INTEGRATION OF COMPUTERS BY INTRODUCTION OF THE VIRTUALIZATION TECHNOLOGY IN SPRING-8

Masahiko Kodera, Taichi Shimizu, Ko Mayama, Masahiko Hanada, Shigeru Yokota, Ryotaro Tanaka
Japan Synchrotron Radiation Research Institute (JASRI/SPring-8)
1-1-1, Kouto, Sayo-cho, Sayo-gun, Hyogo 679-5198

Abstract

We applied virtualization technology for the computers integration in SPring-8. The virtualization technology is useful to maintain the large number of computers with the minimum staffs preventing the computer proliferation in a large accelerator facility. Furthermore, we improved availability of the computer system with less hardware system and the effectively distributed computer resources.

SPring-8における、仮想化技術を用いた計算機統合

１．はじめに

SPring-8の制御・情報系計算機においては、広大な敷地内に存在する多数の計算機メンテナンスを少人数で行い、同時に計算機コストを低減しながら信頼性を向上させるために、仮想化技術を用いた計算機ハードウエアの統合を進めている。

SPring-8ではこれまでに、ビームラインの操作端末の仮想化[1]や、制御アプリケーション開発用計算機の仮想化統合を行なってきた。

本稿では、計算機仮想化における可用性の向上や性能を確保した構成方法、運用方法について述べる。ここで対象とする範囲は、各部門業務用や研究プロジェクトの情報発信用などに用いる、我々が「共通情報計算機」と称している計算機の仮想化統合についてである。

２．高信頼仮想化インフラの構成

2.1 仮想化インフラの設計方針

可用性の向上を目指した計算機仮想化統合の概念を図1に示す。

我々が今回仮想化統合の対象とした計算機はSPring-8ホームページ用サーバーや代表メールサーバーとは異なり、コストの関係などから低価格サーバーやPC機材が多数用いられ、電源、ファン、ディスクなの冗長性が確保されていなかった。

そこで仮想化統合インフラの設計に当たっては、構成全体で冗長性を確保し、1台1台ではハードウエアの信頼性に十分なコストを掛けられなかった計算機が統合することによって高い可用性が得られるように検討を行なった。

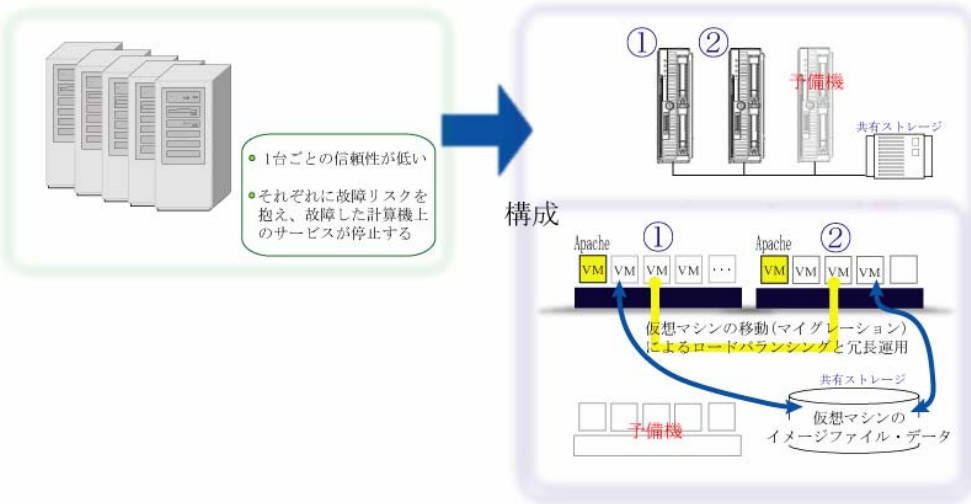

図１：可用性の向上を目指した計算機仮想化統合の概念図

図1に示すとおり、仮想化計算機(以下仮想ゲストと称す)を運用する仮想化サーバー計算機を複数台用意し、1台が運用不可能になった場合は残りのサーバー計算機に仮想ゲストを移動して運用継続可能な構成としている。

複数の仮想化サーバー計算機間で仮想ゲストを移動可能とするために、各サーバーからアクセス可能な共有ストレージも用意した。 共有ストレージの使用は仮想化において単一障害点に成り得るので、特に冗長性、可用性には注意を払った構成とした。

2.2 仮想化ソフトウエアの選定

統合対象計算機のアプリケーションは、主にWeb 、Webベースのアプリケーション、ftpサーバー、グループウエアサーバーなどであった。

そこで仮想化インフラのベースOSにLinux、仮想化ソフトウエアにはXenを採用して、仮想ゲストは原則としてパラバーチャライゼーションモードで動作させる事にした。パラバーチャライゼーションとは、仮想化用に修正したkernelが、CPUに対して直接特権命令を送る方式で、ハードウエアのレイヤーで行う仮想化(VM)ではオーバーヘッドが最小にとどまる。

統合対象の計算機では様々なOSが使用されていたが、Xenのパラバーチャライゼーションモードを主に利用した仮想化統合を行なうためにLinux OSに統一した。これはOSの種類を整理して、管理を容易にする目的も兼ねている。

ただし統合対象の一部には、アプリケーションの関係でWindowsサーバーOSから直ちに移行できない計算機も有った。Xenは、CPUの仮想化支援機能と組み合わせたフルバーチャライゼーションモード(完全仮想化)を使うことも出来る。Xen 3.2.0とIntel-VT(CPUはXeon L5335)の組み合わせをテストしたところ、完全仮想化でWindows OSが安定に動作することが確認できた。この結果を得て、OS移行が困難な一部Windows OSの対象は完全仮想化によって統合を行うことにした。

なお、XenやApacheなど必要なアプリケーションのバージョンや、仮想化サーバーのライセンスで無制限にパラバーチャルの仮想ゲストを使用できることから、SUSE Linux Enterprise Serverを採用している。

2.3 共通情報仮想化インフラの諸元

以上の考え方で構築した仮想化インフラの各諸元を表1に示す。

今回の仮想化統合では、仮想化対象のサーバーのメモリ量を平均1GB(512MB～2GB)、当初見通していた統合対象計算機を約30台と見積っていた。これをもとに、2.1で述べた複数サーバーによる冗長性の確保を必要条件とした。すなわち、仮想化サーバーが1台停止したときに、残りの仮想化サーバーで全ての仮想ゲストを運用することが出来るメモリ容量とCPU能力を算定して仕様を決定した。

統合対象計算機の必要メモリ合計は30GB であり、サーバー自身が必要とするメモリも考慮すると、仮想化サーバー3台構成では1台あたり16GB メモリ(1台停止時、2台で32GB)必要である。一方2台構成の場合は、1台停止時に残り稼動台数が1台だけになるので、通常運用時には16GBのメモリで賄えるところを32GBが必要になる。

サーバー1台停止時に3台構成の残り2台で運用しなければならない仮想ゲストの数は1台あたり15、2台構成の1台停止では、1台で30ゲスト全てを動作させなければならない。

統合対象の稼動状況から、15ゲストまでならCPUコア8で運用可能と判断したが、30ゲスト運用には合計16コアが必要である。本システムを構成した時点ではクアッドコアCPUが一般化したところであり、3台構成が前提の8コアCPU、16GBメモリのサーバーは筐体がコンパクトなハーフハイトブレードや2Uサーバーで構成が容易であった。

一方、2台構成で必要なCPUコア16を確保するにはフルハイトブレード、または4Uラックサイズの4CPUサーバーが必要で、価格も8コア/16GBサーバーの2台分以上となるので断念し、3台構成で行なうことにした。

項目	仕様	備考
筐体	ブレード型ハーフハイト2台、2Uラック型1台(ブレード1台を追加)	ブレードシャーシは他のシステムと共用。
CPU	ブレード型：Xeon L5335 2.0GHz x2 ラック型：E5410 2.33GHz x2 (追加ブレードL5420 2.5GHz x2)	1台あたり、合計8コア。 TDP=50w E5410は TDP=80w
メモリ	16GB FB-DIMM 2x2GBx4組	1台あたり。
ストレージ	NetApp FAS2020c	NAS、NFS
ディスク	VM OSイメージ用：SASディスク VMデータ用：SATAディスク	システムで利用可能なサイズ：SAS 1.5TB、SATA 3.8TB
冗長性	RAID-DP (RAID-6相当)＋スペアディスク SASノードとSATAノードで冗長クラスタを構成。	ノードあたりディスク2台故障と、片ノードのヘッド故障で運用可。

表1：仮想化インフラ主要諸元

ブレード型サーバーとラック型サーバーを混在させているのは、ブレードシステムのメンテナンス時にも必要な仮想ゲストを運用できるように考慮したためである。

構成上必須となった共有ストレージについては、

そのシステムダウンが全仮想システムの停止につながるので特に可用性について注意を払った。検討の結果、加速器制御でも実績の有るNetApp社製のNASシステムを選定した。ディスクのみならず、コントローラー部(NASヘッド)を含んだ冗長構成である。

また別途、ファイルバックアップシステムも構築して、万一のデータ損失も無い様に備えている。

３．仮想化インフラの運用

3.1 共有ストレージ上でのセキュリティ

共有ストレージでは、異なる部門グループや研究プロジェクトのデータをNFS共有するので、各仮想ゲストOS側から他のゲストOSのデータが絶対に見えないよう、NFSのexport設定には厳密なアクセス権限管理を行なっている。特にデータ領域は、仮想ゲストごとにボリュームツリーの細分化を行い、そのゲストだけがボリュームをmountできるように設定している。またツリーごとに、使用できるボリュームサイズの上限(ディスククォータ)も設定している。共用の大きなディスクボリュームの全サイズを仮想ゲストに見せてしまうと、仮想化統合を意識していない一般のユーザーが、自分が独占的に使える空きが大量に有ると誤解して、パブリックなストレージエリアを圧迫する可能性が有るので注意が必要である。

3.2 完全仮想方式ゲスト混在時の注意

本構成の実運用においては、アクセス頻度の低い計算機の統合が主体だったこともあり、Linuxゲストのみを10本以上起動しても安定した性能が得られている。しかしCPUの仮想化支援機能を利用した完全仮想化ゲスト(今回の場合はWindows Server)を起動する場合は、仮想サーバーのCPU利用率が上昇するので注意が必要である。

図2は、1台の仮想サーバー上における完全仮想(Windows)ゲストの数が、Linuxのパラバーチャルゲストで起動したWebサーバー(Apache)にどう影響するかをベンチマークソフトで測定したものである。

一番左の値が、評価用の計算機を仮想でないLinuxで起動して測定したもの、次がXen kernelで起動し、Webサーバーを立ち上げたLinuxゲストが1のみの場合の結果で有る。以後順に、Windowsゲストを1ずつ追加して、Linuxゲスト上のWebサーバーの性能を測定している。

Web性能の測定条件は全て同条件で、apache bench 2を用い、65kBのページに同時100リクエストを5万回繰り返し測定している。Windowsゲスト上では、ベンチマークソフトCrystal Markを用いて、CPU 100%ビジーの状態と、HDD性能測定によるI/OがビジーなCPU状態の2つを作り出している。

評価用の計算機は仮想化サーバー追加用に準備した追加ブレードサーバーで、表2の通りXeon L5420 2.5GHz x2。仮想環境でのCPU割り付け(VCPU)は各2である。

結果、評価用の計算機を仮想化しなかった場合と、Linux仮想のみの場合では1765.51 対 1710.78 Request/secと3%程度の性能低下に収まった。

さらにWindowsゲストを1ずつ追加していくと、4ゲスト追加で顕著な性能低下になり、CPU負荷とI/O負荷の場合でそれぞれ23.4%、37.5%の影響が有った。またWindowsゲストに対してCPUよりI/O負荷が掛かっている状態のほうが、より仮想サーバーと他のゲストへの影響が大きい。

実運用では、Windowsゲスト4VM、Linuxゲスト10VMでメモリを使い切る運用を行なっても、統合前のローコストサーバーより速いので、問題になっていない。しかしより高い性能を提供するためには、完全仮想ゲストのI/Oをパラバーチャライズ化する有償ドライバでサーバー負荷を軽減するか、少なくとも運用面において完全仮想ゲストを各サーバーに分散すべきである。

参考文献

[1] T. Ohata, M. Kodera, M. Ishii, M. Takeuchi and T. Fukui, “VIRTUALIZATION OF OPERATOR CONSOLES ON BEAMLINE CONTROL SYSTEM”, Proceedings of the ICALEPCS07, Knoxville, Tennessee, USA, URL:http://accelconf.web.cern.ch/accelconf/ica07/PAPERS/WPPA18.PDF

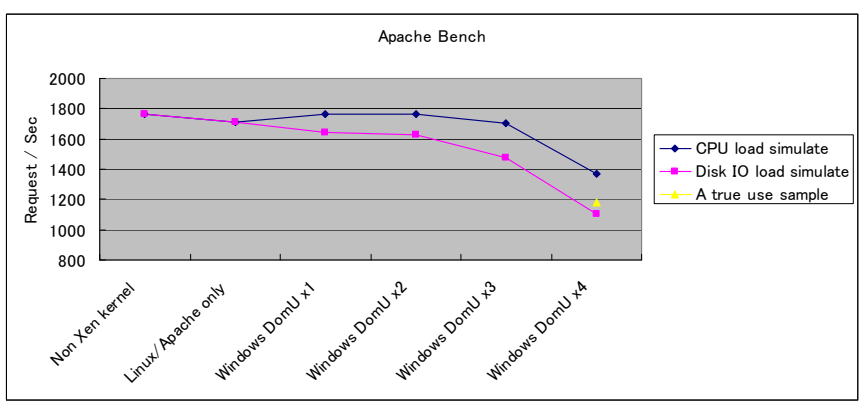

図２：完全仮想ゲストの集中による、他のゲストへの性能圧迫