UNIVERSITÄT LEIPZIG

Mainframe Internet Integration

Prof. Dr. Martin Bogdan
Prof. Dr.-Ing. Wilhelm G. Spruth

SS2013

Virtualisierung Teil 4

Intelligent Resource Director

Intelligent Resource Director
IRD

Der Intelligent Resource Director (IRD) ist eine Erweiterung der LPAR Technology. Die Erweiterungen bestehen im Wesentlichen aus 4 Einzelkomponenten.

- **Dynamische Verwaltung des realen Speichers**
- **LPAR CPU Management**
- **Dynamisches Channel Path Management**
- **Channel Subsystem Priority Queuing**

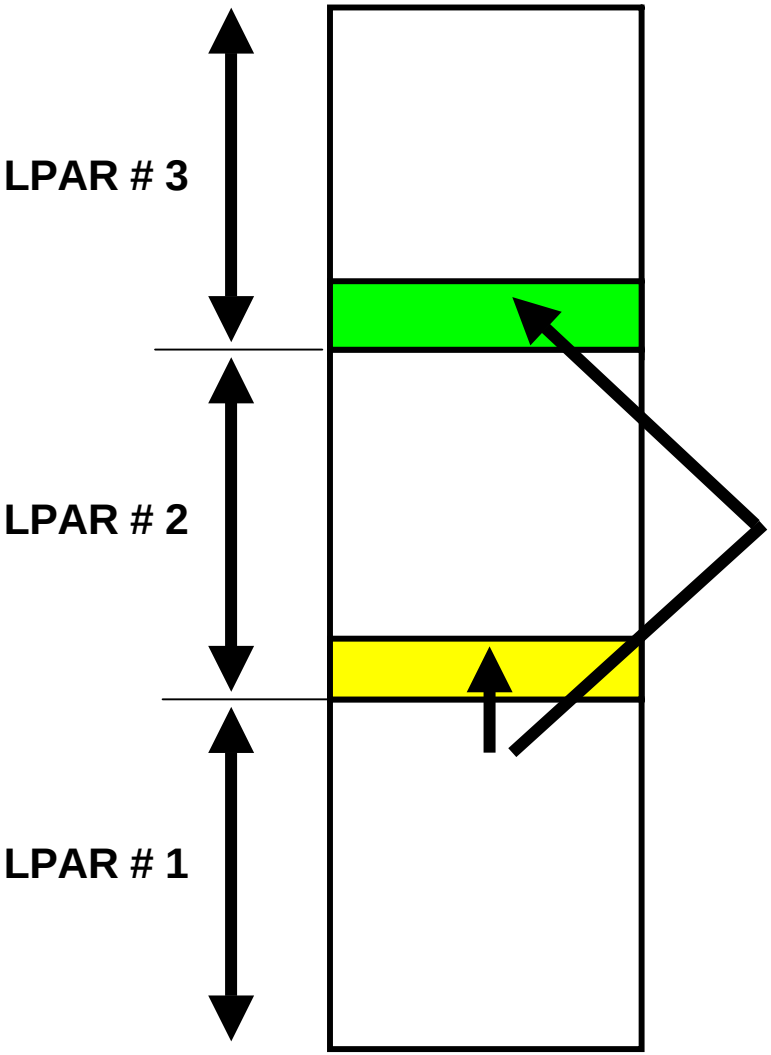
Größe des realen Speichers einer LPAR

Das Betriebssystem einer LPAR nimmt an, dass sein realer Speicher über einen kontinuierlichen linearen Adressenraum, beginnend mit der Adresse Hex 000000, verfügt.

Wird ein Mainframe Rechner neu installiert und konfiguriert, so wird entschieden, wieviele LPARs vorhanden sein sollen, und über wie viel realer Speicher jede LPAR verfügen soll. All diese Parameter werden in einer Konfigurationsdatei abgespeichert und bleiben in einfachen Fällen (z.B. auf unserem eigenen Mainframe Rechner an der Uni Leipzig) während der Lebensdauer unverändert.

Dies bedeutet, dass der LPAR x im physischen Speicher die Adressen von aaaa bis bbbbb zugeteilt werden. Der reale Speicher der LPAR verfügt über den linearen Adressenraum von 000000 bis (bbbb – aaaa). Dieser reale Adressenraum wird linear in den physischen Adressenraum von aaaa bis bbbb abgebildet. Das z/OS Betriebssystem (und auch das Windows Betriebssystem) unterstellt, dass sich bestimmte Teile des Kernels auf bestimmten realen Adressen befinden.

Will man den realen Speicher einer LPAR auf Kosten einer anderen LPAR vergrößern, ist es notwendig, alle LPARs herunterzufahren, die Konfigurationsdatei zu ändern, und die LPARs wieder hochzufahren. Es ist jedoch wünschenswert, dass dies während der z.B. 8 Jahre dauernden Lebensdauer des Rechners nie geschehen muss.



Die mehrfachen realen Speicher der einzelnen LPARs werden in einem einzigen physischen Speicher abgebildet. Jeder LPAR wird ein zusammenhängender (contiguous) Adressenbereich in dem physischen Speicher zugeordnet.

Wenn sich im Laufe eines Tages die Auslastung der LPARs ändert, würde man gerne einer LPAR zusätzlichen physischen Speicher auf Kosten einer anderen LPAR zuordnen.

In dem gezeigten Beispiel wäre es möglich, dass LPAR # 1 auf Kosten der benachbarten LPAR # 2 Speicherplatz erhält. Es wäre nicht möglich, dass LPAR # 1 auf Kosten der nicht benachbarten LPAR # 3 Speicherplatz erhält, weil jede LPAR glaubt, einen eigenen physischen Speicher mit einen linearen Adressenraum mit kontinuierlichen Adressen zu besitzen.

Änderung der Größe des realen Speichers einer LPAR

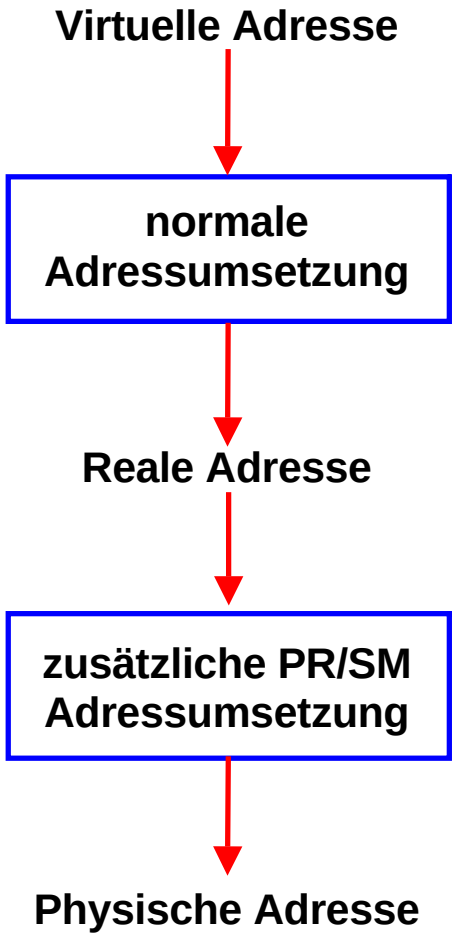
IRD Speicherplatzverwaltung

z/OS Anwendungen verwenden virtuelle Adressen, die mit Hilfe von Segment- und Seitentabellen der normalen virtuellen Adressumsetzung in reale Adressen umgesetzt werden. Die Aufteilung des physischen Speichers auf die einzelnen LPARs erfolgt in 64 MByte großen logischen „Logischen Memory Blöcken“ (LMB).

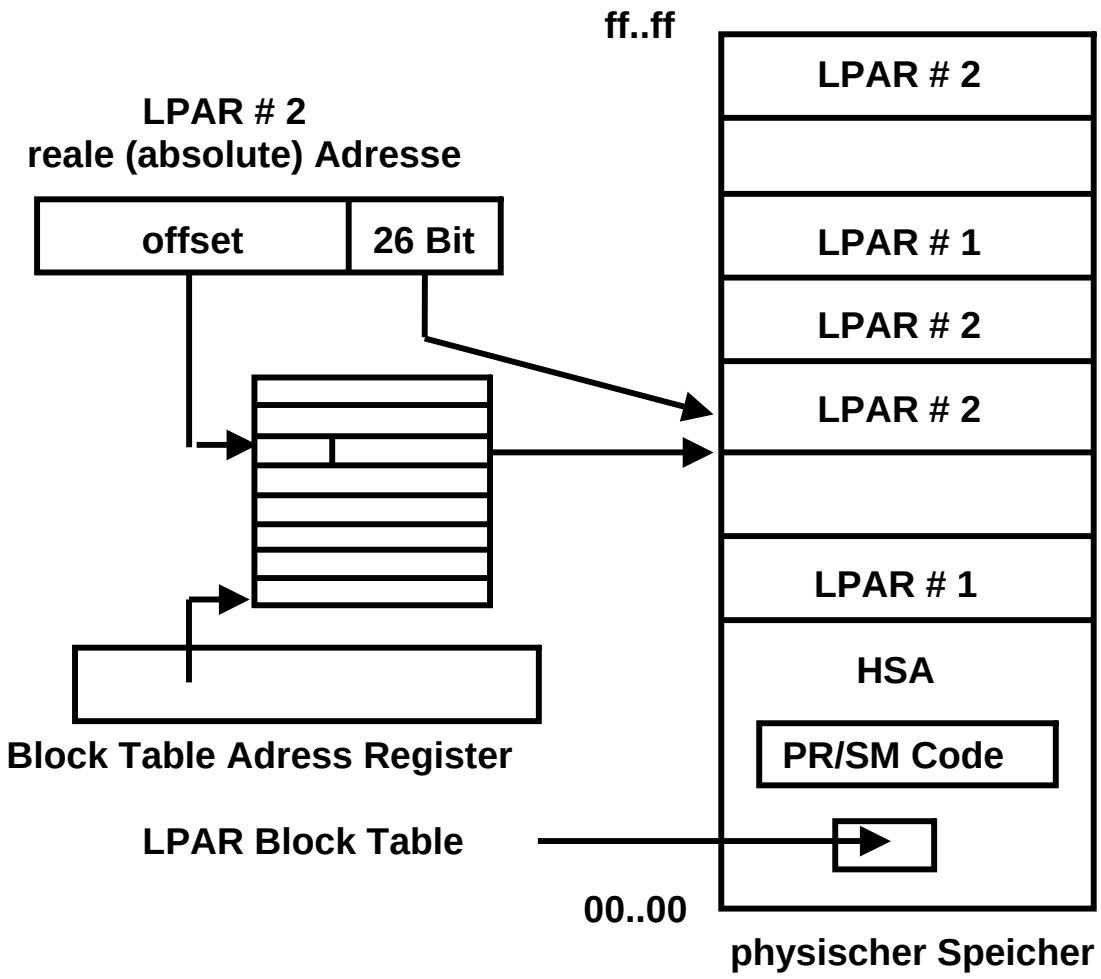
Bei der als IRD (Intelligent Resource Director) bezeichneten Erweiterung von PR/SM kann mit einem zusätzlichen Mechanismus, der ähnlich wie die virtuelle Speicherplatzverwaltung arbeitet, der den einzelnen LPARs zugeteilte Platz dynamisch (während des laufenden Betriebes) in LMB Blöcken von je 64 MByte vergrößert und verkleinert werden.

Hierbei ist nicht mehr sichergestellt, dass sich die Menge der einer LPAR zugeteilten LMBs in einem zusammenhängenden linearen Teil des physischen Speichers befindet. Damit für die LPARs die Illusion eines kontinuierlichen linearen Adressenraums gewährt bleibt, wird ein Ansatz ähnlich der virtuellen Speicherplatzverwaltung benutzt. Spezifisch wird ähnlich wie bei den Seitentabellen der kontinuierliche (contiguous) reale Adressenraum einer LPAR in eine diskontinuierliche Menge von 64 MByte Blöcken innerhalb des physischen Speichers mit Hilfe einer Block Tabelle umgesetzt.

Die Größe des den einzelnen LPARs zugeordneten physikalischen Speicherbereiches kann mit Hilfe einer weiteren Adressenumsetzung dynamisch variiert werden



Die Seiten und Rahmen der zusätzlichen PR/SM Adressumsetzung haben eine Größe von 64 MByte, gegenüber 4 KByte bei der normalen virtuellen Adressumsetzung



IRD Adressumsetzung

Der physische Speicher wird in 64 MByte große LMBs aufgeteilt.

An Stelle des Zone Origin Registers besteht ein Block Table Adress Register. Dieses enthält die Anfangsadresse eines Block Table. Für jede LPAR ist ein derartiger Block Table in der HSA enthalten. Die unteren 26 Bit der von einer LPAR erzeugten realen Adresse zeigen auf ein Datenfeld innehalb eines 64 MByte Blockes. Die oberen Bit der von einer LPAR erzeugten realen Adresse zeigen auf einen Eintrag des LPAR Block Tables. Dieser enthält die Anfangsadresse eines 64 MByte Blockes im physischen Speicher.

Die Funktion ist vergleichbar mit einer einstufigen virtuellen Address Translation. Jede LPAR hat einen linearen Adressenraum für ihre realen Adressen.

Für jeden 64 MByte Block im physischen Speicher ist ein Eintrag im LPAR Block Table vorhanden. Deshalb ist ein Fehlseiten-vergleichbarer Mechanismus nicht erforderlich.

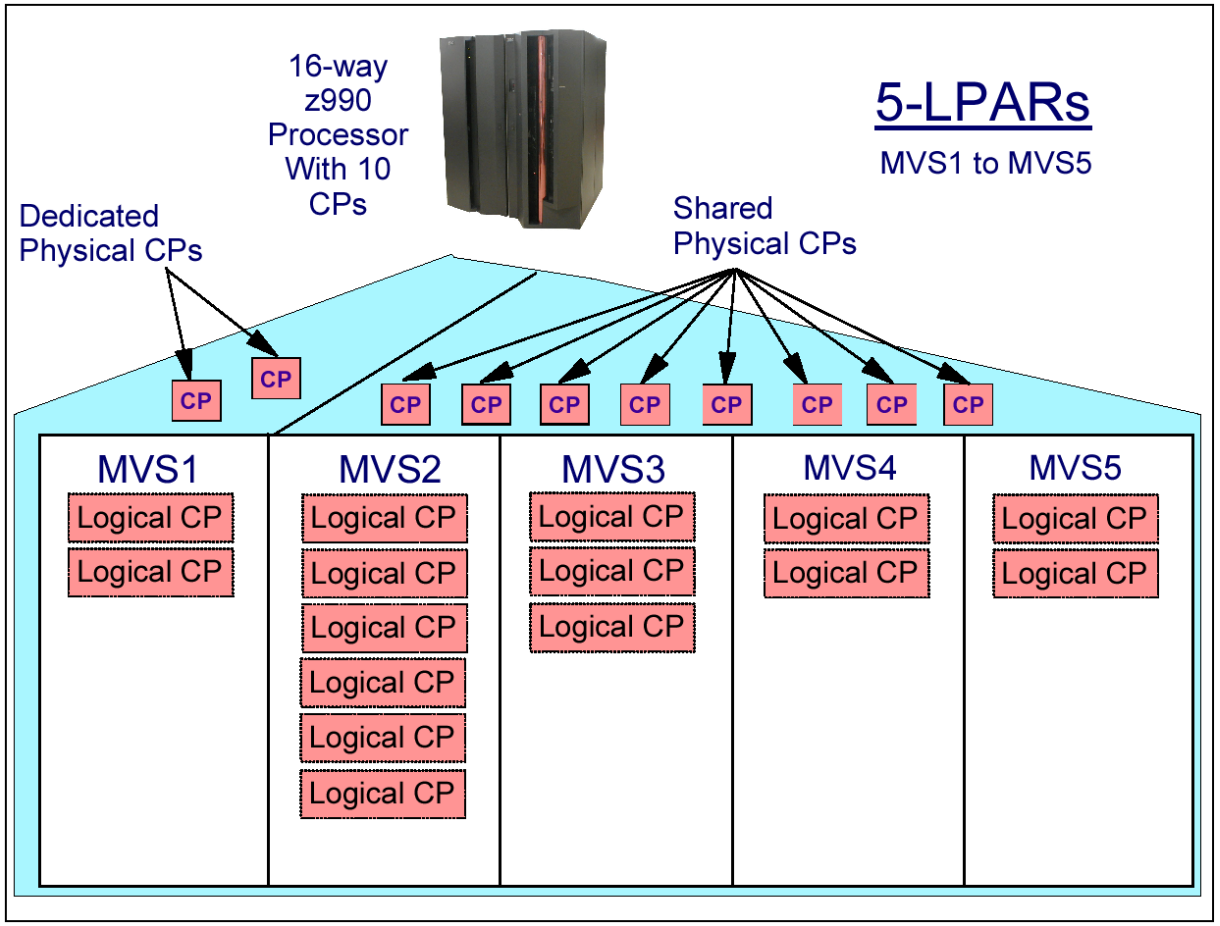
LPAR CPU Management

Bei der Systeminstallation wird ein SMP für eine definierte Anzahl von CPUs definiert. Wird beispielsweise Windows 7 auf einem neuen Quadcore PC installiert, erfolgt die Konfiguration automatisch für 4 CPUs. Bei einer Installation auf einem Dual Core PC erfolgt die Konfiguration automatisch für 2 CPUs.

In den meisten LPARs läuft in der Regel ein SMP mit mehreren CPUs. Im einfachste Fall werden die physischen CPUs auf die einzelnen LPARs fest aufgeteilt, und das Betriebssystem wird für diese Anzahl von CPUs konfiguriert. Es wäre jedoch wünschenswert, die Aufteilung dynamisch änderbar zu machen, um bei Lastschwankungen während des täglichen oder wöchentlichen Betriebs die Auslastung der CPUs zu optimieren.

Dies kann dadurch geschehen, dass man die Betriebssysteme in den einzelnen LPARs für logische CPUs konfiguriert, die dann durch PR/SM auf physische CPUs abgebildet werden. Dabei könnte und würde die Anzahl der logischen CPUs in allen LPARs zusammen die Anzahl der physischen CPUs übersteigen.

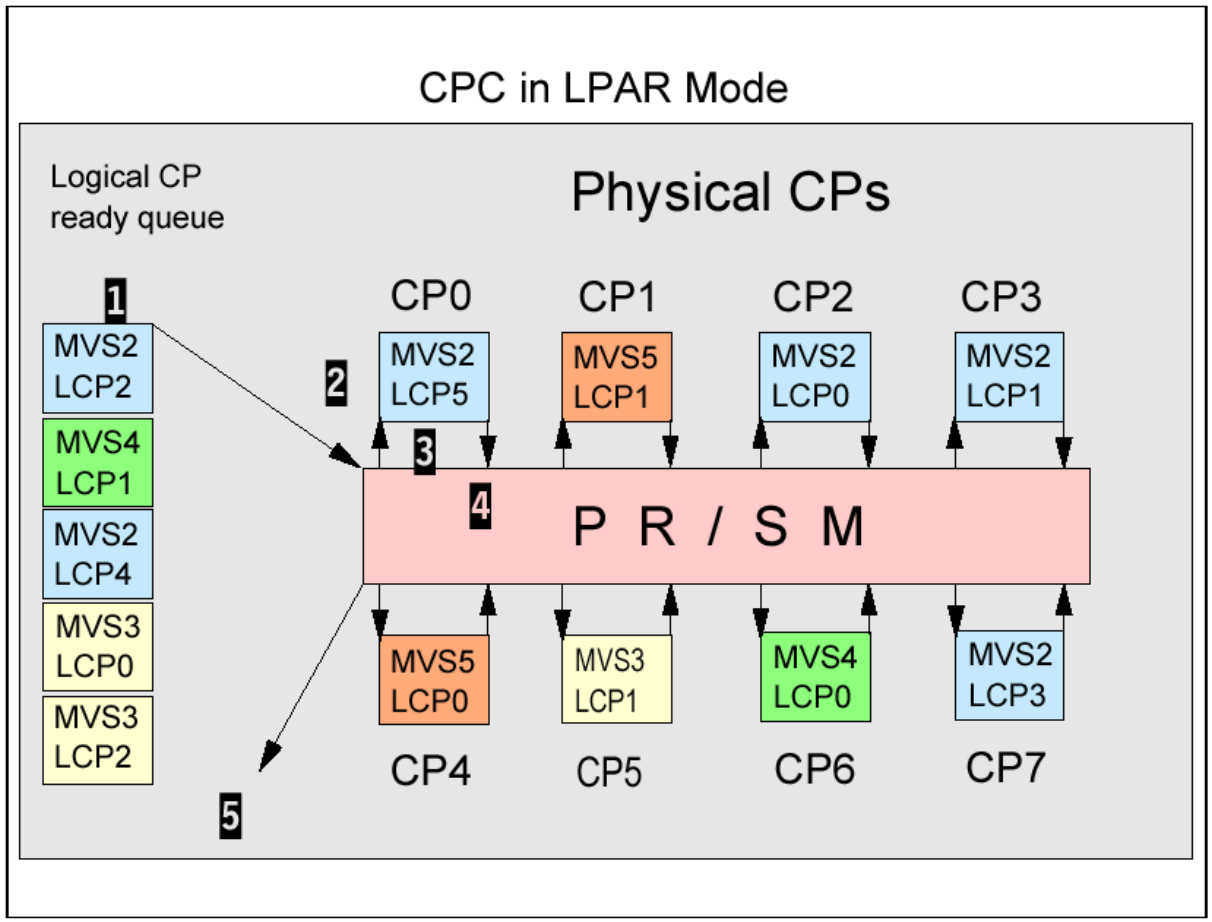
Wenn eine CPU einer anderen LPAR zugeordnet wird, werden in das LPAR Zone Origin Register und Zone Limit Register neue korrekte Adressenwerte geladen. Dies ist eine Aufgabe des PR/SM Hypervisors.



Physische CPUs (CPs) können einer bestimmten LPAR fest zugeordnet, oder von mehreren LPARs gemeinsam genutzt (shared) werden. Bei einer LPAR mit fest zugeordneten CPUs ist eine logische CPU permanent einer physischen CPU zugeordnet. Dies bedeutet weniger Overhead.

Gemeinsam genutzte (shared) physische CPUs erzeugen mehr Overhead. Dies wird überkompensiert, weil eine LPAR CPU Kapazität nutzen kann, die eine andere LPAR gerade nicht benötigt. Wenn ein Betriebssystem in den Warte- (wait) Zustand versetzt wird, gibt es die benutzten CPU(s) frei.

In dem gezeigten Beispiel sind der LPAR MVS1 zwei CPUs fest zugeordnet. Die vier LPARs MVS2, MVS3, MVS4 und MVS5 teilen sich in die Nutzung von acht physischen CPUs.



Den vier LPARs mit den Betriebssystemen MVS2, MVS3, MVS4 und MVS5 stehen 8 physische CPs (CPUs) zur Verfügung (CP0 .. CP7).

Jedes der Betriebssysteme ist als Multiprozessor konfiguriert und glaubt, über eine bestimmte Anzahl (logischer) CPUs zu verfügen. Die Anzahl der logischen CPUs übertrifft die Anzahl der physisch vorhandenen CPUs.

Der PR/SM Hypervisor ordnet die logischen CPs den physisch vorhandenen CPs zu. Er unterhält eine „ready Queue“ der ausführbaren logischen CPUs aller LPARs. Wenn eine physische CPU frei wird, ordnet ihr PR/SM eine logische CPU aus der ready Queue zu.

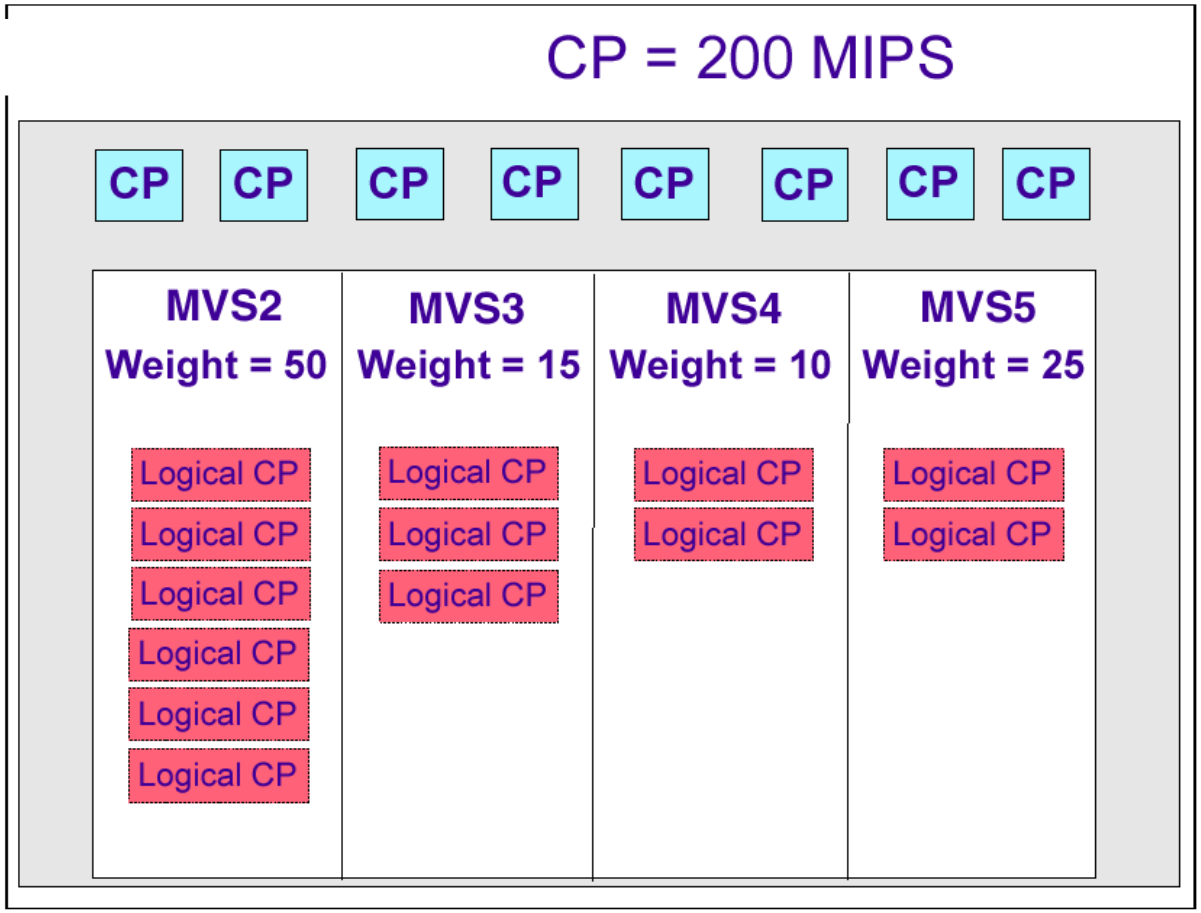
LPAR dispatching

LPAR dispatching

Der LPAR Dispatching Code ist Teil des PR/SM Hypervisor.

Angenommen, eine physische CPU wird frei, z.B. CP0 in dem obigen Beispiel. Das Dispatching einer logischen CPU durch CP0 erfolgt in den folgenden Schritten:

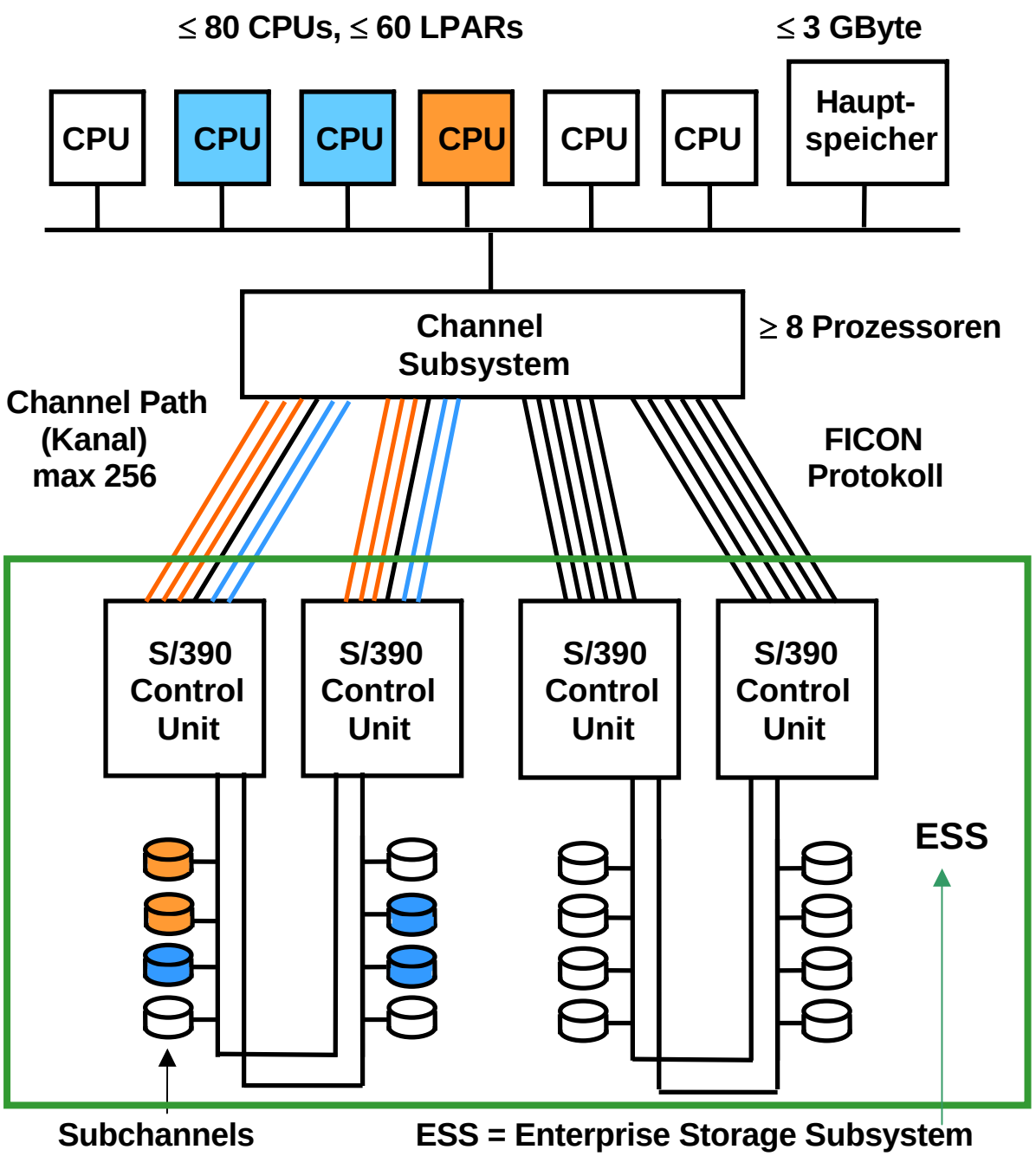
1. **Die als nächstens zu dispatchende logische CPU wird wird von der logischen CP ready Queue ausgewählt, entsprechend dem Gewicht (Weight) der logischen CPU.**
2. **LPAR Firmware (LIC) dispatches die selektierte Logische CPU (LCP5 von MVS2) auf eine physische CPU (CP0 in der obigen Abbildung).**
3. **Die z/OS dispatchable Unit, die auf der logischen CPU (MVS2 logical LCP5) läuft, startet die Ausführung auf der physischen CP0. Sie läuft, bis die Zeitscheibe (typisch zwischen 12.5 und 25 ms) abgelaufen ist, oder ein Wait Ereignis eintritt.**
4. **Bei Zeitscheiben-Ende wird die Laufzeitumgebung der logischen CP5 abgespeichert (saved). Kontrolle geht zurüch an den LPAR LIC, der auf der physischen CP0 wieder startet.**
5. **LPAR LIC ermittelt, warum die logische CPU die Ausführung beendete, und queued diese entsprechend. Wenn LPC5 ausführbar ist, wird sie wieder in die logical CP ready queue eingereiht. Darauf beginnt Schritt 1 von neuem.**



Mit Hilfe von LPAR Gewichten (weights) wird die Zuordnung von logischen CPUs auf physische CPUs gesteuert.

LPAR Gewichte bestimmen das garantierte Minimum an physischen CPU Ressourcen, welche eine logische CPU erhält, falls diese sie anfordert. Dieser garantierte Betrag ist gleichzeitig das Maximum, wenn alle logischen CPUs das garantierte Minimum voll ausschöpfen.

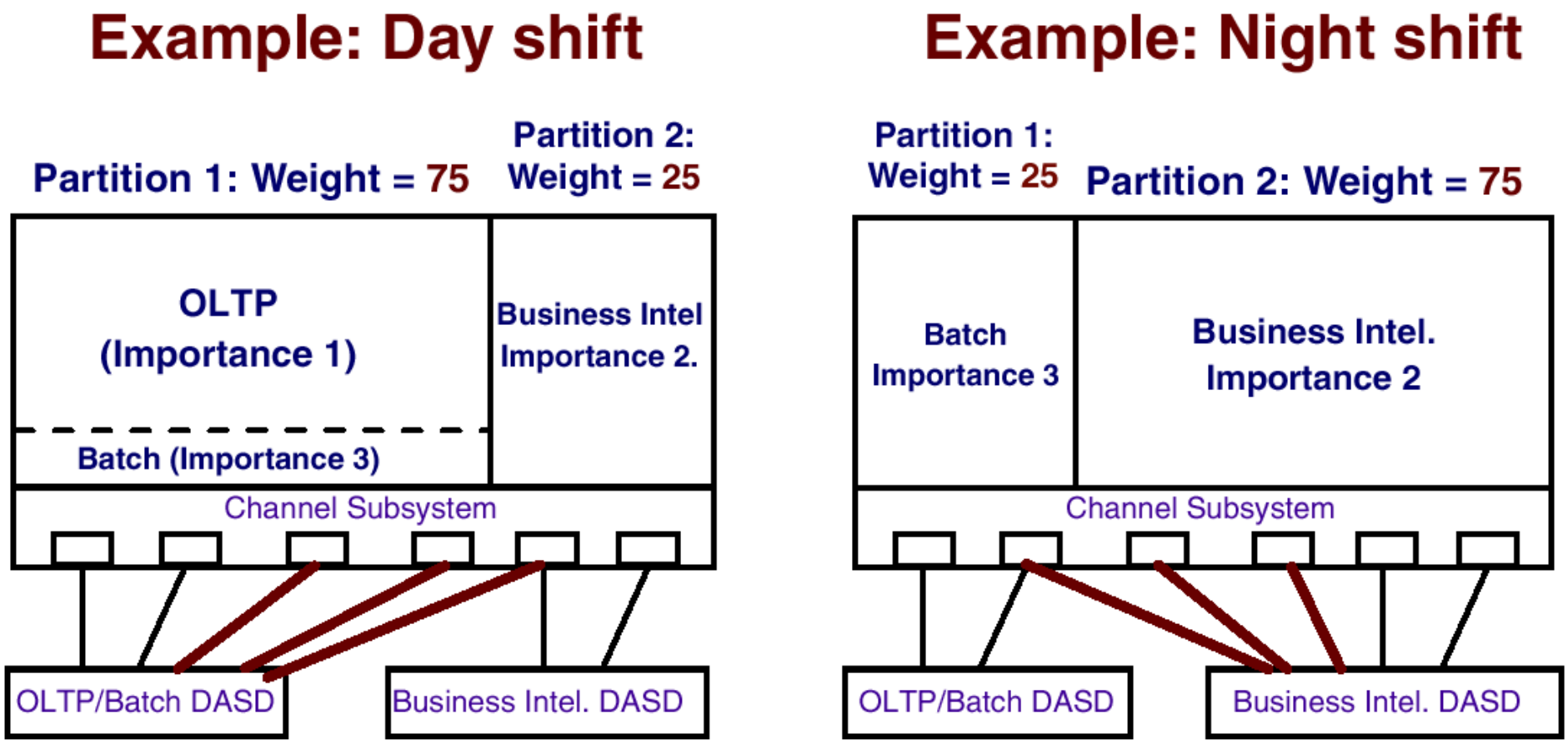
Eine LPAR verbraucht möglicherweise weniger als das garantierte Minimum, wenn nicht ausreichend Arbeit anfällt. In diesem Fall steht den anderen LPARs mehr CPU Leistung zur Verfügung.



Dynamic Channel Path Management (DCM)

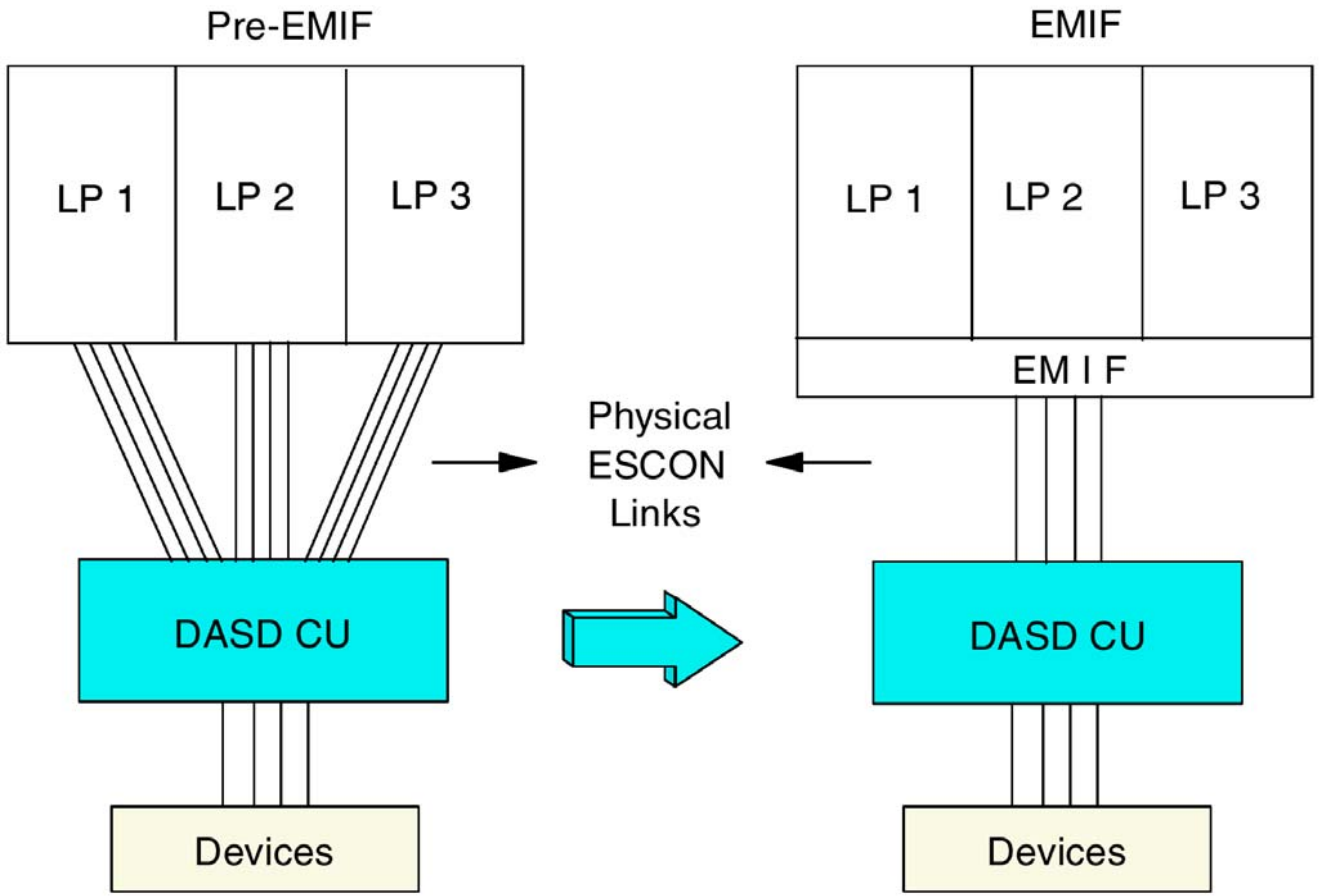
Plattenspeicher werden über ein Enterprise Storage Subsystem (ESS) angeschlossen, welches S/390 Control Units abbildet.

Dynamic Channel Path Management (DCM) ist in der Lage, die Kanal Configuration dynamisch an sich ständig ändernde Auslastungen anzupassen.



Dynamic Channel Path Management (DCM)

In dem gezeigten Beispiel sind während der Tagesschicht die Mehrzahl der Kanäle den OLTP (Online Transaction Processing) Anwendungen zugeordnet. Während der Nachtschicht sind die Mehrzahl der Kanäle einer LPAR für Batch Processing Anwendungen (z.B. Business Intelligence) zugeordnet.



Multiple Image Facility (MIF)

Im einfachsten Fall sind die Kanäle einzelnen LPARs fest zugeordnet.

Mit Hilfe der (extended) Multple Image Facility (EMIF) können Kanäle von mehreren LPARs gemeinsam genutzt werden.

An Stelle der Bezeichnung MIF wurde früher auch die Bezeichnung EMIF verwendet.

Channel subsystem I/O priority queuing

Normalerweise werden I/O Anforderungen vom Channel Subsystem auf einer first-in, first-out Basis abgearbeitet.

Wenn eine Arbeitseinheit mit hoher Priorität ihre Bearbeitungsziele wegen Überlastung der I/O Einrichtungen nicht erreicht, kann der z/OS Work Load Manager (siehe nächstes Thema) dieser Arbeitseinheit eine höhere Priorität als anderen Arbeitseinheiten zuordnen. Die „Channel Subsystem Priority Queuing" Einrichtung des Channel Subsystems wird dann diese Arbeitseinheit bevorzugt abfertigen, indem sie hierfür zusätzliche Bandbreite auf Kosten anderer Arbeitseinheiten mit weniger Priorität verfügbar macht..

Hierfür werden Dynamic Channel Path Management und MIF eingesetzt.

PR/SM and LPAR Sicherheit und Isolation

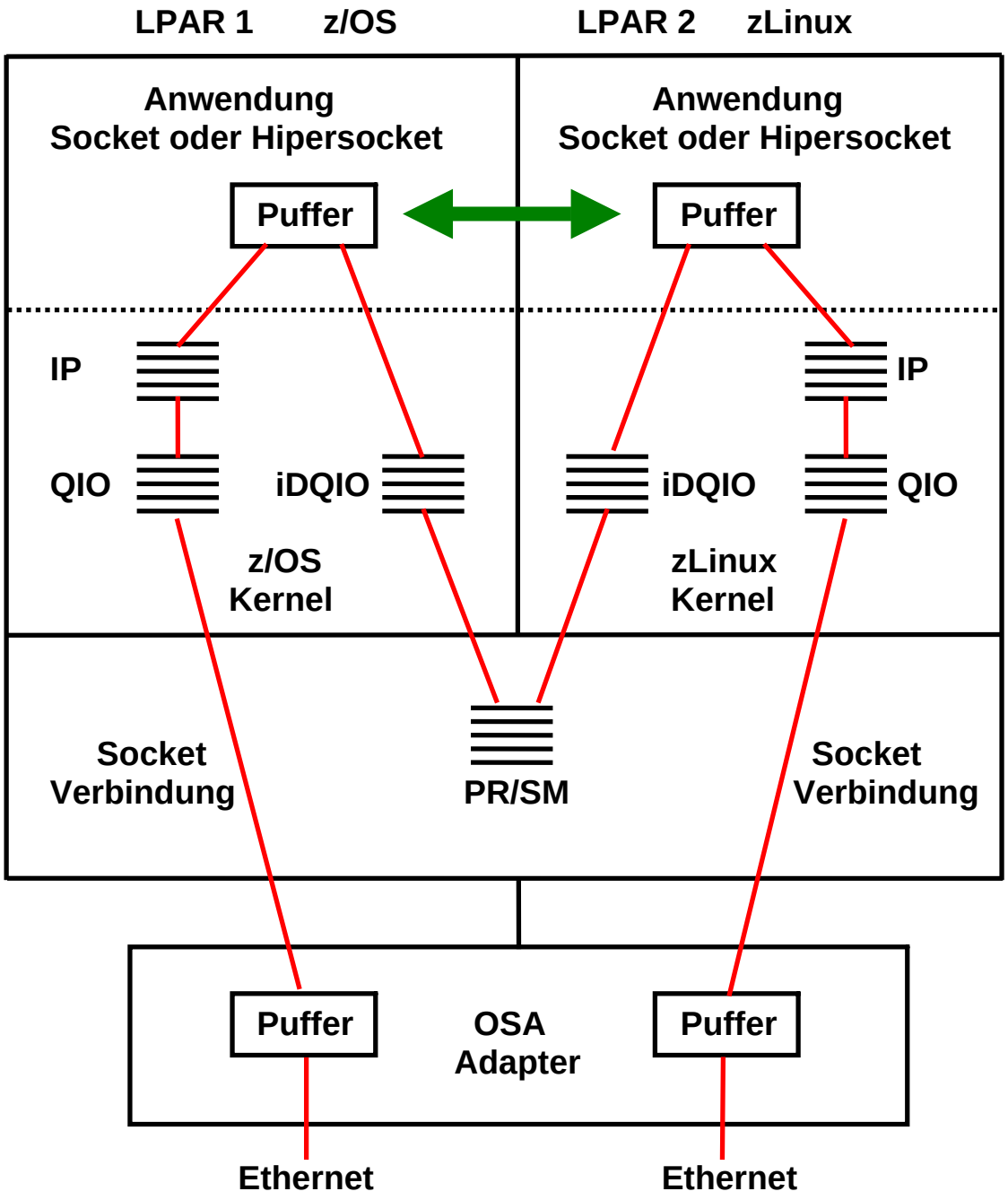
Zertifikat der USA Regierung: LPARs haben äquivalente Sicherheits-Eigenschaften wie physisch getrennte Rechner“.

Das Bundesamt für Sicherheit in der Informationstechnik (BSI) stellt IBM für den Processor Resource/System Manager (PR/SM) des Mainframes 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. Mainframes mit PR/SM waren damals die ersten Server, die nach der Evaluierungsstufe EAL5 für seine Virtualisierungstechnologie zertifiziert wurde.

Die Zertifizierung des BSI bescheinigt, dass Programme, die auf einem Mainframe Server in verschiedenen logischen Partitionen (LPAR) laufen, ebenso gut voneinander isoliert sind, als würden sie auf getrennten physikalischen Servern laufen.

Die Logische 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.

IBM Presseinformation SG/94/2003, CeBIT 2003: 14. März 2003 -



Socket bzw. Hipersocket Verbindung

Dargestellt sind zwei LPARs, die eine Nachricht austauschen möchten.

Da sich die LPARs wie getrennte physische Rechner verhalten, kann man jeder LPAR einen Ethernet Adapter Port zuordnen, und die beiden Ports über ein Ethernet Kabel miteinander verbinden.

System z verfügt über den OSA Adapter, der als eine Reihe von Ethernet Adapter Ports konfiguriert werden kann.

Das z/OS Communication Manager Subsystem unterstützt eine Ethernet Verbindung mit der „Queued I/O Access Method“ (QIO).

LPARS auf dem gleichen Rechner können statt dessen iDQIO verwenden.

HiperSockets

HiperSockets ermöglichen es mehreren LPARs innerhalb des gleichen Rechners miteinander zu kommunizieren, ohne auf ein externes, physisches Netzwerk zurückgreifen zu müssen (LPAR-übergreifende Kommunikation). Mit dieser Funktion lässt sich innerhalb des Systems ohne eine zusätzliche physische Verbindung ein "systeminternes Netz" aufbauen.

Die Hochgeschwindigkeitskommunikation über HiperSockets kann für die Kommunikation zwischen einer Mischung von z/OS, z/VM und zLinux Betriebssystemen in unterschiedlichen LPAR Instanzen eingesetzt werden.

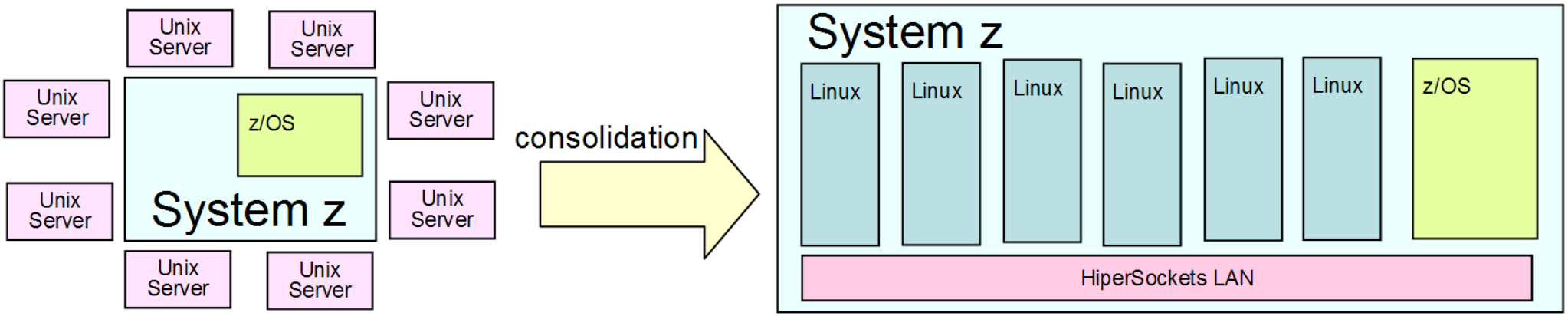
Da die Kommunikation den Hauptspeicher benutzt, erfolgt sie mit Hauptspeicher-Zugriffsgeschwindigkeit. Es bestehen zusätzlich Sicherheitsvorteile, weil die Möglichkeit einer Netzwerk Interception (z.B. Man-of-the-Middle Attack) von außen entfällt. HiperSockets verbessern Reliability und Availability, weil Network Hubs, Routers, Adapters oder Kabelverbindungen alle im Hauptspeicher virtuell dargestellt werden.

Die HiperSockets Implementierung basiert auf der OSA-Express Queued Direct I/O (QDIO)-Protokoll. Daher werden HiperSockets auch als internes QDIO oder IQDIO bezeichnet. Der Microcode des OSA Express Adapters emuliert den Link Control Schicht (Schicht 2 des OSI Stacks) einer OSA-Express QDIO Schnittstelle.

VSWITCH ist eine Firmware Funktion, die unter Nutzung von iQDIO wie ein OSI Layer 2 Switch arbeitet. Es verbindet ein externes Netzwerk über den OSA Adapter mit den virtuellen Gast Maschinen.

HiperSockets ermöglichen in einem Rechner bis zu 16 “virtual” Local Area Networks (LANs) mit deutlich reduziertem System Overhead.

HiperSockets können für die Kommunikation in einem einzigen physischen System oder zwischen Sysplex Instanzen eingesetzt werden. Sie unterstützen ebenfalls die zBX (siehe Thema Hybrid Computing Teil 3).



In vielen Unternehmen und staatlichen Organisationenexistieren neben dem Mainframe hunderte oder tausende von dezentralisierten Servern. Diese Server sind vielfach über ein historisch gewachenes Netzwerk von Ethernet Verbindungen, Switches und Routern miteinander verbunden. Um die Administrationskosten zu senken ist es wünschenswert, diese Agglomeration von Servern zu zentralisieren.

Angenommen, auf allen dezentralisierten Servern läuft Linux als Betriebssystem. Es ist möglich, die dezentralen Server abzubauen und statt dessen eine Reihe von zLinux Instanzen auf einem größeren Mainframe Server einzurichten.

Die CPU Auslastung auf dezentralen Servern liegt häufig unter 20 %. Da zLinux Prozessoren mit einer höheren Auslastung (bis zu 100 %) betrieben werden können, kann die Anzahl der zLinux Instanzen deutlich kleiner als die Anzahl der bisherigen Linux Server sein.

Die Firma IBM macht Reklame mit dem relativ niedrigen Energie Verbrauch und der Umweltfreundlichkeit ihrer Mainframe Server. In Bezug auf Linux Server Konsolidierung ist dies berechtigt.

Rezentralisierung mit zLinux

Mit der Rezentralisierung zahlreicher Linux Server auf dem Mainframe kann die komplexe Ethernet Infrastruktur vereinfacht werden. Da physische Ehernetkabel durch virtuelle Hipersocket Verbindungen ersetzt werden, ist ein Abhören der Verbindungen (Man of the Middle Attack) nicht mehr möglich. Komplexe Verschlüsselungsverfahren können entfallen. Separate Firewall Rechner und mehrfache Demilitaized Zones (DMZ) sind möglicherweise ebenfalls nicht mehr erforderlich. Die physische Ethernet Struktur wird durch virtuelle LANs ersetzt und dabei vermutlich wesentlich vereinfacht.

Über Hipersockets können die zLinux Instanzen Dienste des z/OS Security Servers wie RACF und LDAP nutzen. Es ist möglich, zLinux LPARS auf unterschiedlichen Systemen eines Sysplex über Hipersockets und XCF miteinander zu verbinden.