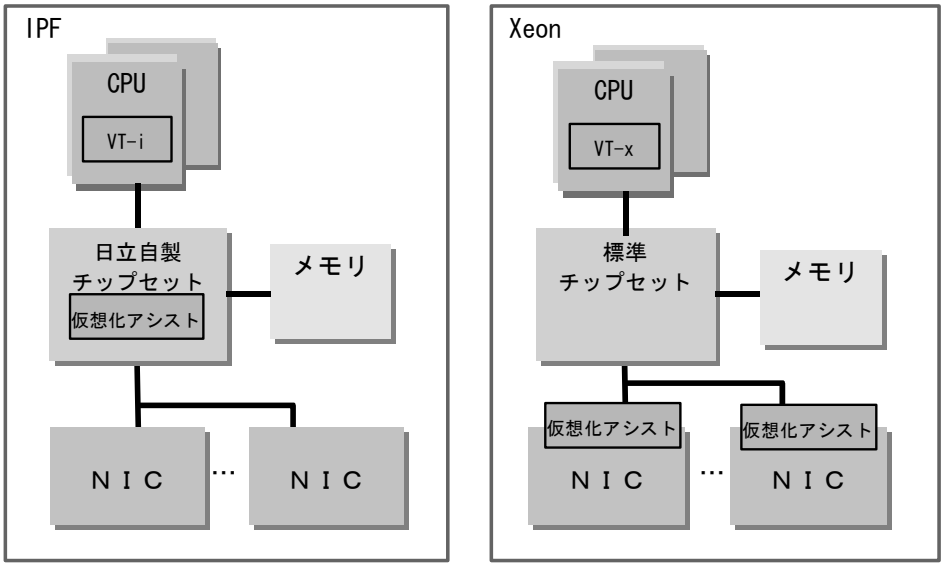
4．BladeSymphony BS1000 での実現方法とサポート結果

Virtage の稼動プラットフォームである日立ブレードサーバ BladeSymphony BS1000 における、ハードウェア透過性の具体的な実現方法をアーキテクチャ別に説明します。

IPF サーバブレードでは、先に述べた DMA アドレスの変換を行なう I/O 仮想化支援機構を自製のチップセットの中に有し、これによって I/O 仮想化のパススルー方式を実現しています(図2(a))。

一方 Xeon サーバブレードでは、自製 HBA に I/O 仮想化支援機構の論理を組み込み、これによって I/O のパススルー方式を実現し、NIC については I/O アクセラレータ(オプション)を用いることによりパススルー方式を実現しています(図2(b))。

このように、方式は異なりますが、Virtage ではアーキテクチャの種別に関係なく、ハードウェア透過性を実現しています。このサポート結果として、2．3節で述べたクラスタシステム・LAN フリーバックアップ・データベースなどの、ハードウェア透過性により実現される高信頼の仮想環境を、物理サーバの場合と同様に構築することを可能にしました。



（a）IPFサーバブレード構成　　（b）Xeonサーバブレード構成

図2 BS1000 のハードウェア構成図

5．まとめ

基幹系システム構築に適したサーバ仮想化機構として、ハードウェア透過性を持つ仮想化制御方式を実現しました。これにより、仮想環境(クラスタシステム・LAN フリーバックアップ・データベースなど)を、物理サーバと同様に構築することが可能になり、物理サーバを用いたシステム構築の経験で得たシステム構築ノウハウを、そのまま仮想環境でも活かすことができます。

VWP-002 2008.09