サーバ仮想化技術を利用した共通プラットフォームの運用事例

テクニカルコンピテンシー部
基盤技術室
シニアITスペシャリスト

高橋　一博

Kazuhiro　Takahashi
kazuhiro-takahashi@exa-corp.co.jp

オープン系サーバ仮想化技術を利用した社内向け共通プラットフォーム開発検証環境を2008年度に構築し、部内および事業部向けのファイル・サーバ、Webサーバ、開発および検証用サーバの運用環境として、仮想マシンを提供してきた。本稿では社内向けという狭い範囲に限定されるが、機能や構成の異なる物理マシンの仮想環境への移行、運用を通して得られた経験をもとに、移行計画および仮想マシンの構成についての留意点と採用した方針を報告する。

1. はじめに

弊社内には部単位で運用しているファイル・サーバやWebサーバおよび開発・検証用サーバなど多数のサーバが存在しており、そのほとんどがx86系サーバである。2008年度にサーバ仮想化技術とブレード・サーバの採用によって共通プラットフォームを構築し、仮想化に適した既存サーバの移行、および新規開発用サーバの提供を進めてきた。その結果、当初の目的であった物理サーバが占有する執務スペースの圧迫、サーバ台数に比例する管理の複雑化、それによって生じる個々人の管理負荷の増加、マシンの老朽化などの問題を解決してきた。

ただし、既存サーバから共通プラットフォームへの移行、および共通プラットフォームの利用環境については、ユーザから多種多様な要望が寄せられ、仮想化ゆえの今までにはなかった課題も浮かび上がった。それらに対応するため、一定方針のもとに集約を図った結果、現時点で仮想サーバ数は70台を超え、共通プラットフォームの増強をなお継続している。サーバ仮想化技術はさらに拡大・深化しているが、社内向けサーバの仮想化という状況下で採用した方針について、共通プラットフォームの改善につなげる意味も含めてその概要を報告する。

2. 共通プラットフォーム

2.1. 共通プラットフォーム概要

サーバ物理統合のため、標準ラックにブレード・シャーシ、ブレード・サーバ、ファイバ・チャネル（以後、FCと略す）RAID、管理サーバ、バックアップ・サーバおよびテープ・ライブラリを搭載した。ブレード・サーバにはFC Host Bus Adapter（以後、HBAと略す）を装着し、FC RAIDによって記憶装置の集中と仮想化を実現した。

実際に採用したブレードは2CPU（8コア）、16GBメモリ、システム・ディスク、NIC（Ethernet 1000Base-T x 2）およびFC HBA（オプション・カード）で構成されている。ブレードを使用した共通プラットフォームは２式あり、１号機は主に当部内で利用中の現行ファイル・サーバやWebサーバの移行用とし、RAIDにはIBM-DS4700（FC 300GB x 16）を採用した。２号機は主に事業部向けの開発および検証用サーバの運用環境であり、RAIDにはIBM-DS4700（FC 300GB x 16およびSATA 500GB x 16）を採用した。

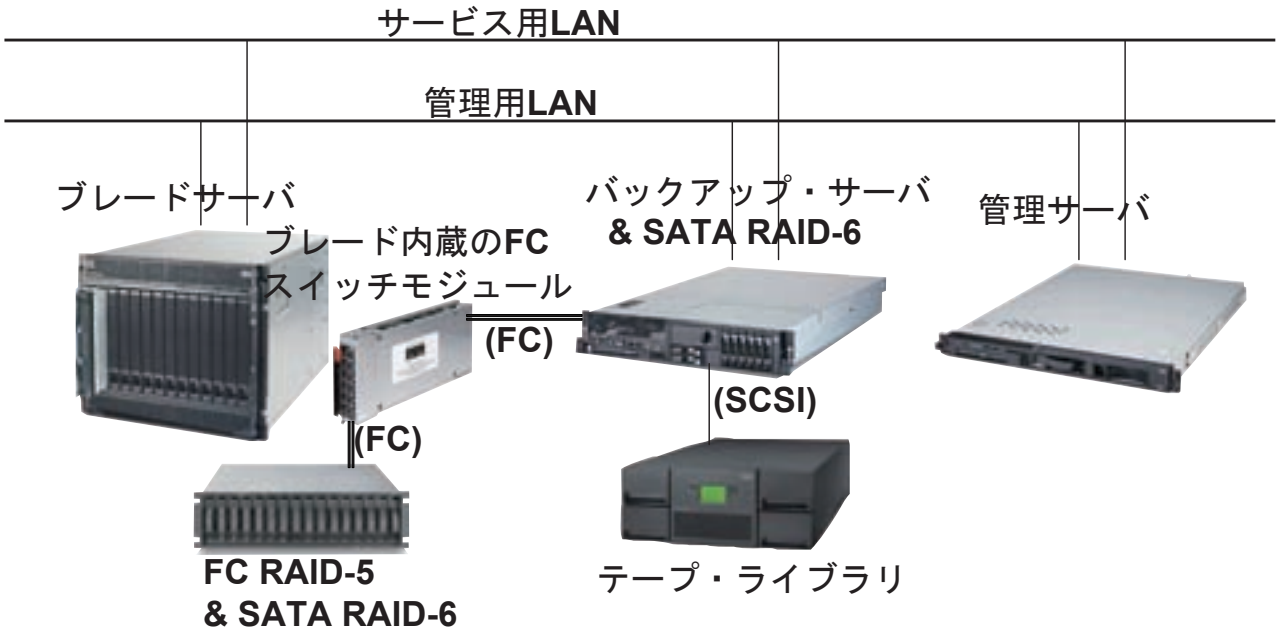

図１　共通プラットフォーム

またサーバ仮想化環境には、VMware Infrastructure 3（以後、VI3と略す）を採用した。本報告も、その利用実績をもとにしていることに注意して欲しい。

2.2. 共通プラットフォームに対する要求事項

共通プラットフォームの利用前・後にユーザから寄せられた主な要望を表１に示す。ユーザには、既存システムの管理者、開発・検証用サーバの管理者およびそれらの利用者が含まれる。

なお、移行した個々の既存サーバには関連性がなく、多種多様な運用・管理がなされていた状況が、表１の要求に反映していると考えられる。

3. 課題への取り組み

セキュリティを重視したため共通プラットフォームへの物理的なアクセスは制限されるが、サーバ仮想化技術を用いた場合、仮想サーバに対しては従来どおりのユーザ・インタフェースが確保される。よって２章に挙げた要求事項のうち、解決すべき課題は下記のとおりとなる。

- 既存サーバの共通プラットフォーム移行に伴うサービス停止期間の短縮
- 性能保証
- システム管理（特にバックアップ）

これら課題に対して共通プラットフォームで採用した方針および施策を以降に記す。

3.1. 既存サーバの移行

3.1.1. 移行ガイドライン

部内の既存ファイル・サーバやWebサーバのシステム管理者およびユーザから、ユーザに対するサービス停止時間が短くなるよう要望を受け、それに対応した移行ガイドラインを設けた。

3.1.2. 留意事項

ガイドラインを設けた2008年はLinuxシステムをサーバ仮想化プラットフォームへ自動的に移行するツール（以後、P2Vツールと略す）が少なく、機能的な選択肢も限られていた。そのためLinuxの場合は特定のツールが移行に失敗した時点で、ゲストOSのインストールを出発点とする再構築を行った。よって本稿ではWindowsの移行について記す。図２に移行の全体的手順を示す。

（1）移行支援ツールの比較

移行先の共通プラットフォームはVI3環境であり、移行ツールにもVMware社製品を使用した。提供される製品に

表１　課題

課題	内容
サービス停止時間	既存システムの共通プラットフォームへの移行に際し、ユーザに対するサービス停止時間を短くする。
ユーザ・インタフェース	従来と同様のアクセス方法とユーザ・インタフェースを提供する。
性能保証	既存システムの仮想化では現行システムと同等以上の性能を、新規システムについても要求仕様を満たす仮想マシンを提供する。
リソース増強	サービス提供を継続する上でリソース不足が懸念される場合は、リソースの追加が可能なこと。
システム管理	現行システムと同じくパッチの適用等を独自に行える。 システムおよびデータのバックアップのために新たにエージェントの追加は行なわない。
その他	DNSやアクティブ・ディレクトリサービスを継続して受けられること。

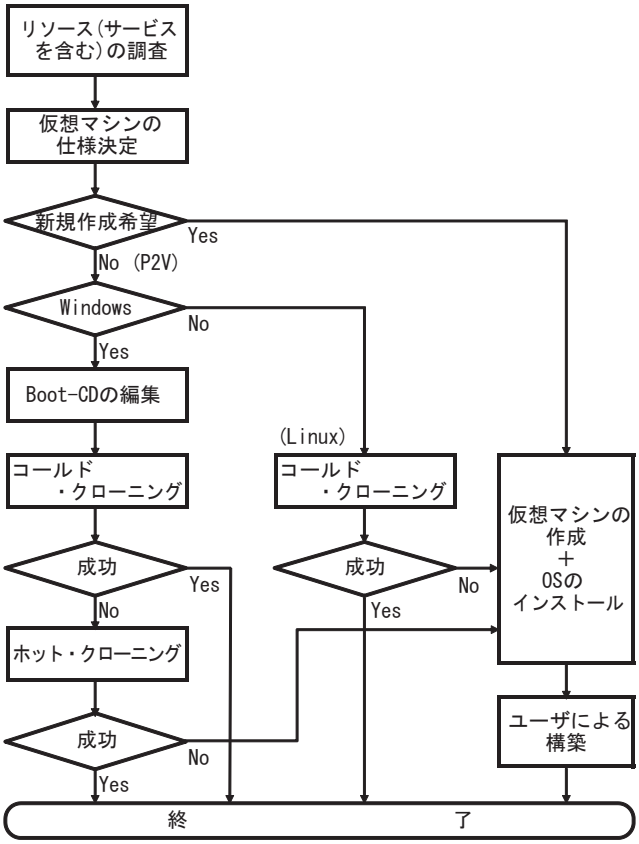

図２　移行の流れ

はコールド・クローニング用のBoot-CDとホット・クローニング用のVMware Converterがあり、その比較を表２に示す。それらに移行元および移行先に関する必要な情報を与えれば、以後はネットワーク経由で自動的にシステムの移行が行われる。

(2) 全体の流れ

今回の移行要件に、サービス停止時間の短縮が挙げられている。コールド・クローニングおよびホット・クローニングとも、正常に作業が終了すれば移行に必要な時間に変わりはなく、ディスク容量にもよるが数時間以上必要なケースが多い。一方、移行作業が失敗したことが判明するのは、コールド・クローニングはブート直後であるのに対して、ホット・クローニングの場合は作業進捗率が95%以上となってからがほとんどである。よって標準手順では、失敗がすぐに判明するコールド・クローニングを行い、次いでホット・クローニングに取り組むこととした。

表２　移行支援ツールの比較

	コールド・クローニング	ホット・クローニング
提供元	VMware, Inc.	
移行対象OS	Microsoft Windows	
特徴	システムのshutdownが必要。	システムの稼働中に静止点が設けられるだけで、shutdownの必要はない。
起動方法	コールド・クローニング専用Boot-CDを用いたboot(Windows PE)後に、アプリケーションとして自動的に起動。	物理サーバ上のOS(Windows)のアプリケーションとして起動させる。
制約事項	専用Boot-CDに、物理サーバに対応したデバイス・ドライバが含まれていない場合、ネットワークが使えない、つまり移行データを送出できない等の不具合が発生する。 VMware提供のpeToolを使用することによって、Boot-CDにドライバの事前追加が可能。	OS上のサービス、例えばセキュリティ監視ツールやP2P監視ツールにより、移行が阻害されることがある。
移行失敗判明の時期	boot直後。	移行終了の最終段階。
移行所要時間	移行対象のファイル・システム(データ)の大きさに比例。	

（3）運用サービスの調査

移行元サーバのリソース調査対象項目にOS上で稼働しているサービスを含める。ホット・クローニング開始時に停止または削除するサービスを事前に決定し、システム管理者の同意を得ておく。なおホット・クローニング開始時にサービスを削除した場合は、移行後のゲストOSに再インストールする。

（4）Boot-CDの編集

移行元サーバのリソース調査対象項目に物理サーバが使用しているデバイスも含めた。特殊なデバイスの場合、デバイス・ドライバが入手できた場合は、あらかじめBoot-CDに組み込んでおく。なお、Boot-CDに含まれるデバイス・ドライバは公式には公開されていないので、特殊なデバイスの判断は経験によるところが大きい。

VMware社およびMicrosoft社はWindows PE用のデバイス・ドライバを提供していない。必要なデバイス・ドライバはハードウェア・ベンダーから入手する。またBoot-CDへの組み込みには、Boot-CDダウンロード時に同時に入手したpeTool.exeを使用する。

（5）コールド・クローニング

Boot-CDによりWindows PEが立ち上がった時点で、必要なデバイス・ドライバが不足していれば移行ステップは停止するので、コールド・クローニング作業を中止し、ホット・クローニングに切り替える。

（6）ホット・クローニング

ホット・クローニング開始前に移行を阻害すると思われるサービスを停止する。移行の失敗が判明するのは終盤であるが、画面に表示される作業進捗率が進まなくなった時点で失敗とみなし作業を停止する。ただし、移行先には仮想マシンが作成されているので、念のために電源を上げて確認を行う。稀にではあるが100%完了していなくとも、仮想マシンとして稼働する場合もあるので、移行元サーバ管理者の確認を受けて、移行完了とするかの判断を下す。移行失敗の場合は、仮想マシンのゲストOSのインストールを含めた再構築に着手する。

3.2. 性能保証

既存システムに対しては、移行前と同等の性能を保証している。そのための仮想マシン配置時の留意点を述べる。

3.2.1. キャパシティ・プランニング

ESXサーバ上への仮想マシン配置決定には、下記のとおりいくつかの観点がある。

- 仮想マシンの管理者・管理組織
- 仮想マシンが提供しているサービス内容
- OSおよびソフトウェアの種類（仮想環境での有効ライセンス数）
- 物理リソース（ESXサーバ）の有効利用

共通プラットフォームでは「物理リソースの有効利用」に重点を置いた。理由は、物理リソースを有効に使いたいということもあるが、物理リソースが不足した場合、仮想マシンのパフォーマンスが極端に低下するため、それを防ぐという意味が大きい。

サーバ仮想化メリットの一つに、ワークロードの異なる仮想マシンを同一のESXサーバ上に適切に配置することによって、物理リソースを有効に利用できることが挙げられる。その結果、個々の仮想マシンが要求するリソースの合計値が物理リソースを上回ることが可能となり、サーバの集約効果が期待できる。そのため既存システムに対しては、物理リソースおよびシステム処理繁忙時を含むリソース調査シートを用いた聞き取り調査を行った。この結果をもとに仮想マシンを配置したが、目安としてはVirtualCenter Management ServerがモニターするESXサーバのCPUおよびメモリの使用率80%を注意レベルとした。またSNMPアラーム通知を利用した監視により、注意レベルの長時間発生を防いでいる。

ただし、今回移行対象となったサーバ群は、基幹システムのようにワークロードのピーク発生が確実に予測可能なものではなく、突然のリソース消費発生を考慮する必要があった。また、ESXサーバ上で稼働するそれぞれの仮想マシンが、CPU、メモリ、ディスク、NIC等を、独自のリソースとして認識かつ使用するためのオーバ・ヘッドも発生しており、ESXサーバの負荷は両者を合計したものになる。オーバ・ヘッド項目のうち「メモリの管理」、「ディスク

の管理」においては予想外の負荷が発生する。以降に具体例を示すが、その特徴を捉えることによってパフォーマンス低下の原因を特定し、共通プラットフォーム内での対応につなげている。

3.2.2. メモリの管理

ESXサーバには通常の物理サーバにおける、

物理マシン　～　OS　～　アプリケーション

による負荷に加えて、サーバ仮想化特有のオーバ・ヘッドがある。VirtualCenter Management Serverがモニターするサーバの各パフォーマンス値が注意レベルに達した場合には、ゲストOS上のアプリケーション負荷が高くなったのかを確認するだけでなく、以降に記す仮想化特有の要因も検討に加え性能の回復に努めた。

CPUをオーバコミットしている場合、物理サーバのリソースはシェア値による重み付けによって仮想マシンに配分される。NICは直接の重み付けはされないが、仮想スイッチのポート単位でトラフィック・シェーピングが適用され、表面上はパフォーマンスの低下となって処理される。これらのリソースは時分割によってシェアされるが、メモリは時分割できないことに注意が必要である。

メモリのアドレス変換は、

ゲストOS上の仮想アドレス　～　仮想マシン上の実アドレス　～　物理マシン上の実アドレス

の二段階となっているが、ESXサーバは[シャドウ・ページ・テーブル]と呼ばれる

ゲストOS上の仮想アドレス　～　物理マシン上の実アドレス

を一段階で行うマッピング・テーブルを作成し、プロセッサ上のTLB（仮想アドレスから物理アドレスへの高速変換を目的としたCPU内キャッシュ）がキャッシュすることによって、アドレスのマッピング・オーバヘッドを回避している。しかし下記の場合は、シャドウ・ページ・テーブル再作成のためにCPUサイクルを費やすため、物理マシン上でOSを直接稼働させる場合に比較して、大きなパフォーマンスの低下を引き起こす。

- 仮想マシンのオペレーティング・システムが、ゲストOS上の仮想アドレス　～　仮想マシン上の実アドレスのマッピングを変更する。
- 仮想マシンのオペレーティング・システムが、コンテキスト・スイッチを行う。

この現象は各仮想マシンのワークロードを重ね合わせただけでは推測が困難であり、サーバ仮想化環境で顕在化するので注意が必要である。

3.2.3. ディスクの管理

仮想マシンにおけるディスクI/O割り込み処理などをエミュレーションし、仮想マシンと物理マシンを橋渡しするバイナリ・トランスレーションの中心となるVMXプロセスについては、下記のとおりである。

① VMXは仮想マシン毎にVMkernelのアプリケーションとして起動する。
② VMXは仮想マシンのCPUとは別のCPUが割り当てられる。
③ 仮想マシンのI/O処理はVMM経由でVMXが行う（バイナリ・トランスレーション）。ただしこのときに、VMX用のCPUが仮想マシンのCPUと同時に割り当てられる必要はない。

よって、仮想マシンがI/O処理を終了するためには、仮想マシン用のCPUの他に1CPU必要になる。

このディスクI/Oが実際に、ゲストOSで処理されるまでの経路を下記に記す。

ゲストOSのドライバ
↓
VMM提供のSCSIコントローラ
↓
仮想SCSIレイヤ・スイッチ
↓
VMFS（通常の物理サーバでは、ここで作業完了だが、VI3ではパス・スルーしている）
↓
VMXプロセス（各仮想マシンにインフラを提供）
↓
VMkernelのキューに登録

仮想マシン群がFCディスクを共有している場合、I/Oの

競合によってパフォーマンスの低下が生じるが、割り込み処理のための大量のバイナリ・トランスレーションは、CPUにおいて予想外の大きな負荷を発生させる。

3.3. 仮想マシンおよびデータのバックアップ

移行対象の既存サーバは統合管理対象となっていないため、個々のサーバ管理者によって独自の方法で管理されており、バックアップも各々の状況に従って行われていた。今回はほとんどの既存サーバ管理者、つまり仮想マシンのサーバ管理者が対応可能なシステム運用環境の提供を目標として、下記のバックアップ方式とした。

3.3.1. 仮想マシンのシステム・バックアップ

サーバ管理者が既存サーバの移行完了を確認した時点で、下記のバックアップを行った。

- リストア操作に複数のステップが必要なVMware Consolidated Backup（以後、VCBと略す）は原則として使用しない。仮想マシンを停止できない状態が長く続くと判定したときだけ、VCBによる暫定バックアップを行う。
- VI3管理者がVI3の機能を使用して、該当仮想マシンのテンプレートを作成。
- 書き込み先はバックアップ・サーバのRAID-6構成のディスク。
- VI3管理者がバックアップ用ソフトウェアを使用して、同テンプレートをテープにバックアップ。
- テンプレートは一世代のみ保管。新規作成依頼時には旧テンプレートを削除（ディスク容量の確保が目的）。

テンプレートの作成は、同時には二台以上行わない。クローン作成やテンプレート作成は共有ディスクに対するI/Oレートが高く、常時発生しているゲストOS群のディスクI/O処理量によってはレスポンス・タイムアウトを起こし、作業が失敗につながるためである。

また、テンプレートおよびテープの管理はVI3管理者が行い、サーバ管理者からの要求に従ってリストアし、操作ミスによるシステムの回復などに対応している。

3.3.2. データのバックアップ

データのバックアップも仮想マシンのシステム・バックアップに近い考え方である。

- 個々のサーバ管理者に対して、ネットワーク上にデータ・バックアップ用のディスク領域を提供。
- バックアップ先に対するアクセス権はサーバ管理者に対してのみ付与。
- 書き込み先はバックアップ・サーバのRAID-6構成のディスク。
- サーバ管理者がバックアップ・データを指定された場所にコピー。
- バックアップ・データの選定およびスケジュールはサーバ管理者が決定し、VI3管理者は関与しない。
- VI3管理者はバックアップ用ソフトウェアを使用して、バックアップ用ディスクからテープにバックアップ。
- テープは二世代まで作成。それ以前のリストアは不可能。

4. おわりに

サーバ仮想化技術を利用した共通プラットフォームを2008年度に構築した。今回は既存サーバの移行および新規開発環境の提供時に、管理者およびユーザの要求を実現するために採った方針・方式を報告した。サーバ仮想化環境管理の一助となれば幸いである。なお、Intelから出荷されたXeon 5500番台は拡張ページ・テーブル（EPT）機能が実装され、ここに述べたシャドウ・ページ・テーブルとは異なったメモリ管理も行っているため、今回記した内容が全てに適用できないことに注意していただきたい。

また、サーバ仮想化ソフトウェアとしては従来のVI3やXenだけでなく、Intel VTやAMD-Vを前提としたMicrosoft Hyper-V、Red Hatが採用を打ち出したLinuxカーネル仮想化基盤のKernel-based Virtual Machine（KVM）なども続々と登場している。今後はそれぞれの短長所を把握し、仮想マシンを適切な環境で稼働させることが重要になると思われる。

参考文献

1) VMware徹底入門 ㈱翔泳社 2008/11
2) リソース管理ガイド VMware, Inc.
3) 基本システム管理 VMware, Inc.
4) サーバ仮想化技術を利用した社内サーバの集約事例 エクサ・レビュー 2008/11
