サイエンティフィック・システム研究会
2007年度システム技術分科会第2回会合資料　　2008年1月30日

溢れるサーバを減らすために
？ 農林水産研究基盤への仮想化適用事例 ？

農林水産省農林水産技術会議事務局
筑波事務所情報システム課
（農林水産研究計算センター）
江　口　　尚
eguchi@maffin.ad.jp

［アブストラクト］
当所では、約１万人の利用者に科学技術計算、Internetサービス、DB、電子図書館などの多様な研究支援のサービスを行っている。システムの安定運用には、機能別のサーバ構築と冗長化が不可欠であるが、溢れるようなサーバ台数では管理負荷を増大させるばかりである。
一方で、近年のIA[1]サーバの処理能力が向上している割に負荷が少なく処理能力に余裕がでてきている。そのため当所では、早い段階から仮想化によるサーバ集約に着目し、実機、仮想化の両方を組み合わせたシステム運用を行ってきた。当所の幾つかの事例をもとに仮想化の利点と課題について本報告で紹介する。

［キーワード］
仮想化、VM、IAサーバ、効率化

1. はじめに

農林水産技術会議事務局筑波事務所は、農林水産研究計算センター／研究情報センター(以下、計算センターという)を運営し、全国80箇所以上に点在する農林水産研究に関連する農林水産省並びに管轄する独立行政法人研究機関をネットワーク(MAFFIN)で結び、科学技術計算、文献や数値等の各種データベース、インターネットサービス、電子図書館機能など、研究に必要な各種情報基盤としてのサービスを提供している。

各サービスでは、システム負荷や冗長性、セキュリティ等を考慮して複数のサーバを利用しており、近年の提供サービスの多様化に伴い、管理するサーバが増加傾向にある。そのため、設置スペースの確保、電力消費の増大、サーバ保守や人的な管理負荷の増大が問題となっている。これらの問題を解決する方法として、少ないハードウェア資源を有効に活用する仮想化技術に焦点をあてて、この数年の計算センターの取り組みについて報告する。

2. 仮想化への初期の取り組み

計算センターでは大型計算機を完全撤去して11年が経過しているが、その当時の大型計算機の運用では、OSのバージョンアップ時の検証等にVM[2]の利用は欠かすことはできなかった。当時の大型計算機(ACOS930，OS:ACOS-6/MVXII)には、VM運用のためのシステム操作卓が通常のシステム操作卓とは別に用意されており、OSへの新規機能追加の挙動確認などに1993年頃まで利用していたことを記憶している。(その後、UNIXサーバを中心としたシステムに転換し、大型計算機は文献DB用途や移行処理用に限定して利用していたためVMは使用していない。)

---

[1] IA : Intel Architecture
[2] VM : Virtual Machine 仮想マシン

VMはこの計算機だけでなく、どのメーカーの大型計算機でも、ごく一般的に備えていた機能であった。そのため大型計算機の運用者からみると、システム仮想化は特に違和感なく利用していた一般的な技術との認識であった。
その後のUNIX系サーバではVM機能が実装されておらず、また特段必要としなかったが、PCの性能が飛躍的に向上してきたと、CPU自身での仮想化は無いものVMwareにより仮想化が身近に利用できるようになったことに興味を持った。Web閲覧、ワープロやメール利用などでは、CPUの処理に十分な余力があることから、2001年の部署内のPCの購入の際に職員の日常利用や実験環境の構築をかねて以下の試みを行った。

・比較的高性能なPCを調達し、職員がLinux関連の技術の修得を行うために必要な環境を構築する。
・複数のLinux-PCを相互に接続し、通信、認証等の相互連携の評価を行う。
・夜間等に各Linux-PCを利用して並列計算を行うなどの評価を行う。
・一つのハードウェア上で複数のOSの評価を行う環境を構築する。
・Windowsでなければ利用できないアプリケーションはVM上で運用する。

（動作環境）
・PC　Xeon 2GHz　2cpu、主記憶1GB　5式
・TurboLinux ＋ VMware2.0
・ゲストOSとしてWindows2000 Professional ＋ MS Office
その他のLinuxディストリビューション、BSD等

しかし仮想I/Oの限界から、ハードウェアのグラフィック性能を全く生かし切れず、ゲストOS上でのPowerPointの矩形描画などが遅い、Wordではテキスト入力に問題はないものの画面スクロールや各種機能が緩慢な動きとなるなど惨憺たる結果であった。結果的に導入したPCでのVM利用をやめて、ごく一般的なWindowやLinux-PCと化したが、この失敗から、

・仮想化されたI/Oのチューニングが向上し、ハードウェアの性能を引き出せるようになれば利用用途が開けてくる。
・CPUの処理能力が今後飛躍的に向上し、通常用途ではCPU資源に余力が生じてくるのは間違いなく、独立した複数のサーバを連携させてサービスを提供するシステムでは、システム負荷が少ないサーバをVMで稼働させてハードウェアを集約することができる可能性がある。
・描画処理を伴わない比較的負荷が軽い処理や、OS、アプリケーションの機能評価には現状でも十分に利用できる。
・個人環境においても、一人で手足の数より多いPCを持たなくても、主として利用するOS以外はVMで対応することにより管理の手間を大幅に軽減できる。

等のPC上のVMの将来性を感じることができた。そのため、今後のシステム設計の際には、この失敗に臆することなく、時期をみて導入していきたいと考えた。
その後、製品紹介のために来訪してくるエンジニアが、ノートPCにVMwareを搭載して得意げに複数のOS上のアプリケーションを連携させたデモを行う光景を目にすることが多くなり、VMware自体も性能が向上してきたこと、VMが徐々に普及がすすんでいることを実感するようになった。

3. 2004年システム更新に向けて

計算センターでは4年に一度のペースでシステム更新を行っており、オリンピックの開催年がちょうどシステム更新時期にあたる。システム設計はその1年前から開始し、科学技術計算系、ネットワークサービス系、DB系、図書館系に分けて設計を行っている。ここでは、1万人の利用者に対して、メールサービスを含むインターネットサービス、共通認証基盤、参加研究機関の公式Webサーバ機能を提供するネットワークサービス系の設計について述べる。

その前の2000年システムでは、PRIMPOWER400(Solaris)22式とSIMS(Sun Internet Mail Server)を中心としたメールシステムを運用してきたが、日々進化するSpam、DoS[3]等の対策に柔軟に対処することが専用システムでは難しくなっていた。そのため、フロントにSpamAssassinの運用などのためにLinuxサーバを複数設置するなどの対応を取ったが、それに合わせてサーバ数が徐々に増加していった。そのため2004年システムでは、インターネットの進展に合わせて柔軟に構成できるように以下の考えを基に設計を行った。

- オープンソースをベースにしたシステムとし、今後のインターネットセキュリティやDoS等の攻撃に対して、柔軟に組み替え、機能追加等が行えるシステム構造とする。
- 各機能の冗長化を高め、サービスダウンを極力排除する。
- 多機能サーバは障害管理、セキュリティ管理が煩雑になることから、機能毎の複数サーバに分離し、単体毎の機能向上が図れるような構造とし、障害時においても原因の迅速な特定、切り分けが容易な構成とする。
- メール配送を外部→内部向け、内部→外部、再送専用、内部利用者のPOP･SMTPアクセス等に機能分類とサーバ割り当てを行い、DoS等の影響が通常のメール配送に支障が出ない構造とする。
- Linuxを搭載したIAサーバ標準構成としてコスト削減を図る。
- 高価な市販アプリケーションに経費をかけない分、インターネットの進展に柔軟に対応するための運用管理や利用者のニーズにマッチしたシステム開発に経費を振り向けることとし、研究需要に最適化したシステムを構築する。

これまでのメール流量やシステム負荷から大手プロバイダ並の処理能力が必要であることが判明しており、各種機能の提供や冗長化を図ったサーバの配置について構成図を描いてみると、必要とするサーバの数は溢れんばかりの量となった。

ネットワークサービス系では、自家用電気工作物の点検時においても、無停止で電源を供給しサービスを継続できる構造となっている。そのCVCF[4]容量は30KVAであり、システムは、この範囲内で設計することが求められている。サーバが増加すると電源だけでなく、機器点数の増加は機器障害の増加につながること、そしてなにより運用管理を担うSEにとっても負担が増すことから、物理的なサーバ数を押さえるシステム設計が鍵となった。

4. 仮想化の検討

多数のサーバを導入する際は、各サーバの負荷にバラツキをあることを承知の上で、管理運営やその後の柔軟な組み替えが行いやすいように、多少余力のある機種に統一して導入を行うことが一般的である。このとき、全体のサーバの余力を各サービスの負荷に応じて適切に配分できる機能があれば、サーバ運用を最適な台数を減らすことが可能となる。また障害時に代替え機能があり、負荷が増大したときにサーバを追加することで、全体の処理能力を向上できる機能があると理想である。

---

[3] DoS : Denial of Services

[4] CVCF : Constant Voltage Constant Frequency (定電圧定周波数装置) :無停電電源装置

そこで、システムの一部に仮想化技術を取り入れることで、このような考えに近い機能を実現することができるのではないかと考えた。負荷が増大している仮想サーバでは、同時に運用する仮想サーバ数を調整したり、実サーバに移行するなど、運用期間中に柔軟な構成変更が可能である。また高負荷となる機能の提供は、仮想サーバを使用することなく最初から実サーバ上で運用すればよい。これらの仮想化機能の導入についての検討と平行して、富士通からも同様の提案があり、双方は一致した考えを持っていたことになる。

これらの検討と以前の教訓から、仮想サーバ、実サーバを組み合わせて運用することとし、サーバの負荷が増大した場合、同時稼働する仮想サーバの個数見直しや仮想化にこだわることなく、以下のように実サーバに割り当てを行う方針とした。

(1)実サーバ1 2CPU　主記憶6GB	CPU、メモリアクセスの負荷が高いサーバ、利用者の負荷が集中するPOP、IMAP、SMTP等の用途
(2)実サーバ2 2CPU　主記憶2GB	CPU、メモリアクセスの負荷が高いウィルス検査用サーバ、迷惑メール判定機能を有したメール中継等の用途
(3)実サーバ3 1CPU　主記憶1GB	DNS等の負荷が少なくても単体で安定した動作が必要な機能、負荷の少ないメール中継、利用者管理DB等の用途
(4)仮想サーバ(VM) VM用サーバ5式上で運用 基本性能は実サーバ1と同じ。	比較的負荷が少ない機能やWebサーバの集約の用途

5. 導入

導入した機種構成は以下のとおりである。

実サーバ　46式	Fujitsu Primagy RX300 2CPU　Xeon(HT)3.2GHz　主記憶6GB RX200 1CPU、2CPU　Xeon(HT)3.2GHz　主記憶1? 6GB
VM用サーバ　5式	Fujitsu Primagy RX300 2CPU　Xeon(HT)3.2GHz　主記憶6GB VMware ESX Server 2　(現在の運用は2.1.2)
その他	NFSサーバ　Fujitsu NR1000 F500 クラスタ構成（合計10TB） ストレージ　Fujitsu ETERNUS3000　他 負荷分散装置　BigIP 5110SC ネットワーク機器　等

OS	実サーバ及びVM上のゲストともにRedHat Eenterprise Linux 3.0
VM上の 主なソフトウェア	Apache-2.0.61, postfix-2.1.5, fml-4.0,　MHonArc-2.6.10, samba-3.0.9-1.3E.13.2, proftpd-1.2.10-1, qmail-1.03, ucspi-tcp-0.88, stone-2.2c, jakarta-tomcat-5.0.28 F-Secure 有償製品として fsavsamba-4.63、virusgw-2.3.1　等

(1)VM用サーバ1、2　(図1)

研究機関の公的情報を公開するWebサーバは、BSEや鳥インフルエンザ等の問題が発生した時などに多数のアクセスが集中するものの、通常ではそれほどの負荷は発生していない。ただし、研究機関ごとの柔軟な運用を行うために、機関数(大別して10組織に分類)と冗長化を考慮したサーバの台数(20式)を必要とする。これを2つのVM用サーバに分散し、相互のサーバに待機系を設定することで冗長化と大幅な実サーバの削減を実現している。

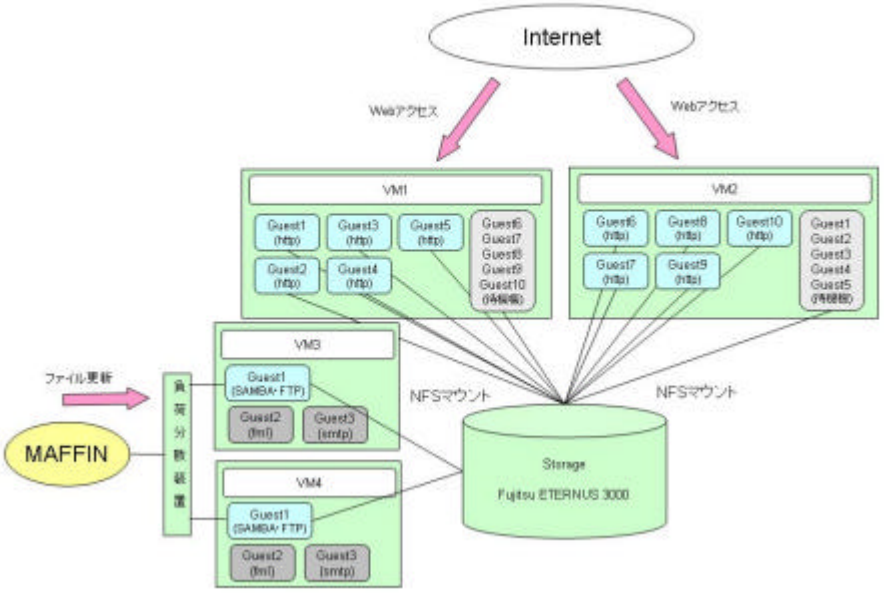

図１

(2)VM用サーバ3、4：(図2)

　メーリングリストは2,000以上稼働しており、利用者相互の情報共有に欠かせないものとなっている。また、利用者向けファイルサーバも提供しており、ftp、SMBのアクセスを提供している。これらの処理負荷は軽いものの安定したサービスが求められるサービスである。これも同様に2つのVM用サーバに分散し、負荷分散装置を利用した冗長化と負荷配分を行い、実サーバの削減を行っている。

(3)VM用サーバ5：(図3)

　船舶とのメール送受信は、専用の配送システムを運用する必要があるが、定期的に衛星回線を通じて交信を行うため、それ以外の時間では負荷は極めて少ない。同様に比較的アクセス負荷の少ない携帯電話向けWebメールサービス、定期的にログを収集サーバ、SEの作業用、評価用途などを1つのVM用サーバに取りまとめている。

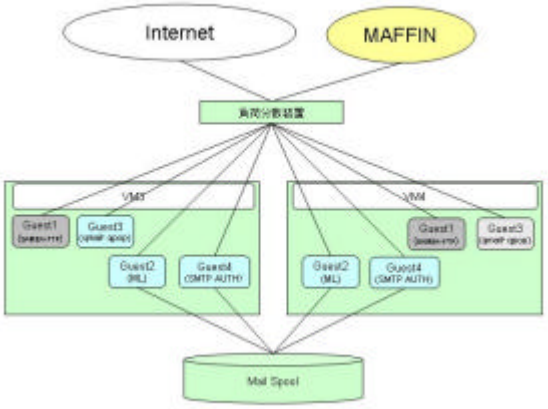

図2

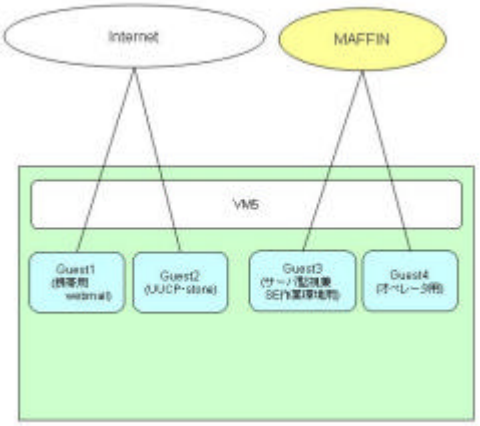

図3

VMotionは評価用に導入しているが実利用は行っていない。2004年当時では、SAN[5]の構築にコストがかかることと信頼性に不安があったためである。(2000年システムではSANの同期に障害が発生し、1週間徹夜の復旧を経験している。)

6. 効果

小型サーバ(1U, 1CPU)を20台程度削減できているが、高性能サーバを5式とVMwareを導入したことにから、4年間の保守料等を含めてみると導入経費は未導入に比べても殆ど差はない。しかし、以下の4つの利点は、それ以上の価値を生みだしている。

第1に電力について、4年間で電気代が計算上200万円程度節減となっている。これは微々たるものであるが、既存の30KVAのCVCF容量に収まり、電力設備の増強が不要となったことは非常に大きな利点である。

第2にシステムの柔軟性にある。問題事項の解決や新たな機能追加を行う際に検証のための機器設置やネットワークを準備しなくても、速やかに試験環境の構築が可能となる。別途サーバが必要な時に機器の調達期間を気にせずに迅速に対応できるため、「ちょっとシステムを立ち上げて実験してみよう」という気にさせることは、日常のシステム改善にとって極めて大きな利点である。

第3に新旧システムの移行が容易な点がある。機器の定期更新にあわせて、各研究機関のWebサーバの移行を迅速に行わなければならないが、新OS等やライブラリとCGIの整合性の検証等の様々な問題が発生して、新機器へ移行がなかなか進まないことが多い。このとき、とりあえずVMごと新機器に移行し、徐々にコンテンツをVM上の新OSに切り替えていけば、旧機器の撤去は容易に行えるはずである。

第4に運用管理の負荷軽減がある。サーバ等の機器構成点数の削減は障害発生を減少させることとなる。またVMwareはバックアップ、リストア作業が極めて容易になることが上げられる。

逆に問題点として、先進的な導入時に往々にしてあることだが、ハードウェアが正式にVMware対応となっていない点がある。そのため、VMwareのバージョンアップ等の対応を難しくしている。これについては次のシステム更新前までにはVMwareがサポートするSCSIインターフェースに変更するなどの機器側の対応を行い、移行が円滑に行えるように準備する予定である。

これ以外にも仮想I/Oであるため実サーバに比べて入出力が多発するような処理がどうしても遅くなることが上げられる。SEからは、システムバージョンアップで一度にコンパイルを実施すると極端にレスポンスが悪くなるとの報告が上がっている。
また当然のことながらハードウェア障害時に一度に複数のVMが停止するため、複数のVM用サーバの筐体にて冗長化を図らなければならない点である。しかしこれらは、実サーバとの棲み分けや冗長化などの対応策を講じておけばそれほど問題とならない。

2004年から3年間の運用を続けているが、VMwareが起因する障害は皆無であり、信頼できるシステムといえよう。計算センターでは、本報告で紹介したシステム以外にも個別研究プロジェクトを対象とした貸しサーバ機能を提供しており、今年度の機器増設においてもVMwareを積極的に利用する形で構築を進めている。

7. 今後の期待

システムの仮想化は、様々な分野にも適用できるのではないかと期待している。以下は科学技術計算系システムへの適用した場合の例であり、今後様々な検討を行なうべき事項であることを了解いただいた上で、幾つかの可能性を示してみたい。

---

[5] SAN : Storage Area Network

・初心者や負荷の少ない利用者に作業用、ログインノードとしてVMを一人に1台を割り当てると、実際にはCPU資源を共有しているにも関わらず、あたかも自分一人で占有しているように見える。初心者ではあるが研究者として他人から何を行っているのか実行状況を見られたくないという要求を満たすことが可能となる。小規模演算やスクリプト処理程度ならVM上で十分処理可能と考える。
・計算パワーを必要とするようになれば、VMから実サーバに利用環境を移行させればよい。
・クラスタノードへの計算処理の投入は、VM、実サーバの双方から同等に実行できる環境を提供する。
・古いカーネルやライブラリに依存するアプリケーションの運用についても、そのOSをVM上で稼働させることで既存機器からの移行が容易となる。ハードウェア能力が飛躍的に向上することから、VM上でもある程度の性能向上を期待できるのではないだろうか。これにより既存機器の撤去が可能となり、その分管理コストが低減する。

近年、スパコン等の科学技術計算において、利用者がパワーユーザに偏る傾向にある。これから利用を試みる利用者に対して、他の利用者に気兼ねなく練習・評価が行える環境作りは利用者の拡大のため必要と考える。また、カーネルやライブラリ変更等の特定用途向けの要望もあるであろう。これらの利用者に対して柔軟に対応できれば、新規利用者の拡大につながるのではないだろうか。VM上で資源配分管理は、利用者のレベルや用途に合わせたCPU、メモリ配分が行えることから、バッチジョブ管理とは異なる考え方での資源配分が可能となる。

利用者すべてが、ハードウェア限界の性能を必要としているのではない。VMによるオーバーヘッドに目を瞑っても利用者層が広がれば、計算機を利用した成果の拡大につながり、計算センターの価値を高めることになる。もちろんVMと実サーバの混在と最適な資源配分が必要であるが、システムインテグレーションを行うベンダー側からも、計算処理速度だけではない、発想の転換をした大胆な提案を期待したいところである。

また一方で、実サーバ上でなければ、アプリケーションやドライバ等の動作保証、ソフトウェア保守にソフトウェアベンダーが対応しない場合がある。今後、仮想化を広く普及させていく上でも、各ベンダーに改善を望みたい事項である。

8. 最後に

現在、計算センターでは、2008年システム調達に向けて設計の準備を進めており、これまでの仮想化技術の運用実績をもとに、さらに利用用途を拡大していきたいと考えている。マルチコア化で処理性能が飛躍的に向上したサーバを有効活用するためには、仮想化は欠かすことのできない技術であること、VT[6]などのCPU上での仮想化への対応がなされてきたことから、今後、サーバ仮想化を利用したシステム構築が広く普及してくるものと確信している。本報告内容がシステム設計の検討の際に一助となれば幸いである。

謝辞

本報告に使用した図は、日常のシステムの運用に利用しているものをベースに作成しました。協力いただいた飯島氏ほか、富士通エンジニアの皆様にこの場を借りて御礼を申し上げます。

[6] VT：Virtualization Technology