仮想化技術の現状と方向性

須崎有康
産業技術総合研究所

発表概要

- 仮想マシンとは？
- 仮想化の問題点とその解決法
  - ソフトウェアのみによる解決とハードウェア支援(Intel VT, AMD-V)による解決
- 仮想化の応用・方向性
- まとめ

＊Intel VT : Intel Virtualization Technology
AMD-V : AMD Virtualization
＊以下、用語説明文責は事務局にあります。

仮想マシンソフトウェアはいくつあるか

- 挙げてみると
  - VMware
  - Xen
  - VirtualPC
  - KVM
  - Parallels
  - VirtualBox
  - QEMU
  - Hyper-V
  - LilyVM
  - Bochs, Plex86, coLinux, sHype, zVM, UserModeLinux, Solaris containers (zones), FreeBSD jail, Linux VServer, Virtuozzo, Cygwin, OpenVZ, SoftwarePod, pearPC, Wine, Tango-VP, Quick Transitive, FAT-Binary, SimOS, Simics, IBM LPAR, ……

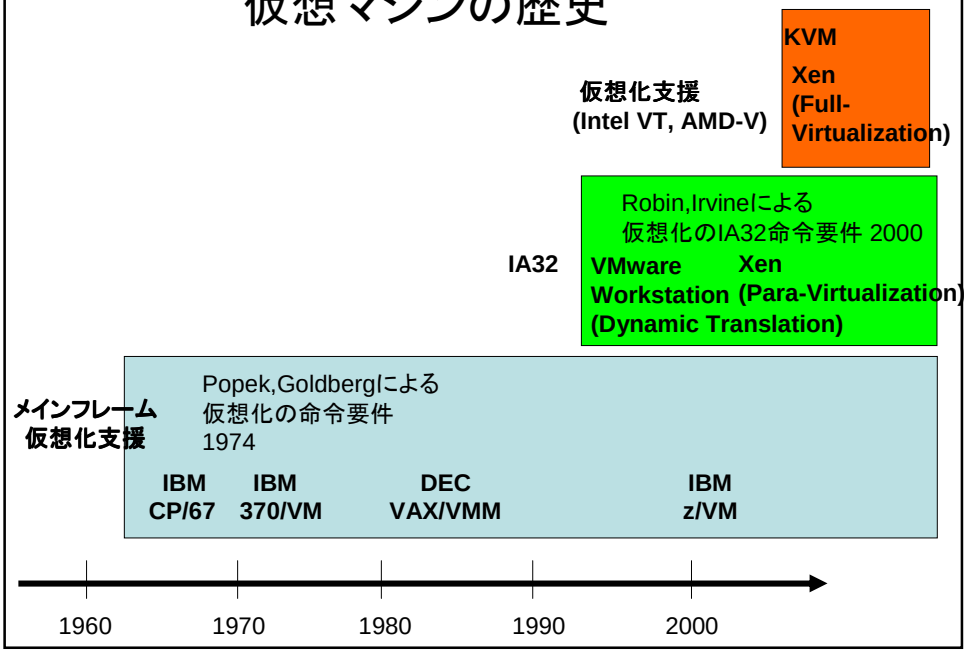
仮想マシンとは

- 仮想機械（かそうきかい）とは、コンピュータのCPUや記憶装置等のリソースを仮想化した上で、その仮想化したコンピュータを実行するソフトウェアである。バーチャルマシン（英：Virtual Machine）とも称される。 (Ja.Wikipedia.orgより)

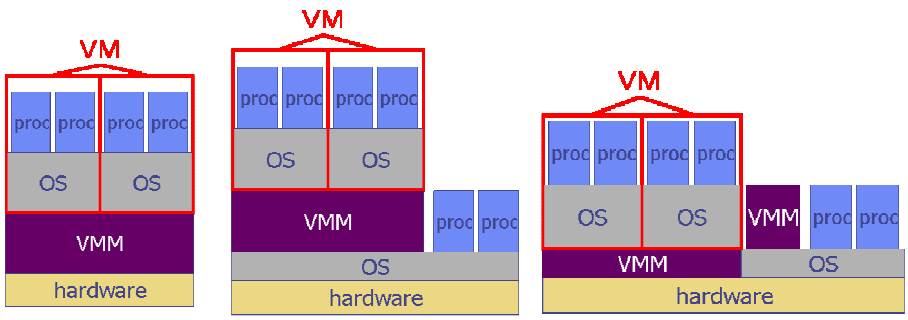
- 全命令、全デバイスをソフトウェアでエミュレートする？
- 「できるだけ実ハードウェアを提供し、できるだけ効率的に仮想化を実現するか」が鍵。
  - 仮想マシンモニタ(VMM, Hypervisor)

＊ VMM : Virtual Machine Monitor

仮想マシンの歴史



仮想マシンモニタ(VMM)の分類



Type Ⅰ
(例:Xen)

Type Ⅱ
(例:QEMU)

Hybrid
(例:VMware Workstation)

IA32での仮想化の問題点

- 仮想化できない命令
  - Sensitiveで且つNon-privileged の命令
    - 状態を保存し、以降の命令実効に影響を与える。
    - 例外で捕捉できない。
- TLBの仮想化
  - TLB flashを制御できない。
  - TLBはハードウェア制御であり、且つ、アドレス空間にID タグが付けられない。
    - SPARCでは仮想化可能。
- IOの仮想化

＊ TLB : Translation Lookaside Buffers

Sensitive&Non-privileged命令の仮想化問題

- Sensitiveな命令
  - CPUの真の姿が見えてしまう命令
    - GDT, LDT, IDTの読出し
  - 以後の実行に影響を与える命令
    - POPF, PUSHF
- Privilegedな命令ならトラップで捕捉することができるが、IA32ではSensitiveで且つNon-privileged命令がある。

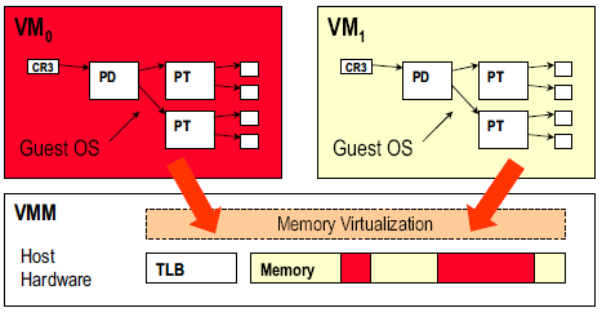
＊ GDT : Global Descriptor Table
LDT : Local Descriptor Table
IDT : Interrupt Descriptor Table
POPF : Pop Flags
PUSHF : Push Flags

Sensitive&Non-privileged命令の解決手法

- ソフトウェアによる解決
  - 静的に命令を置き換える
    - Para-Virtualization (Xen)
  - 動的に命令を置き換える
    - Dynamic Translation (VMWare)
- ハードウェアの補助による解決
  - 仮想化モード対応CPUを使う
    - Intel VT, AMD-V
    - Sensitive、Privileged命令が実行されるとモードが切り替わり、処理をVMMが行なう。

TLBの仮想化問題

- TLB flashを制御できない。
- TLBはハードウェア制御であり、且つ、アドレス空間にIDタグが付けられない。
  - SPARCでは仮想化可能。



TLB問題の解決法

- ソフトウェアによる解決
  - Shadow Paging
    - ゲストOSが利用するページテーブルに対応するページテーブル(Shadow Paging)を用意し、TLBの変換が起こる際にVMMを呼び出し、実ページテーブルとのマッピング、更新管理を行なう。
- ハードウェアの補助による解決
  - TLB切り替え時にフラッシュが起こるのを防ぐ機構。ページテーブルにタグをつけて、ゲストOSの利用を認識できる。
    - RVI: Rapid Virtualization Indexing (AMD)
      - NPT: Nested Page Table (AMD)と呼ばれていた。
    - EPT: Expended Page Table (Intel)

IOの仮想化の問題点

- 効率的なIOを実現できない。
  - Block Device, Network, Video Cardなど効率を要求されるデバイスでは仮想化のオーバーヘッドが無視できない。

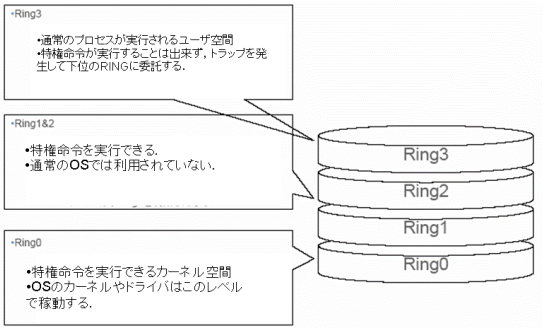
IO問題の解決法

- ソフトウェアによる解決
  - Para-Virtual ドライバ, Passthrough
    - ゲストOSに新たなドライバを入れ、デバイスをダイレクトにアクセスすることができる。
      - VMGL: トロント大学で公開されているOpenGL (ATI, NVIDIA)用のParaVirtualドライバ
- ハードウェアの補助による解決
  - Intel VT-d, AMD-IOMMU
    - ゲスト物理アドレスと実物理アドレスの不一致を解消するDMAのリマッピング機能を提供
    - CPUのみではなく、DMAをコントロールするためチップセットの補助が必要。デバイス側でもリマッピングを認識する必要があるため、PCI-Expressで仕様強化を行なっている。

＊ Intel VT-d : Intel Virtualization Technology for Directed I/O
AMD-IOMMU : AMD I/O Virtualization Technology

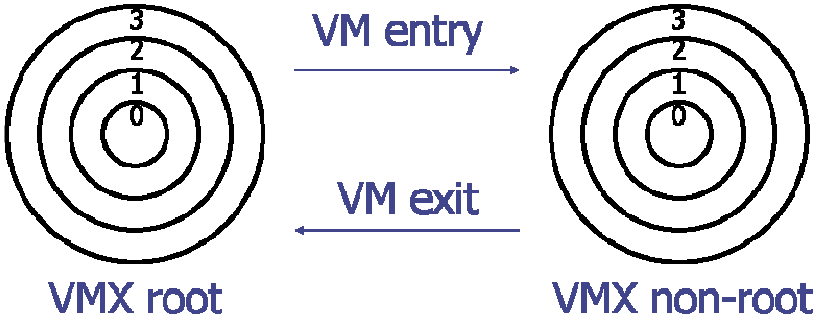
CPUの仮想化対応

- Intel VT, AMD-Vが使えない場合はring構造をしていた。



CPUの仮想化拡張

- 仮想化モードの追加



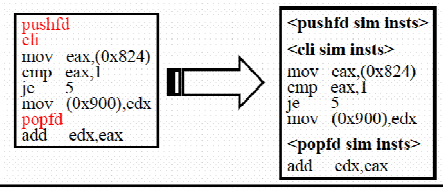
よく使われる仮想マシンソフトウェア

- VMWare
- Xen
- KVM
- QEMU

VMware

- 現在のVM界の巨人
  - Founder: Mendel Rosenblum (Stanford)
  - VMware workstationの成功
  - Dynamic Translation技術を使った仮想マシンモニタ
    - 異なる特権レベルとセグメントを利用したフォールトハンドラ経由の命令変換。仮想化で問題になる特権命令やセンシティブ命令を置き換える変換は再利用される。効率が良い実装。
      - 米国特許：6,397,242, Virtualization system including a virtual machine monitor for a computer with a segmented architecture, (Date of Patent: May 28, 2002, Filed: Oct 26, 1998)
  - VMotion (ESX)
  - バルーンメモリ

命令変換



Xen

- Para-Virtualization (準仮想化)の導入
  - Leader: Ian Pratt (Cambridge)
  - ゲストOSのカーネル内でCPUに影響を与える処理を隠し、VMMに処理を依頼する。
  - Live Migration
- Full-Virtualization
  - Intel VT/ AMD-SVMを使う。
  - 高速IOにはpara virtual driverを使う。
    - Intel VT-d, AMD-IOMMUにも対応

QEMU

- 汎用プロセッサエミュレータ
  - Main developer: Fabrice Bellard
  - 二つのエミュレーション
    - システムエミュレーション
      - 中間コードを介した動的コンパイルを行うことにより、x86、PowerPC、SPARCなどエミュレート可能
    - ユーザエミュレーション
      - 非特権モードのエミュレーションおよび、Lシステムコール命令をネイティブのシステムコール
  - アクセラレータ技術(KQEMU)
    - カーネルモードの仮想化モニタとして動作

KVM: Kernel-based Virtual Machine

- Linux カーネルベースの仮想化技術
  - Main developer: Avi Kivity (Qumranet)
  - Linuxのドライバとして実現
    - /dev/kvm
- Intel VT, AMD-SVMを使う。
- QEMUをインターフェースとして使う

＊ AMD-SVM : AMD Virtualization

仮想化技術を使った応用

- サーバの集約、システムの一元管理
- セキュリティの確保
- 実行環境の移動(マイグレーション)

サーバ集約・一元管理

- 複数のサーバをVMを使うことで1台に集約したり、逆に複数のマシンを1台と仮想化することで効率的な運用を目指す。
  - キャパシティプランニングが必要。

- Xen ベースの開発
  - Virtual Iron
    - 負荷に合わせてリソースキャパシティの供給 。
  - Ganeti by google
    - Cluster VM
  - OracleVM
    - 仮想化環境を含めたデータベース、ミドルウェア、アプリケーションの一元管理。

セキュリティの強化

- VMによる隔離(isolataion)
- 2つの意味の隔離
  1. 汚染が他に影響しない、実行の監視
     - IBMのsHypeはVM間の資源共有、情報の流れをポリシーで制御。
  2. VM内は強固なセキュリティ確保
     - HPのNetTop は機密情報扱うためにOSを隔離。
     - Intel TXT (trusted Execution Technology)はVM起動をTrustedなVMに限定する技術。

＊ NetTop : NetTop is an NSA project to run Multiple Single-Level systems with a Security-Enhanced Linux host running VMware with Windows as a guest operating system. (出典:Wikipedia)

組込みへの仮想化技術的用

- セキュリティの確保の観点から組込み分野にも仮想化技術が開発されている。
  - ARM9用Xen Samsung
  - ARM11用Xen MontaVista

＊ ARM : ARM社開発のAcorn RISC Machine

OSの移動(マイグレーション)

- Liveマイグレーション
  - 実行途中の状態を他へ移動。
    - VMwareのVMotion, XenのLive Migration などVMのサポートが必要。
    - サーバのメンテナンス時退避、ディザスタリカバリなどに役立つ。
    - CMUのISR(Internet Suspend/Resume)、ColumibaのZAPなどの研究もあり。
- Coldマイグレーション
  - **ソフトウェアの新しい配信方法**
  - インストール作業無しに仮想ディスクを取得するだけで色々なOSを試すことができる。
  - サーバでOSのセキュリティアップデート、メンテナンスが可能。

ソフトウェアの配信

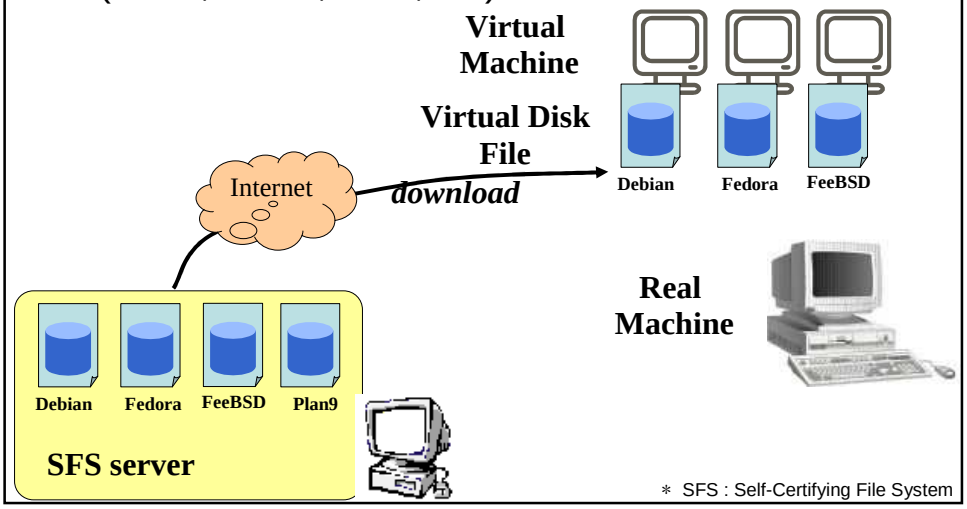
	Requirement	Merit	Size	Flexibility
Virtual Disk	(Xen, KVM, QEMU, etc) Virtual Machine		GB	
Package	Package Manager (yum, apt-get, etc)	**eliminate installation & maintenance? (human power)**	?	?
Source Code	Tool Chain (gcc, make, header files, etc)	**eliminate compilation (CPU power)**	KB	

Cold マイグレーションのプロジェクト

- OS Zoo
  - QEMUの仮想ディスクファイル(Linux, Minix, Plan9, OpenSolaris, etc.)を配布。
- Collective at Stanford
  - VMwareのCOW(CopyOnWrite)ファイルで差分更新を配布。
- Moka5のLivePC
  - Stanford のCollectiveグループが設立したベンチャー企業
  - VMWare Playerの仮想ディスクをストリームダウンロード。
    - HTTPベース、マルチコネクション
- OS Circular
  - 産総研 で開発されているOS配信プロジェクト。
  - Xen/QEMU/KVMを対象にしたネットワーク仮想ディスク ***Trusted HTTP-FUSE CLOOP***を開発
  - ディスクイメージの必要な部分のみをダウンロード。差分更新も可能。

仮想ディスクの共有(OS Zoo)

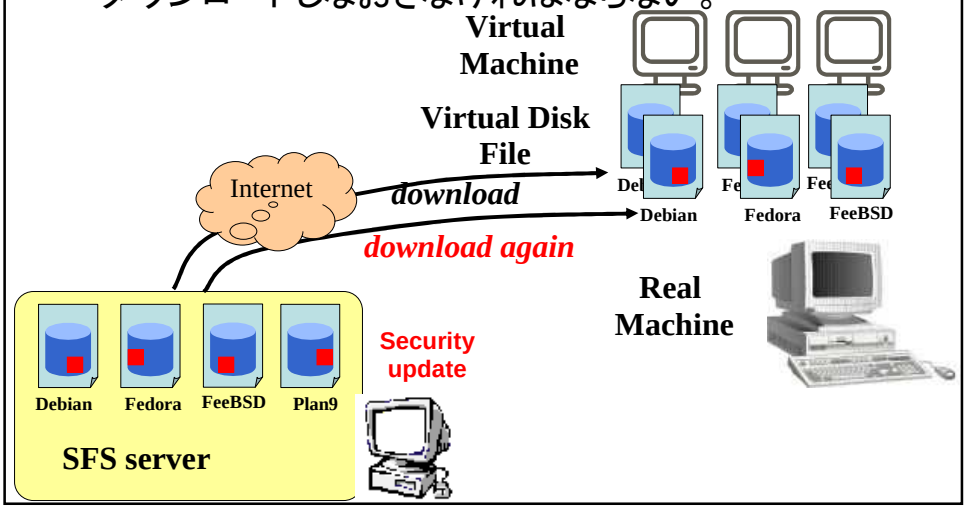
- *OS Zoo project* ではQEMU用のディスクファイル(Linux, Plan9, BSD, etc)を配布。



＊ SFS : Self-Certifying File System

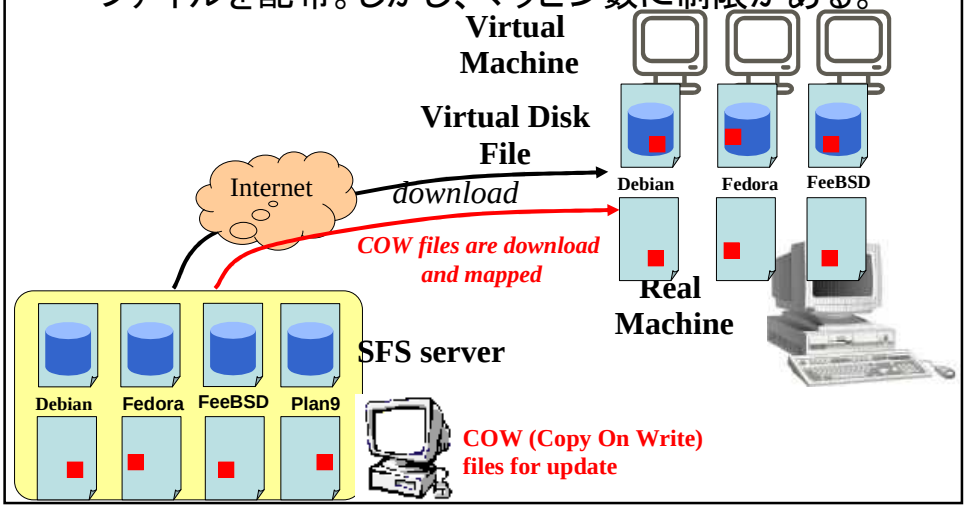
仮想ディスクの共有(OS Zoo)

- このタイプでは更新があった場合、全てのファイルをダウンロードしなおさなければならない。



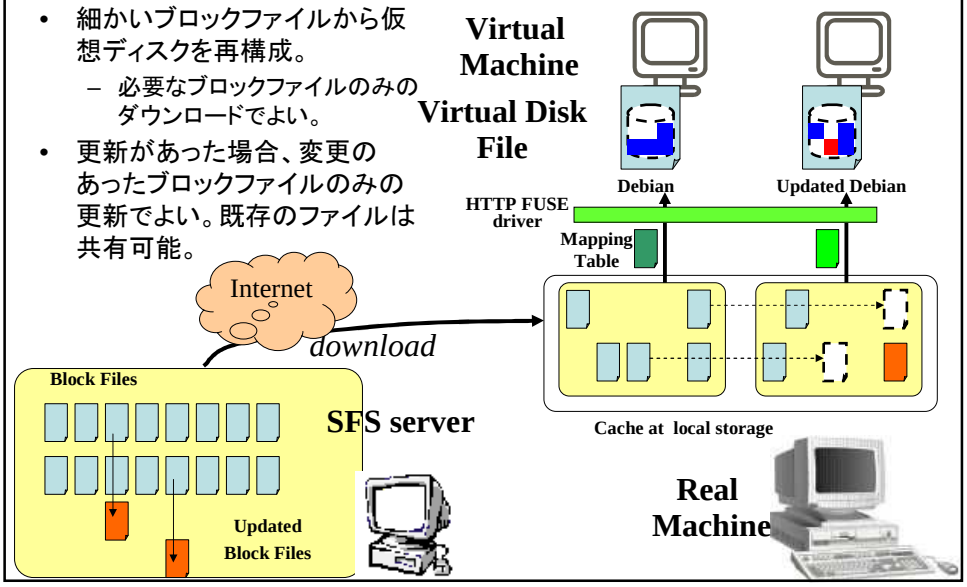
仮想ディスクの共有(Collective)

- *Collective project* ではアップデートのためにCOWファイルを配布。しかし、マッピン数に制限がある。



仮想ディスクの共有(OS Circular)

- 細かいブロックファイルから仮想ディスクを再構成。
  - 必要なブロックファイルのみのダウンロードでよい。
- 更新があった場合、変更のあったブロックファイルのみの更新でよい。既存のファイルは共有可能。



その他の仮想化のメリット

- 手元のPCには無い機能を試す
  - BIOS
    - EFI (Extensible Firmware Interface)
      - 次期BIOS規格。Intel Macで利用。
      - QEMU(IA32), Xen(IA64)用のEFI はTristan Gingold が移植。
    - LinuxBIOS, OpenFirmware
  - TPMによるTrusted Computing
    - Trusted BootはBIOSに依存する
    - XenではTPM-Emulator を利用した Trusted Boot環境を提供。
      - TCG BIOSはIBMのStefan Berger が作成。

国内の動向

- SecureVM (科学技術振興調整費 H18-H20)
  - 政府機関(内閣官房情報セキュリティセンターなど)を対象とし、現在のシステム環境(Windows)を維持しつつ、高度なセキュリティ機能が確保するVMMの開発。
- セキュアプラットフォームプロジェクト(経済産業省 H19-H21)
  - VMを使ったサーバ集約による省エネ、セキュリティ向上のための統合アクセス制御機構の開発

＊ SecureVMプロジェクト：
http://www.securevm.org/
セキュアプラットフォーム推進コンソーシアム
http://spf.jeita.or.jp/library_files/01-071019.pdf

抑えておきたい国際会議

- XenSummit
  - 年2回(Spring, Fall)
  - AMD, Intel, IBM, SUNなどがホストを持ち回り
- Linux関連
  - Kenel Summit (09は日本)
  - Linux.Conf.AU (オーストラリア)
  - Ottawa Linux Symposium (カナダ)
  - LinuxKonogress (ドイツ)
- ACM/IEEE
  - VEE: International Conference on Virtual Execution Environments
  - OSDI, SOSP

まとめ

- 仮想化技術はハードウェア、ソフトウェア双方から開発が進められ、現在ではかなり効率的なものになっている。
- 仮想化によりサーバの集約、セキュリティの確保、実行環境の移動、システムの一元管理などの新しい応用が進められている。