

Internet Anwendungen unter z/OS und OS/390

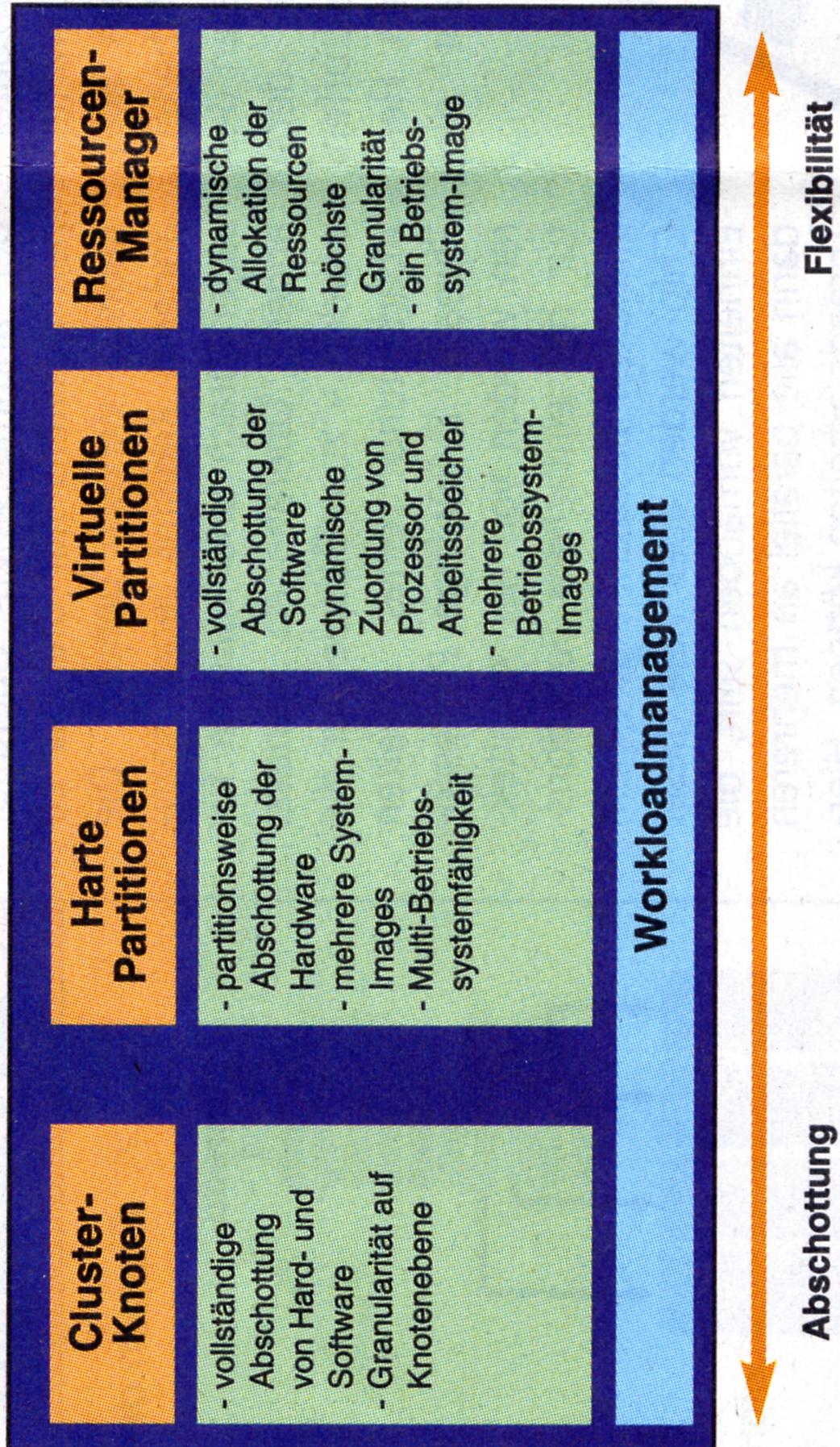
**Dr. rer. nat. Paul Herrmannn
Prof. Dr.rer.nat. Udo Kebschull
Prof. Dr.-Ing. Wilhelm G. Spruth**

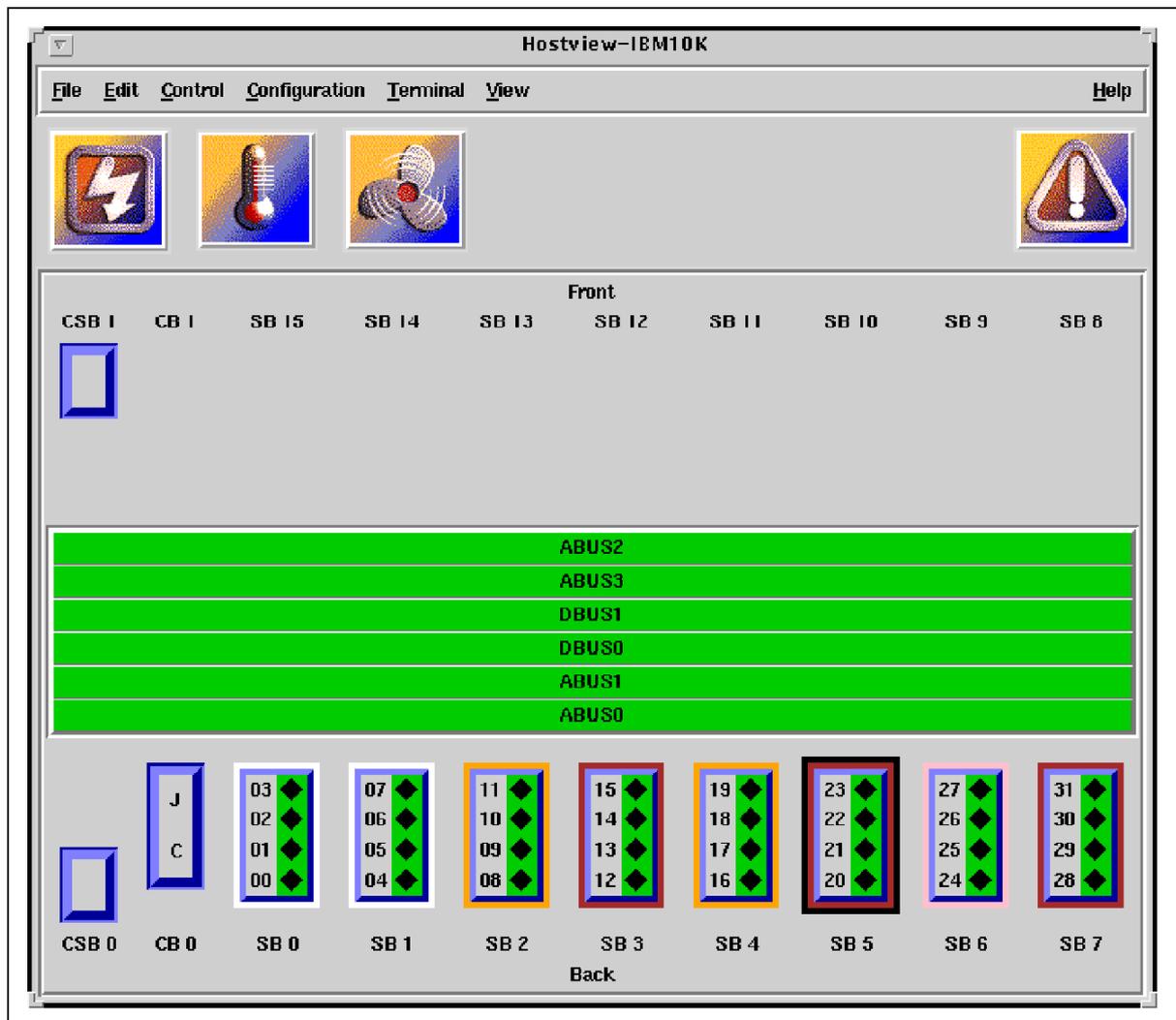
WS 2004/2005

Teil 4

Virtuelle Maschinen, Partitionierung

Partitionskonzepte im Vergleich





Sun E 10 000 Administrator Konsole

Dargestellt sind 8 Prozessor Boards SB0 .. SB7.

Cluster: SB0, SB1
 SP2, SB4
 SP3, SB7

Emulator

Auf einem Rechner mit der Hardware-Architektur x (Host) wird ein Rechner mit der Hardware-Architektur y (Gast) emuliert.

Beispiele:

Hercules und *FLEX-ES* emulieren einen zSeries Rechner mit dem z/OS Betriebssystem auf einem Intel/AMD Windows oder Linux Rechner.

Microsoft VirtualPC Typ 1 emuliert einen Intel/AMD Windows Rechner auf einem Apple MAC PowerPC Rechner.

Bochs ist ein in C++ geschriebener Open Source Emulator, der die Intel/AMD Architektur auf vielen anderen Plattformen emuliert.

Mehrere Gast Rechner auf einem Host Rechner sind möglich, aber nicht üblich.

Virtuelle Maschine

Auf einem Host Rechner mit der Hardware-Architektur x wird ein (oder mehrere) Gast Rechner der gleichen Architektur abgebildet.

Beispiele:

VM/370, *z/VM* und *PR/SM* für die S/390 und zSeries Hardware-Architektur.

VMWare und *Microsoft VirtualPC* Typ 2 bilden mehrere Windows oder Linux Gast Maschinen auf einem Windows oder Linux Host ab.

Paravirtualization wird von *Xen* und *Denali* implementiert.

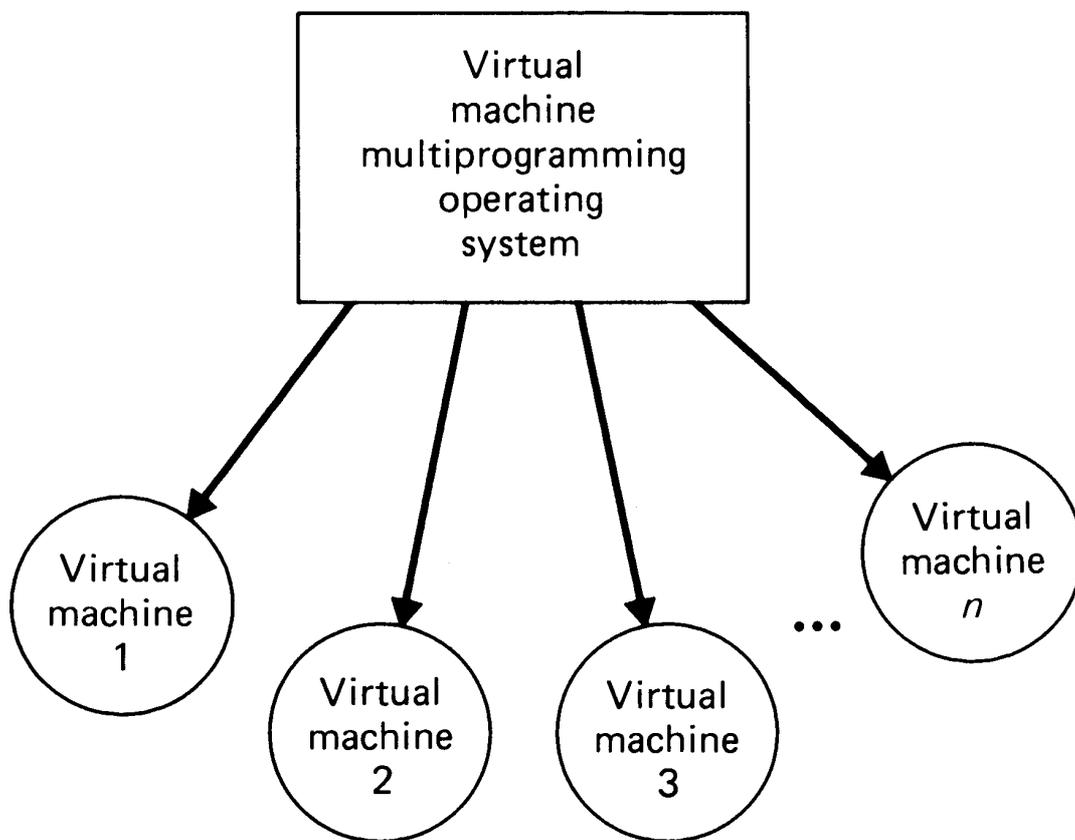
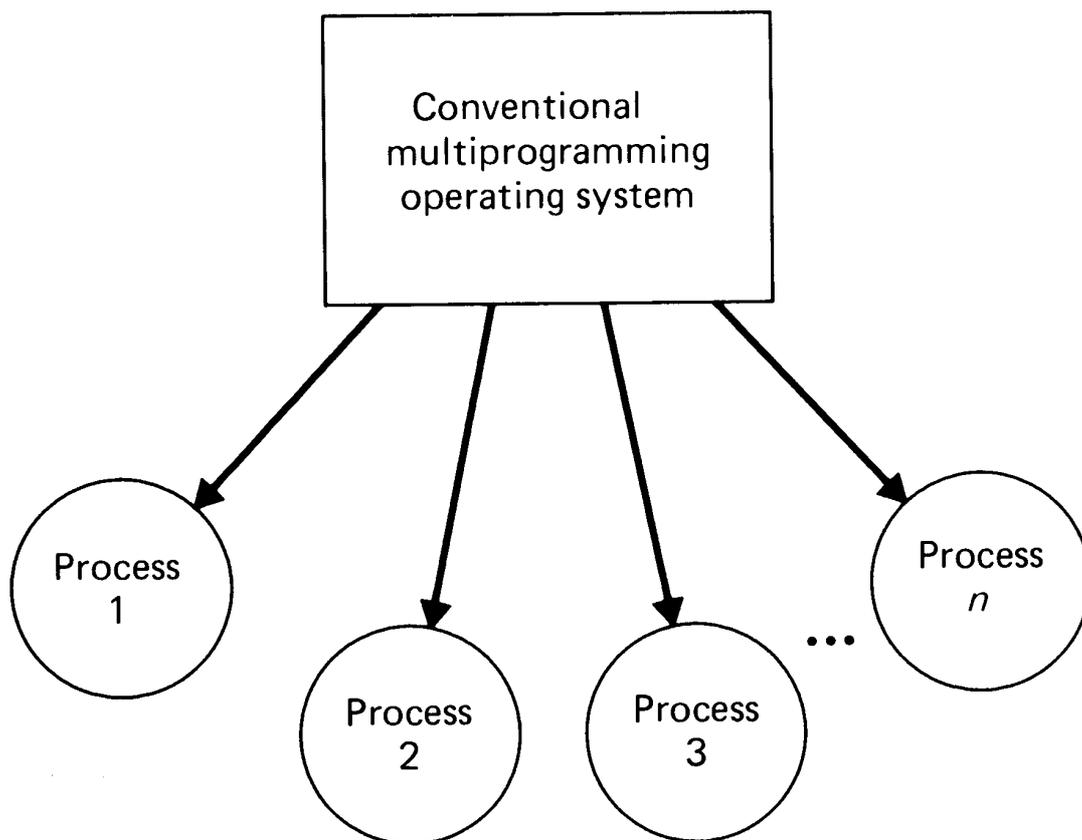
Kommentar

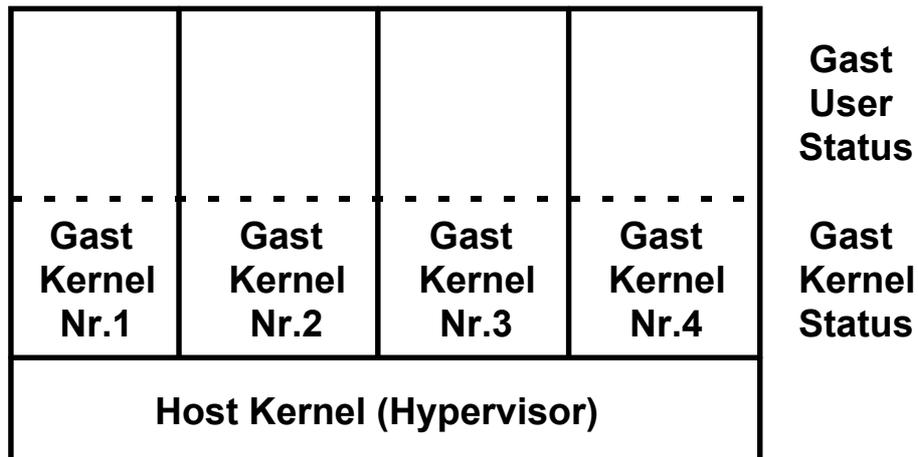
Jurassic Park mit Zukunft

Wenn schon als Techno-Dinosaurier verspottet, dann kann man seine Mainframe-Produkte auch gleich nach den Weltherrschern von einst benennen – meint IBM und wählt mit Raptor und T-Rex besonders furchteinflößende Vertreter. Big Blue kann sich das kecke Spiel mit dem eigenen Image leisten: Immer deutlicher wird, dass die Architektur der Zeit nicht hinterher hinkt, sondern ihr sogar voraus eilt. Bei der logischen Partitionierung etwa geben Experten ihr einen zehnjährigen Entwicklungsvorsprung, und der Workload-Manager, der Ressourcen ziel- und situationsabhängig nutzt, fällt bei Standard-Servern viel primitiver aus. Eben solche Techniken sind aber die Voraussetzung für das On-Demand-Computing – die Nutzung von Rechenleistung ganz nach Bedarf. Neidische Blicke auf die Privilegierten, die sich einen T-Rex leisten können, sind aber unnötig: Stets ist der Open-Systems-Tross durch die technologischen Breschen, welche die Rechner-Dinos schlugen, nachgefolgt. Frank-Michael Kieß

**Bei der logischen
Partitionierung geben
Experten ihr (zSeries) einen
zehnjährigen
Entwicklungsvorsprung**

**Computer Zeitung
19.5.2003, S. 1**





Gleichzeitiger Betrieb mehrerer Betriebssysteme auf einem Rechner

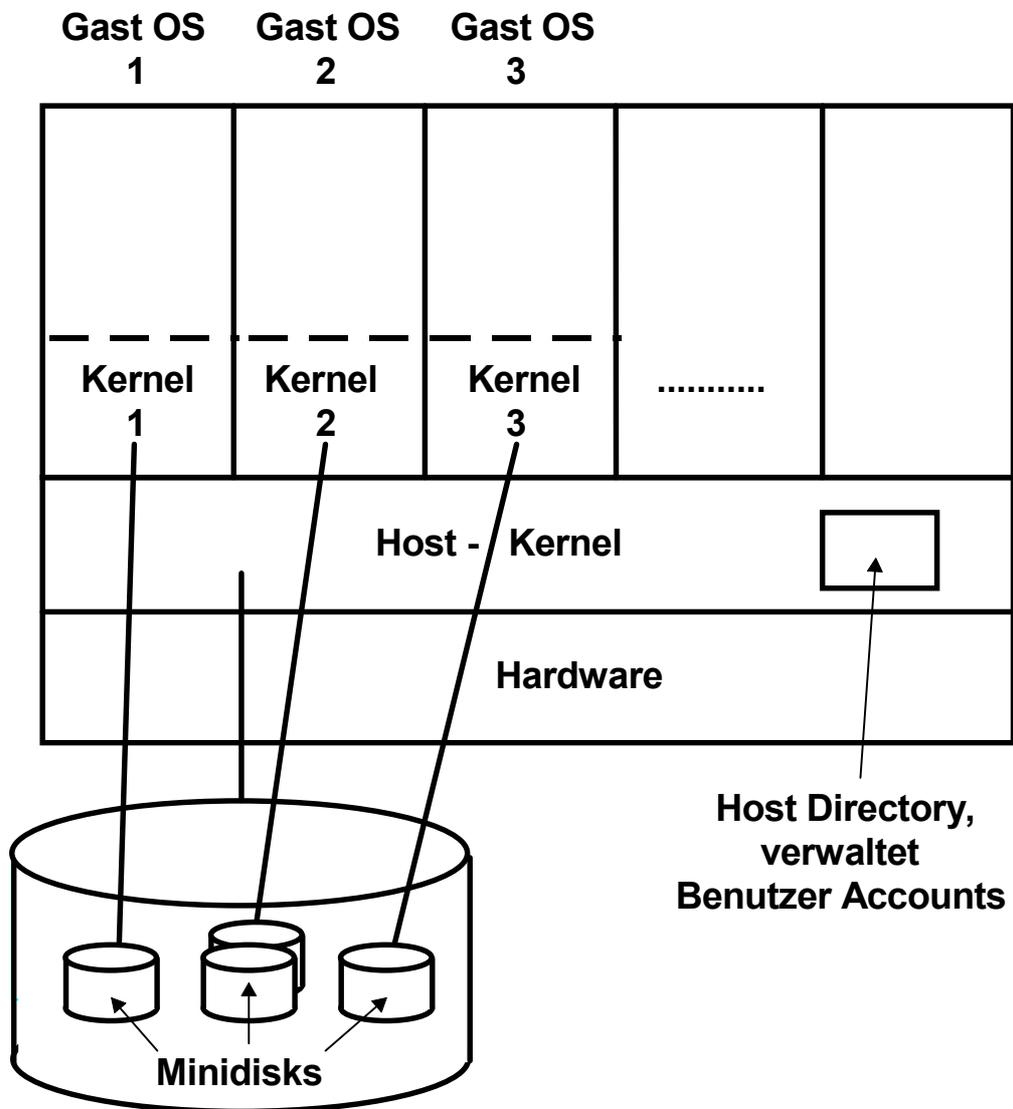
Ansatz: mehrere *Gast*-Betriebssysteme unter einem *Host*-Betriebssystem betreiben.

Dieser Vorgang wird als Partitionierung bezeichnet

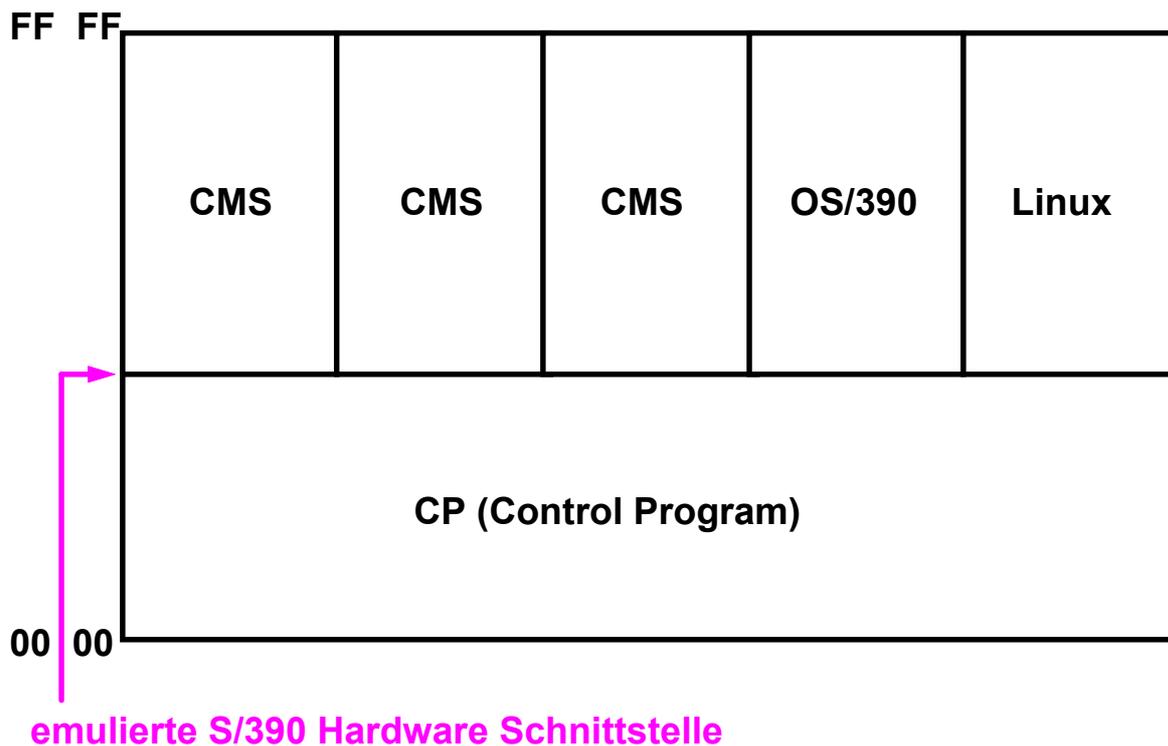
Das Host Betriebssystem verfügt über einen *Host Kernel*
 Das Gast Betriebssystem verfügt über einen *Gast Kernel*

Implementierungsbeispiele:

- VM/370
- VMware
- MS VirtualPC Typ 2
- z/VM
- PR/SM



Den Gast Betriebssystemen müssen Ressourcen wie CPU Zeit, Aufteilung auf mehrere CPUs in einem Mehrfachrechner, Hauptspeicher, Ein-/Ausgabe-Geräte und -Anschlüsse zugeordnet werden.



VM/ESA Betriebssystem

CP läuft im Überwacherstatus.

CMS und alle anderen Gast Betriebssysteme (einschließlich ihrer Kernel Funktionen) laufen im Problemstatus. Privilegierte Maschinenbefehle (z.B. E/A) werden von CP abgefangen und interpretativ abgearbeitet.

Volle S/390 Kompatibilität für alle Gast Betriebssysteme. Geringer Performance Verlust (< 5 %).

CMS (Conversational Monitor Program) ist ein besonders für die Software Entwicklung ausgelegtes Einzelplatz Betriebssystem. Für 1000 gleichzeitige CMS Benutzer werden 1000 CMS Instanzen angelegt. Ähnliches ist mit Linux/390 möglich.

Plattenspeicherplatz wird allen Gastbetriebssystemen in der Form virtueller „Minidisks“ statisch zugeordnet. Hauptspeicherplatz wird dynamisch verwaltet.

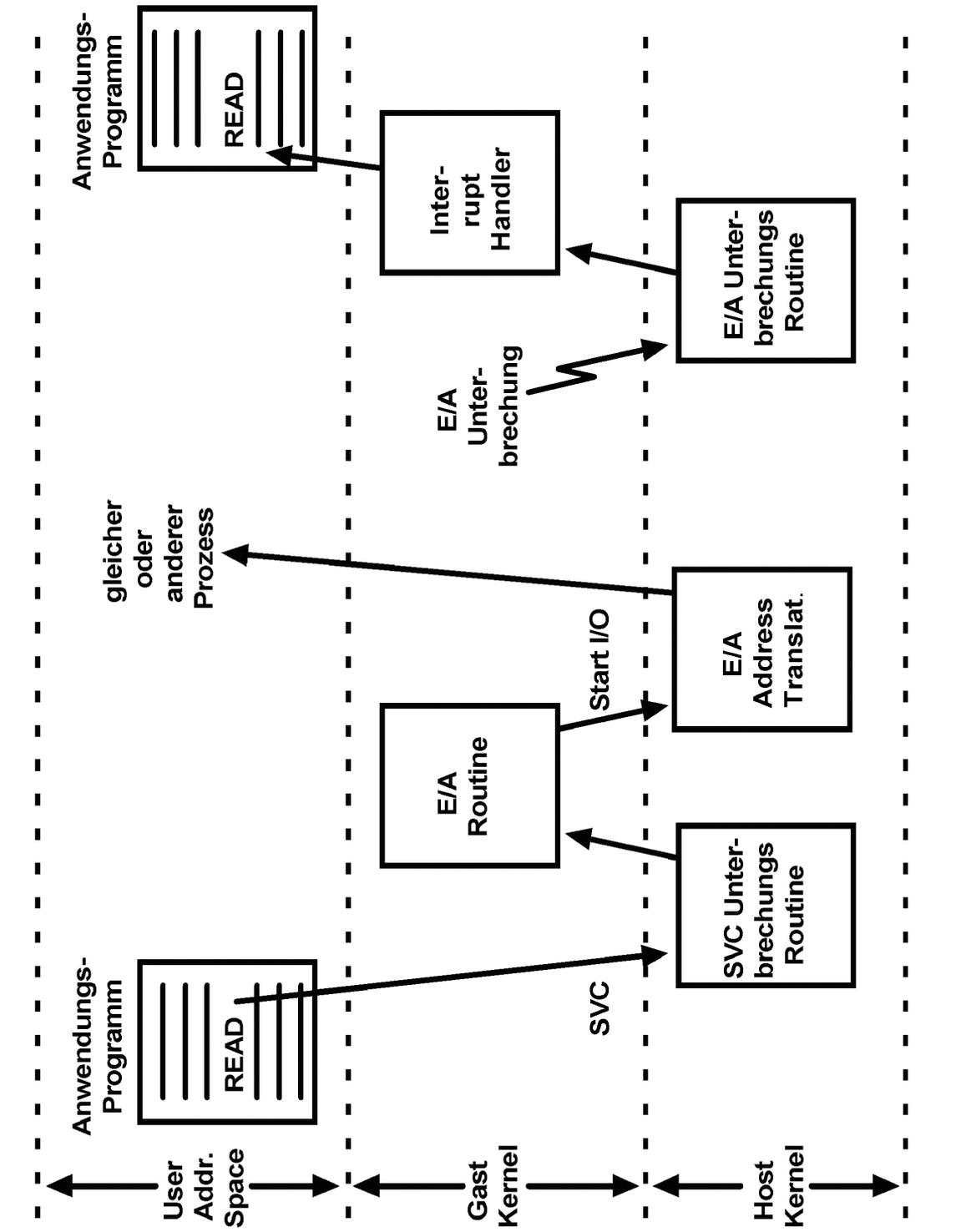
Steuerung der virtuellen Maschine

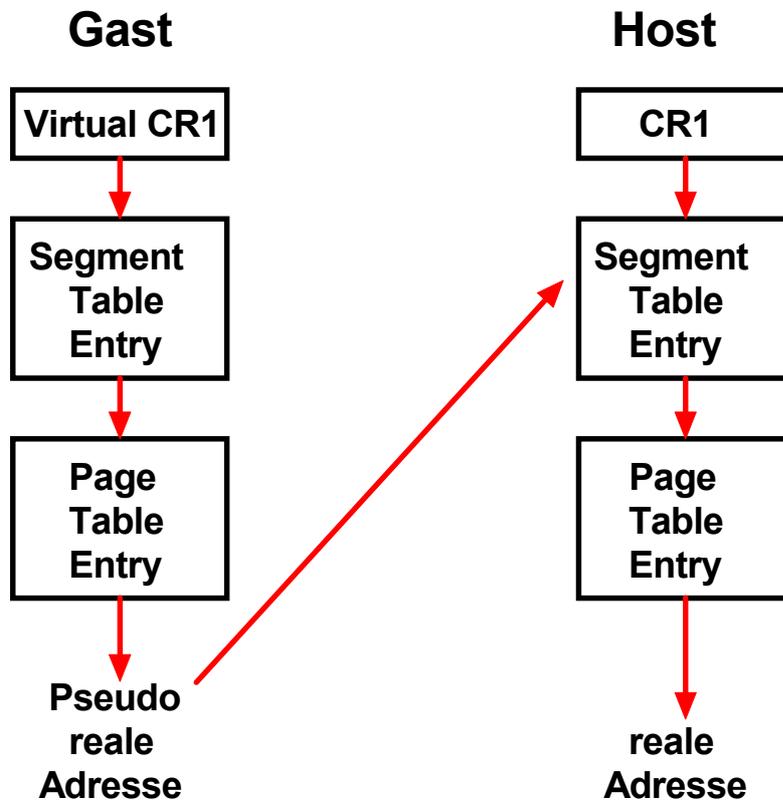
Alle Gast Maschinen laufen in einem eigenen virtuellen Adressenraum

Der Host Kernel Zeitscheiben-Scheduler übergibt die Kontrolle über die CPU einer Gast Maschine.

Der Kernel der Gast Maschine läuft im User Mode (Problem Mode). Wenn das Programm des Gast Betriebssystems versucht, einen privilegierten Maschinenbefehl auszuführen, führt dies zu einer Programmunterbrechung.

Die Programmunterbrechungsroutine des Host Kernels interpretiert den privilegierten Maschinenbefehl soweit als erforderlich und übergibt die Kontrolle zurück an den Kernel des Gastbetriebssystems



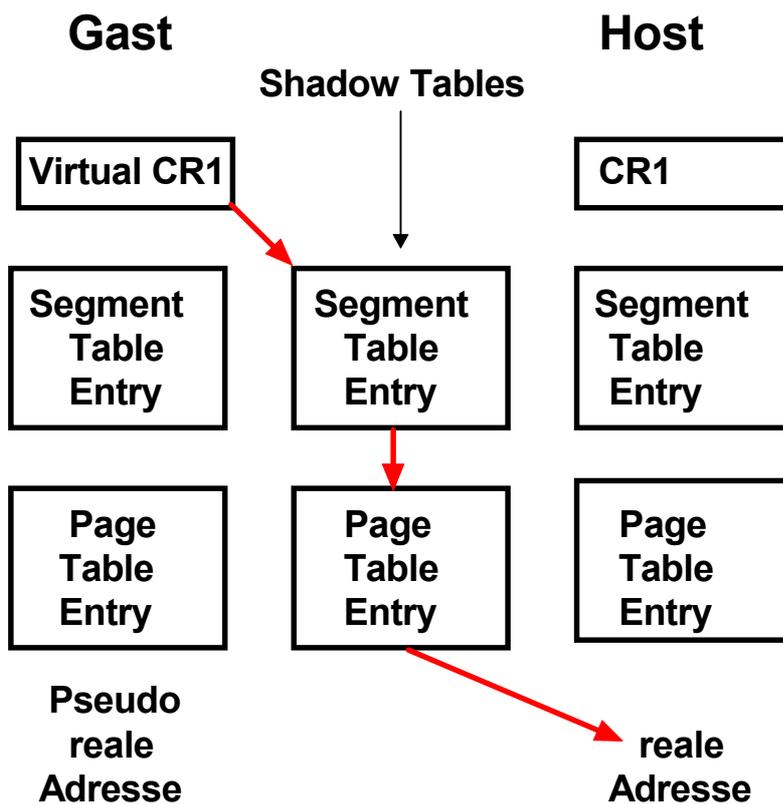


Problem der Adressübersetzung für die Virtuelle Maschine

Ein Gast Betriebssystem verfügt in seinem Kernel Bereich über eigene Segment- und Seitentabellen, mit denen es seine Adressübersetzung beschreibt.

Der Adressraum, den der Gast als real ansieht, ist jedoch in Wirklichkeit virtuell aus Sicht des Host Kernels und wird von diesem ebenfalls über Segment- und Seitentabellen beschrieben.

In den Seitentabellen des Gastes stehen die falschen Werte.



Shadow Page Tables unter VM/370 und VMware

VM/370 und VMware erstellen anhand der Gast- und ihrer eigenen Tabellen sogenannte *Shadow Tables*, die direkt die virtuellen Adressen des Gastes auf reale Adressen des Hosts abbilden. Diese Tabellen liegen im Speicherbereich des Host-Kernels

Probleme der IA32 Architektur

Im Vergleich zu VM/370 sind der ESX Server und VMware benachteiligt, weil einige kritische Eigenschaften in der IA32 Architektur fehlen. Für den Betrieb von Gast-Maschinen ist es erforderlich, dass alle Maschinenbefehle, welche den privilegierten Maschinenstatus abändern oder auch nur lesen, nur im Kernel Status ausgeführt werden können.

Wenn ein Gast ein Kontrollregister schreibt, muss der Host Kernel diese Instruktion abfangen, damit nicht das reale Kontrollregister des Hosts verändert wird. Der Host Kernel wird jetzt nur die Effekte der Instruktion für diesen Gast simulieren. Liest der Gast anschließend diese Kontrollregister wieder aus, so muss diese Instruktion ebenfalls abgefangen werden, damit der Gast wieder den Wert sieht, den er vorher in das Register geschrieben hat (und nicht etwa den realen Wert des Kontrollregisters, der nur für den Host sichtbar ist).

Da die IA32 Architektur diese Bedingung nicht erfüllt, ist es nicht möglich, wie unter VM/370 alle Maschinenbefehle einfach im User Mode auszuführen, und auf Programmunterbrechungen zu vertrauen wenn auf privilegierten Maschinenstatus Information zugegriffen wird. Beispielsweise:

Many models of Intel's machines allow user code to read registers and get the value that the privileged code put there instead of the value that the privileged code wishes the user code to see.

G.J. Popek, R.P. Goldberg: Formal Requirements for Virtualizable Third Generation Architectures. Comm. ACM, Vol. 17, Nr. 7, Juli 1974, S. 412-421.

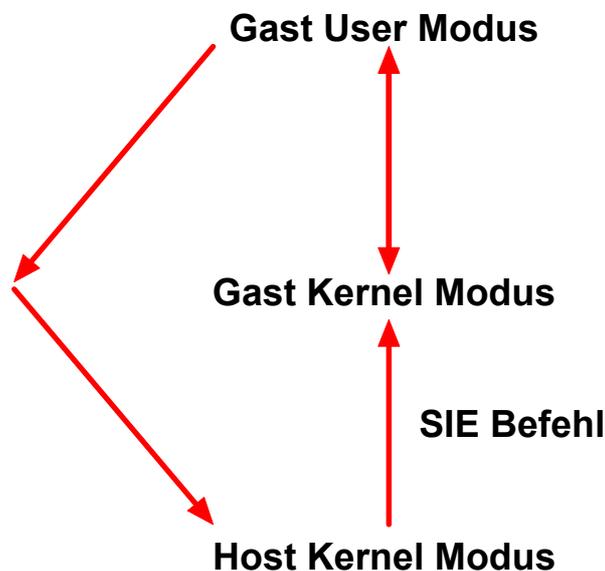
<http://www.cap-lore.com/CP.html>

	Kernel Modus	User Modus
Host Modus	+	—
Gast Modus	+	+

**Mögliche Zustände beim Einsatz des
Host/Gast Modus**

Interpretive Execution Facility

Einführung eines Host/Gast Modus
zusätzlich zum Kernel/User Modus.

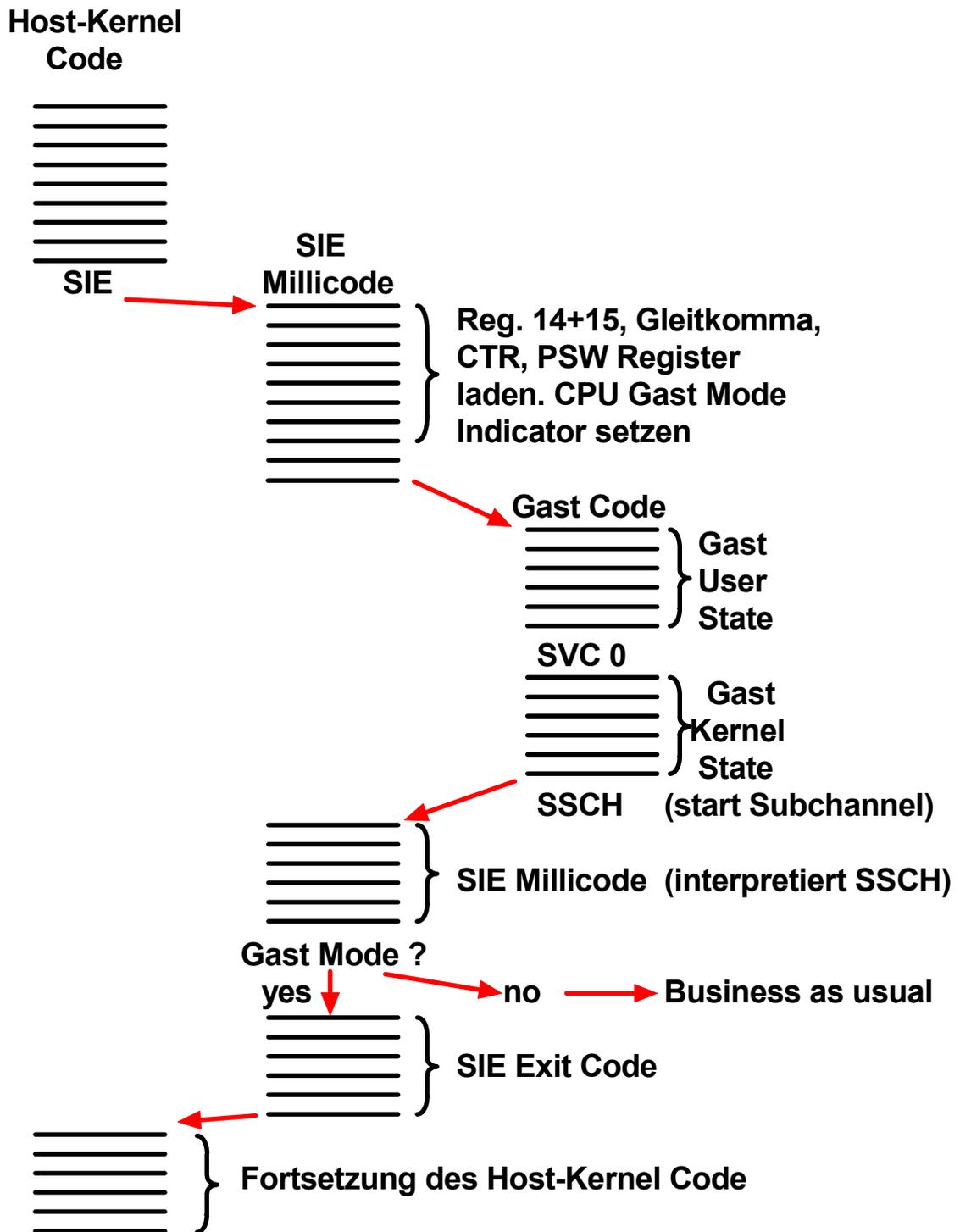


Die nicht-privilegierten Maschinenbefehle des Gastes werden vom Gast im User Modus direkt ausgeführt.

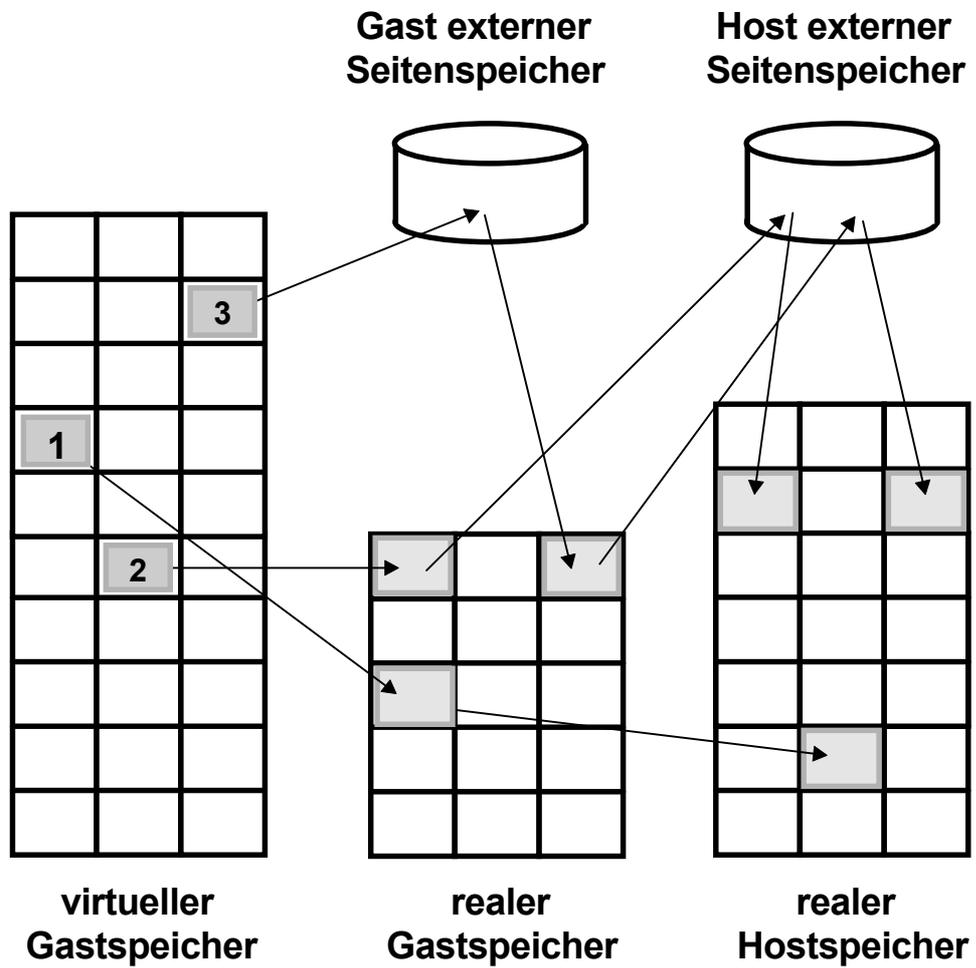
Die privilegierten Maschinenbefehle des Gastes werden in zwei Gruppen geteilt. Ein State Descriptor spezifiziert die Aufteilung.

Die meisten privilegierten Maschinenbefehle des Gast-Kernels werden innerhalb des Gast-Modus ausgeführt. Maschinenbefehle oder Zustände, die weiterhin die Unterstützung durch den Host-Kernel erfordern, führen zu einer *Interception*. Diese bewirkt einen Transfer in den Host-Kernel Modus.

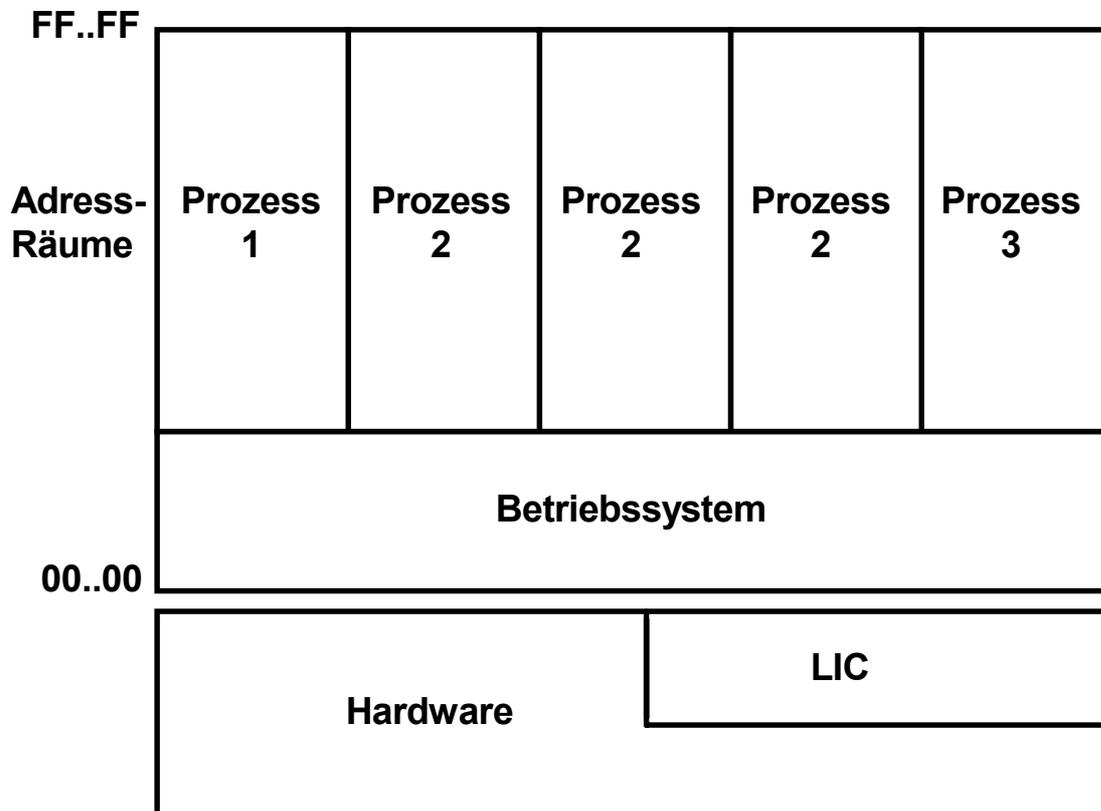
Der Host-User Modus hat keine Bedeutung.



Ausführung der SIE Instruktion



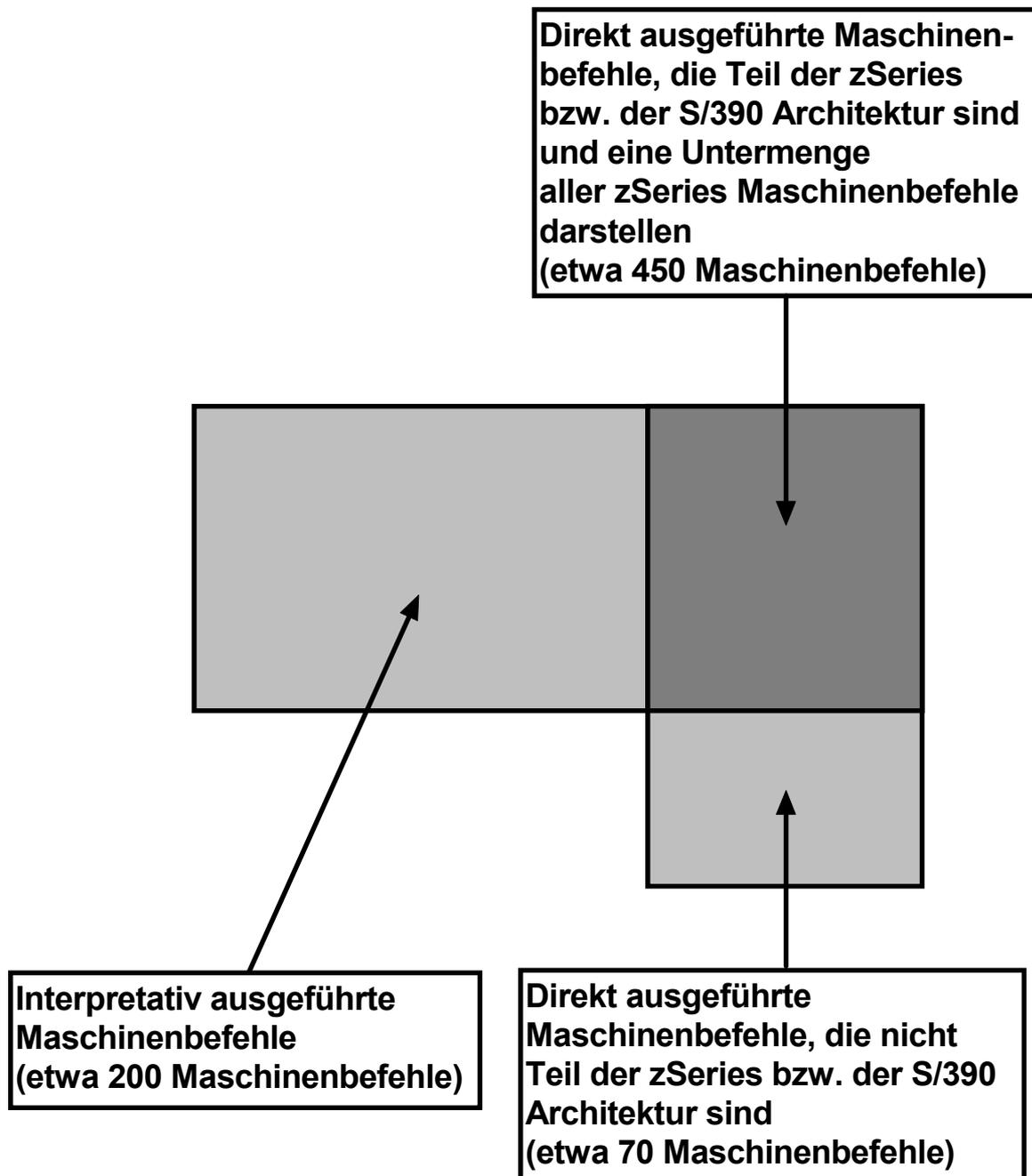
Adressumsetzung zwischen Gast-Kernel und Host-Kernel



S/390 Struktur

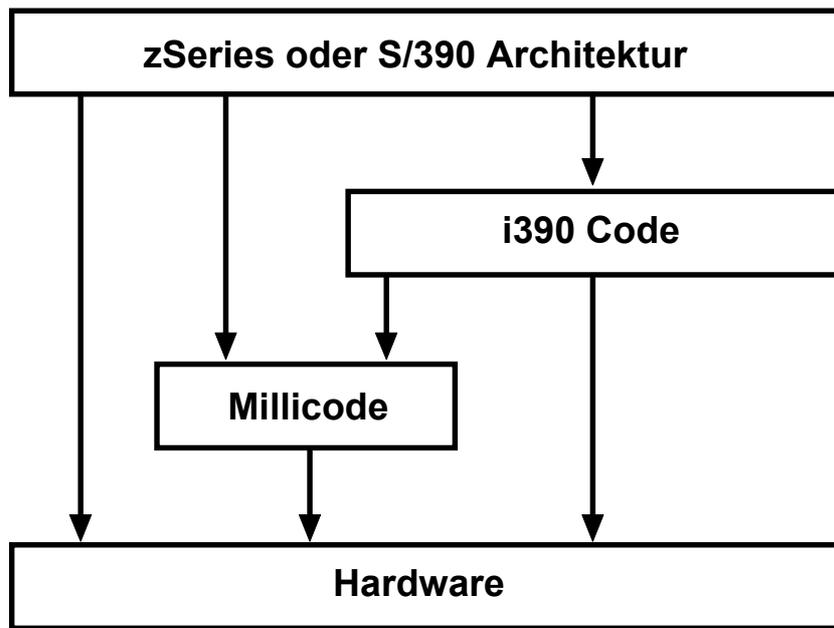
LIC (Licensed Internal Code) ist Maschinencode, der von der CPU ausgeführt wird, aber in einem separaten Speicher, außerhalb des architekturmäßig zugänglichen realen Hauptspeichers untergebracht und ausgeführt wird.

Häufig bestehen Teile aus regulären S/390 Maschinenbefehlen

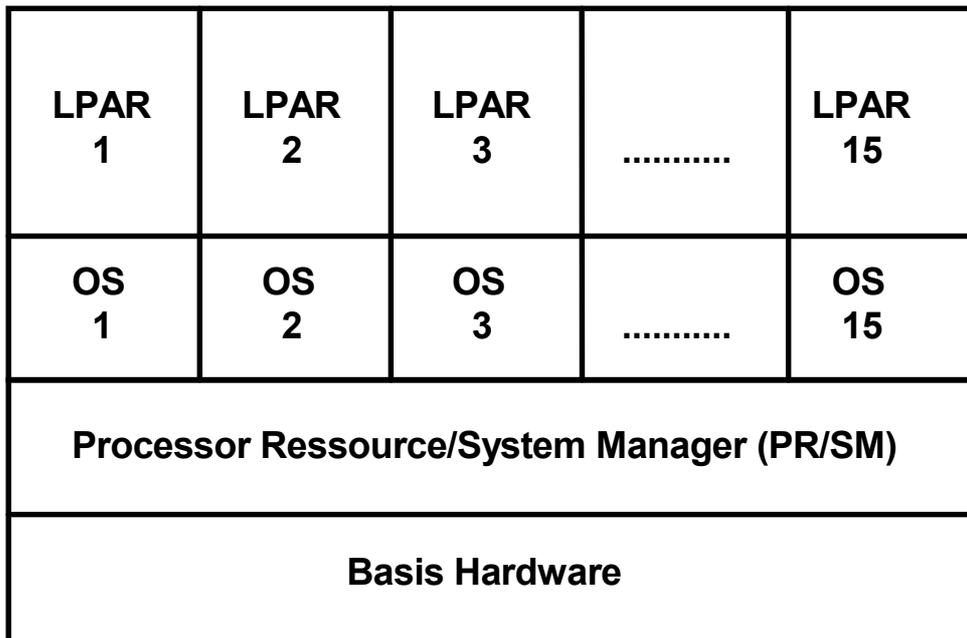


Die mehrstufige Instruction Execution Unit der zSeries (Pipeline) führt diese Maschinenbefehle direkt aus.

zSeries Maschinenbefehle



**Implementierung der zSeries oder S/390
Maschinenbefehle**



LPAR Logical Partition
OS Operating System

IBM PR/SM und LPAR

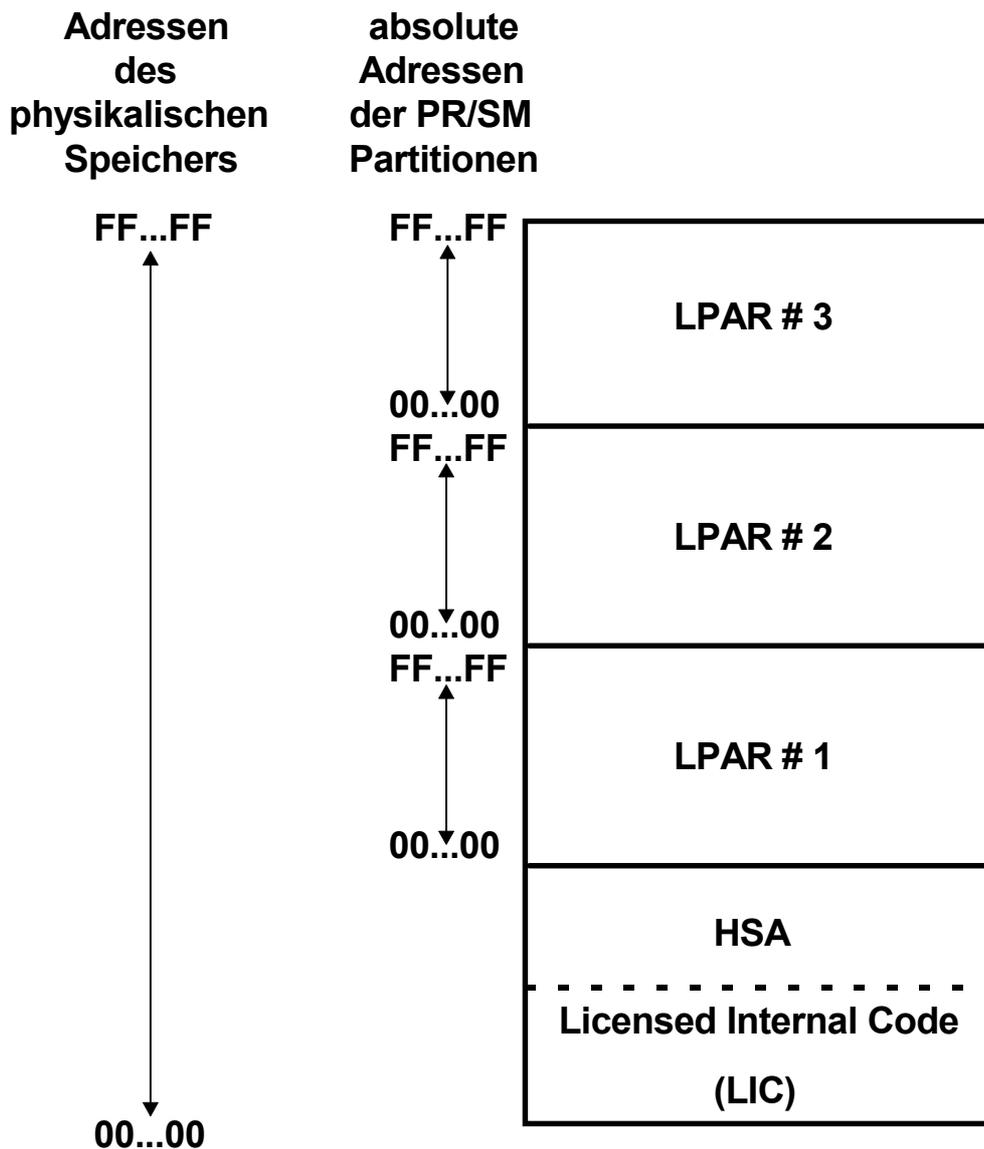
PR/SM ist eine Hardware (Microcode) Einrichtung, welche die Partitionierung eines physikalischen Rechners in mehrere logische Rechner (LPAR. Logical Partition) ermöglicht. Jeder logische Rechner hat sein eigenes Betriebssystem, seinen eigenen unabhängigen realen Hauptspeicherbereich und seine eigenen Kanäle und I/O Geräte. Gemeinsame Nutzung von Krypto Coprozessoren und I/O Geräten durch mehrere LPAR's ist möglich (EMIF).

PR/SM schedules die einzelnen Betriebssysteme auf die einzelnen CPU's eines SMP.

Mit PR/SM vergleichbare Möglichkeiten sind unter dem VM/390 Betriebssystem vorhanden.

S/390 Rechner anderer Hersteller verfügen über ähnliche Einrichtungen: Hitachi MLPF, Amdahl Multiple Domain Facility.

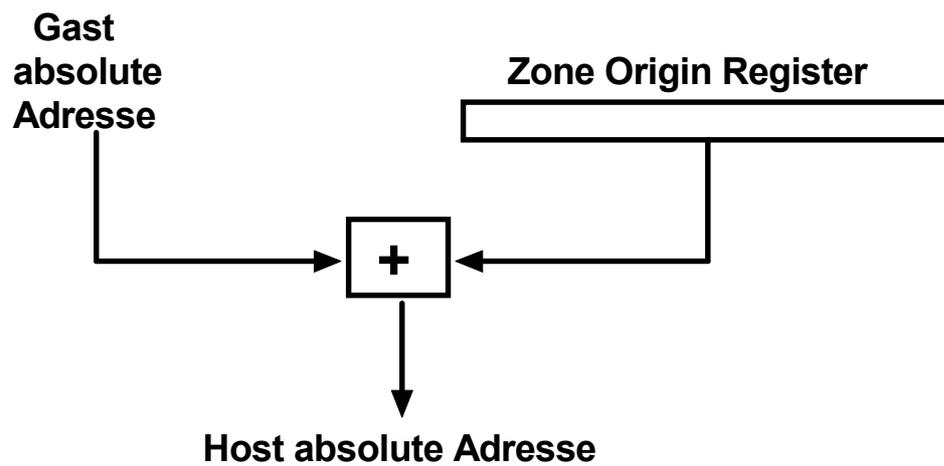
Für die Intel Architektur bietet die Firma vmware ein Softwareprodukt mit (nicht ganz) vergleichbarer Funktionalität.



Aufteilung des physikalischen Speichers in mehrere reale Speicher

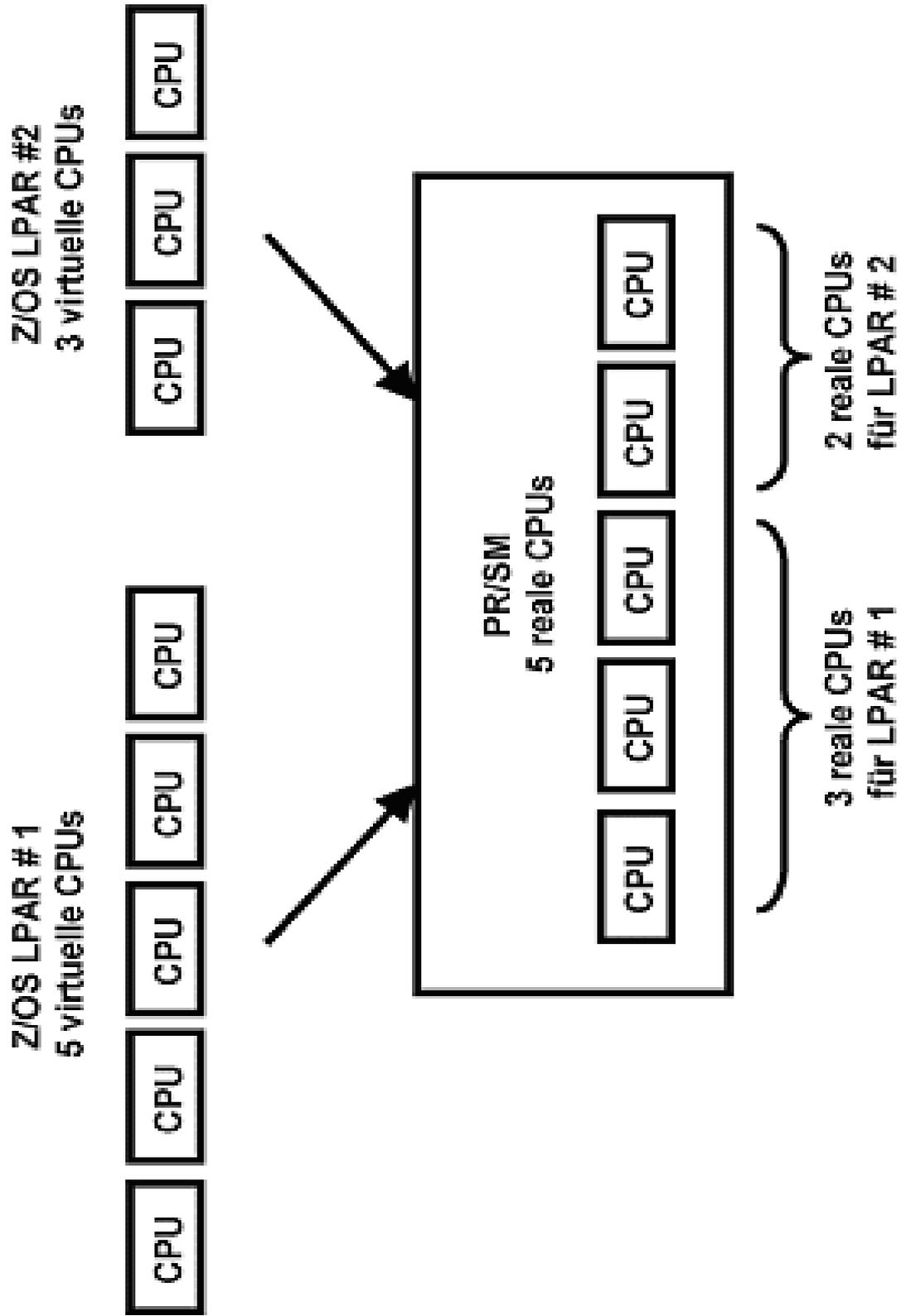
(in der zSeries Architektur besteht ein (kleiner) Unterschied zwischen realen und absoluten Adressen)

(zSeries Principles of Operations, IBM Form No. SA22-7832-00, Abb. 3-15)



Ein weiteres Zone Limit Register stellt sicher, dass der Gast innerhalb des ihm zugewiesenen physikalischen Adressenbereiches bleibt

Umsetzung der Gast Adressen in physikalische Adressen



PR/SM and LPAR

Zertifikat der USA Regierung:

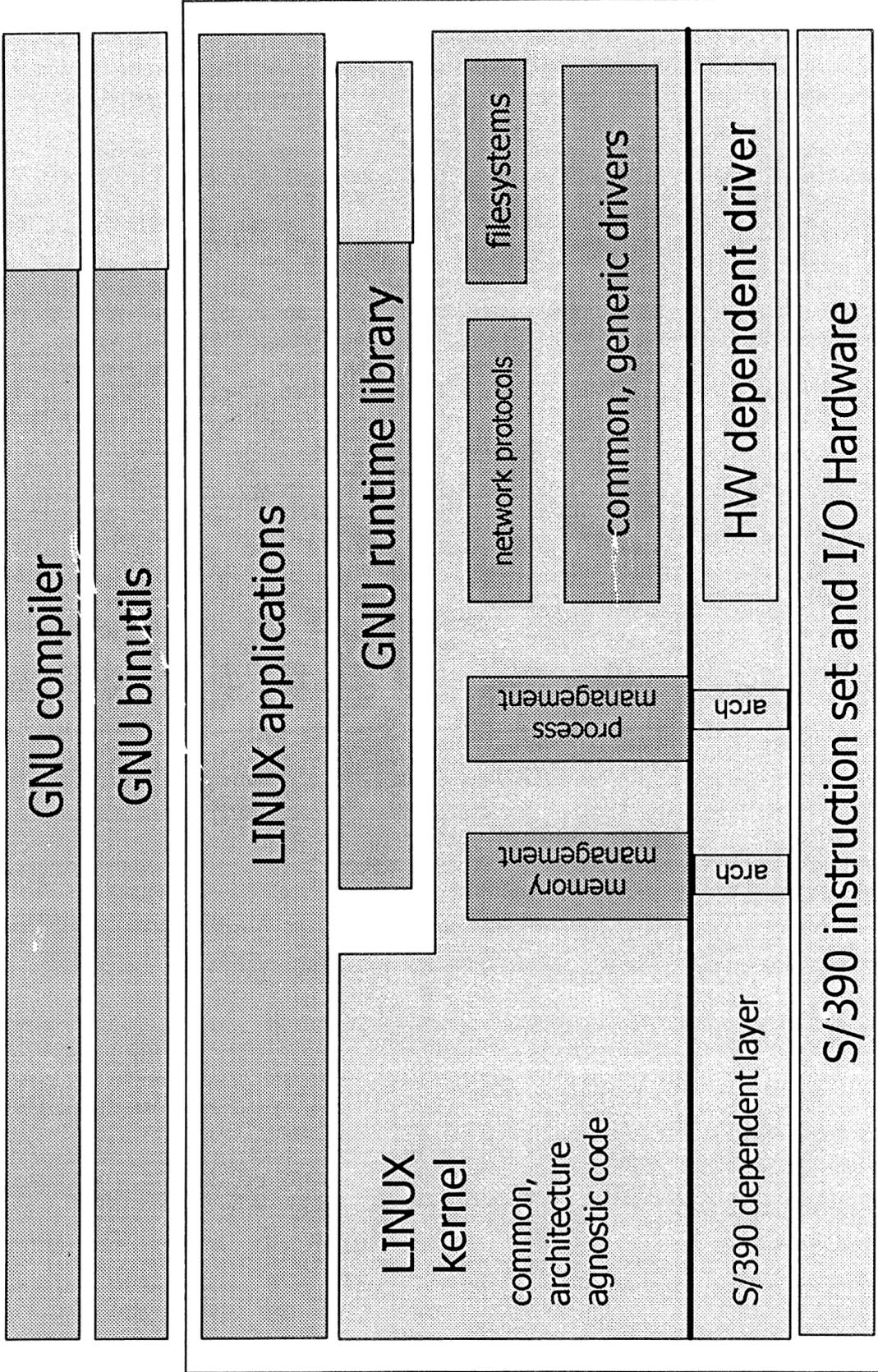
LPARs haben äquivalente Sicherheits-Eigenschaften wie physikalisch getrennte Rechner.

Der z900 Rechner verfügt über die "Virtual Image Facility". Dies ist eine PR/SM ähnliche Einrichtung, die > 16 LPARserlaubt, z.B. für S/390 Linux.

Das Bundesamt für Sicherheit in der Informationstechnik (BSI) stellt IBM für den Processor Resource/System Manager (PR/SM) des Mainframes z900 das weltweit höchste Sicherheitszertifikat für einen Server aus. Die Bescheinigung nach dem internationalen Standard Common Criteria (CC) für die Stufen EAL4 und EAL5 wurde auf der CeBIT 2003 an IBM verliehen. Der z900 ist der erste Server, der nach der Evaluierungsstufe EAL5 für seine Virtualisierungstechnologie zertifiziert wurde.

Die Zertifizierung des BSI bescheinigt, dass Programme, die auf einem IBM eServer zSeries z900 in verschiedenen logischen Partitionen (LPAR) laufen, ebenso gut voneinander isoliert sind, wie auf einem separaten physikalischen Server.

Die Partitionierung weist einzelnen Applikationen und Workloads unterschiedliche Bereiche auf dem Server zu und kann diese komplett voneinander abschirmen. So können beispielsweise Web-Anwendungen und Produktionsanwendungen, die in getrennten logischen Partitionen laufen, komplett voneinander isoliert betrieben werden, obwohl sie die physikalischen Ressourcen des zSeries Servers gemeinsam nutzen.



Linux for S/390 structure

S/390 Linux

Entwicklung TSO	Lohn/Gehalt Stapel	SAP/R3 USS	Linux Anwendung	User Status
z/OS LPAR # 1			Linux LPAR # 2	Kernel Status
PR/SM				Micro Code
Hardware				

**Unix System Services (USS) z/OS Nutzung:
Sysplex, WLM, E/A
aber begrenzter Funktionsumfang**

Linux

Reichhaltige Anwendungsumgebung

CMS	CMS	z/OS	Linux	Linux	Linux
VM - CP					

Sun-Server abgelöst

Telekom konsolidiert mit Linux-Mainframe

MÜNCHEN (CW) – Die Telekom-Festnetzsparte T-Com hat 25 Sun-Solaris-Server durch einen IBM-Großrechner mit Linux-Anwendungen ersetzt. Laut einem Systemverantwortlichen waren Hochverfügbarkeit und Skalierbarkeit wichtige Kriterien für die Server-Konsolidierung. Zentrale Serviceapplikationen wie Mail-, News- und Backup-Server arbeiten nun in virtuellen Linux-Partitionen unter einer Distribution der Suse Linux AG. Die neue IT-Struktur benötige weniger Stellfläche und senke den Energieverbrauch um die Hälfte. (*wb*) ←

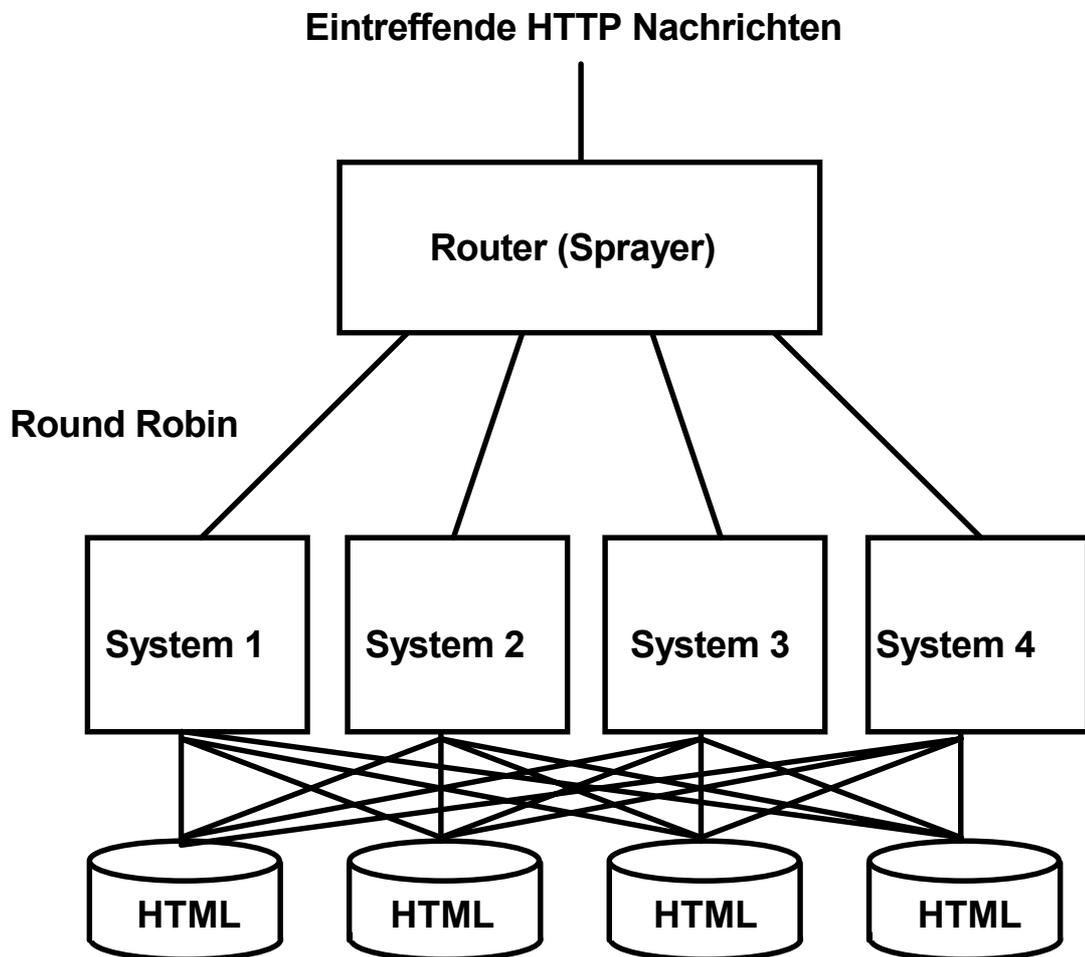
Weltweites zLinux SAP-System auf 36 zSeries 990 Prozessoren

Rhein/Stuttgart, 11. August 2004. Endress+Hauser, Anbieter von Messgeräten und Automatisierungslösungen für die industrielle Verfahrenstechnik, hat eine der größten Linux/Großrechner-Plattformen in Deutschland installiert. Die Installation, die IBM eServer zSeries Großrechner einsetzt, unterstützt die weltweiten betriebswirtschaftlichen Anwendungen des Unternehmens. Endress+Hauser hat dabei seine SAP-Plattform auf einer zentralen Serverlösung am Standort Weil am Rhein konsolidiert. Die Firmengruppe stellt damit eine noch höhere Verfügbarkeit, Sicherheit und Wirtschaftlichkeit seiner SAP-Systeme sicher. Um die Kostenstruktur weiter zu verbessern, hat Endress+Hauser darüber hinaus die ehemals auf Unix-Plattformen betriebenen SAP R/3-Systeme auf die Linux-Großrechner-Kombination migriert.

Endress+Hauser hat insgesamt zwei IBM eServer zSeries 990 Mainframes mit zusammen 328 GB Hauptspeicher in seinem Rechenzentrum in Weil am Rhein installiert.

Höchste Verfügbarkeit der Anwendungen und Geringhaltung der Kosten waren die Hauptkriterien. Die Installation von Linux unter dem Betriebssystem z/VM bietet in Kombination mit der IBM zSeries eine flexible Lösung nicht nur für unsere heutigen Anforderungen, sondern auch für zukünftige Entwicklungen.

Die SAP Systeme der Firmengruppe bedienen derzeit rund 3.500 Anwender in aller Welt und sind auf Zuwachs bemessen. Auf der IBM eServer zSeries 990 werden alle 19 produktiven SAP-Systeme der Gruppe mit dem kompletten, von SAP angebotenen Funktionsumfang betrieben, welche auf 14 logische Partitionen (LPARs) verteilt sind. Die angebundene SAP-Datenbanken auf Basis von IBM DB2 sind dabei auf 6 logische Partitionen verteilt. Die Anwendungsserver nutzen insgesamt 36 zSeries Prozessoren (IFL, Integrated Facility for Linux), die unter Linux in den zwei Großrechnersystemen laufen.



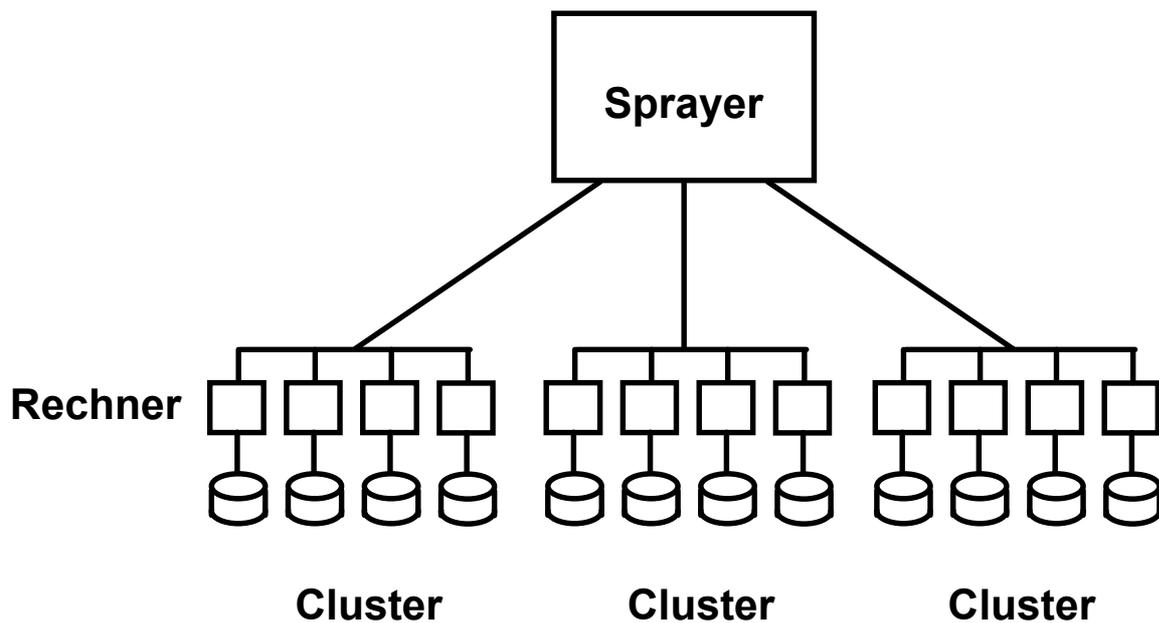
Work Load Management für einen WWW Cluster

Bei großen Providern kann die Web Site aus hunderten oder tausenden von Systemen bestehen

Vorteil: Einheitliche Anwendung, nur Lesezugriffe zu den HTML Daten

www.google.com

Google unterhält in 5 Rechenzentren ca. 10 000 Rechner. Mehrere Cluster in jedem Rechenzentrum. Jeder Cluster dupliziert den ganzen Google Datenbestand.



je 30 - 50 Rechner pro Cluster
je 3 - 5 Tbyte pro Cluster
Kopie aller Daten auf jedem Cluster

Sprayer verteilt Anfragen auf die einzelnen Cluster. Jeder Cluster ist in der Lage, jede Art von Anfrage zu bearbeiten.

Einfacher Workload Algorithmus.

WLM

OS/390 Work Load Manager (1)

Traditionelle Verfahren - viele Einstellungen:

- **Run-to-Completion**
- **Zeitscheibensteuerung, z.B. exponentiell**
- **Anzahl aktiver Prozesse - Größe des Working Sets**
- **Multithreading, Anzahl von Subsystem Instanzen**
- **Zuordnung von Anwendungen auf physikalische Server**
- **Zuordnung der E/A Kanäle**
- **Prioritäten**
- **etc.**

Komplexität wächst mit der Systemgröße

- **Stapel - interaktive Verarbeitung**
- **Belastungsschwankungen (Während des Tages, Woche, Jahr)**
- **Affinität Prozesse - Daten**
- **Zuordnung Prozesse - reale CPU's**
- **Unterschiedliche E/A Anforderungen/Belastungen**

Zwei Alternativen

1. Unterschiedlichen Anwendungen unabhängige physikalische Server zuordnen; schlechte Auslastung akzeptieren (Hardware ist billig)

- **Viele Server**
- **Heterogene Landschaft**
- **Komplexe LAN strukturen**
- **Hoher Administrationsaufwand**

2. Workload Manager

WLM

OS/390 Work Load Manager (2)

Einstellungen justieren:

- **automatisch**
- **dynamisch**

Leistungsverhalten der Installation durch Vorgaben für „Business Goals“ festlegen

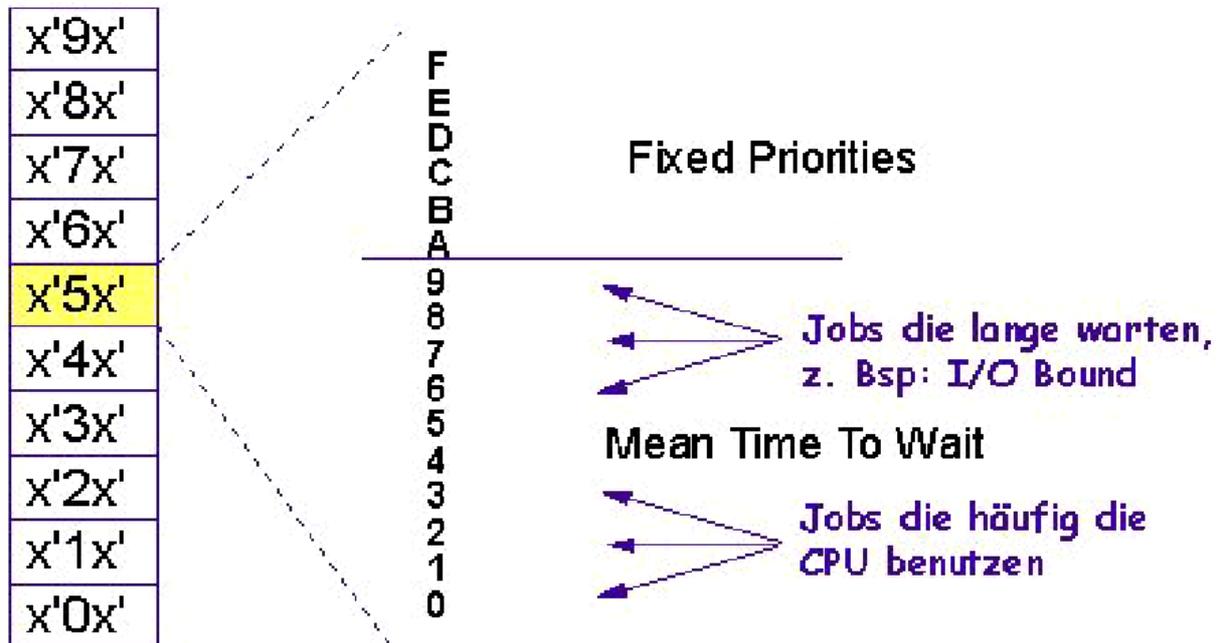
Work Load Manager setzt Vorgaben in Einstellungen um

Bei der Vielzahl der Einstellungsmöglichkeiten ist es denkbar, daß eine Änderung keine Verbesserung, sondern eine Verschlechterung bringt.

WLM

- **reduziert Aufwand für System Administration**
- **reduziert Notwendigkeit für Detailwissen**

Prioritäten von Prozessen



Für das Scheduling werden alle Prozesse 10 Gruppen zugeordnet, wobei jede Gruppe noch einmal in 16 Prioritäten unterteilt ist.

Die unteren 10 Prioritäten werden für einen *Mean-Time-To-Wait* Algorithmus verwendet. Es wird der Tatsache Rechnung getragen, dass Workloads unterschiedliches Verhalten bei der Benutzung der Prozessoren und des I/O-Subsystems zeigen. Dieses Verhalten ändert sich mit der Abarbeitungszeit von Programmen und trifft insbesondere auf langlaufende Batch-Jobs zu, die unter Umständen eine längere Zeit Daten ein-oder auslesen, im Wechsel mit Perioden, in denen sie Rechenvorgänge ausführen. Um dies zu berücksichtigen, kann eine Installation eine Dispatching Priority festlegen, die durch den Buchstaben M und eine Zahl charakterisiert wird.

SRM beobachtet die Address Spaces in diesen Gruppen und verändert die Dispatching Priorities über einen Alterungsprozess, abhängig davon, ob die Address Spaces die Prozessoren benutzen oder auf die Abarbeitung von I/O-Operationen warten. Address Spaces, die lange Zeit auf I/O-Operation warten, bekommen eine Dispatching Priority am oberen Ende der Gruppe, und solche, die häufig die Prozessoren benutzen, liegen am unteren Ende. Für alle anderen Arten von Arbeit werden feste Dispatching Priorities vergeben, wobei eine hohe Zahl einen besseren Zugang zum Prozessor bedeutet. Aus diesem Grund

WLM

OS/390 Work Load Manager (3)

Menge aller möglichen Arbeitsanforderungen in Dienstklassen (Service Classes) einordnen (klassifizieren).

Erfolgt nach Attributen wie

- **User ID**
- **Accounting Information**
- **Art des aufgerufenen Prozesses (Transaktionstyp, Stapel,)**
- **.....**

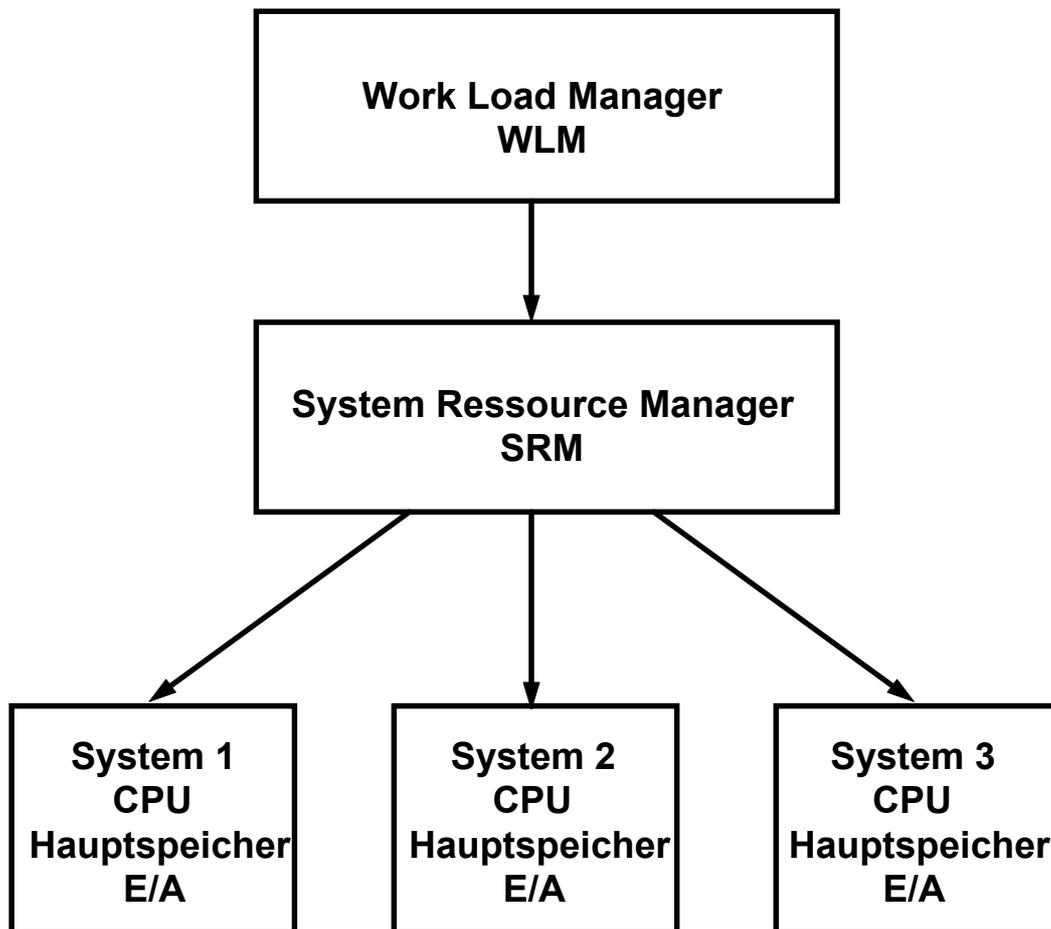
Jeder Dienstklasse sind „Ziele“ zugeordnet

- **Antwortzeit (Response Time)**
- **Geschwindigkeit (Velocity)**
- **Stellenwert (Importance)**
- **andere (discretionary)**

Jede Dienstklasse besteht aus Perioden

- **Während einer Periode begrenzte Ressourcen verfügbar (CPU Zyklen, E/A Zugriffe,)**
- **Nach Ablauf der Periode i Migration nach i+1**
- **Unterschiedliche Ziele in jeder Periode**

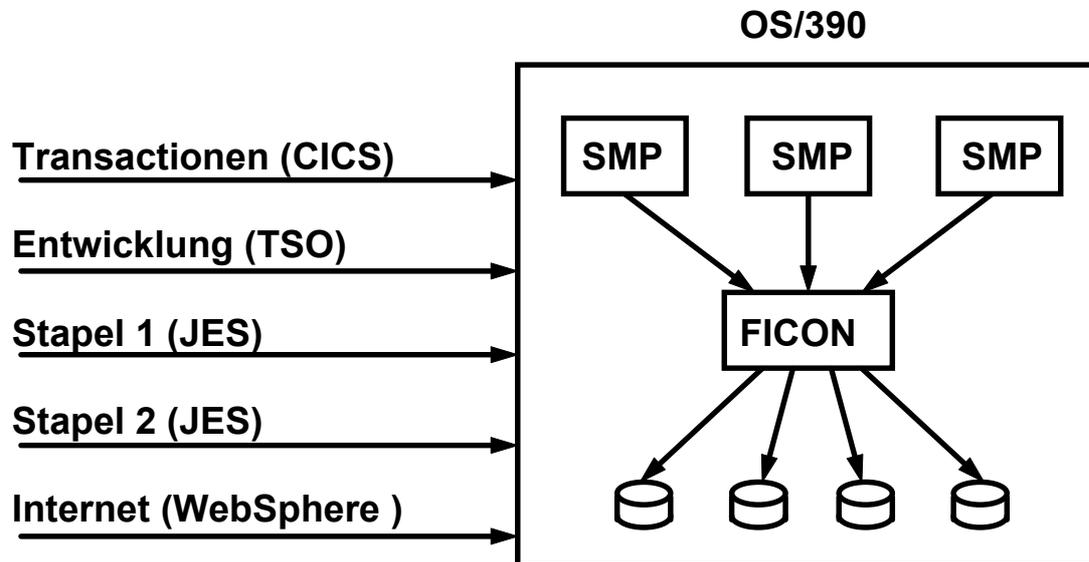
WLM plaziert Arbeitsanforderungen so, daß die Wahrscheinlichkeit alle Ziele zu erreichen optimiert wird



System Ressource Manager

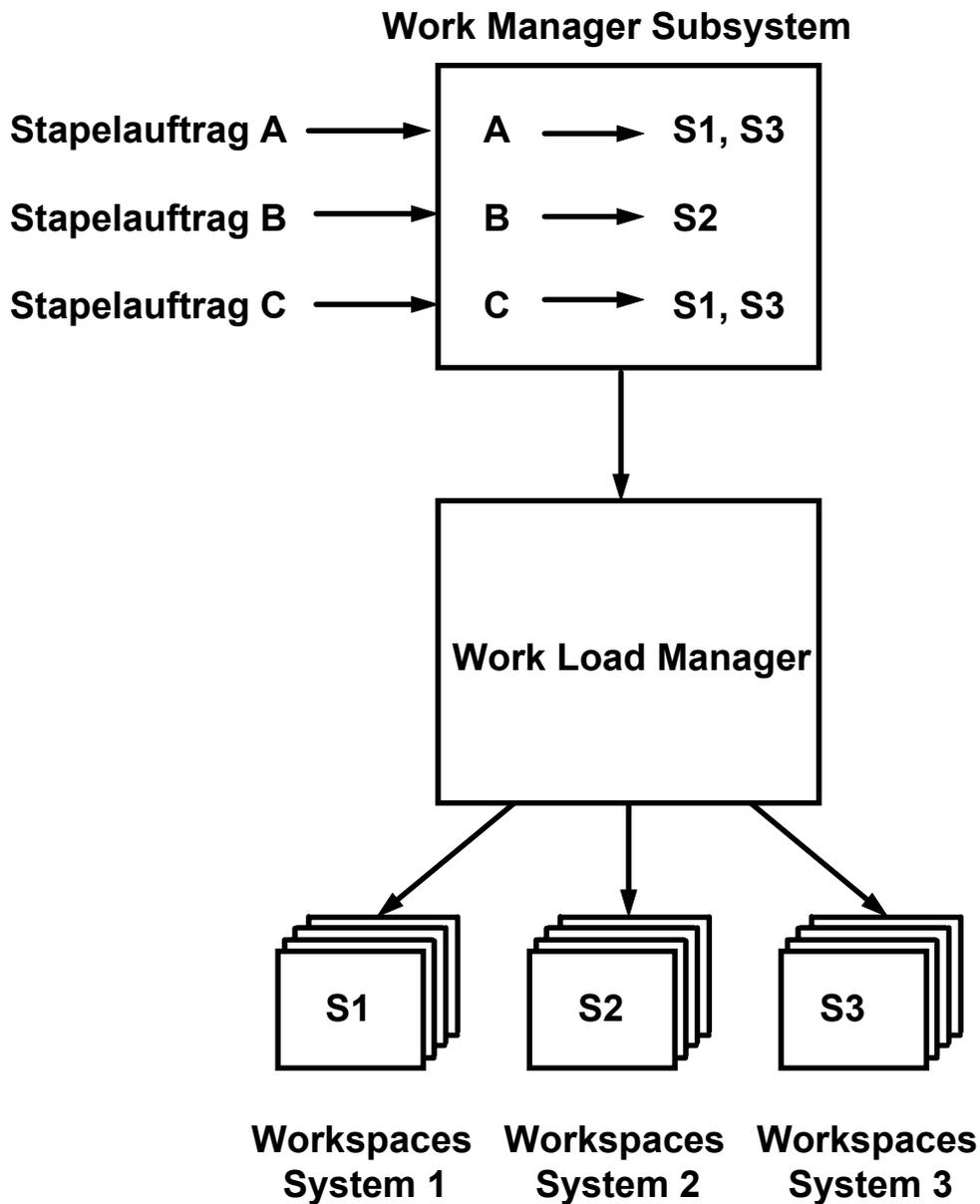
Die System Ressource Manager Komponente des Work Load Managers beobachtet für alle angeschlossenen Systeme:

- CPU Auslastung
- Hauptspeicher Nutzung
- E/A Belastung



Work Load Manager Classification Rules

Unterschiedliche Ziele für Service Classes:	Wichtigkeit
Transaktionen, 90 % Antwortzeit < 0,3 s	1
Stapel 1, 90 % complete < 3 Stunden	4
Stapel 2, optimiert für E/A Operationen	2
Internet (WebSphere) , 90 % Antwortzeit < 5 s	3
Entwicklung (TSO), 90 % Antwortzeit < 2 s	5



Zuordnung von Aufträgen zu Systemen

Einstellungen der einzelnen Systeme dynamisch an sich ändernde Belastungen anpassen und automatisch justieren

Bei widersprüchlichen Anforderungen (Regelfall) verfügt der WLM über Algorithmen, einen möglichst optimalen Kompromiss zu erreichen

Service Level Agreement (SLA)

Kontrakt zwischen einem Dienstanbieter und einem Dienstabnehmer.

In der IT-Welt häufig ein Kontrakt zwischen Unternehmensbereichen und der zentralen Unternehmens - IT (Rechenzentrum).

Mögliche Vereinbarungen:

- **Plattenspeicherplatz**
- **Rechenzyklen auf den Systemen**
- **Auslastung des Netzwerks**
- **Antwortzeiten**
- **Verfügbarkeitskriterien**

Die Aufgabe des Rechenzentrums besteht darin, die vorhandenen Systeme so abzustimmen, dass die Service Level Agreements eingehalten werden.

Das Einhalten von Service Level Agreements kann vereinfacht werden, wenn die Systeme Definitionen unterstützen, die in ihrer Form den Definitionen eines Service Level Agreements entsprechen oder nahe kommen. (Abilden von SLAs auf Zieldefinitionen - Goals - des z/OS Workload Manager).