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

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

組織紹介　農林水産技術会議事務局筑波事務所

農林水産研究にかかる各種共同利用種施設、
サービスを提供

全国規模

農林水産研究計算センター

科学技術系計算機利用、ネットワーク（MAFFIN）の提供

農林水産研究情報センター

各種データベース、電子図書館機能の提供
国立国会図書館支部分館

農林交流センター

産学官連携、セミナー、ワークショップ開催等

筑波農林研究団地

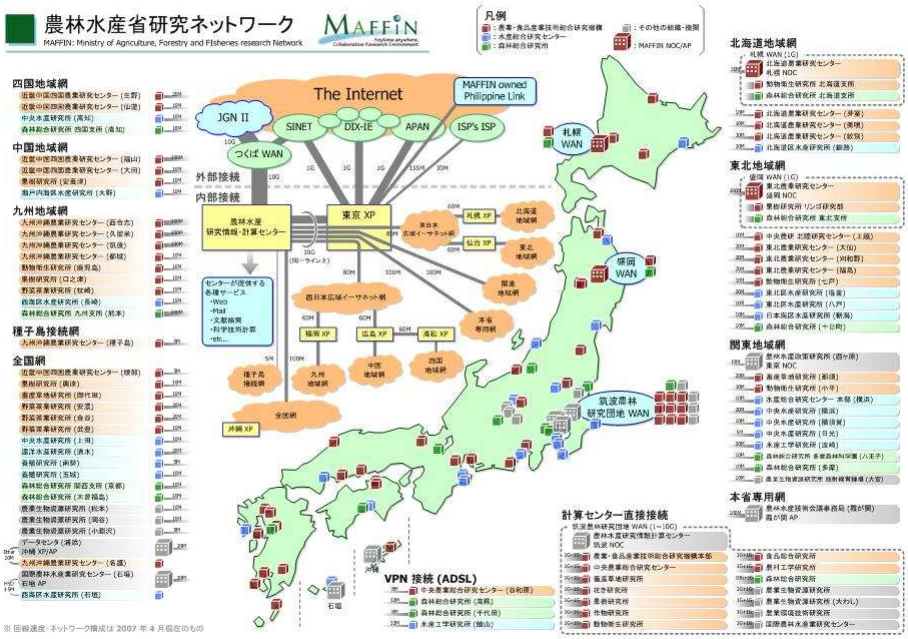
特別高圧受変電設備、エネルギーセンター、実験排水処理施設、各種共同利用施設
診療所、会議室、ホール、食堂、体育館等の福利厚生施設、研修施設

組織紹介　農林水産研究計算センター／情報センター

MAFFIN
Anytime anywhere, Collaborative Research Environment

科学技術計算
文献データベース
地球観測、気象、農林水産統計等各種データベース
インターネットサービス
電子図書館機能
その他研究に必要な各種情報基盤　を提供

- 農林水産研究に関わる国、独立行政法人研究機関が全国80カ所以上に点在
  - 農林水産分野の研究領域は非常に幅広く、気候、地理的な条件から全国各地で研究を行っている
- 全国の各研究所の情報格差の解消
  - 研究者が数人の小さな研究支所であっても、大規模な研究所と同様の情報資源を必要とする。
- 全国を広域イーサネット網で接続
  - 筑波農林研究団地内は、Gigabitイーサネット接続（2Gbps－10Gbps）
  - 光ファイバー網を平成7年に整備



センター運営の課題

- システムの複雑化
- システム負荷や冗長性、セキュリティへの対応等
- 提供サービスの多様化

☆　管理するサーバが増加していく

センター運営の課題

- 設置スペース
- 電力消費の増大
- サーバ保守や人的な管理負荷の増大

☆　問題解決のため、さまざまな技術検討

仮想化への初期の取り組み

VMといえば

- レガシーな世界では一般的な技術
  - 1993年頃までVM（仮想マシン）を利用
  - 当時使用の大型計算機（ACOS930, OS:ACOS-6/MVXII）
  - VM運用のためのシステム操作卓、入出力装置
  - 1993年頃まで新規機能追加の挙動、確認利用OSのバージョンアップ時の検証等にVM（仮想マシン）の利用
- ただし、1996年に大型計算機を完全撤去
  - »（写真は最終モデルACOS3700）

☆　ある年代以上の管理者にとってVMに親近感？

PCへのVM環境導入（失敗事例）

2001年の部署内のPC購入

目的
- Linuxを含めた様々な関連技術修得の環境構築
- PCを相互に接続し、通信、認証等の相互連携を評価
- 空き時間利用　Linux-PCによる並列計算

☆一つのハードウェア上で複数のOSの評価環境を構築
Windowsでなければ利用できないアプリケーションはVM上で

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

性能が出ない

- 仮想I/Oの限界
- ハードウェアのグラフィック性能を全く生かし切れない
- ゲストOS上でのPowerPointの矩形描画などが遅い
- 画面スクロールや各種機能が緩慢な動き

結果
- 導入したPCでのVM利用とりやめ
- ごく一般的なWindowsやLinux-PCと化した

しかし、手応えが

2001年当時でも十分に利用可能なもの
- 描画処理を伴わない比較的負荷が軽い処理
- OS、アプリケーションの機能評価
- 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で対応すれば管理の手間を軽減

今後
- 仮想化されたI/O性能が向上、CPUの処理能力が飛躍的に向上すれば通常用途ではCPUやI/O資源に余力が生じる
- システム負荷が少ないサーバをVMで集約が可能に

☆　VMの将来性を感じる

計算センターシステム更新

2004年システム更新

4年に一度システム更新

- オリンピックの開催年がちょうどシステム更新時期
- システム設計はその1年前から開始
- 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

ネットワークサービス系

- 1万人の利用者
- メールサービスを含むインターネットサービス
- 情報共有サービス
- 共通認証基盤（シングルサインオン）
- 参加研究機関の公式Webサーバ機能　等々

旧2000年システム

- PRIMPOWER400（Solaris）22式
- SIMS（Sun Internet Mail Server ）
- 日々進化するSpam、DOS等の対策が専用システムでは難しくなっていた。
  - SpamAssassinの運用など
  - Linuxサーバを複数増設

    ⇨ サーバ数が徐々に増加

☆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変更、機能追加が可能な構造を目指す

2004年システム基本方針

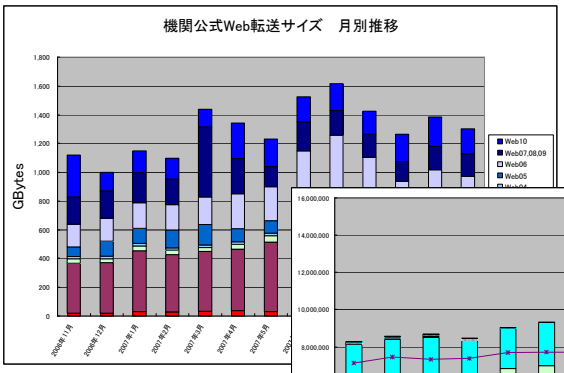
- オープンソースをベース
- インターネットセキュリティの進展、DoS等の攻撃に対応
  - 柔軟に組み替え、機能追加等が行えるシステム構造
- 各機能の冗長化
- 機能毎に複数サーバに分離
  - 単体毎の機能向上
  - 障害時の原因の迅速な特定、切り分けが容易な構成
- DoS等の影響が通常のメール配送に支障が出ない構造
  - メール配送を外部→内部向け、内部→外部、再送専用
  - 内部利用者のPOP・SMTPアクセス等に機能分類とサーバ割り当て

予算の配分

- 人件費に重点配分
  「箱はケチっても、人はケチるな」
  - LinuxとIAサーバの標準構成でコスト削減
  - 高価な市販アプリケーションは極力導入しない
- 運用管理経費を増額
  - インターネットの進展と研究需要に柔軟に対応するために

システム規模

- 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

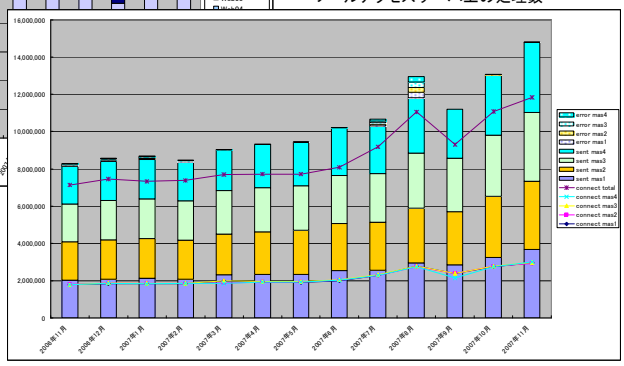


機関公式Webサーバのデータアクセス量

メール流量　1470万通／月
Webアクセス　1.3TB／月

（但し2007年11月分統計による）

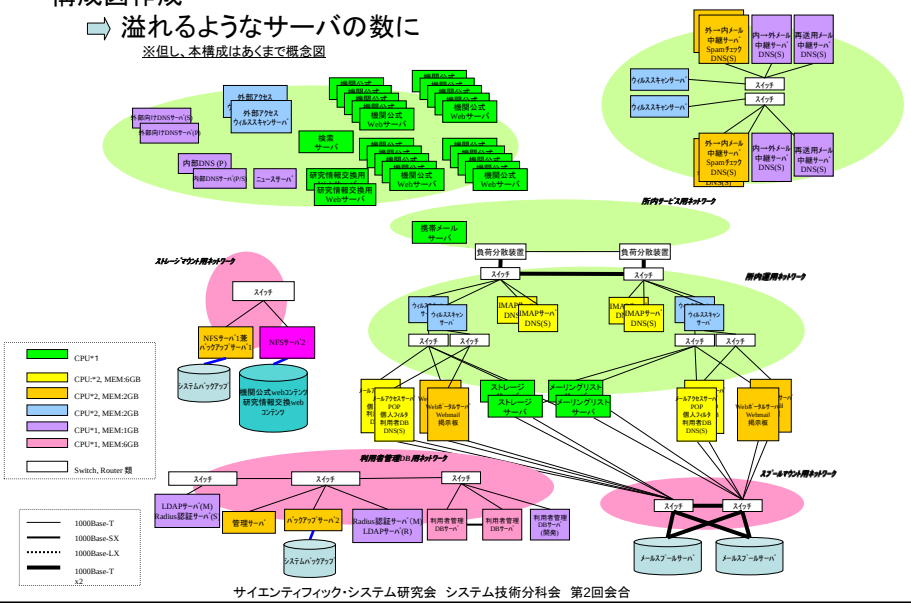
メールアクセスサーバ上の処理数



システム規模

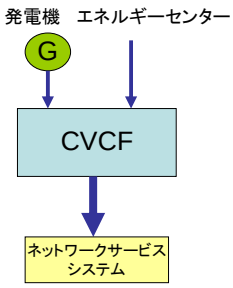
- 構成図作成
  - ⇨ 溢れるようなサーバの数に

※但し、本構成はあくまで概念図



制限事項

- 完全無停止のCVCF容量は30KVA
  - エネルギーセンターから供給と発電機により自家用電気工作物の点検時においても無停止運用



- 機器点数の増加は機器障害増加に
  - SEの管理負担増大

☆　物理的なサーバ数を押さえるシステム設計が鍵

サーバの要件

多数のサーバを導入

- 各サーバの負荷にバラツキ
- 管理運営や組み替えを考慮し、多少余力のある機種に統一

理想

- 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
- 障害時の代替え機能
- 負荷増大に応じてサーバ追加により全体の処理能力を向上

☆ サーバ運用を最適な台数を減らすことが可能

仮想化技術の検討

- 負荷が増大に対して、同時運用の仮想サーバ数、CPU負荷配分を調整
- 高負荷機能は、実サーバ上で運用

⇩

☆ 以前の教訓から、仮想サーバ、実サーバを組み合わせた運用に

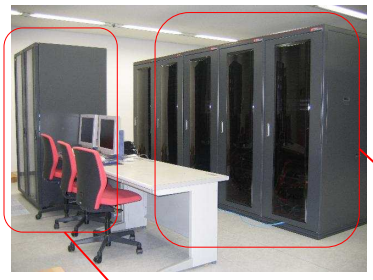
サーバ基本構成

- 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
- 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
- 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
- 仮想サーバ （VM用サーバ上に構築）
  - 比較的負荷が少ない機能やWebサーバの集約の用途

導入機器

- 実サーバ　46式
  - Fujitsu Primagy RX300 2CPU　Xeon(HT)3.2GHz　主記憶6GB
  - 同　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、postfix、fml、MHonArc、samba、proftpd、qmail 等
  - F-Secure有償製品として　fsavsamba、virusgw　等

導入機器



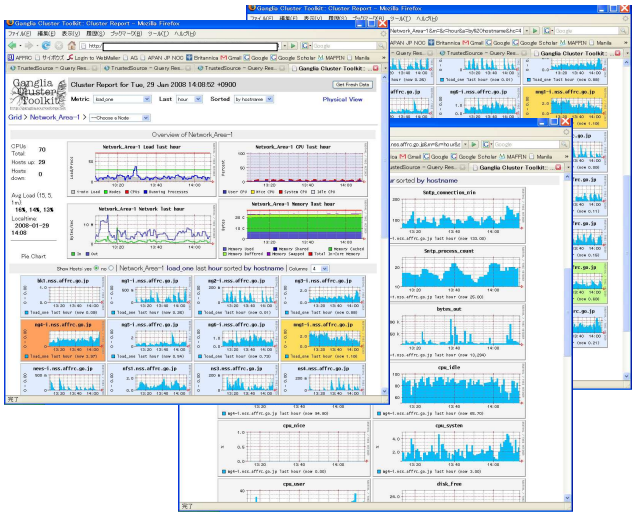
ネットワークサービスシステム

Fujitsu NR1000 F500

Fujitsu Primagy RX300/RX200

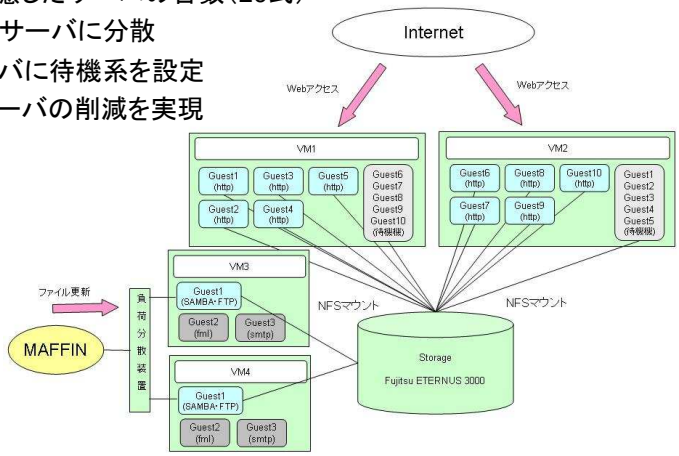
機器管理

Ganglialによる負荷監視



VM用サーバ1、2

- 研究機関の公式Webサーバ
  - 機関数（大別して10組織に分類）
  - 冗長化を考慮したサーバの台数（20式）
  - 2つのVM用サーバに分散
  - 相互のサーバに待機系を設定
  - 大幅な実サーバの削減を実現



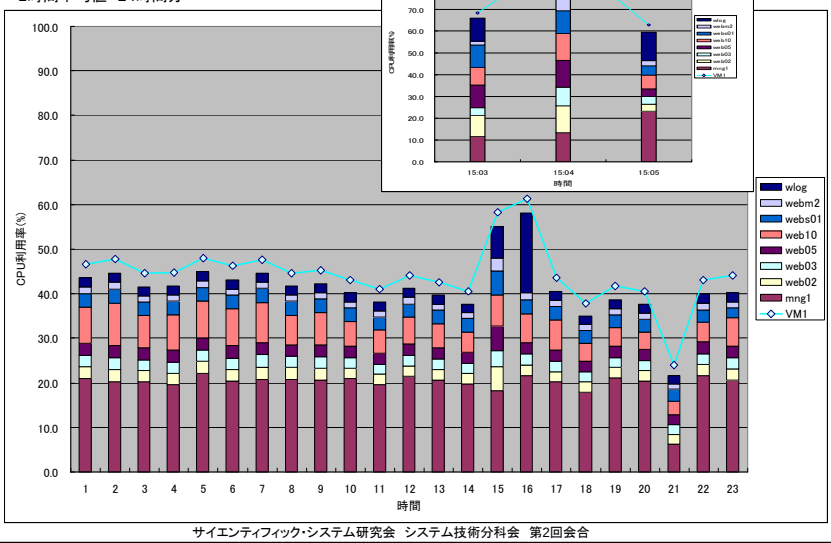
利用統計：VM1

サンプリング日時：2008/01/16

設定：全Guest OS共通
CPU MAX利用率：160%（最大値：200）
CPU負荷バランシング：均等

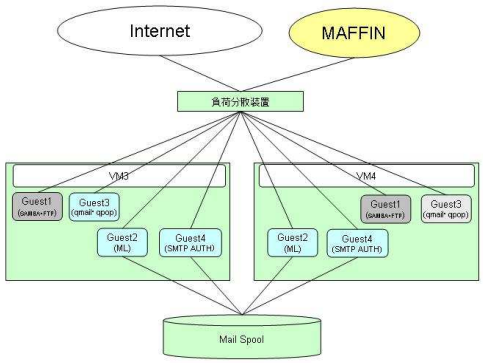
1時間平均値　24時間分

15時台のピーク部分を拡大



VM用サーバ3、4

- メーリングリスト運用
  - 2,000以上稼働
- 利用者向けファイルサーバ
  - ftp、SMBの提供
- 2つのVM用サーバに分散
- 負荷分散装置を利用した冗長化と負荷配分

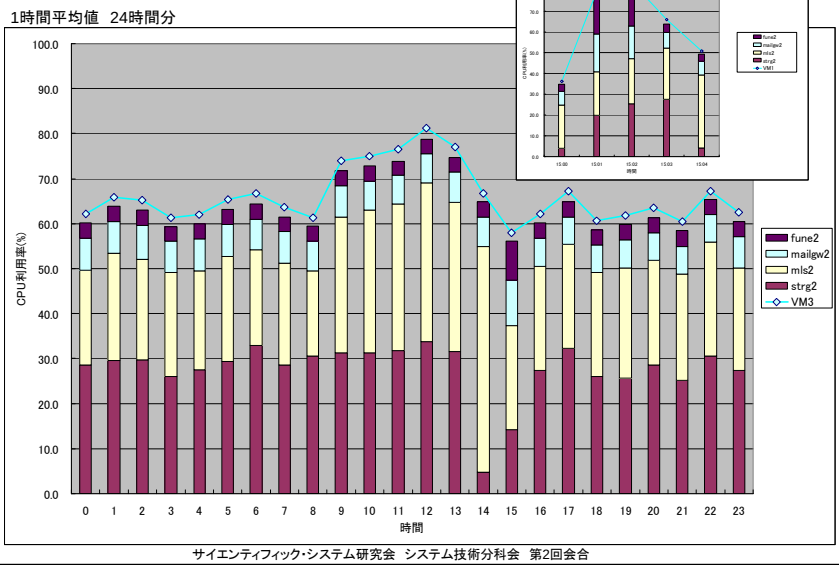


利用統計：VM3

サンプリング日時：2008/01/16

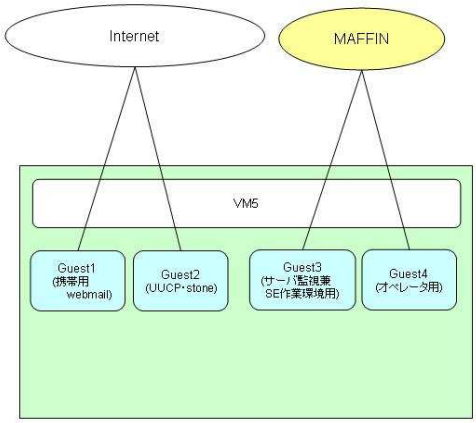
設定：全Guest OS共通
CPU MAX利用率：160%（最大値：200）
CPU負荷バランシング：均等

1時間平均値　24時間分



VM用サーバ5

- 負荷が少ないサーバの集約
  - 船舶とのメール送受信
  - 携帯電話向けWebメールサービス
  - ログ収集サーバ
  - SEの作業用、評価用途など



制限事項

- VMotionは使用せず
  - 2004年当時SAN（Storage Area Network ）の構築コスト、信頼性に不安
  - SANサポート機器が一部のメーカーに限られている

まとめ

効果　　導入時のコスト

- サーバ数削減
  - 小型サーバ（1U、1CPU）を20台程度削減
- 高性能サーバを5式とVMwareを導入
- 4年間の保守料等

☆　VM導入、未導入は導入時のコストに、ほとんど差がない

効果　運用上の利点（1）

MAFFIN Anytime anywhere, Collaborative Research Environment

電力

- 4年間で電気代が計算上200万円程度節減
- 既存30KVAのCVCF容量に収まり電力設備の増強が不要

システムの柔軟性

- 問題事項の解決や新機能追加の検証
- 新たに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

☆「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

効果　運用上の利点（2）

MAFFIN Anytime anywhere, Collaborative Research Environment

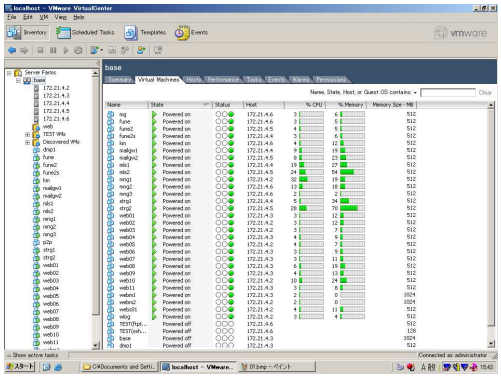
システム更新

- 新旧システムの移行が容易
  - 機器の更新時に新OS、ライブラリ、CGIの整合性の検証等の様々な問題が発生
  - 新サーバへ移行がなかなか進まない
- とりあえずVMごと新機器に移行
- 徐々にコンテンツをVM上の新OSに切り替え
- 旧機器の撤去が容易に

効果　　運用上の利点(3)

<u>運用管理の負荷軽減</u>

- サーバ等の機器構成点数の削減
- 障害発生を減少
- VMwareはバックアップ、リストア作業が極めて容易
  ※ただし、現VMwareバージョンでは享受できていない



問題点

- ハードウェアが正式にVMware対応していない
  - VMwareのバージョンアップ等の対応を難しい
  - VMwareがサポートするSCSIインターフェースに変更を検討
- 実サーバに比べて入出力が多発するような処理が遅い
- ハードウェア障害時に一度に複数のVMが一度に停止
  - 実サーバとの棲み分けや冗長化などで対処可能

安定性

- 2004年から3年間運用
- Vmwareに起因する障害は皆無

⇩

☆　信頼できるシステム

※　個別研究プロジェクトを対象とした貸しサーバ
　　機能等にWMwareを積極的に利用を拡大予定

今後の期待

システムの仮想化は、様々な分野にも適用でないか

例えば科学技術計算分野

- 初心者や負荷の少ない利用者用途
- 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
- 自分一人で占有しているように見える
- 実際にはCPU資源を共有

⇩

※　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、わがままな要求に対応

- 計算パワーが必要となれば、VMから実サーバに利用環境を移行
- 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
- 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
  - 既存機器の撤去が可能
  - 管理コスト低減
- VM上でバッチジョブ管理とは異なる考え方での資源配分

今後の期待

- 利用者すべてが、ハードウェア限界の性能を必要としてない
- 利用者層が広げ、計算機を利用した研究成果の拡大は、計算センター利用の価値を高める

☆ 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構想を期待したい

要望　仮想化普及にむけて

<u>仮想化ベンダー</u>
- 対応機種を増やす努力
- ソフトウェアベンダーとの連携

<u>ソフトウェアベンダー</u>
- VM上でのアプリケーションやドライバ等の動作保証
- ソフトウェア保守対応

<u>システムベンダー</u>
- 仮想化技術を様々な分野に応用した提案

2008年システムに向けて

- これまでの仮想化技術の運用実績
- マルチコア化でサーバ処理性能が飛躍的に向上
- VTなどのCPU上での仮想化対応

⇩

☆ さらに利用用途を広げたシステム設計を開始

☆サーバ仮想化は広く普及してくるものと確信します

☆ユーザ、ベンダー、ともに新たな活用方法を考えて
いきましょう。

ご参加ありがとうございました。