

Lernziele dieses Foliensatzes

- Am Ende dieses Foliensatzes kennen/verstehen Sie...
 - was **Virtualisierung** ist, ihre Vorteile, Nachteile und Grenzen
 - den Unterschied zwischen **Hardware-Emulation** und **Virtualisierung**
 - verschiedene **Virtualisierungskonzepte**:
 - **Partitionierung**
 - **Anwendungsvirtualisierung**
 - **Vollständige Virtualisierung** (Virtueller Maschinen-Monitor)
 - **Paravirtualisierung** (Hypervisor)
 - **Hardware-Virtualisierung**
 - **Betriebssystem-Virtualisierung** bzw. **Container** bzw. **Jails**
 - **Speichervirtualisierung (SAN)**
 - **Netzwerkvirtualisierung (VLAN)**

Übungsblatt 10 wiederholt die für die Lernziele relevanten Inhalte dieses Foliensatzes



Ursprung der Virtualisierung

- Virtualisierung ist kein neues Konzept
 - Einführung in den 1960er-Jahren bei Großrechnern
- 1970/71: IBM stellt die Virtual Machine Facility/370, kurz VM/370 vor
 - Auf dieser Plattform wurde Mehrbenutzerbetrieb realisiert, indem mehrere Einzelbenutzerbetriebinstanzen in virtuellen Maschinen ausgeführt wurden
 - Jede VM stellte eine vollständige Nachbildung der darunter liegenden, physischen Hardware dar

Quellen

- Creasy RJ. **The origin of the VM/370 time-sharing system.** IBM Journal of Research and Development 25 (1981), No. 5, 483–490
- Amit Singh. **An Introduction to Virtualization.** 2004
<http://www.kernelthread.com/publications/virtualization/>

Beispiel für Partitionierung – Watson (1/2)

- Sieger beim US-Quiz *Jeopardy Challenge* im Februar 2011 war *Watson*
 - Watson ist ein Cluster aus 90 IBM Power 750 Servern mit 2.880 Power7 CPU-Kernen (je 8 Kerne pro CPU) und 16 TB RAM



Bildquelle (Watson stage replica in Jeopardy! contest, Mountain View, California): Atomic Taco. flickr.com (CC-BY-SA-2.0)
 Bildquelle (Interns demonstrating Watson capabilities in Jeopardy! exhibition match): Rosemaryetoufee. Wikimedia (CC-BY-SA-4.0)

Beispiel für Partitionierung – Watson (2/2)

- Auf den 90 Knoten können Partitionen erstellt werden
 - In jeder Partition kann ein AIX, Linux oder IBM i (früher OS/400) laufen
 - Die Partitionen sind unabhängige Installationen
 - In jeder Partition kann ein unterschiedliches Betriebssystem laufen
- Auf jedem Knoten läuft ein *POWER Hypervisor*
 - Er regelt den Hardwarezugriff
- Seit Power6 kann man laufende Partitionen ohne Unterbrechung auf andere physische Server umziehen (\implies Live Partition Mobility)
- Man kann Partitionen auch erlauben, dass sie sich Hauptspeicher teilen (\implies Active Memory Sharing)
 - Active Memory Expansion kann Speicherseiten komprimieren
 - Je nach Anwendung geht das schneller als Verschieben oder Auslagern



← → ↻ bellard.org/jslinux/

JSLinux

Run Linux or other Operating Systems in your browser!

The following emulated systems are available:

CPU	OS	User Interface	YFsync access	Startup Link	TEMU Config	Comment
x86	Alpine Linux 3.12.0	Console	Yes	click here	url	
x86	Alpine Linux 3.12.0	X Window	Yes	click here	url	Right mouse button for the menu.
x86	Windows 2000	Graphical	No	click here	url	Disclaimer .
x86	FreeDOS	VGA Text	No	click here	url	
risev64	Buildroot (Linux)	Console	Yes	click here	url	
risev64	Buildroot (Linux)	X Window	Yes	click here	url	Right mouse button for the menu.
risev64	Fedora 29 (Linux)	Console	Yes	click here	url	Warning: longer boot time.
risev64	Fedora 29 (Linux)	X Window	Yes	click here	url	Warning: longer boot time. Right mouse button for the menu.

© 2011-2020 Fabrice Bellard - [News](#) - [VM list](#) - [FAQ](#) - [Technical notes](#)

← → ↻ bellard.org/jslinux/vm.html?url=alpine-x86.cfg&mem=192

```

Loading...
Welcome to JS/Linux (i586)

Use 'vflogin username' to connect to your account.
You can create a new account at https://vfsync.org/signup .
Use 'export file filename' to export a file to your computer.
Imported files are written to the home directory.

localhost:~# uname -a
Linux localhost 4.12.0-rc6-g48ec1f0-dirty #21 Fri Aug 4 21:02:28 CEST 2017 i586
Linux
localhost:~# █
  
```

- Seit 2011 hat der Autor von JSLinux den Funktionsumfang stark erweitert

Bild oben rechts: FreeDOS 1.2 (x86)

Bild unten links: Alpine Linux 3.12.0 (x86)

Bild unten rechts: Windows 2000 (x86)

← → ↻ bellard.org/jslinux/vm.html?url=freedos.cfg&mem=64&graphic=1&w=720&h=400

```

Type HELP to get support on commands and navigation.

Welcome to the FreeDOS 1.2 operating system (http://www.freedos.org)

Use KEYB to set the keyboard mapping (e.g. KEYB FR for a French keyboard)

C:\>mem

Memory Type          Total      Used      Free
-----
Conventional          639K       33K       606K
Upper                 0K         0K         0K
Reserved             385K       385K       0K
Extended (XMS)       64,512K    320K      64,192K
-----
Total memory         65,536K    746K     64,790K

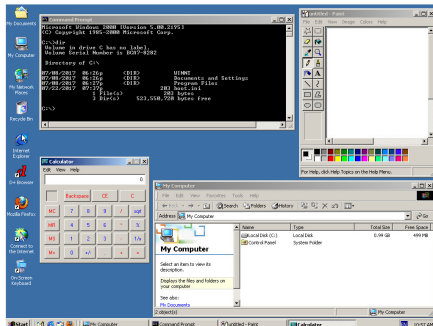
Total under 1 MB     639K       33K       606K

Total Expanded (EMS)                8,688K (8,896,512 bytes)
Free Expanded (EMS)                8,192K (8,388,608 bytes)

Largest executable program size      606K (620,080 bytes)
FreeDOS is resident in the high memory area.

C:\>
  
```

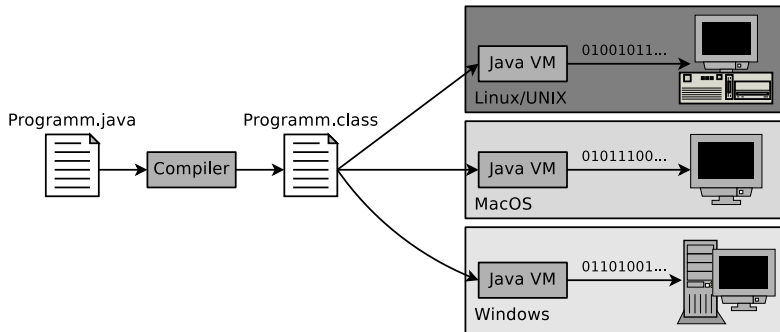
← → ↻ bellard.org/jslinux/vm.html?url=win2k.cfg&mem=192&graphic=1&w=1024&h=768



Anwendungsvirtualisierung

- Anwendungen werden unter Verwendung lokaler Ressourcen in einer virtuellen Umgebung ausgeführt, die alle Komponenten bereitstellt, die die Anwendung benötigt
 - VM befindet sich zwischen der auszuführenden Anwendung und dem Betriebssystem
- Populäres Beispiel: Java Virtual Machine (JVM)
 - Die JVM ist der Teil der Java-Laufzeitumgebung (JRE), der für die Ausführung des Java-Bytecodes verantwortlich ist
 - Die JVM ist für Java-Programme die Schnittstelle zum Rechnersystem und dessen Betriebssystem
- Vorteil: Plattformunabhängigkeit
- Nachteil: Geringere Ausführungsgeschwindigkeit gegenüber nativer Programmausführung

Prinzip der Java Virtual Machine (JVM)



- Der Compiler `javac` übersetzt Quellcode in architektur-unabhängige `.class`-Dateien, die Bytecode enthalten, der in der Java VM lauffähig ist
- Das `java`-Programm startet eine Java-Anwendung in einer Instanz der Java VM

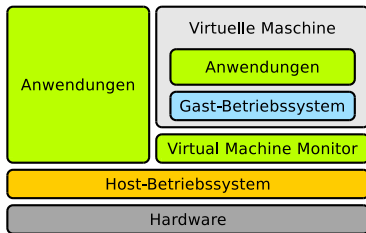
VMware ThinApp

<http://www.vmware.com/products/thinapp/>

- Weiteres Beispiel für Anwendungsvirtualisierung: VMware ThinApp
 - Bis 2008 unter dem Namen Thinstall bekannt
- Eine Windows-Anwendung wird in eine einzelne .exe-Datei gepackt
- Die Anwendung wird dadurch portabel und kann ohne lokale Installation verwendet werden
 - Die Anwendung kann u.a. auf einem USB-Stick ausgeführt werden
- Es erfolgen keine Einträge in der Windows Registry. Es werden auch keine Umgebungsvariablen und DLL-Dateien auf dem System erstellt
- Benutzereinstellungen und erstellte Dokumente werden in einer eigenen Sandbox gespeichert
- Nachteil: Funktioniert ausschließlich mit Microsoft Windows

Vollständige Virtualisierung (1/3)

- Vollständige Virtualisierungslösungen bieten einer VM eine vollständige, virtuelle PC-Umgebung inklusive eigenem BIOS
 - Jedes Gastbetriebssystem erhält eine eigene VM mit virtuellen Ressourcen (u.a. CPU, Hauptspeicher, Laufwerken, Netzwerkkarten)
- Es kommt ein **Virtueller Maschinen-Monitor (VMM)** zum Einsatz
 - Der VMM heißt auch **Typ-2-Hypervisor**
 - Der VMM läuft *hosted* als Anwendung im Host-Betriebssystem
 - Der VMM verteilt Hardwareressourcen an VMs
- Teilweise emuliert der VMM Hardware, die nicht für den gleichzeitigen Zugriff mehrerer Betriebssysteme ausgelegt ist
 - Ein Beispiel sind Netzwerkkarten
 - Die Emulation populärer Hardware vermeidet Treiberprobleme



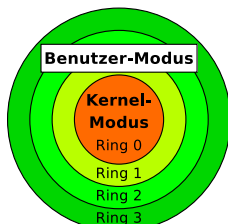
Virtualisierungsgrundlagen der x86-Architektur (1/2)

- x86-kompatible CPUs enthalten 4 Privilegienstufen
 - Ziel: Stabilität und Sicherheit verbessern
 - Jeder Prozess wird in einem Ring ausgeführt und kann sich nicht selbstständig aus diesem befreien

Realisierung der Privilegienstufen

- Das Register CPL (Current Privilege Level) speichert die aktuelle Privilegienstufe
- Quelle: Intel 80386 Programmer's Reference Manual 1986
<http://css.csail.mit.edu/6.858/2012/readings/i386.pdf>

- In Ring 0 (= **Kernelmodus**) läuft der Betriebssystemkern
 - Hier haben Prozesse vollen Hardwarezugriff
 - Der Kern kann physischen Speicher (\implies Real Mode) adressieren
- In Ring 3 (= **Benutzermodus**) laufen die Anwendungen
 - Hier arbeiten Prozesse nur mit virtuellem Speicher



Keine Virtualisierung



Systemaufruf →

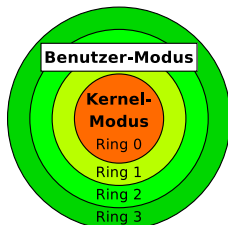
Virtualisierungsgrundlagen der x86-Architektur (2/2)

Moderne Betriebssysteme verwenden nur 2 Privilegienstufen (Ringe)

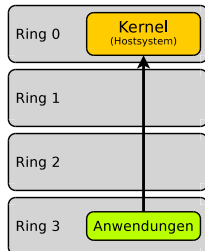
- Grund: Einige Hardware-Architekturen (z.B. Alpha, PowerPC, MIPS) unterstützen nur 2 Stufen
- Ausnahme: OS/2 nutzt Ring 2 für Anwendungen, die auf Hardware und Eingabe-/Ausgabeschnittstellen zugreifen dürfen (z.B. Grafiktreiber)

- Muss ein Prozess im Benutzermodus eine höher privilegierte Aufgabe erfüllen (z.B. Zugriff auf Hardware), kann er das dem Kernel durch einen **Systemaufruf** (\implies Foliensatz 7) mitteilen

- Der Prozess im Benutzermodus erzeugt eine Exception, die in Ring 1 abgefangen und dort behandelt wird



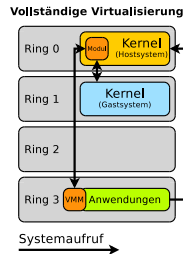
Keine Virtualisierung



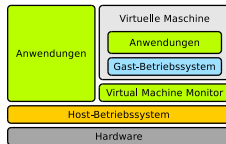
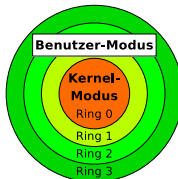
Systemaufruf →

Vollständige Virtualisierung (2/3)

- Vollständige Virtualisierung nutzt die Tatsache, dass x86-Systeme meist nur 2 Privilegienstufen verwenden
 - Der VMM läuft wie die Anwendungen in Ring 3
 - VMs befinden sich im weniger privilegierten Ring 1



- Der VMM enthält für jede Ausnahme eine Behandlung, die privilegierte Operationen der Gastbetriebssysteme abfängt, interpretiert und ausführt



- VMs erhalten nur über den VMM Zugriff auf die Hardware
 - Garantiert kontrollierten Zugriff auf gemeinsam genutzte Systemressourcen

Vollständige Virtualisierung (3/3)

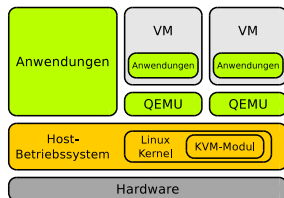
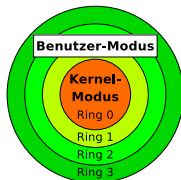
- Vorteile:
 - Kaum Änderungen an Host- und Gast-Betriebssystemen erforderlich
 - Zugriff auf die wichtigsten Ressourcen wird nur durchgereicht
⇒ Fast native Verarbeitungsgeschwindigkeit der Gast-Betriebssysteme
 - Jedes Gast-Betriebssystem hat seinen eigenen Kernel
⇒ Hohe Flexibilität
- Nachteile:
 - Wechsel zwischen den Ringen erfordern einen Prozesswechsel
⇒ Jeder Prozesswechsel verbraucht Rechenzeit
 - Fordert eine Anwendung im Gast-Betriebssystem die Ausführung eines privilegierten Befehls an, liefert der VMM eine Ersatzfunktion und diese weist die Ausführung des Befehls über die Kernel-API des Host-Betriebssystems an
⇒ Geschwindigkeitseinbußen

Beispiele für Vollständige Virtualisierung

- Beispiele für Virtualisierungslösungen, die auf dem Konzept des VMM basieren:
 - VMware Server, VMware Workstation und VMware Fusion
 - Microsoft Virtual PC (in der Version für x86)
 - Parallels Desktop und Parallels Workstation
 - VirtualBox
 - Kernel-based Virtual Machine (KVM)
 - Mac-on-Linux (MoL)

Kernel-based Virtual Machine (KVM)

- KVM ist als Modul direkt im Linux-Kernel integriert
 - KVM-Basismodul: `kvm.ko`
 - Hardwarespezifische Module: `kvm-intel.ko` und `kvm-amd.ko`



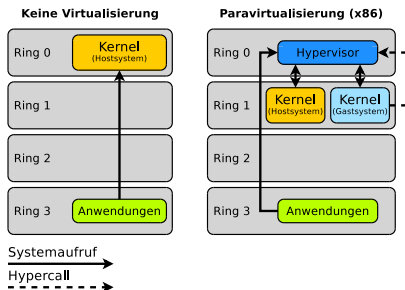
- Nach dem Laden der Module arbeitet der Kernel selbst als Hypervisor
- KVM kann nur mit CPUs mit Hardwarevirtualisierung arbeiten
 - Dadurch braucht KVM weniger Quellcode als z.B. Xen
- Neben den Kernelmodulen enthält KVM den Emulator QEMU
 - KVM stellt keine virtuelle Hardware zur Verfügung. Das macht QEMU
 - CPU-Virtualisierung stellt der Prozessor bereit (Intel VT oder AMD-V)
 - Hauptspeicher und Blockspeicher wird durch KVM virtualisiert
 - E/A wird durch einen QEMU-Prozess pro Gastsystem virtualisiert

Paravirtualisierung (1/4)

- Es wird keine Hardware virtualisiert oder emuliert
 - Gast-Betriebssystemen steht keine emulierte Hardwareebene, sondern eine API zur Verfügung
- Gast-Betriebssysteme verwenden eine abstrakte Verwaltungsschicht (⇒ **Hypervisor**), um auf physische Ressourcen zuzugreifen
 - Hypervisor ist ein auf ein Minimum reduziertes **Metabetriebssystem**
 - Der Hypervisor verteilt Hardwareressourcen unter den Gastsystemen, so wie ein Betriebssystem dieses unter den laufenden Prozessen tut
 - Der Hypervisor ist ein **Typ-1-Hypervisor** und läuft *bare metal*
 - Ein Metabetriebssystem ermöglicht den unabhängigen Betrieb unterschiedlicher Anwendungen und Betriebssysteme auf einer CPU
- Der Hypervisor läuft im privilegierten Ring 0
 - Das Host-Betriebssystem läuft im weniger privilegierten Ring 1
 - Ein Host-Betriebssystem ist wegen der Gerätetreiber nötig

Paravirtualisierung (2/4)

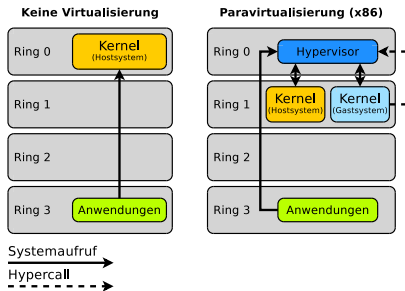
- Das Host-Betriebssystem läuft nicht mehr in Ring 0, sondern in Ring 1
 - Darum kann der Kernel keine privilegierten Anweisungen ausführen
 - Lösung: Der Hypervisor stellt **Hypercalls** zur Verfügung



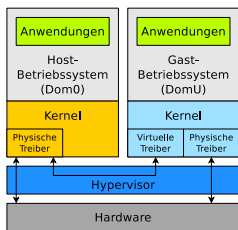
- Hypercalls sind vergleichbar mit Systemaufrufen (System Calls)
 - Die Interrupt-Nummern sind verschieden
 - Fordert eine Anwendung die Ausführung eines Systemaufrufs an, wird eine Ersatzfunktion im Hypervisor aufgerufen
 - Der Hypervisor weist die Ausführung des Systemaufrufs über die Kernel-API des Betriebssystems an

Paravirtualisierung (3/4)

- In den Betriebssystemkernen der Gast-Betriebssysteme müssen alle Systemaufrufe für Hardware-Zugriffe durch die entsprechenden Hypercall-Aufrufe ersetzt werden
- Das Abfangen und Prüfen der Systemaufrufe durch den Hypervisor führt nur zu geringen Geschwindigkeitseinbußen
- Beispiele: Xen, Citrix Xenserver, Virtual Iron, VMware ESX Server



Paravirtualisierung (4/4)



- VMs heißen **unprivilegierte Domain (DomU)**
- Der Hypervisor ersetzt das Host-Betriebssystem
 - Die Entwickler können aber nicht alle Treiber selbst schreiben und pflegen
 - Darum startet der Hypervisor eine (Linux-)Instanz mit ihren Treibern und leiht sich diese Treiber
 - Diese spezielle Instanz heißt Domain0 (Dom0)

• Nachteile:

- Kernel der Gast-Betriebssysteme müssen für den Betrieb im paravirtualisierten Kontext angepasst sein
 - Rechteinhaber proprietärer Betriebssysteme lehnen eine Anpassung aus strategischen Gründen häufig ab
 - ⇒ Funktioniert häufig nur mit OpenSource-Betriebssystemen

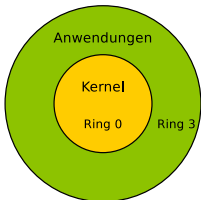
• Vorteil:

- Geschwindigkeitseinbußen, die beim VMM entstehen, werden vermieden

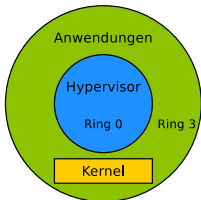
Problem: x86-64-Architektur

- Die x86-64-Architektur (z.B. IA64) verzichtet auf die Ringe 1 und 2

Keine Virtualisierung

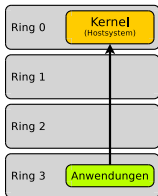


Paravirtualisierung (IA64)

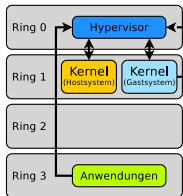


- Der Hypervisor befindet sich wie bei der x86-32-Architektur in Ring 0
- Der Betriebssystemkern wird bei der x86-64-Architektur in Ring 3 zu den Anwendungen verschoben

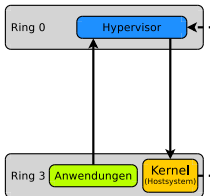
Keine Virtualisierung



Paravirtualisierung (x86)



Paravirtualisierung (IA64)

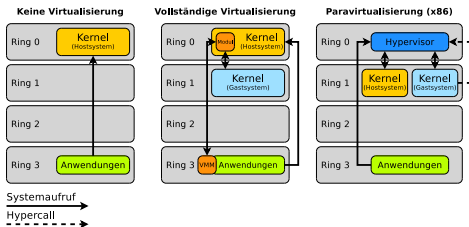


- Der Betrieb der Hardwaretreiber und Anwendungen in einem Ring ist tendenziell unsicher

Systemaufruf →
Hypercall - - - - ->

Zusammenfassung: Voll- vs. Paravirtualisierung

- **Paravirtualisierung** erfordert angepasste Gastsysteme
 - Typ-1-Hypervisor läuft *bare metal* (= ersetzt das Host-Betriebssystem)
 - Hypervisor läuft in Ring 0 und hat vollen Zugriff auf die Hardware
 - Beispiele: VMware ESX(i), Xen, Microsoft Hyper-V
- **Vollvirtualisierung** ermöglicht unveränderte Gastsysteme
 - VMM (Typ-2-Hypervisor) läuft *hosted* als Anwendung im Host-Betriebssystem
 - VMM läuft in Ring 3 auf der Ebene der Anwendungen
 - Beispiele: VMware Workstation, KVM, Oracle VirtualBox, Parallels

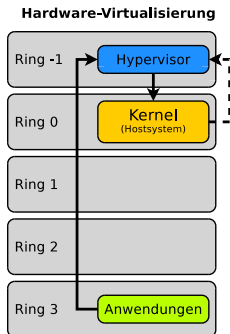


Hardware-Virtualisierung (1/2)

- Aktuelle Prozessoren von Intel und AMD enthalten Erweiterungen um Hardware-Virtualisierung zu ermöglichen
 - Vorteil: Unveränderte Betriebssysteme können als Gast-Systeme ausgeführt werden
 - Die Lösungen von Intel und AMD sind ähnlich aber inkompatibel
 - Seit 2006 enthalten AMD64 CPUs den Secure-Virtual-Machine-Befehlssatz (**SVM**)
 - Die Lösung heißt **AMD-V** und war vorher als **Pacifica** bekannt
 - Die Lösung von Intel heißt **VT-x** für IA32-CPU's und **VT-i** für Itanium
 - Intels Lösung lief vormals unter dem Stichwort **Vanderpool**
- Xen unterstützt ab Version 3 Hardware-Virtualisierung
 - Auch Windows Server 2008 (Hyper-V) nutzt Hardwarevirtualisierung
 - VirtualBox unterstützt Hardware-Virtualisierung
 - KVM kann nur mit CPU's mit Hardwarevirtualisierung arbeiten

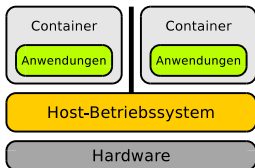
Hardware-Virtualisierung (2/2)

- Kern der Neuerung ist eine Überarbeitung der Privilegienstruktur
- Ein neuer Ring (\implies Ring -1) für den Hypervisor ist hinzugefügt
 - Der Hypervisor bzw. VMM läuft im Ring -1 und besitzt jederzeit die volle Kontrolle über die CPU und die Ressourcen, da damit ein höheres Privileg als Ring 0 implementiert ist
- VMs laufen in Ring 0 und heißen HVM
 - HVM = Hardware Virtual Machine
- Vorteile:
 - Gastbetriebssysteme müssen nicht angepasst sein
 - Auch proprietäre Betriebssysteme (z.B. Windows) laufen als Gastsysteme
 - Der Kernel läuft nicht wie bei Paravirtualisierung (IA64) auf der Privilegienstufe der Anwendungen



Betriebssystem-Virtualisierung / Container / Jails (1/2)

- Unter ein und demselben Kernel laufen mehrere voneinander abgeschottete identische Systemumgebungen
 - Es wird kein zusätzliches Betriebssystem gestartet
 - Es wird eine isolierte Laufzeitumgebung erzeugt
 - Alle laufenden Anwendungen verwenden denselben Kernel
 - Betriebssystem-Virtualisierung heißt **Container** in SUN/Oracle Solaris
 - Betriebssystem-Virtualisierung heißt **Jails** in BSD



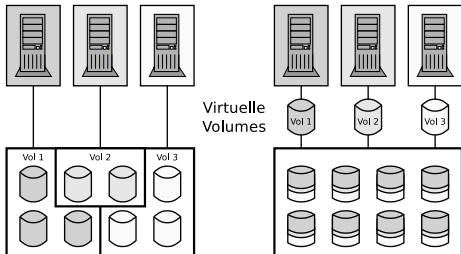
- Anwendungen sehen nur Anwendungen in der gleichen virtuellen Umgebung
- Ein Vorteil ist der geringe Overhead, da der Kernel in gewohnter Weise die Hardware verwaltet

- Nachteil: Alle virtuellen Umgebungen nutzen den gleichen Kernel
 - Es werden nur unabhängige Instanzen eines Betriebssystems gestartet
 - Verschiedene Betriebssysteme können nicht gleichzeitig verwendet werden

Betriebssystem-Virtualisierung / Container / Jails (2/2)

- Diese Art der Virtualisierung nutzt man, um Anwendungen in isolierten Umgebungen mit hoher Sicherheit zu betreiben
- Besonders Internet-Service-Provider, die (virtuelle) Root-Server oder Webdienste auf Mehrkernprozessorarchitekturen anbieten, nutzen diese Form der Virtualisierung
 - Wenig Performance-Verlust, hoher Grad an Sicherheit
- Beispiele:
 - SUN/Oracle Solaris (2005)
 - OpenVZ für Linux (2005)
 - Linux-VServer (2001)
 - FreeBSD Jails(1998)
 - Parallels Virtuozzo (2001, kommerzielle Variante von OpenVZ)
 - FreeVPS
 - Docker (2013)
 - chroot (1982)

Speichervirtualisierung



- Vorteile:

- Nutzer sind nicht an die physischen Grenzen der Laufwerke gebunden
- Physischen Speicher umstrukturieren/erweitern stört die Nutzer nicht
- Redundantes Vorhalten erfolgt transparent im Hintergrund
- Besserer Auslastungsgrad, da der physische Speicher effektiver auf die Benutzer aufgeteilt werden kann

- Nachteil: Professionelle Lösungen sind teuer

- Bekannte Anbieter: EMC, HP, IBM, LSI und SUN/Oracle

- Speicher wird in Form virtueller Laufwerke (*Volumes*) den Benutzern zur Verfügung gestellt
- Logischer Speicher wird vom physischen Speicher getrennt

