仮想化技術の概説と効用
VMwareの実装例

ヴイエムウェア株式会社 システムエンジニア
竹洞陽一郎
ytakehora@vmware.com

vmware

ヴイエムウェア社について

企業概要:

- スタンフォード大学内の研究所にて研究されていたテクノロジ
- インテルCPU用のメインフレームクラスの仮想化技術を開発
- 1998年にVMware社設立

主な研究開発分野:

- インテルアーキテクチャ上で複数のオペレーティングシステムを動作させる仮想マシン技術
- 従業員の50%以上を研究開発(R&D)分野にアサイン

方向性:

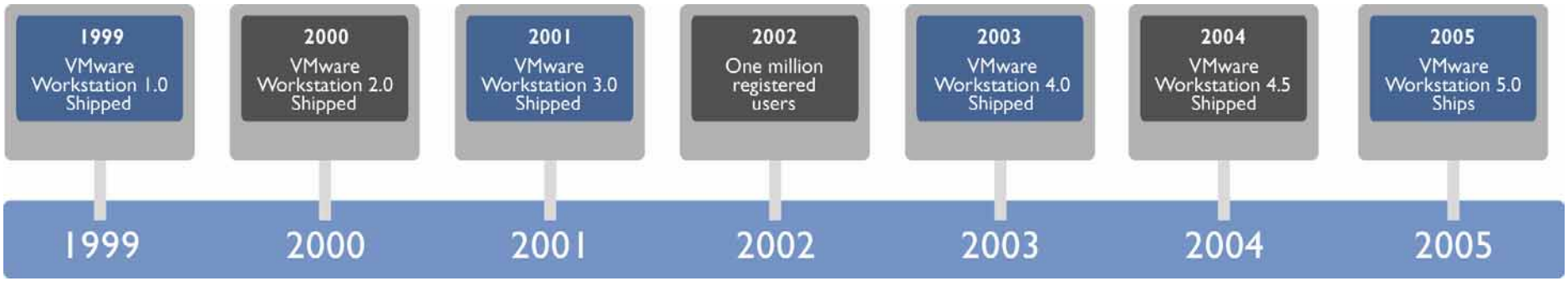
- Workstation製品を1999年にリリース、GSX Server™ を2001にリリース、ESX Server™ を2001にリリース
- 社内テスト、ベータプログラムの徹底 – 品質を最重点項目

会社概要:

- 本社は、カリフォルニア・パロアルト
- 従業員　900人以上
- 健全な財務基盤
- フォーチューン100企業の805%以上がVMware製品のユーザー
- 200万人以上の登録ユーザー
- 100以上の国々に、10,000社以上の企業ユーザー

VMware Workstation

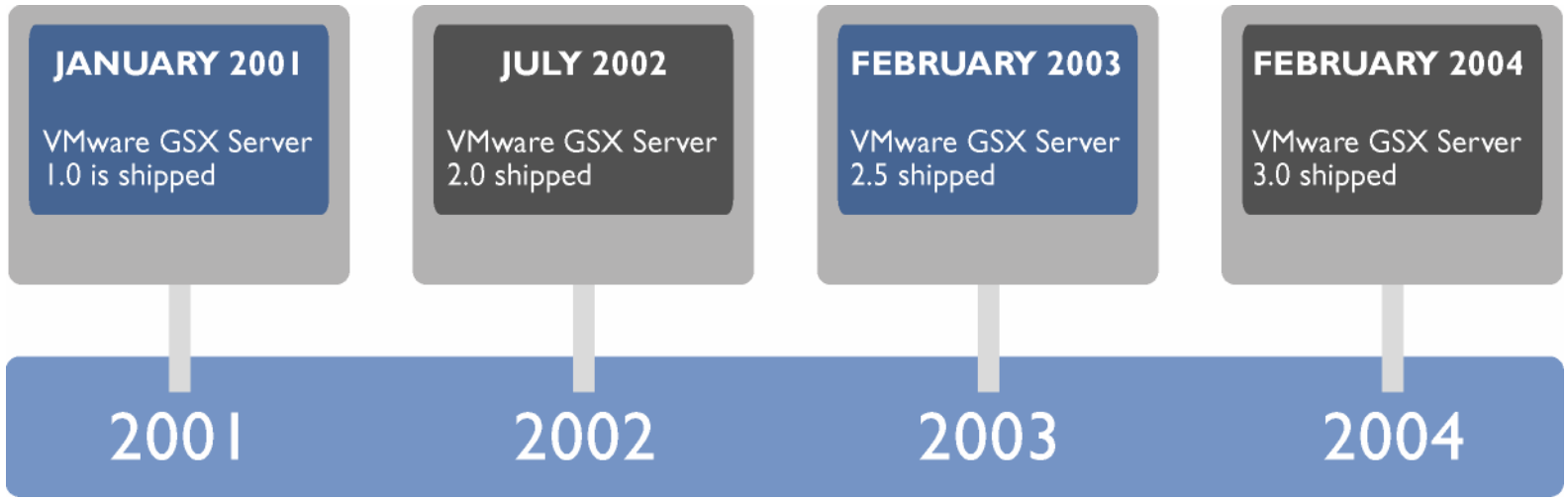
- 1999年に発売されたLinux上で稼動する、
  x86アーキテクチャの初の仮想化技術の実装製品



vmware

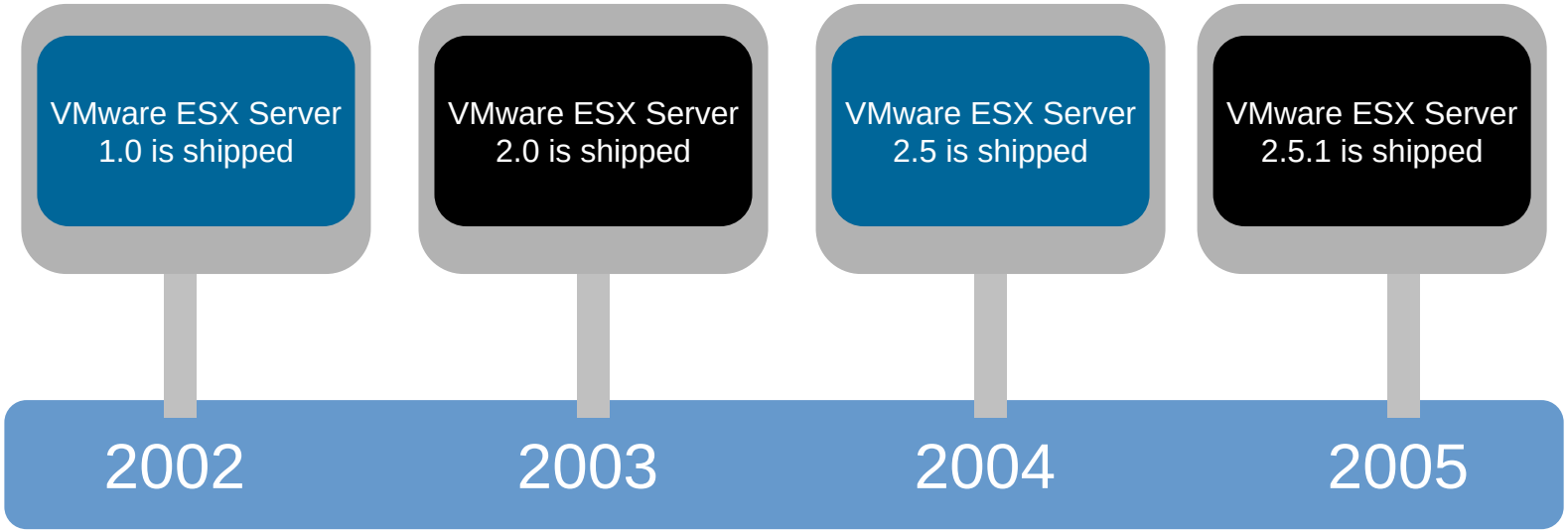
VMware GSX Server

- Windows、Linux**上で稼動する、**
  x86アーキテクチャのサーバ用仮想化技術製品



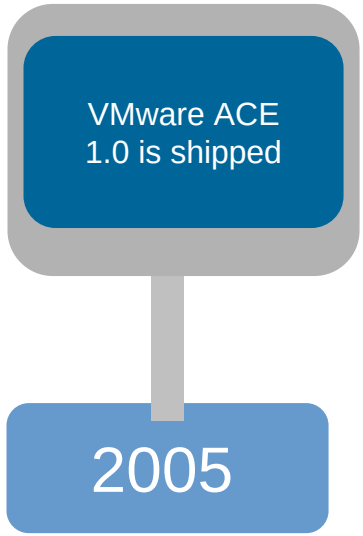
VMware ESX Server

- IBMのOS/370をモデルにした専用カーネルで稼動する、x86アーキテクチャのサーバ用仮想化技術製品



VMware ACE

- Workstationから派生した、エンドユーザ向け
  配布専用のクライアント用仮想化技術製品

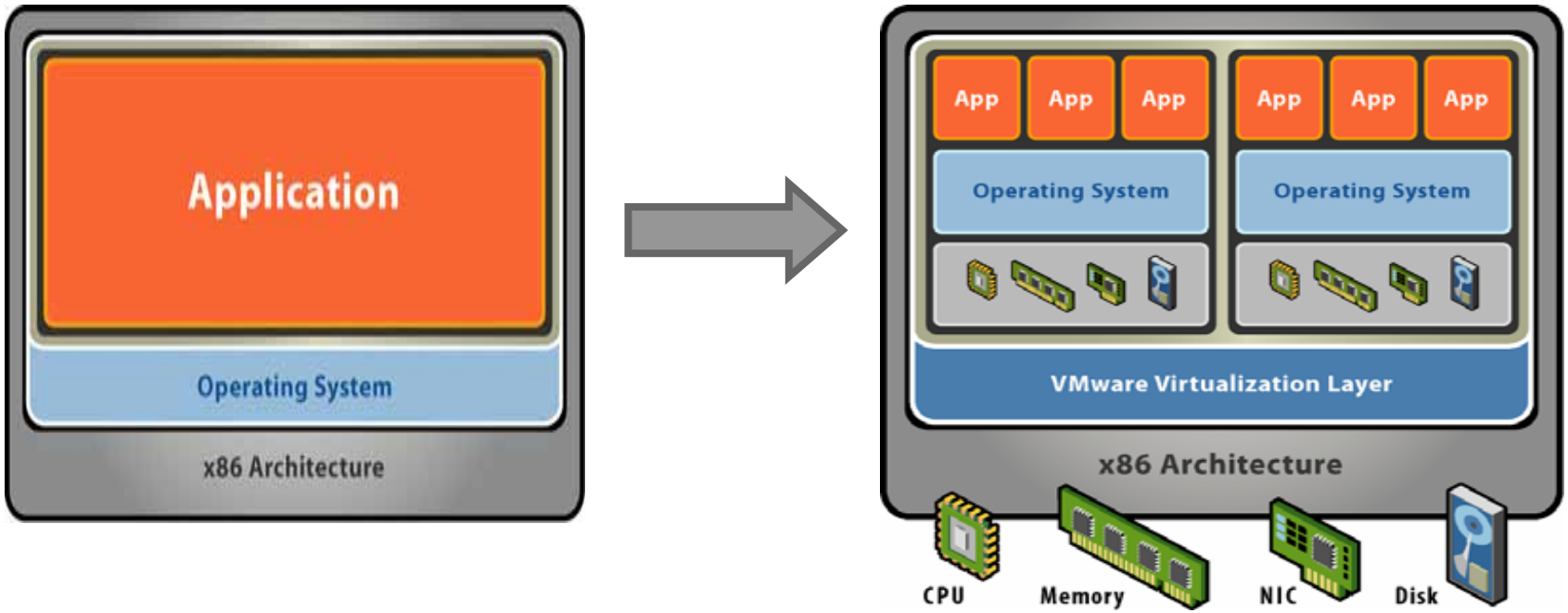


仮想化技術とは?

vmware

仮想化(Virtualization)技術とは?

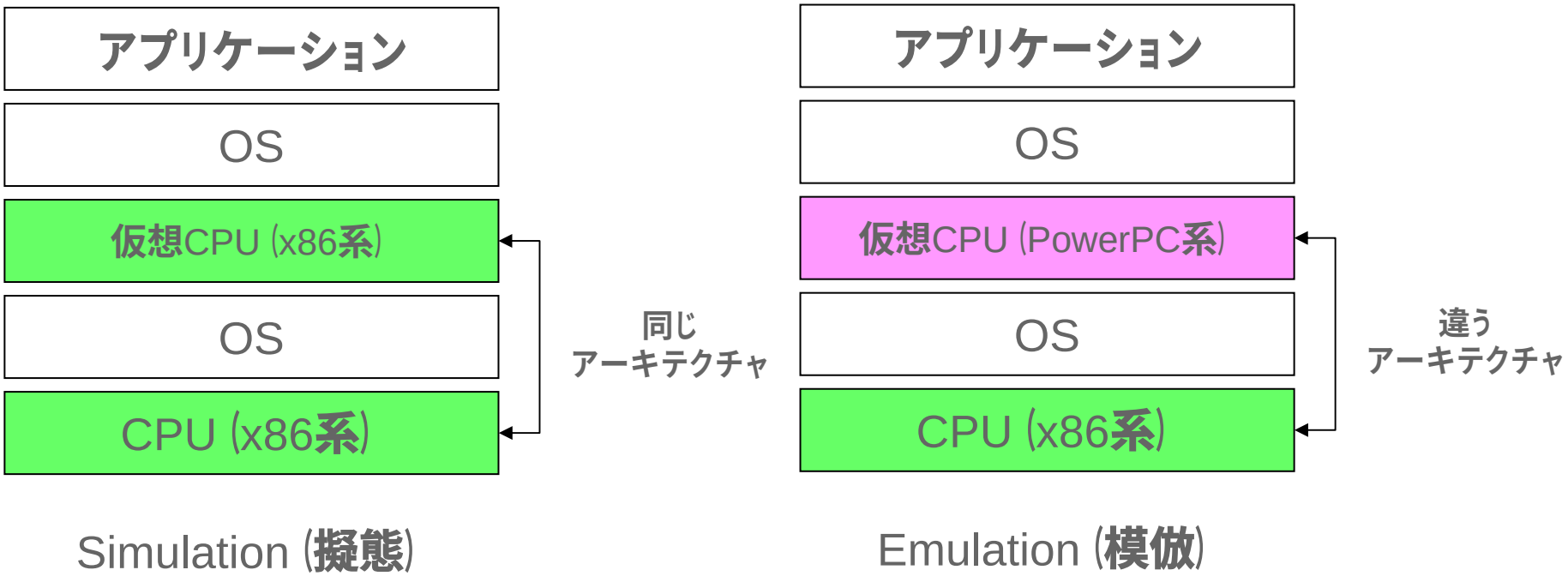
- 通常は、1台のコンピュータに、1つのOSが稼動
- 仮想化技術とは?
  - マシンが提供するデバイス群 (CPU、メモリ、HDD、NIC等) をソフトウェア ***シミュレート***、もしくはソフトウェア ***エミュレート*** して、仮想マシンを稼動させ、1台のコンピュータ上で複数のOSを稼動させる技術



vmware

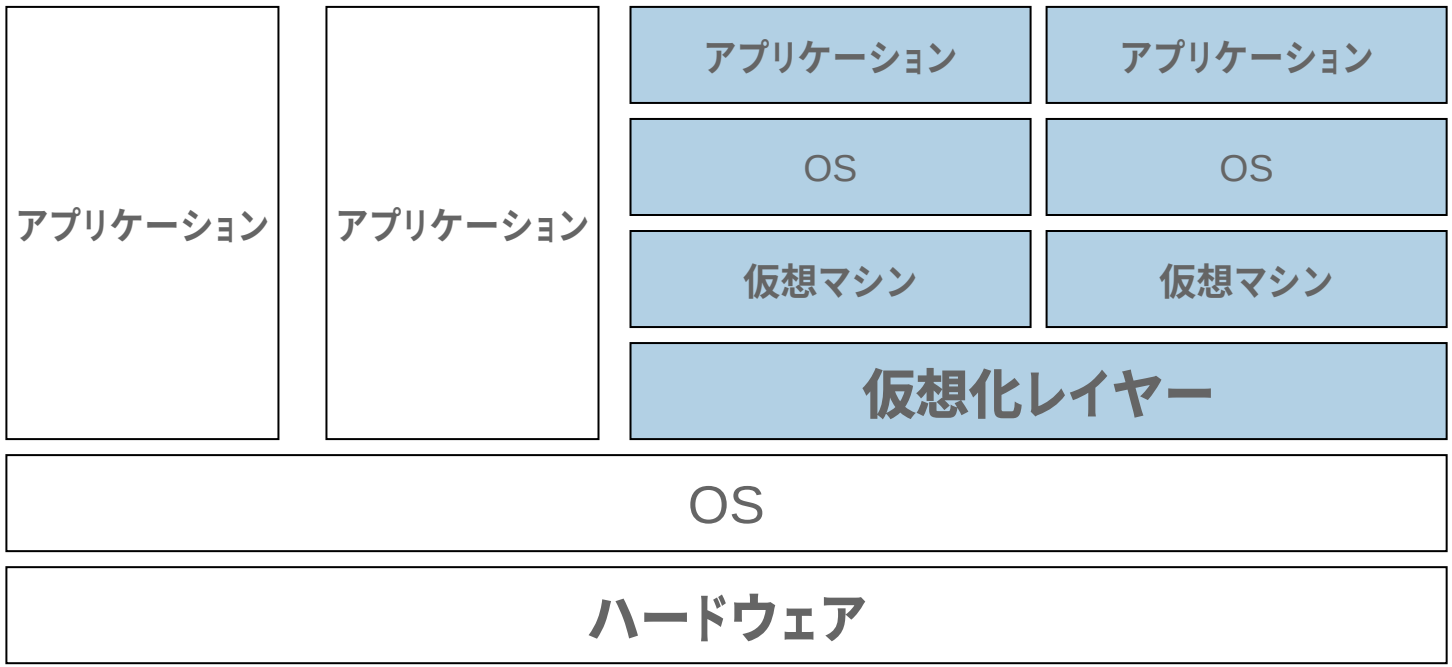
SimulationとEmulation

- シミュレーション（擬態）
  - 同じアーキテクチャのものをソフトウェア的に提供する
- エミュレーション（模倣）
  - 違うアーキテクチャのものをソフトウェア的に提供する



仮想化技術のアーキテクチャ 1~汎用OSを利用する

- **汎用OS (Windows、Linux) 上に仮想化レイヤーをユーザプロセスとして稼動させて、仮想マシンを稼動させる**
- **汎用OSで利用できるデバイス (デバイスドライバがある) であれば、仮想マシンでも基本的に利用できる**
- VMware Workstation、GSX Server、Microsoft Virtual PC、Virtual Server、Open SourceのBochs、PearPC

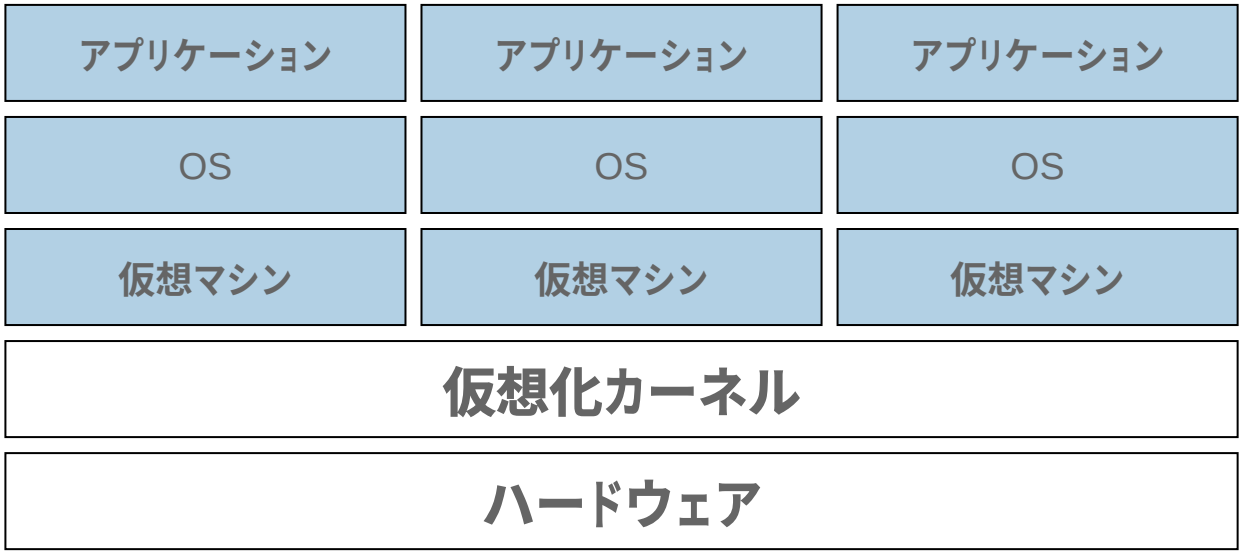


仮想化技術のアーキテクチャ 2~OSを仮想化して利用する

- 特定のOSに特化して仮想化することで、処理能力を向上させる
- 仮想OSはユーザープロセスとして稼動
- OS側で利用できるデバイスであれば、仮想OS上でも基本的に利用可能
- Open SourceのUser Mode Linux、Cooperative Linux (coLinux)、JavaVM

仮想化技術のアーキテクチャ 3~専用カーネルを利用する

- 専用カーネル上に仮想化マシンを稼動させて、処理能力を向上させる
- 専用カーネルを利用するため、専用カーネルが対応しているハードウェアでなければ、デバイスは利用できない
- VMware ESX Server、Open SourceのXen



VMwareの仮想化技術

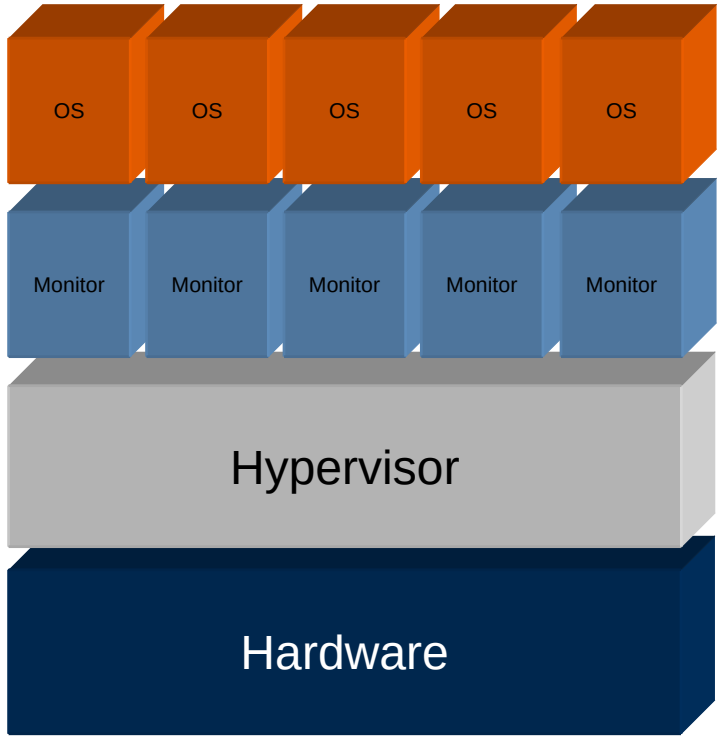
vmware

仮想化技術の原点

- 仮想化技術を最初に実装したのはIBMのOS370
- 1960年代後半にMITがIBMのメインフレーム上で仮想化の技術を実装
- 仮想化の背景
  - メインフレームのOSはシングルユーザだった
  - 当時、メインフレームのCPU処理能力は急速に向上しており、仮想化によって複数のOSを同時に稼動させることにより、高額且つシングルユーザしか使えないリソースの使用率を高めることが目的だった

OS370

- Hypervisor
  - ハードウェアデバイスをマルチタスクで使用し、“Monitor”を生成･稼動させる
  - ハードウェアを直接操作
- Monitor
  - 仮想マシンを管理する親プロセス
  - 個々の仮想マシン毎にMonitorが用意される
  - MonitorのインスタンスはHypervisorによって生成される
- OSのインスタンスはMonitorによって生成された仮想マシン上にインストールされる

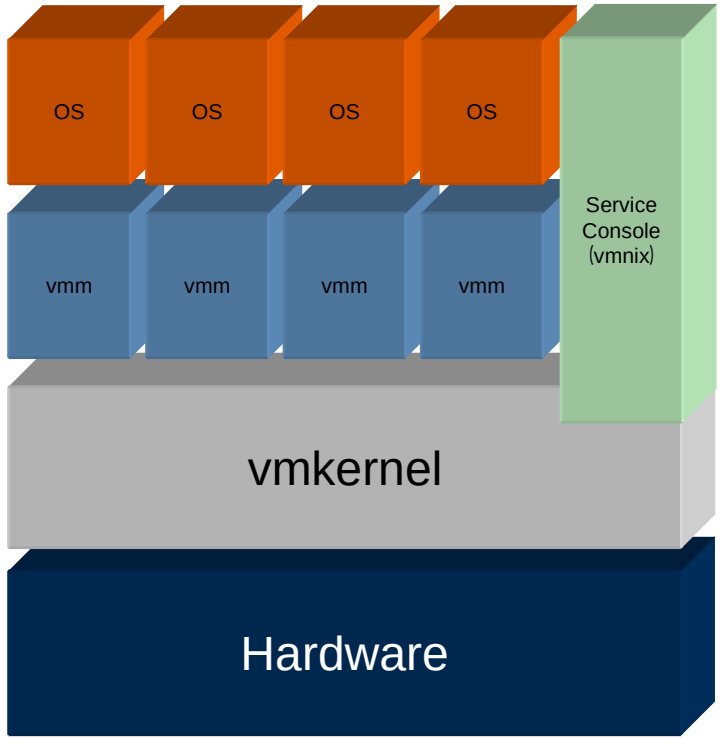


VMwareが仮想化技術製品を開発した背景

- X86アーキテクチャも今やメインフレームが1960年代に経験したようなハードウェアリソースを使い切れない状況を迎えている
  - 市場競争に晒されて、高機能化・高速化・低価格化したCPU
  - OS上で1つの“*アプリケーション*”が稼動している状況
- メインフレームの世界で実績が認められている同じ種類の多重化方法をVMwareの技術でx86アーキテクチャ上で実現
  - 1台の物理サーバ上で複数のOSが稼動することができるようにする

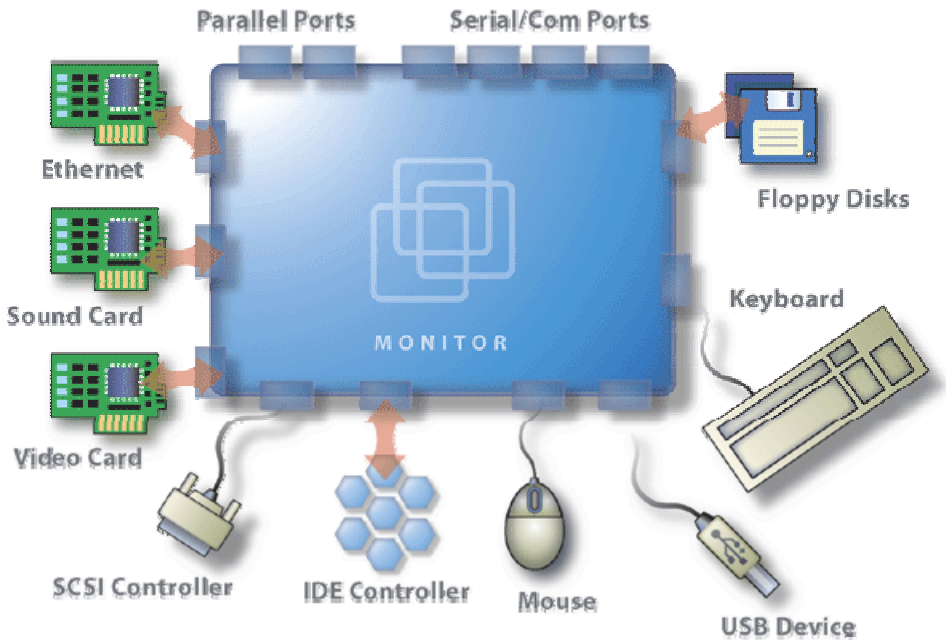
ESX Server

- VMkernel
  - 各仮想マシンからVMMを経由して出てくる命令をキャプチャして、スケジューリングし、処理する
  - ハードウェアを直接操作
  - 300K LOCの非常にコンパクト且つ堅牢なカーネル
- Service Console
  - 間接的に命令をVMkernelに受け渡してESX Serverを管理するコンソール
- vmm
  - 仮想マシンから出てくるバイナリ命令を監視するモニターが仮想マシン毎に生成される
  - vmmはマシンの全てのオペレーションをコントロールする
    - キーボード/グラフィックス/マウス
    - ネットワークカード
    - SCSIコントローラ



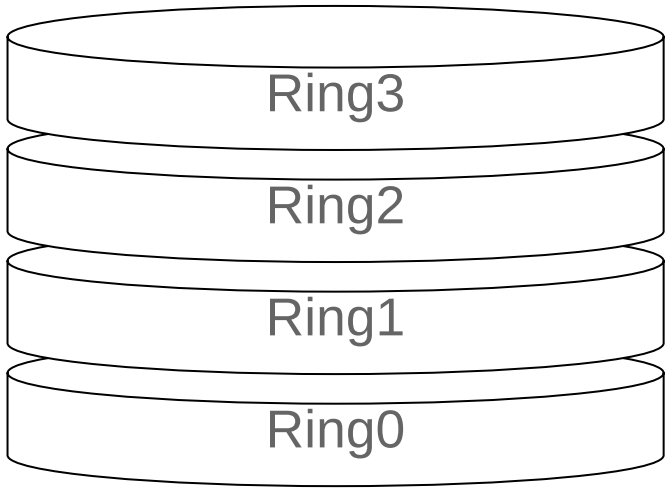
ESXが提供する仮想化ハードウェア

- 仮想化ハードウェアのスペック
  - 440vxチップセット
  - AMD PCIネットワークカード(vlance)
  - LSI LogicもしくはBus Logic SCSIアダプタ
  - VMware独自のネットワークカード (vmxnet)
  - VMware独自のグラフィックスカード
    - パフォーマンスの問題から



Processor Ring

- Ring3
  - User Landと呼ばれる
  - 一般のプロセスはこのRingで実行される
  - ハードウェアに対する命令は直接実行することができず、デバイスドライバと通信を行うAPIを通じてRing0へ引き渡される
- Ring1&2
  - Ring1の命令セットはRing0と同じ動作であるが、実行される保証がない
  - 現在のOSでは、Ring1とRing2は利用されておらず、モデルとしてのみ存在が残っている
  - IBMのOS/2は、Ring1を利用した珍しいOS
- Ring0
  - Privileged（特権）
  - Kernel Land/Kernel Spaceと呼ばれる
  - ハードウェアに直接命令を下せる
  - OSのカーネルやデバイスドライバはこのレベルで稼動

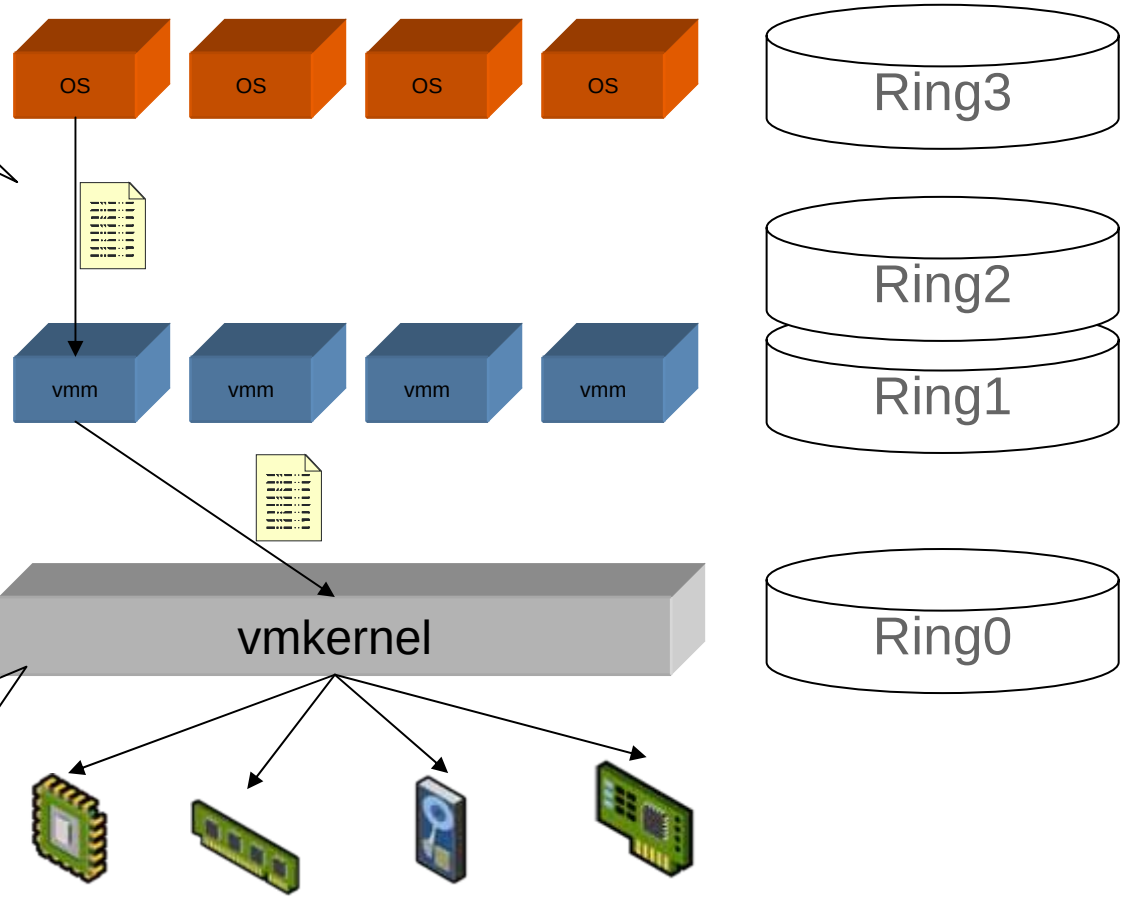


バイナリトランスレーション

1. OSから出た命令はvmmによってキャプチャされ、VMkernelへ引き渡される

2. VMkernelはOSから出た命令（バイナリ）を解釈（トランスレーション）し、処理に必要なデバイスに命令を出す
この処理を高速化するため、VMkernelは各OSのバイナリがどのような意味を持つかをテンプレートとして持っている（トランスレーションキャッシュ）

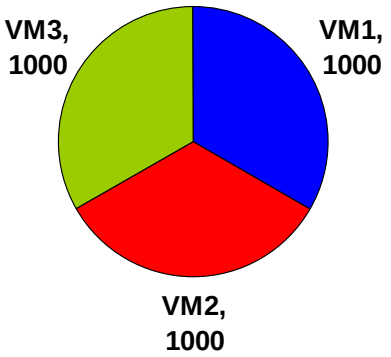
3. デバイスから処理の実行結果をVMkernelが受け取り、OSへあたかもハードウェアが実行処理した結果であるが如く返す

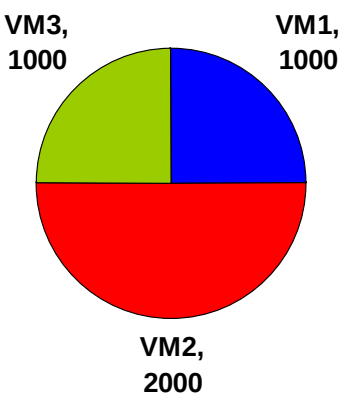


ESXで行えるサーバリソース管理

- 以下のサーバリソースを、各仮想マシンに割り当てることが可能
  - CPU
    - Hyperthreadingに対応～論理CPU単位で仮想マシンにアサインできる
    - 仮想SMP機能～SMPを仮想化して仮想マシンに組み込むことが可能（現在2CPUまで）
  - メモリ
    - 最小メモリ量と最大メモリ量を割り当てることが可能
  - NIC
    - トラフィックシェーピング方式で帯域管理が可能
    - 物理NICを束ねて1つのNICに見せる、NIC Teamingの機能を搭載
  - ディスク帯域
    - ディスクの帯域使用割合を設定可能
  - ディスク容量
    - 仮想化ディスクのサイズ変更が可能（OS側でパーティションマジックなどを使って、パーティションを拡大させる作業は必要）
    - 仮想化ディスクの4つのモード
      - Persisitent～コンピュータ上の従来のディスクドライブとまったく同じ様に動作。Persistent モードのディスクに書き込まれたデータはすべて、ゲストOS がデータの書き込みを行った時点でディスクに恒久的に保存される
      - Nonpersistent～ディスクへの変更は仮想マシンを電源オフにすると全て破棄される
      - Undoable～仮想マシンの電源オンから電源オフまでの変更を保存するか破棄するか選択可能
      - Append～変更は継続的にRedoログに追加され、Redoログファイルを削除すれば変更は破棄できる。Commit すれば、恒久的に保存される

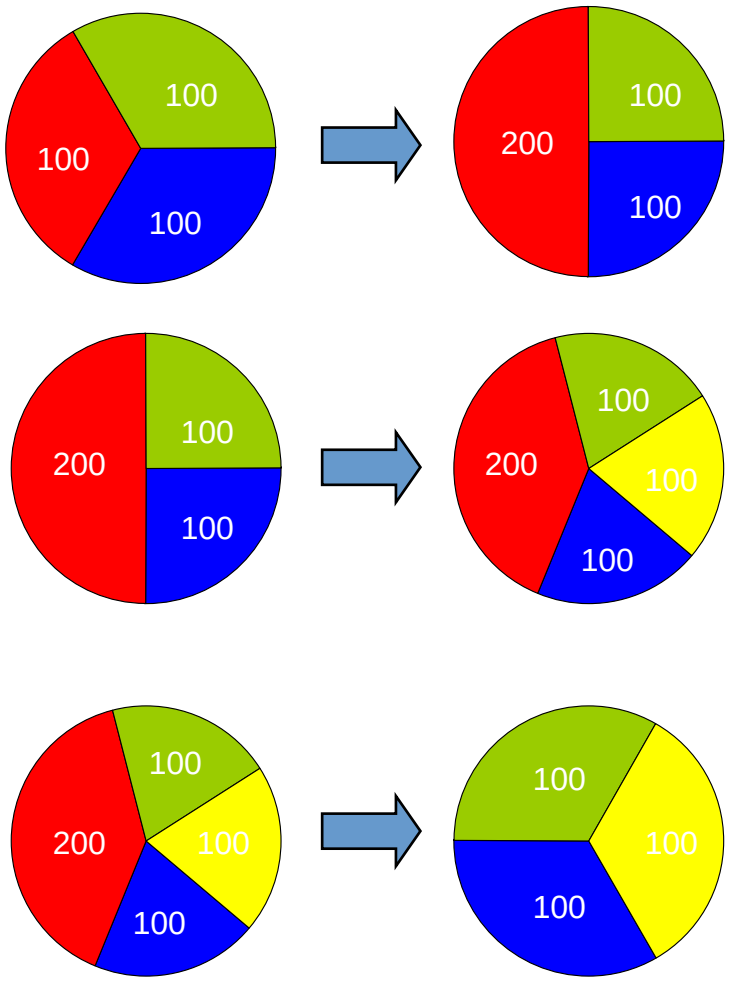
リソースの割当方法～シェアという考え方





- VM1～3でディスク帯域幅のシェア値を1000にする
- 160MB/秒のディスク帯域幅がある場合、それぞれのVMは53MB/秒でディスクアクセスを行う
- VM2のディスクアクセスを優先させるために、シェア値を2000に増加
  - 1000：2000：1000 = 1:2:1の比率となり、VM1とVM3は、160MB×1/4=40MB/秒のディスクアクセス
  - VM2は160MB×2/4=80MB/秒のディスクアクセス

CPUスケジューリング～シェア値とMin/Maxの設定



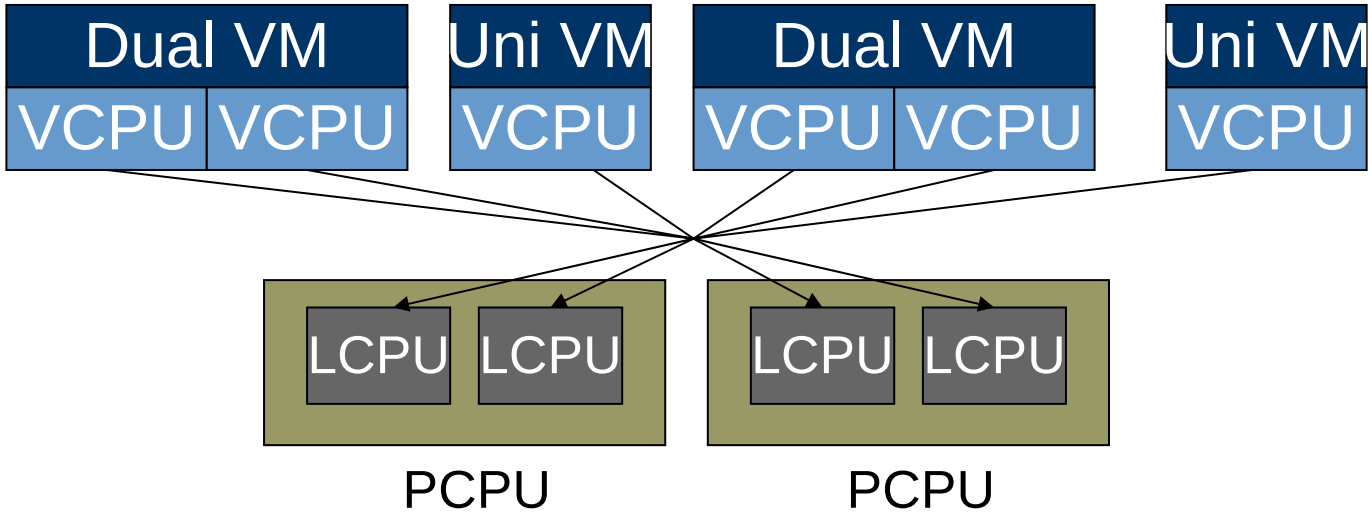
- VM シェアを変更
- ダイナミック リアロケーション

- VM を追加
- シェア相対値は変わらない

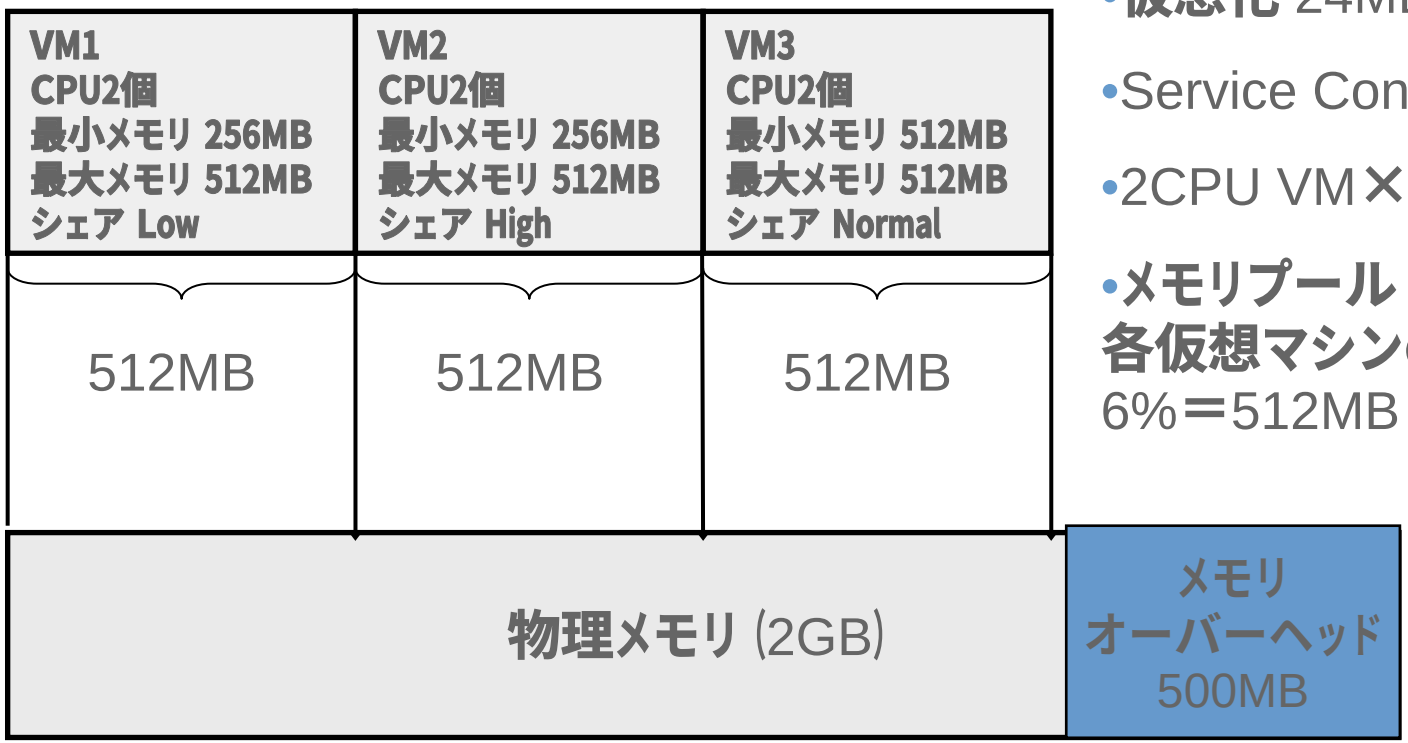
- VM を削除
- リソースの浪費はなし

ハイパースレッド サポート

- VMkernelは、VMを論理CPU上にインテリジェントにスケジューリングする
  - 仮想CPUは、VMkernelのスケジューリング毎に別の物理CPUで動作
- VMは論理CPU (HT)にバインドされる
  - HTを利用し、仮想CPU 0 と 1 は物理CPU0のそれぞれの論理CPUにマッピングされる
  - 注意：負荷の高いVM同士を同じ物理CPUにマッピングすると速度低下の可能性がある
- VMはHTから独立させることも出来る
  - CPU依存、キャッシュ依存のワークロードに最適



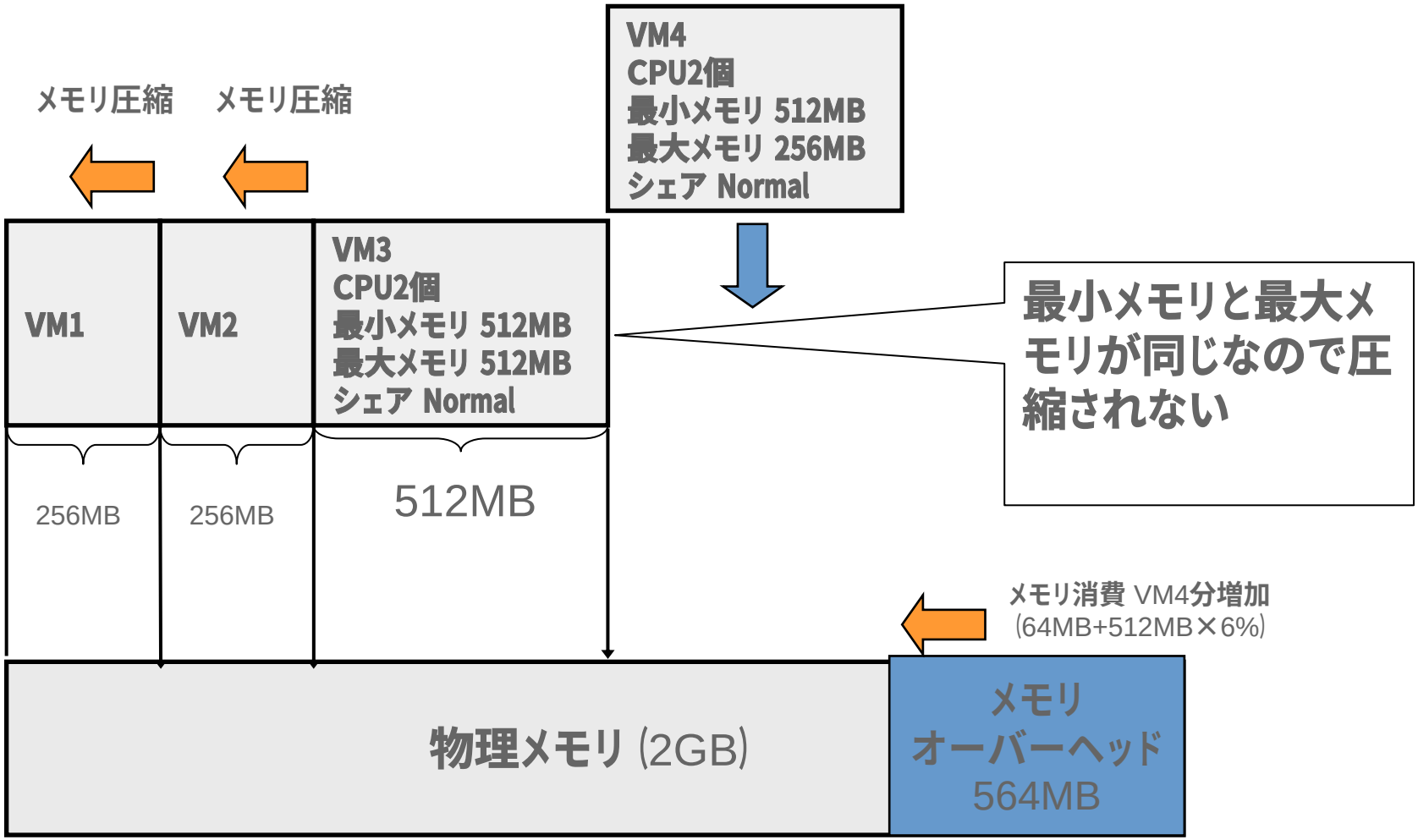
メモリのシェア割当1



- **仮想化** 24MB
- Service Console 192MB
- 2CPU VM×3＝64MB×3
- **メモリプール**
  各仮想マシンの最大メモリの
  6%＝512MB×6%×3

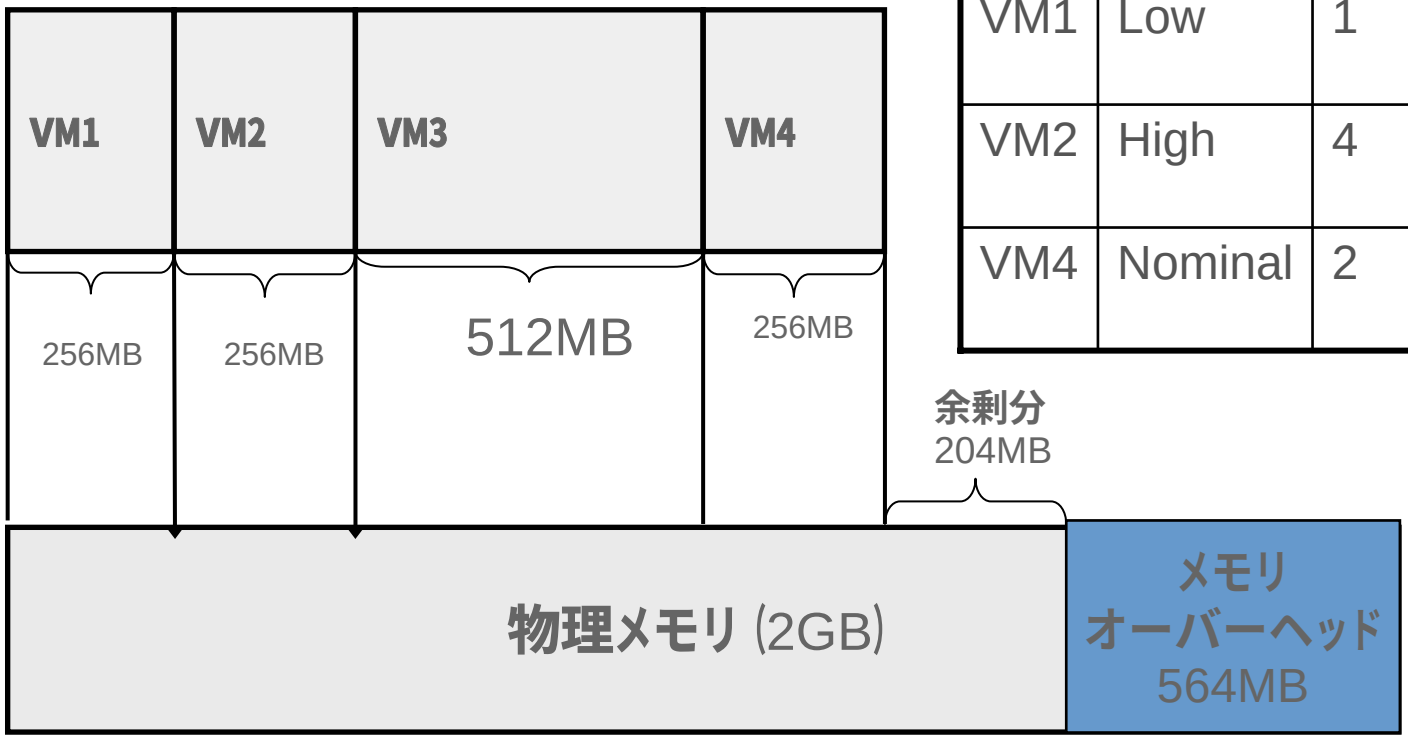
物理メモリに余裕がある内は、各々の仮想マシンに最大メモリ量を割り当てる

メモリのシェア割当2



新たにVM4を起動しようとすると、メモリが足りないので、VM1とVM2については割り当てメモリが圧縮される

メモリのシェア割当3

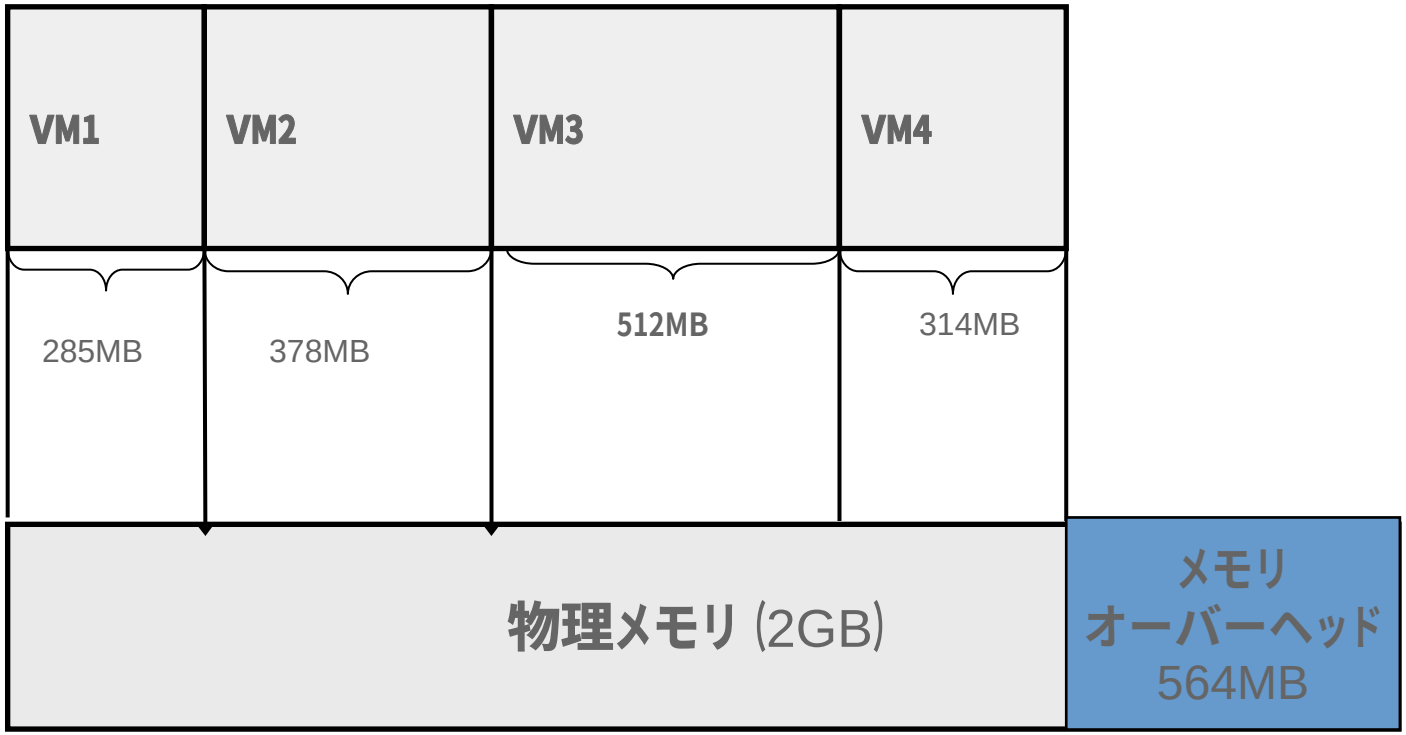


	シェア	比率	割り当てメモリ量
VM1	Low	1	204×1/7 =29MB
VM2	High	4	204×4/7 =116MB
VM4	Nominal	2	204×2/7 =58MB

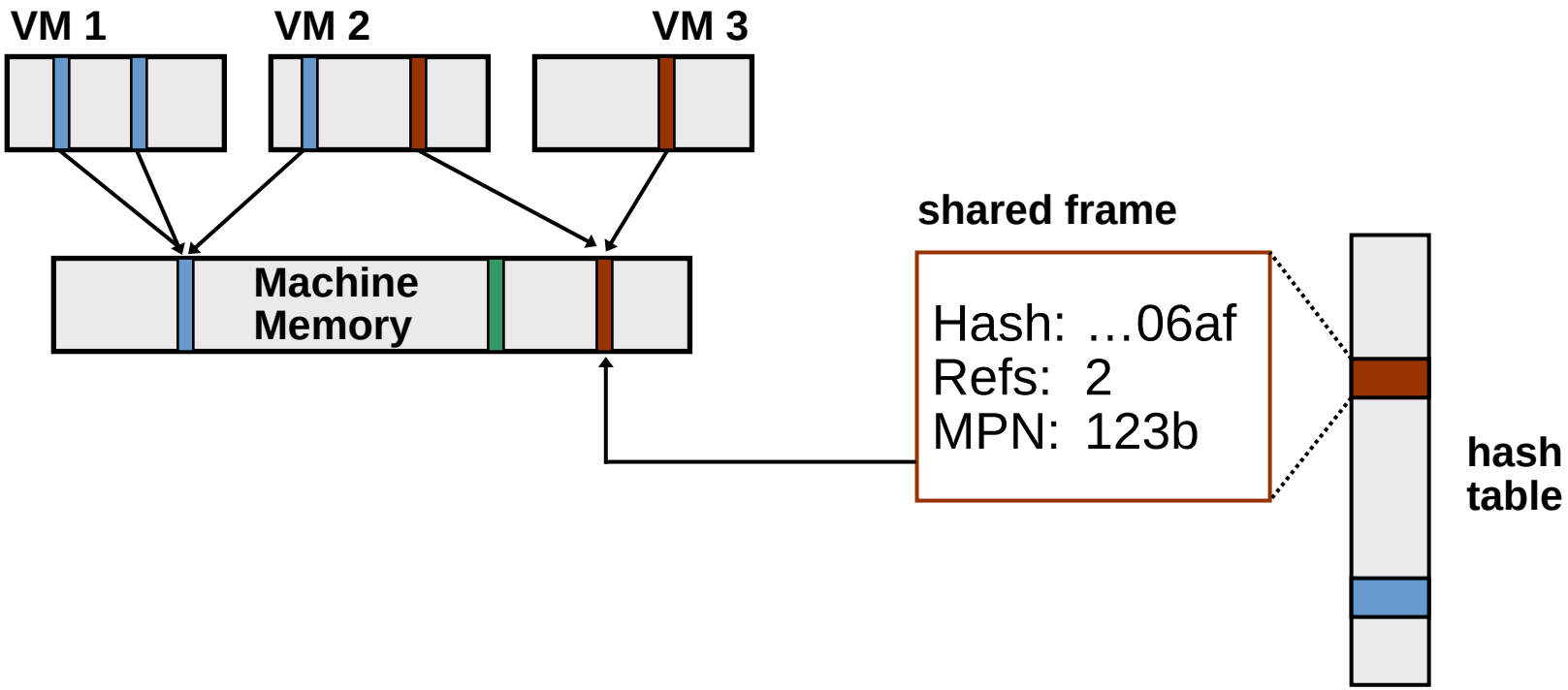
最小メモリでの割り当て状態のVM

メモリのシェア割当4

メモリの再配分後の各VM

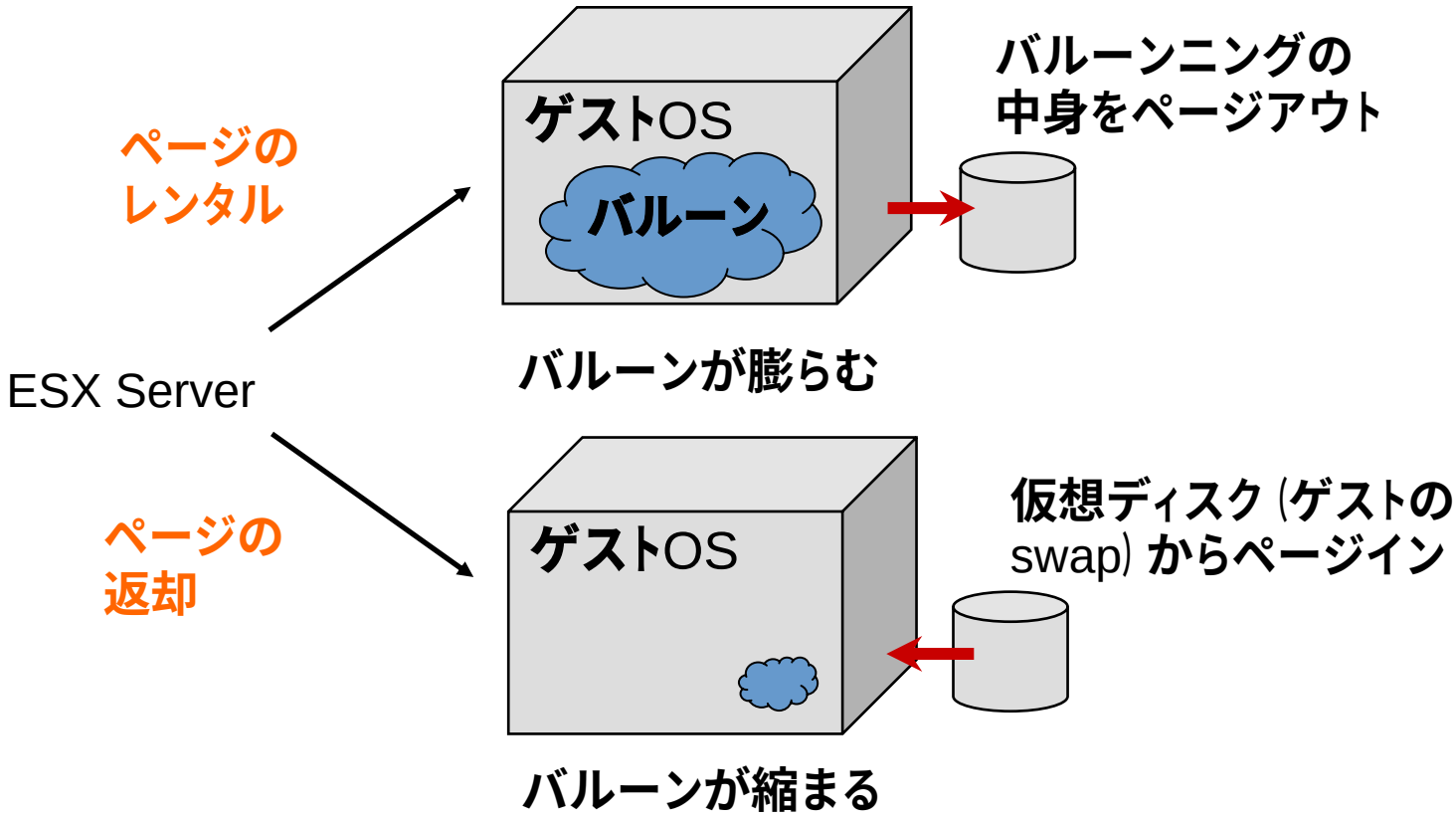


仮想化によるメモリの有効活用1：ページシェアリング



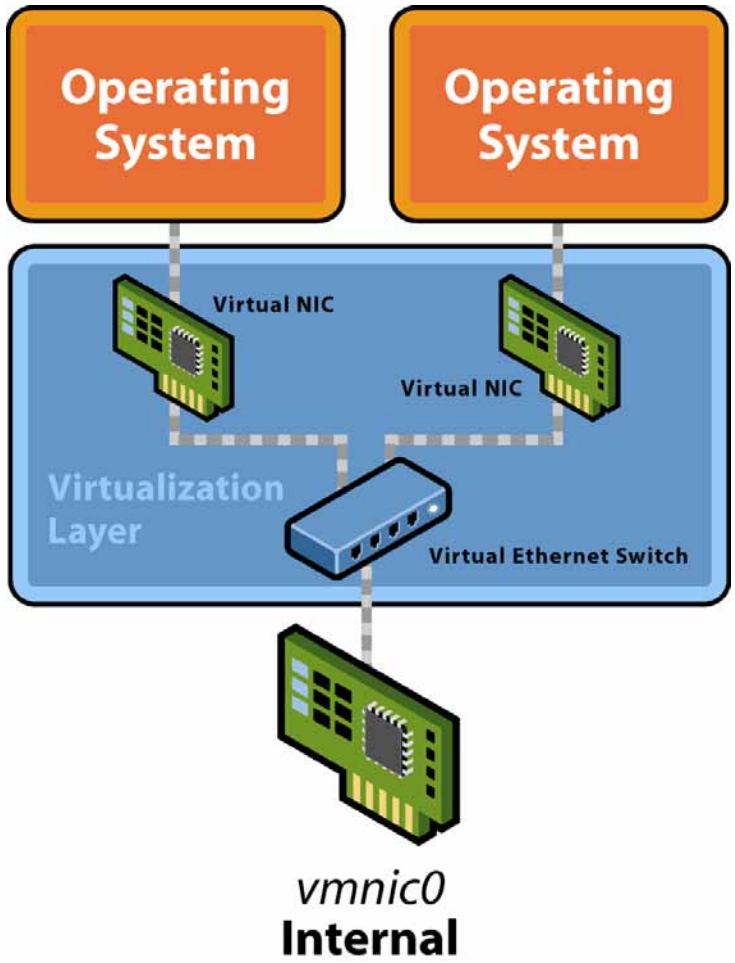
- プログラムコードページやゼロページの共有
- 標準的には10% - 20% のメモリ節約

仮想化によるメモリの有効活用：バルーニング



ESXサーバーが引き起こすメモリ・プレッシャー

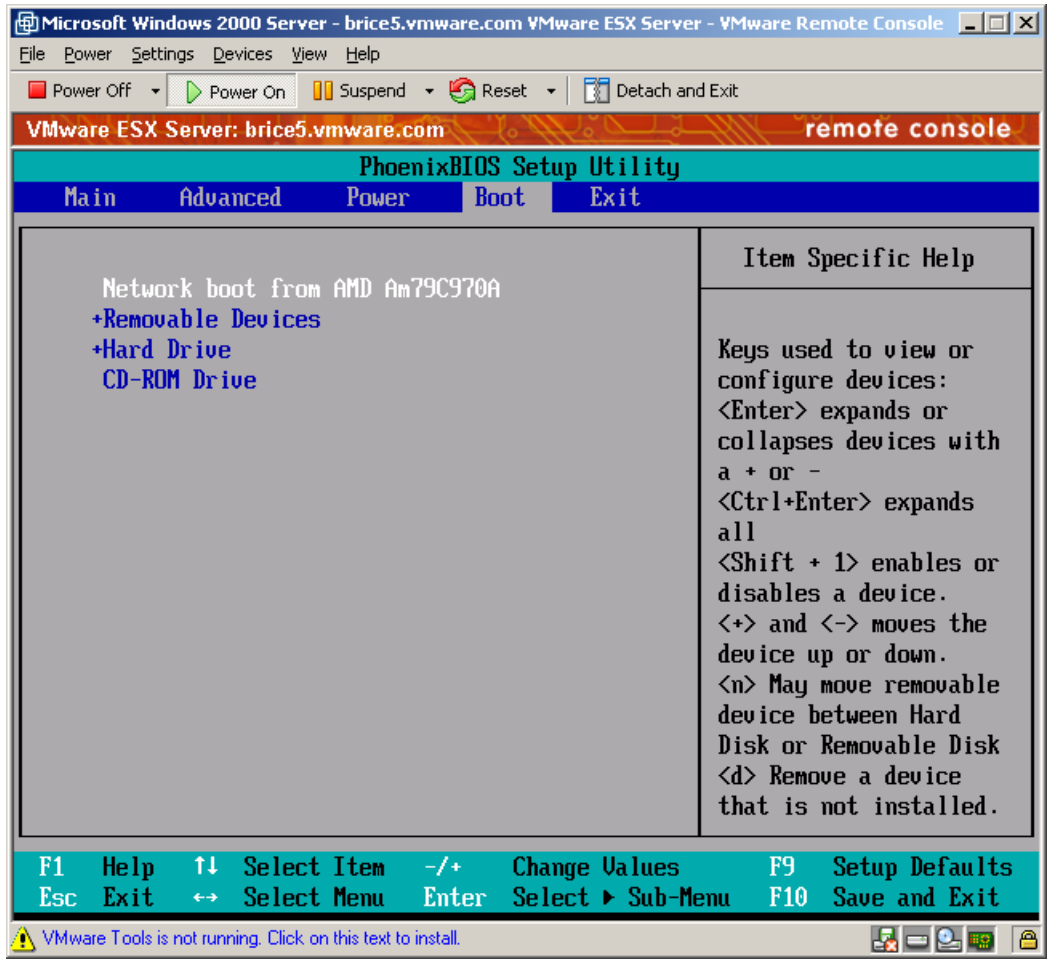
仮想ネットワーク



- 仮想マシンに対して、最大4枚までの仮想NICを設定可能
  - 各々の仮想NICは、一意のMACアドレスを持つ
- 複数の仮想スイッチをサーバー内に構成可能
- 仮想スイッチを経由し、外部接続、loopバックなど様々なネットワーク構成がサーバー内で可能
- 仮想NICの帯域指定が可能
  - 標準・最小・最大
- NICチーミングにより、外部接続を強化
  - 帯域の拡大
  - NICフェイルオーバー
- VLANサポート

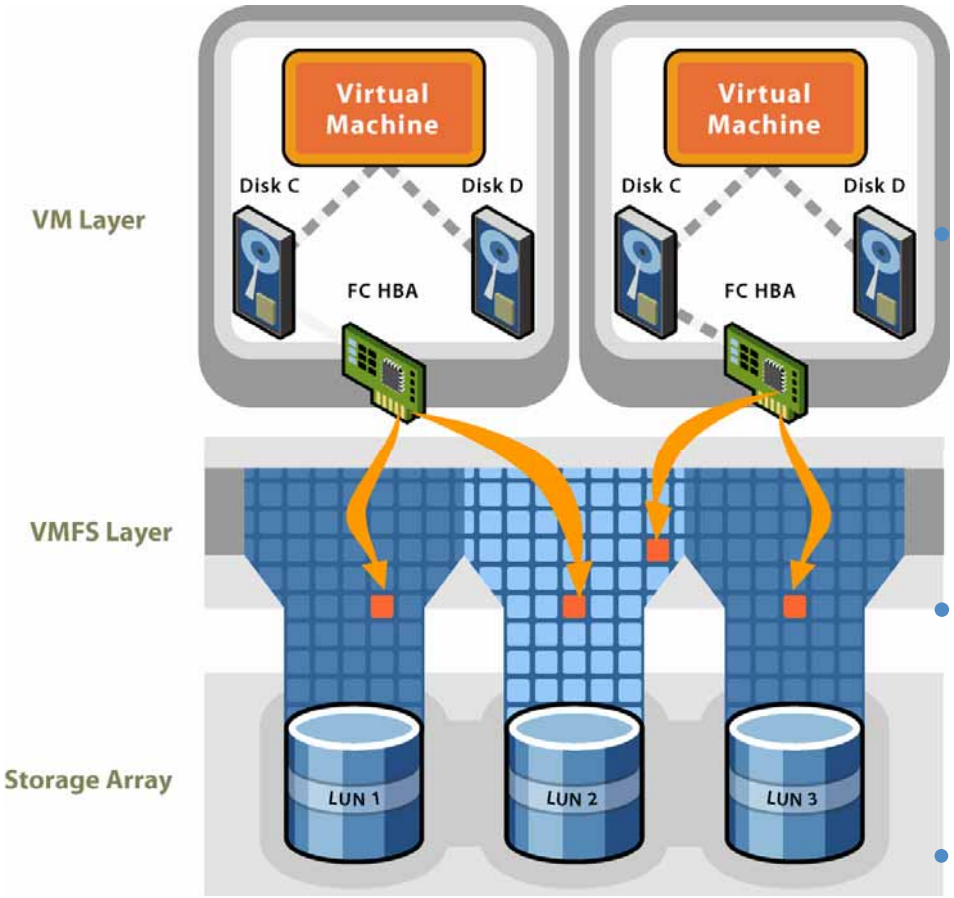
VM PXE Boot のサポート

- Intel の Preboot Execution Environment (“PXE”) に対応。
- リモートシステムデプロイが可能に
  - NICからのトリガでVMがブート可能
  - VMのリモート管理がより容易に



vmware

仮想ストレージ



- **仮想ディスク**
  - 仮想マシン内では仮想SCSIアダプタ経由のディスクとして認識
  - vmfs内のファイル、またはRawパーティションを使用可能
- Virtual Machine File System (VMFS)
  - エクステントベースのファイルシステム
  - 複数の物理パーティションを使用し、論理パーティションを作成可能
  - (ダイナミックディスク)
- **マルチパスフェイルオーバー**
  - ESX標準ドライバにて、マルチパス・フェイルオーバーが可能
  - ベンダからの追加ドライバの必要無
- VM動作中にバックアップ可能
  - 24x7サービスのバックアップが簡単に

vmware

仮想化技術の効用

vmware

増殖し続けるIAサーバ

- 市場競争の波に晒されて、低価格化、高性能化が進むIAサーバ
  - プロジェクトや部門毎にIAサーバの導入が進む
    - 場所の問題～増殖し続けるIAサーバの設置スペースの確保
    - リソース使用率の問題～ピーク時を考慮してリソースのサイジングをしなければいけないので、台数は減らせない
    - 安定性･安全性の問題～ミッションクリティカルな業務への適用も広がる中、OSの安定性やセキュリティの強度を考慮してシステムを構築しなければならない

場所の問題

- タワー型サーバからラックマウント型サーバ、そしてブレード型サーバへと、サーバの省スペース化・高集積化により、サーバスペースを節約
  - ブレード型サーバは1サーバ1アプリケーションで稼動するのが最適
  - 高負荷がかかる、サーバリソースを豊富に必要とするシステムには適用できない
  - ブレード型サーバに物理的に集約しても、それぞれのサーバのCPU使用率は低いまま

タワー型サーバ

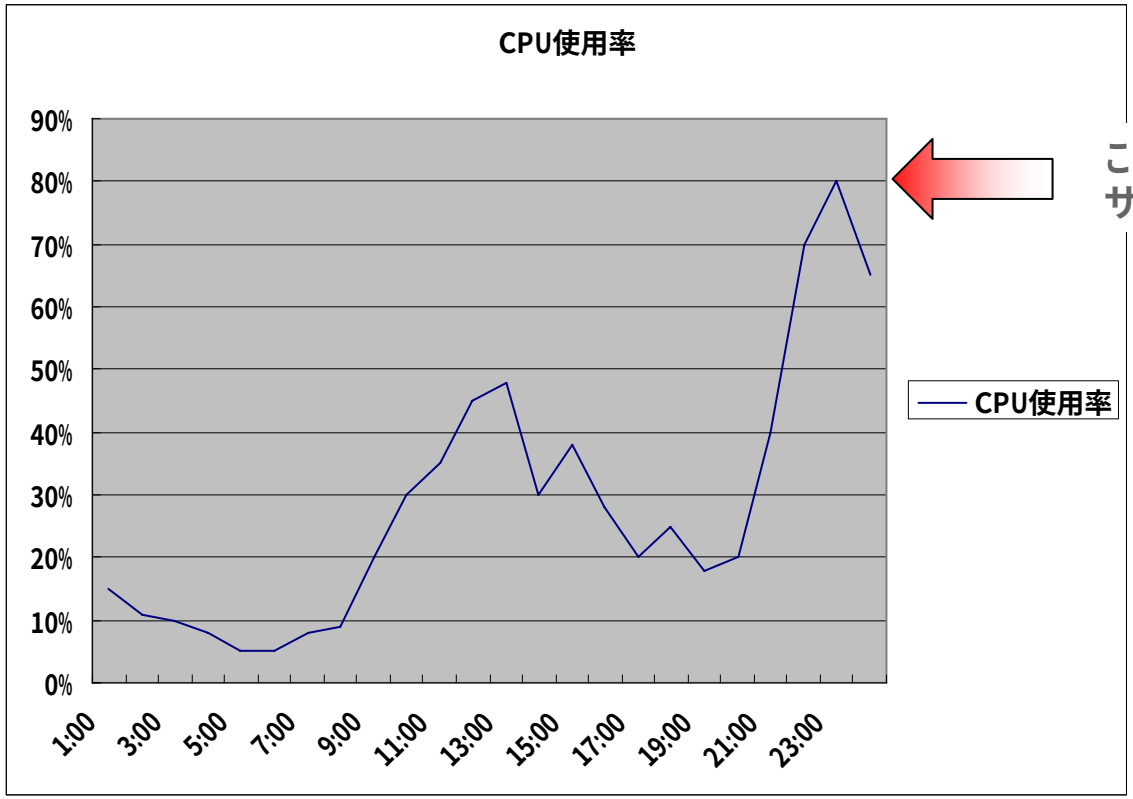
ラックマウント型サーバ

ブレード型サーバ

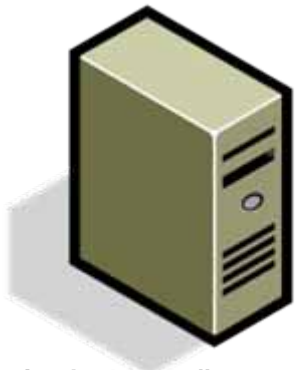
省スペース化・高集積化

リソース使用率の問題

- ピーク時を想定してサーバのサイジングが行われる
  - サーバの平均稼働率は、サーバのキャパシティに比べて遥かに低い
  - ビジネスの状況次第で、想定していたピーク時の負荷を大きく上回ったり下回ったりする



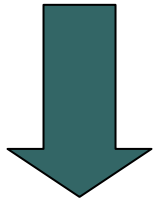
この時の負荷に耐えるようにサイジング



ピーク時を想定して購入したWebサーバ

工場の機械の操業率が10%だとしたら・・・

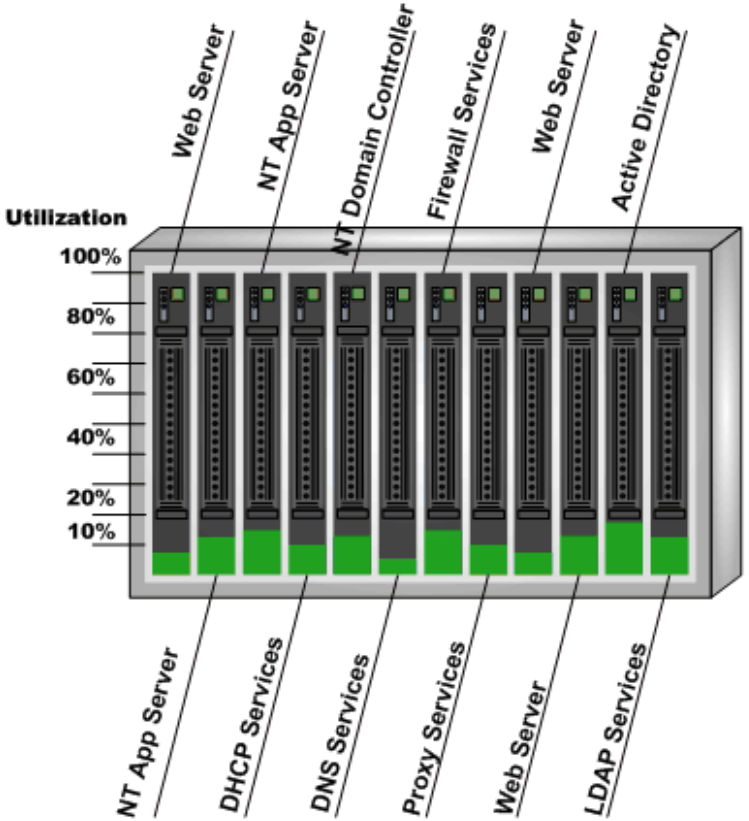
- もし工場の機械の操業率が10%前後しかないとしたら、あなたが会社の経営者なら許しますか？



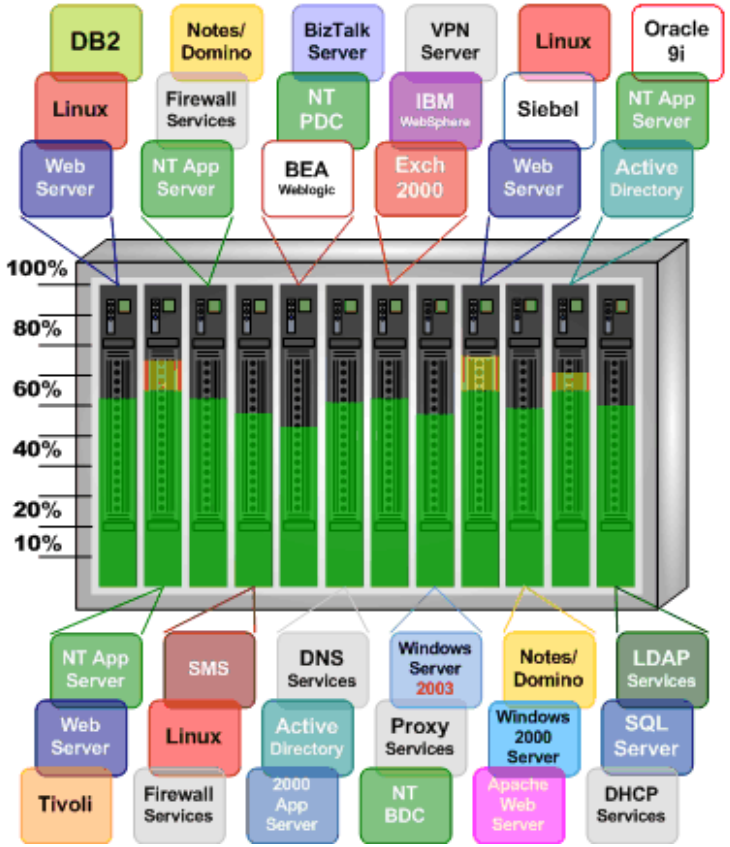
- サーバのリソース使用率は10%前後
- 1000万円のサーバを購入したのであれば、900万円は無駄な投資になる
- 管理コストも含めると、IT投資が有効に使われていない額はもっと大きくなる

仮想化で効率的なサーバー利用
1台のサーバー（シャーシー）で4台分のワークロードに対応

Blade Servers without VMware VirtualCenter



Blade Servers with VMware VirtualCenter



vmware

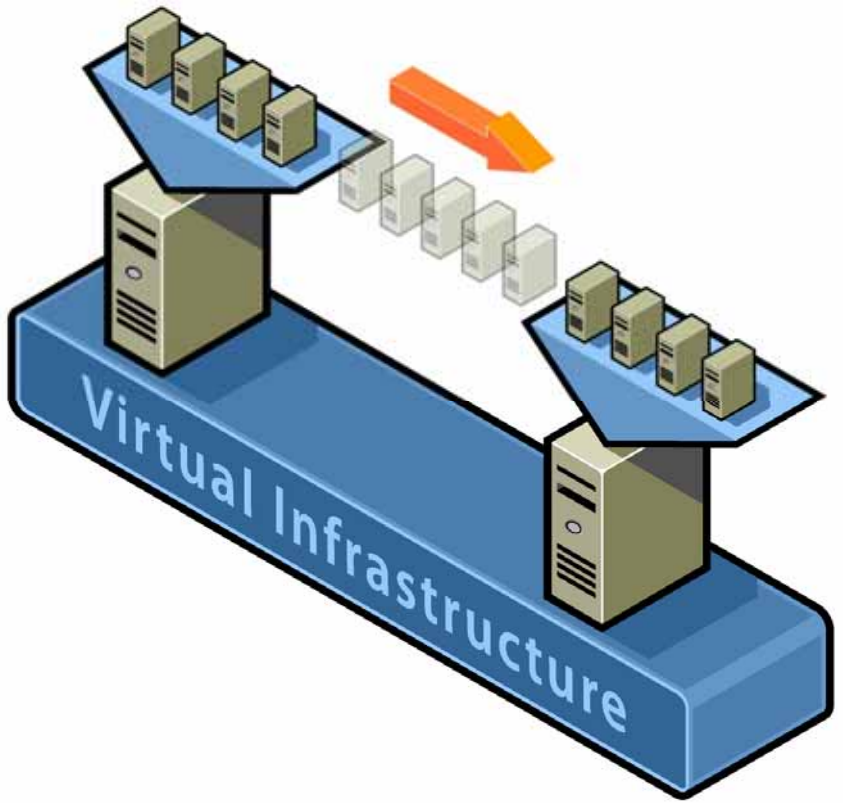
ピーク時にどう対処する？

- 例え、平時のリソース使用率が10%前後だとしても、ピーク時にはどうする？
- ビジネスが上手く行って、平均リソース使用率が上がったらどうする？

VMotion™ テクノロジ（in VirtualCenter）

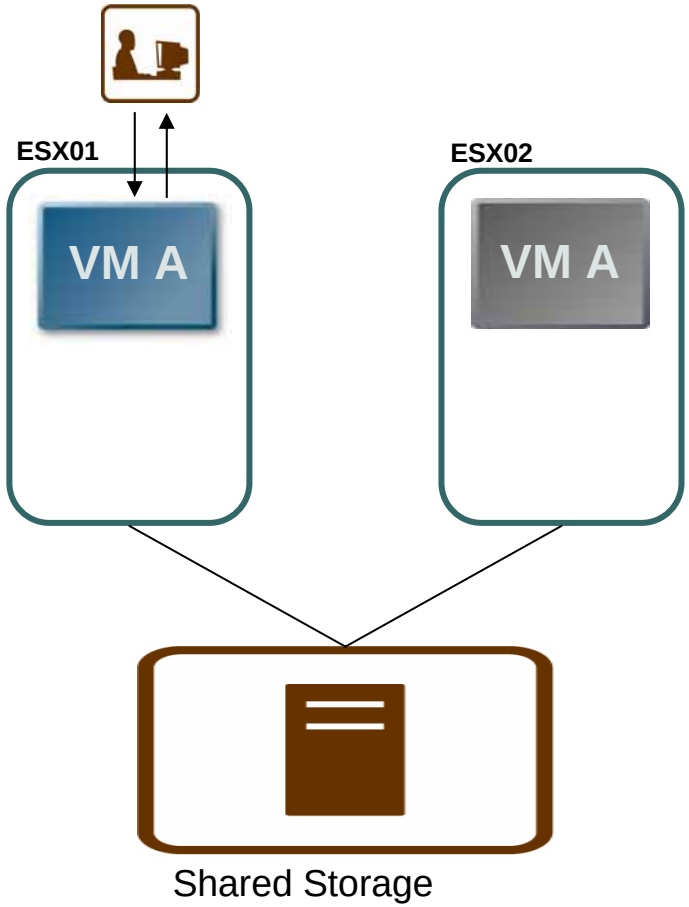
スケールアウトソリューションのキラーテクノロジ

VMを、稼動中OSを停止させることなく、物理サーバー間で移動させることが可能



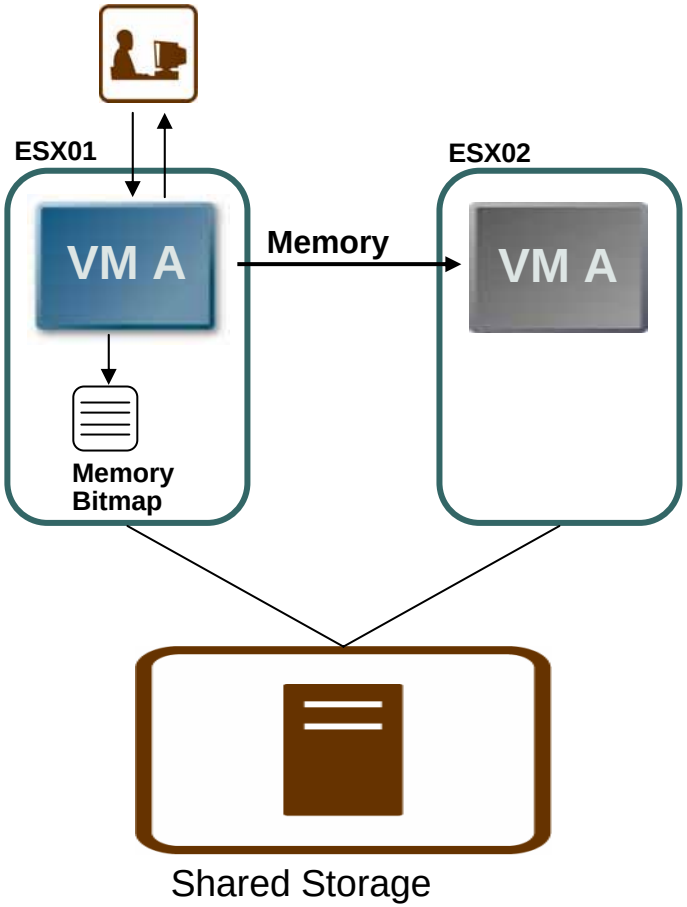
- ダウンタイム無のハードウェアアップグレード
  - VMのシステム稼動状態を維持したままサーバーを移動させ、ハードウェアメンテナンス、アップグレードが可能
- ダウンタイム無で、サーバーの使用目的を変更
  - ブレードを交換、抜き出しする場合に、ブレード上で動作しているVMを全て移動させる
  - VMを移動させサーバーの負荷を最適化
- 効率的なサーバー管理
  - 障害分析をベースに、事前にVMを移動可能

VMotion: 動作説明



1) 新VMを新しいサーバー上に準備

VMotion:動作説明

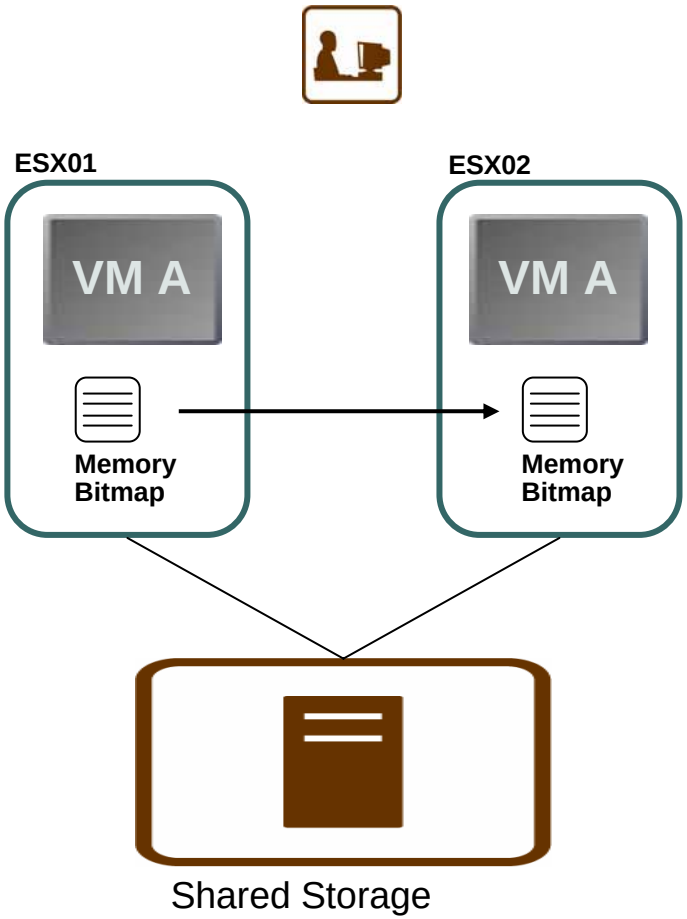


1) 新VMを新しいサーバー上に準備

2) ソースVMからターゲットVMにメモリーのスナップショットをコピー

スナップショット以降のメモリアクセスは、メモリビットマップに記録される。(ダーティページの記憶)

VMotion:動作説明



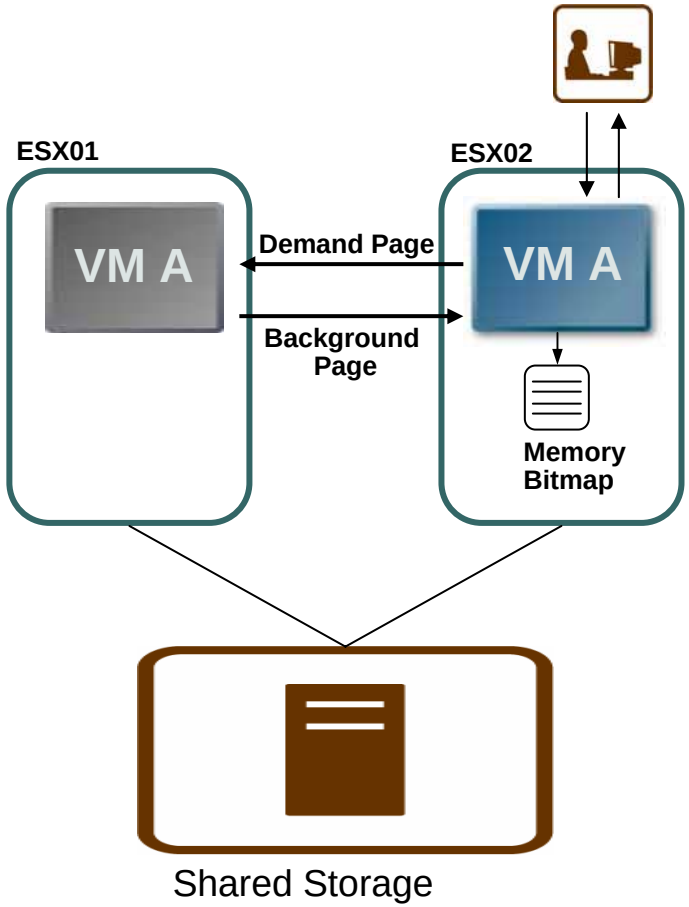
1) 新VMを新しいサーバー上に準備

2) ソースVMからターゲットVMにメモリーのスナップショットをコピー

スナップショット以降のメモリアクセスは、メモリビットマップに記録される。(ダーティページの記憶)

3) ソースVMを動作させたまま、次にメモリビットマップをターゲットVMを稼動させるホストにコピー

VMotion:動作説明



1) 新VMを新しいサーバー上に準備

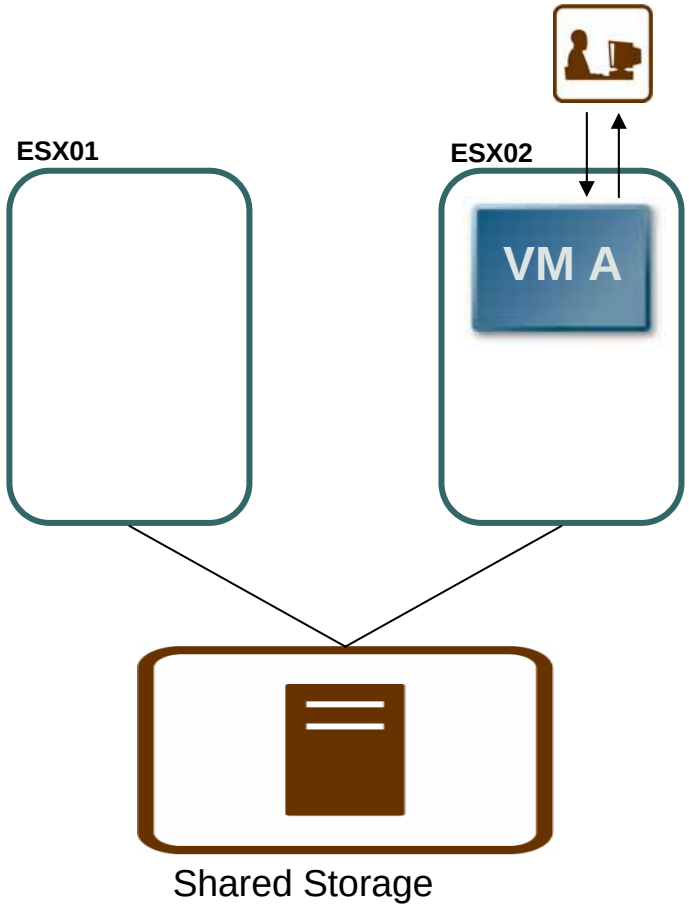
2) ソースVMからターゲットVMにメモリーのスナップショットをコピー
スナップショット以降のメモリアクセスは、メモリビットマップに記録される。(ダーティページの記憶)

3) ソースVMを動作させたまま、次にメモリビットマップをターゲットVMを稼動させるホストにコピー

4) ターゲットホストでVMを稼動

5) VMの稼動に伴いダーティページをアクセスした場合は、オンデマンドに該当ページを旧VMからコピー

VMotion:動作説明



1) 新VMを新しいサーバー上に準備

2) ソースVMからターゲットVMにメモリーのスナップショットをコピー

スナップショット以降のメモリアクセスは、メモリビットマップに記録される。(ダーティページの記憶)

3) ソースVMを動作させたまま、次にメモリビットマップをターゲットVMを稼動させるホストにコピー

4) ターゲットホストでVMを稼動

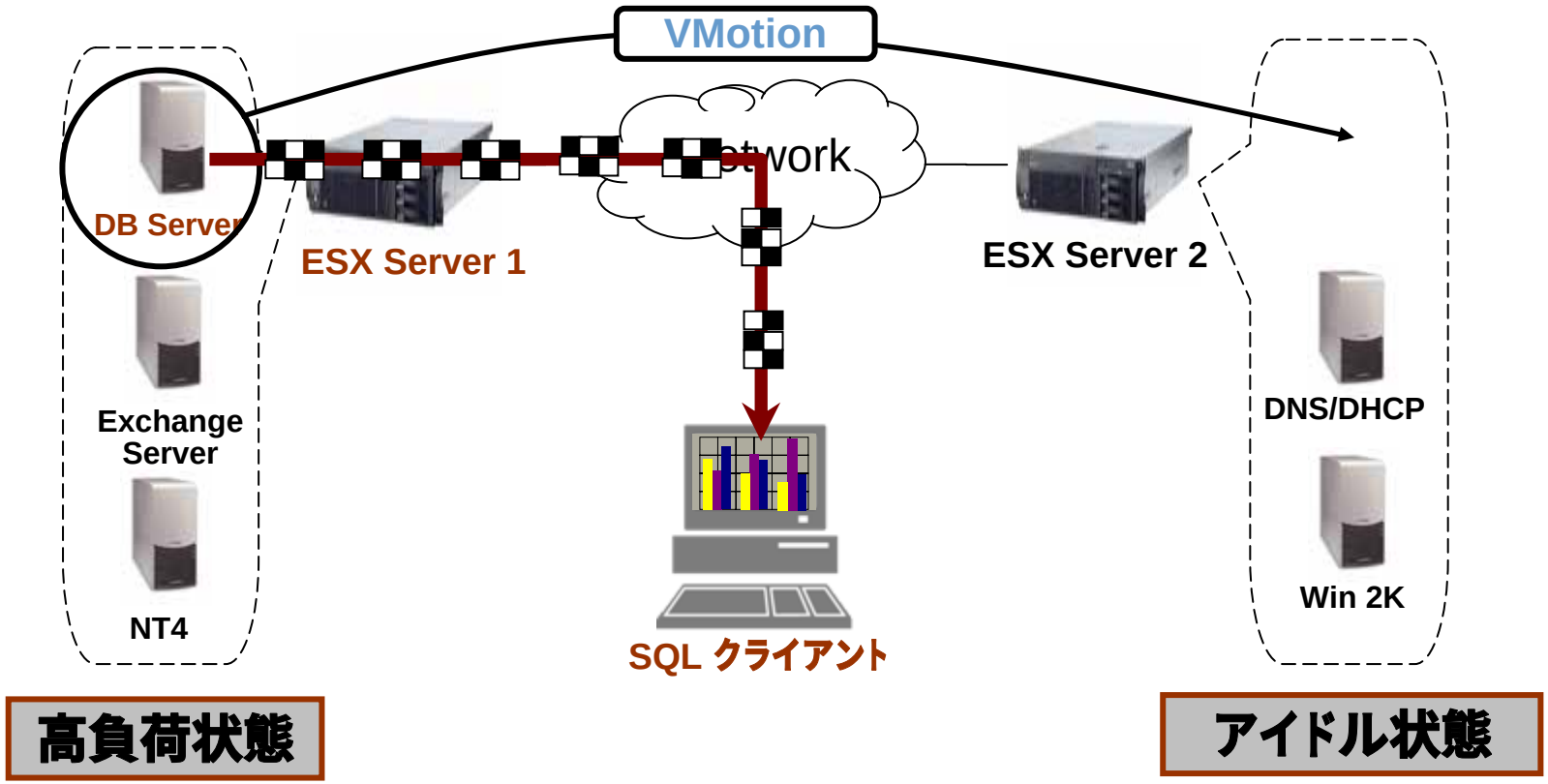
5) VMの稼動に伴いダーティページをアクセスした場合は、オンデマンドに該当ページを旧VMからコピー

6) ソースホストからVMを削除

7) Arpプロトコルを流し、経路情報を更新

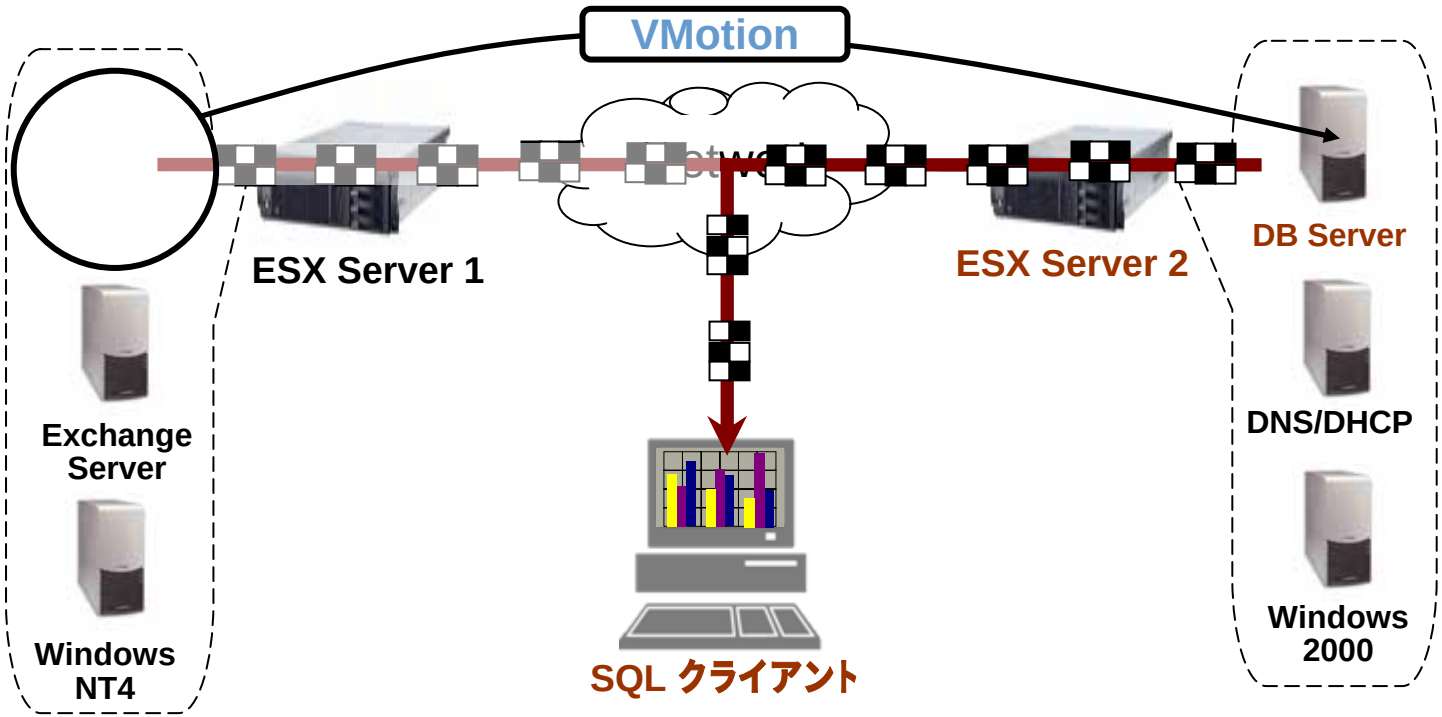
VMotion: DB Serverホットマイグレーション

サービスの停止、トランザクションの停止無し



Demo: Migrate SQL Server

サービスの停止、トランザクションの停止無し

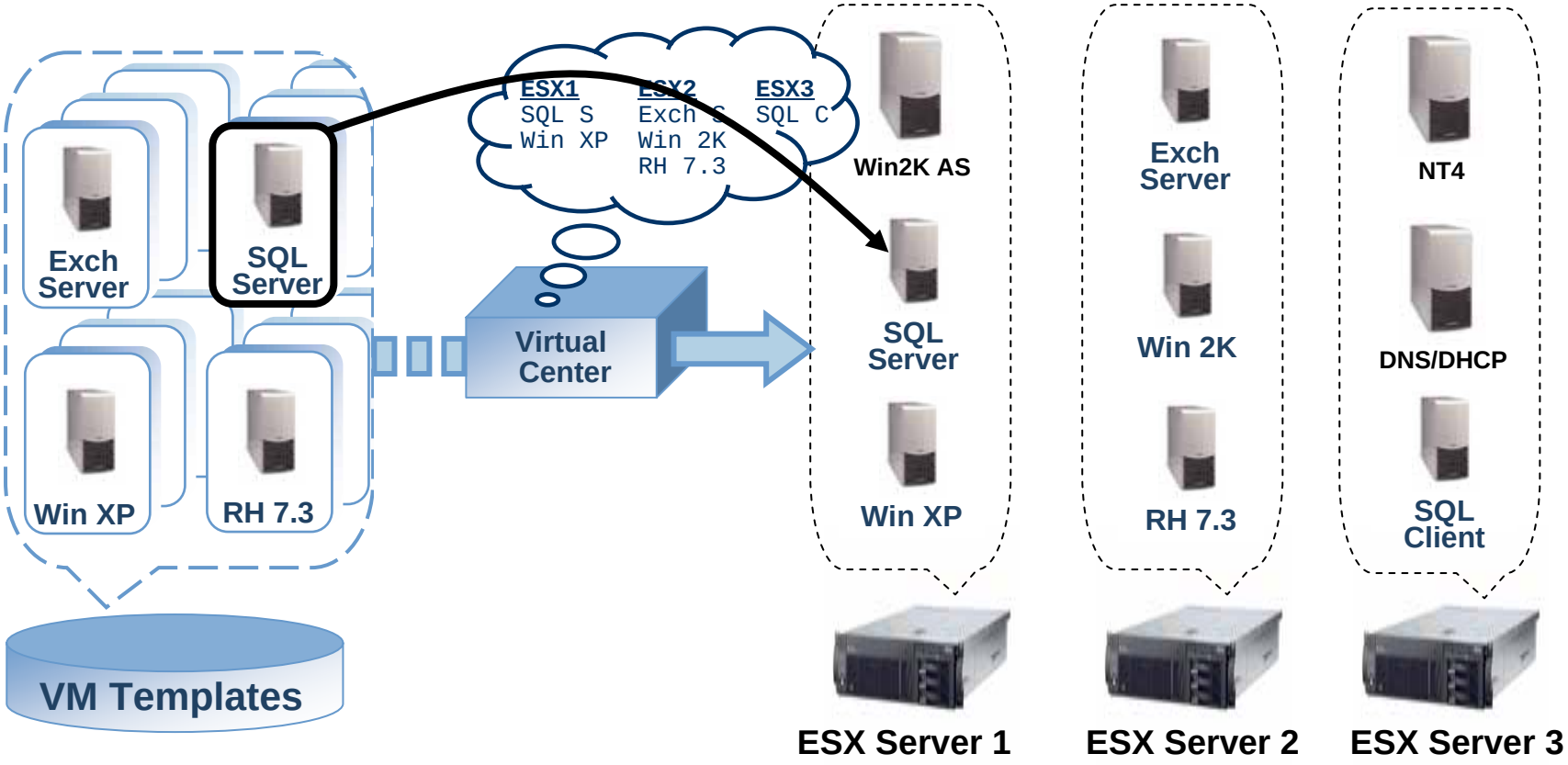


IT投資に求められる迅速性

- 新規システムを立ち上げなければいけないタイミングは、誰にも予測不能
  - 会社のビジネス戦略の方向転換
  - 社内の需要の増大
  - ビジネス環境や法規制の変化
- 通常、新規ハードウェア購入には時間がかかる
  - 購入予算の部門内承認
  - IT部門の承認
- プロビジョニング作業には時間がかかる
  - 社内標準システム規定の則ったプロビジョニング作業
  - セキュリティ対策
  - ライセンス管理

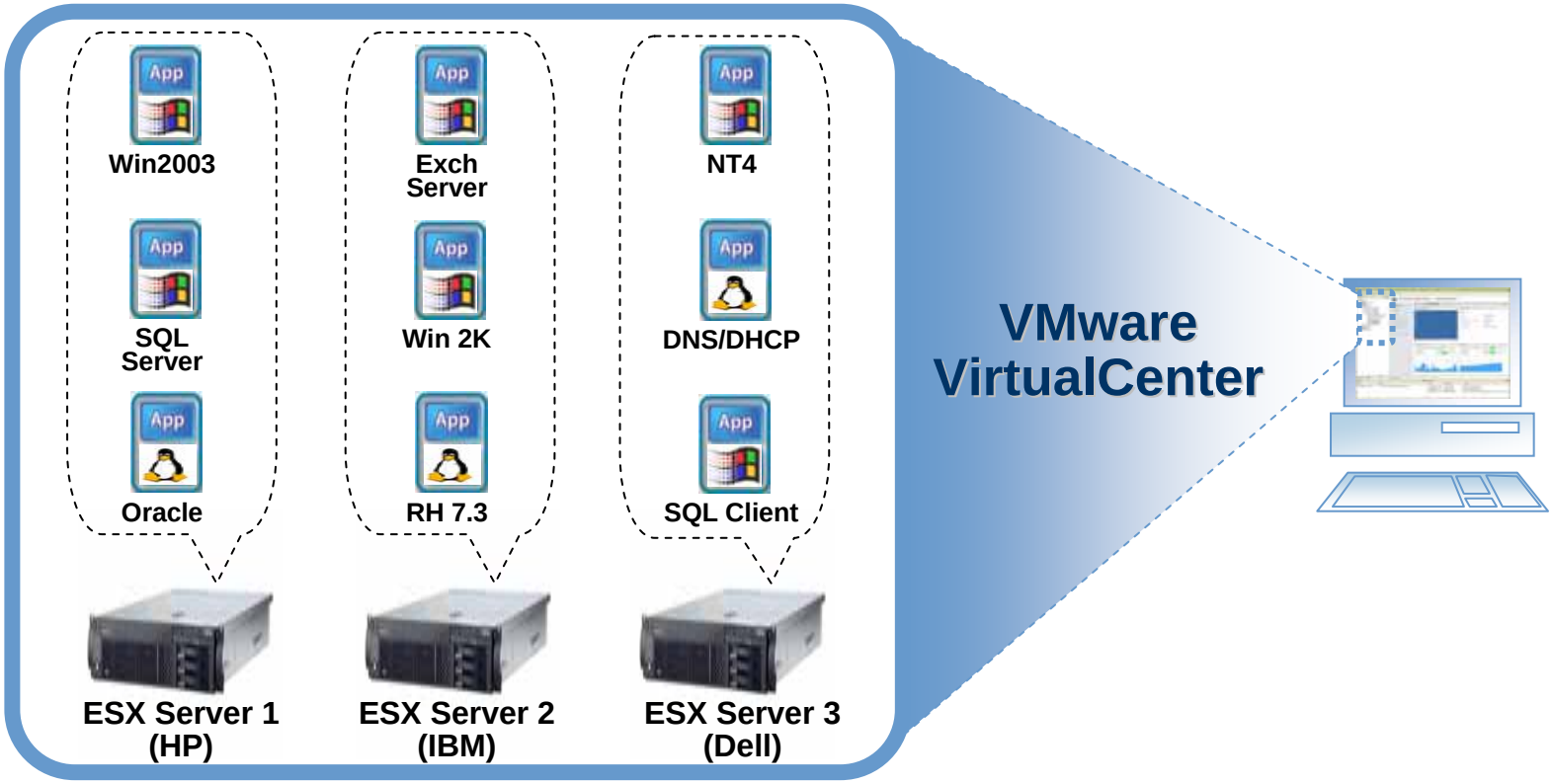
サーバープロビジョニングの自動化

- VirtualCenter 仮想マシンリポジトリから、適切な仮想マシンを選び、自動的に実体化
- サーバープロビジョニングのAutmation化、動的にサーバーを作成･削除可能



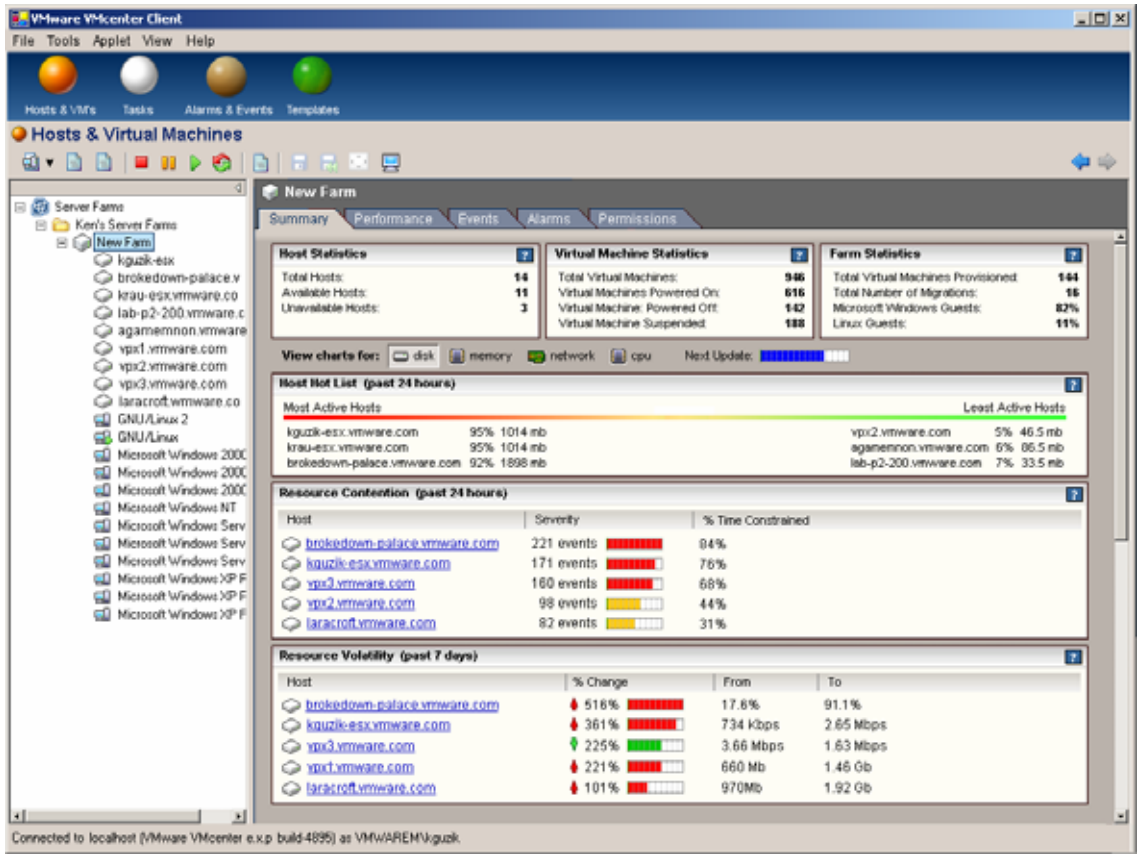
ESXサーバー群、VM群の統合管理

- 一台の端末PC上から、全てのESX Server、及び仮想マシンのコンソールにアクセス
- ヘテロジーニアスな混在環境での統合監視・管理が可能



VMware VirtualCenter の概要

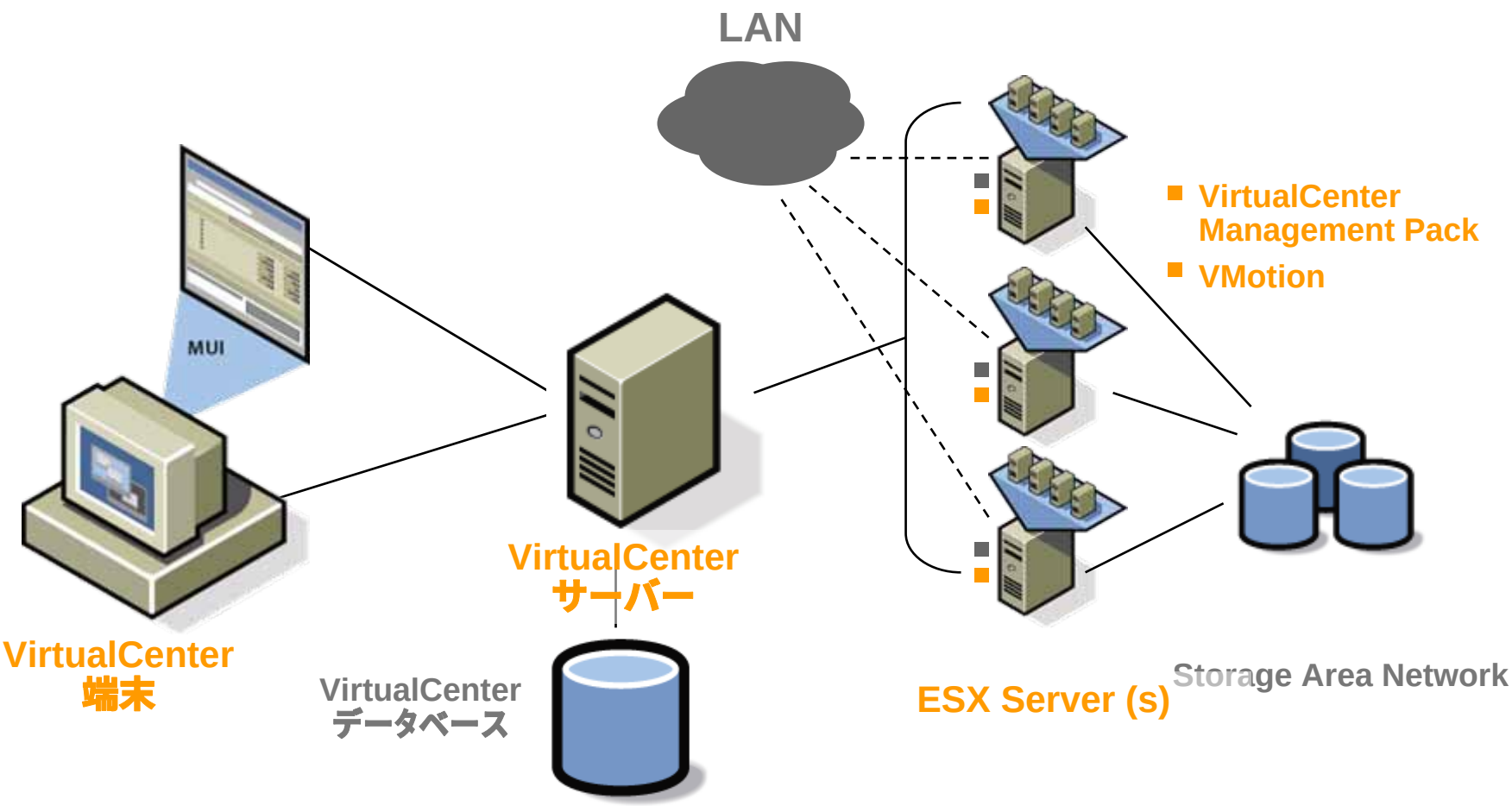
VMware VirtualCenter は、Virtual Machive テクノロジを利用したサーバー群をベースに、真のユーティリティコンピューティングを実現します：



- 分散環境の中で、サーバーの負荷をダイナミックに移動させる
- 主だったサーバー群の運用評価と管理
- 複数のIntelハードウェアを、VMへの単一のリソースプールとして管理
- システムのアベイラビリティやパフォーマンスを計測

vmware

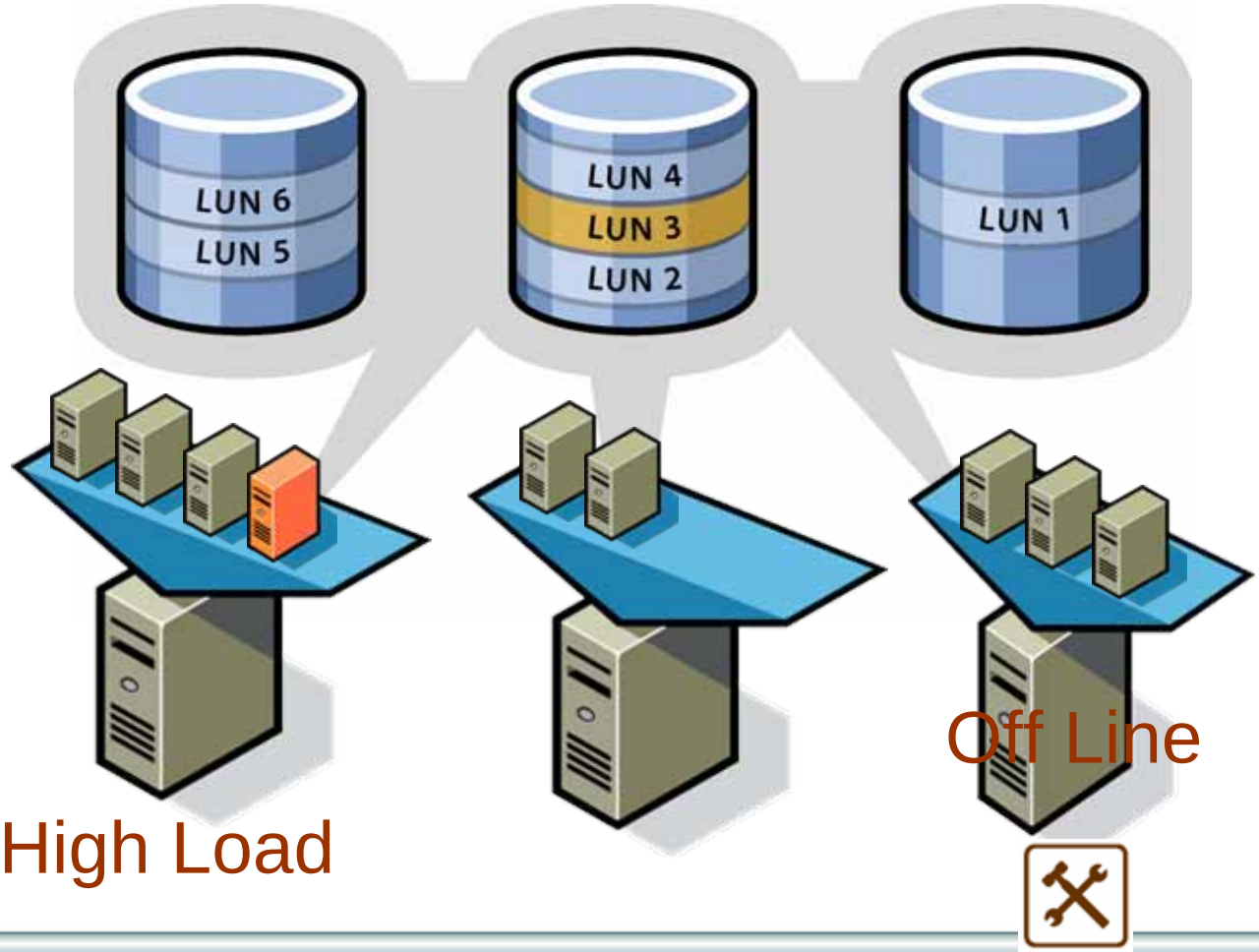
VirtualCenter 構成例



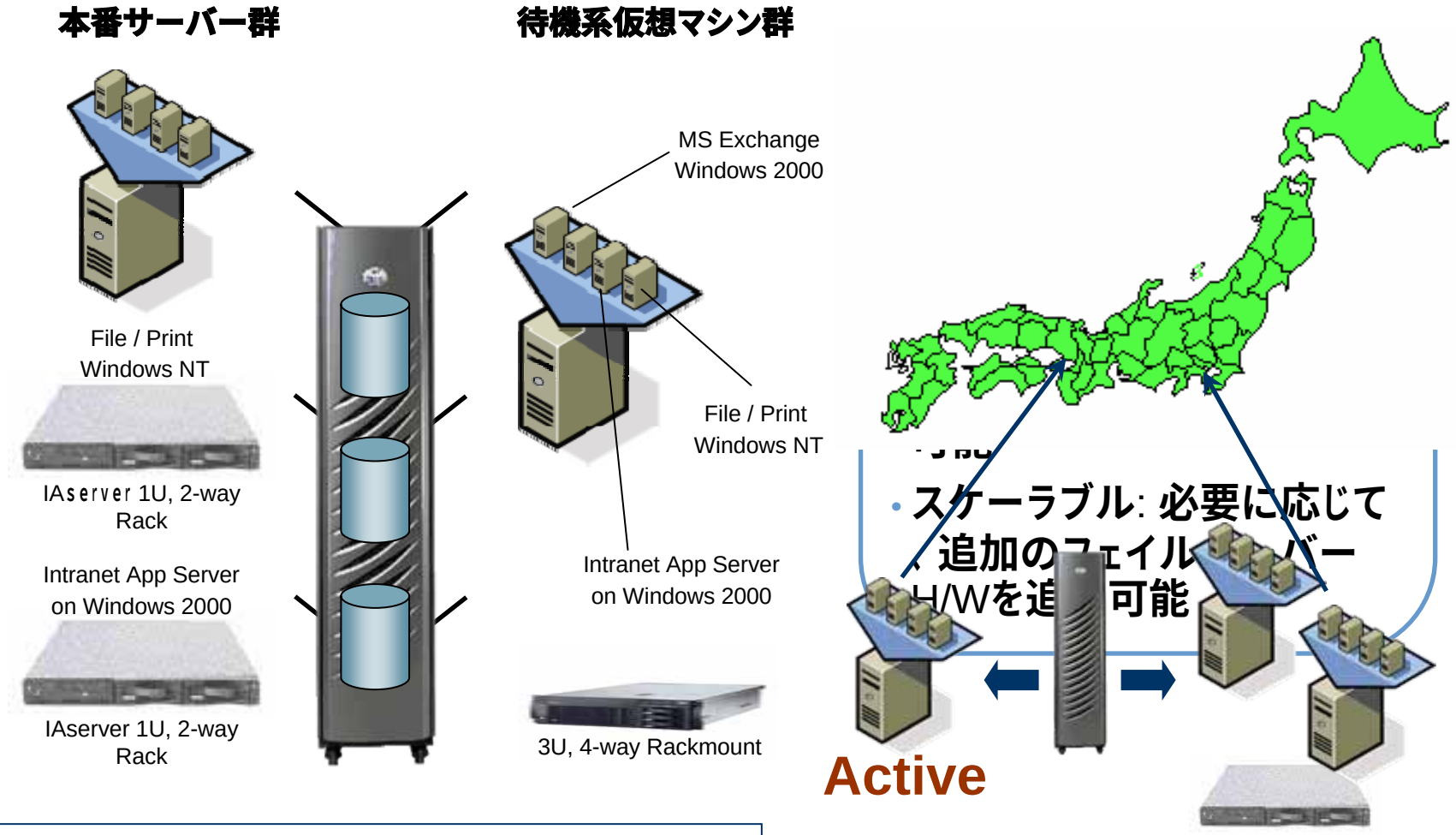
サーバーワークロード 自動管理

- VMware SDK / VMotion でサーバーワークロードの自動管理が可能

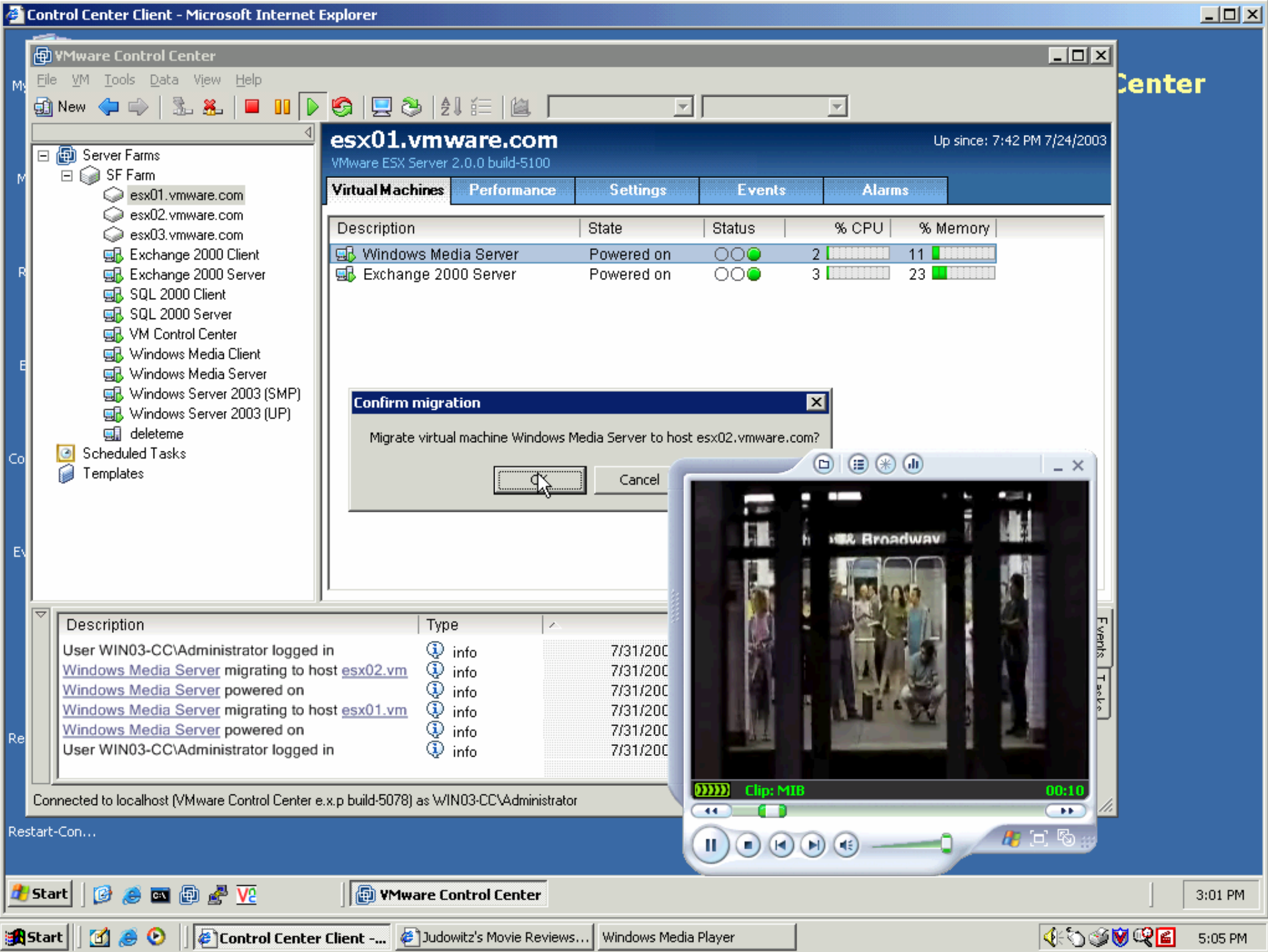
- サービスを停止させることなく、物理サーバーのワークロードを平均化
- 物理サーバー リソースの有効利用
- サービスと非同期にHWメンテナンス作業が可能



VMクラスタ・ディザスタリカバリ構成



- データだけでなくVM自体を遠隔地へミラー


Control Center Client - Microsoft Internet Explorer
VMware Control Center
File VM Tools Data View Help
New
Server Farms
SF Farm
esx01.vmware.com
esx02.vmware.com
esx03.vmware.com
Exchange 2000 Client
Exchange 2000 Server
SQL 2000 Client
SQL 2000 Server
VM Control Center
Windows Media Client
Windows Media Server
Windows Server 2003 (SMP)
Windows Server 2003 (UP)
deleteme
Scheduled Tasks
Templates
esx01.vmware.com
VMware ESX Server 2.0.0 build-5100
Up since: 7:42 PM 7/24/2003
Virtual Machines
Performance
Settings
Events
Alarms
Description
State
Status
% CPU
% Memory
Windows Media Server
Powered on
2
11
Exchange 2000 Server
Powered on
3
23
Confirm migration
Migrate virtual machine Windows Media Server to host esx02.vmware.com?
OK
Cancel
Description
Type
User WIN03-CC\Administrator logged in
info
Windows Media Server migrating to host esx02.vm
Windows Media Server powered on
Windows Media Server migrating to host esx01.vm
Windows Media Server powered on
User WIN03-CC\Administrator logged in
Events
Tasks
Connected to localhost (VMware Control Center e.x.p build-5078) as WIN03-CC\Administrator
Clip: MIB
00:10
Start
VMware Control Center
3:01 PM
Start
Control Center Client -...
Judowitz's Movie Reviews...
Windows Media Player
5:05 PM

Q&A

vmware