Modulo 5 : La Virtualizzazione Tramite Hardware

Obiettivi

- Descrivere dispositivo denominato PR/SM
- Definire le caratteristiche di funzionamento del PR/SM
- Descrivere le modalita' di condivisione delle CPU e di dispositivi di I/O

Premessa

La Virtualizzazione di Risorse Informatiche tramite Hardware e' una caratteristica specifica dei Sistemi IBM ed in particolare dei Sistemi Centrali caratterizzati dalla z/Architecture .

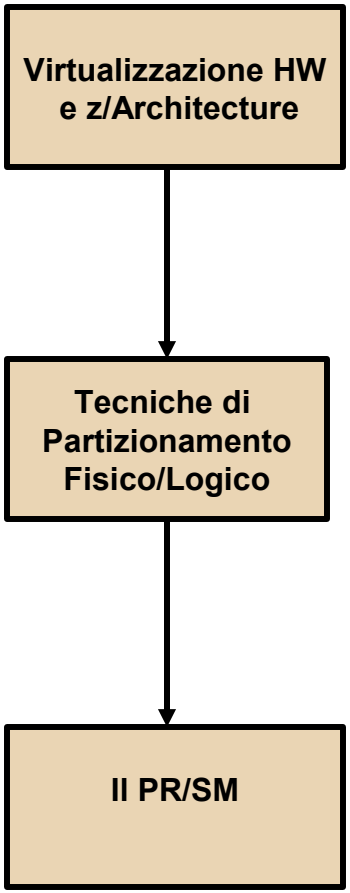
IL Processor Resource & Systems Management (PR/SM) introdotto da IBM nel 1986 sui calcolatori IBM 3090 della famiglia S/370 e poi riportato come un dispositivo standard sulle piu' recenti Famiglie S/390, ES9000, zSeries e zSystems.

Si tratta di un **Firmware microcodificato** e non modificabile dall'utente

Per l'efficienza e le caratteristiche e' stato riproposto con le dovute modifiche anche su altre Famiglie di elaboratori con diversa architettura (RISC) detti pSeries dove viene chiamato **Micro-Partitioning**

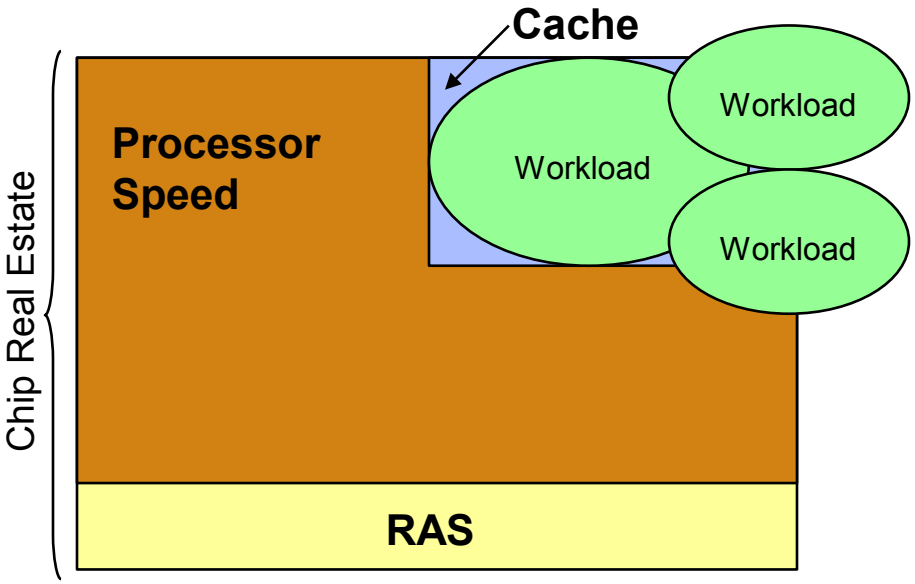
Il PR/SM rimane ad oggi una caratteristica **Unica** ed Univoca dei Sistemi Centrali **IBM** e non e' stato ancora eguagliato in efficienza e funzioni da nessuna altra implementazione.

le tecniche di Partizionamento piu' comuni per approdare a passi successivi al modello PR/SM che ne rappresenta la implementazione piu' avanzata e che possiede altresi' caratteristiche di Virtualizzazione.



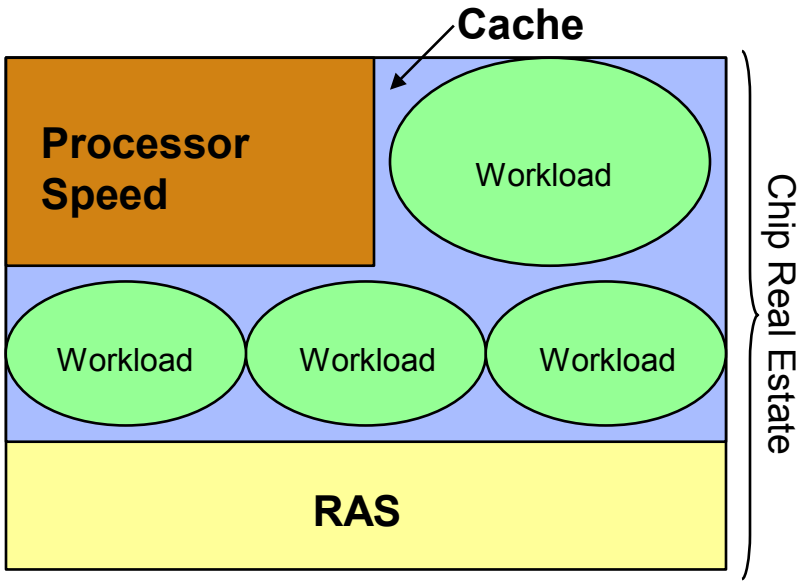
Il Design dei Chips influisce sulle capacità di Virtualizzazione

Replicated Server Chip Design



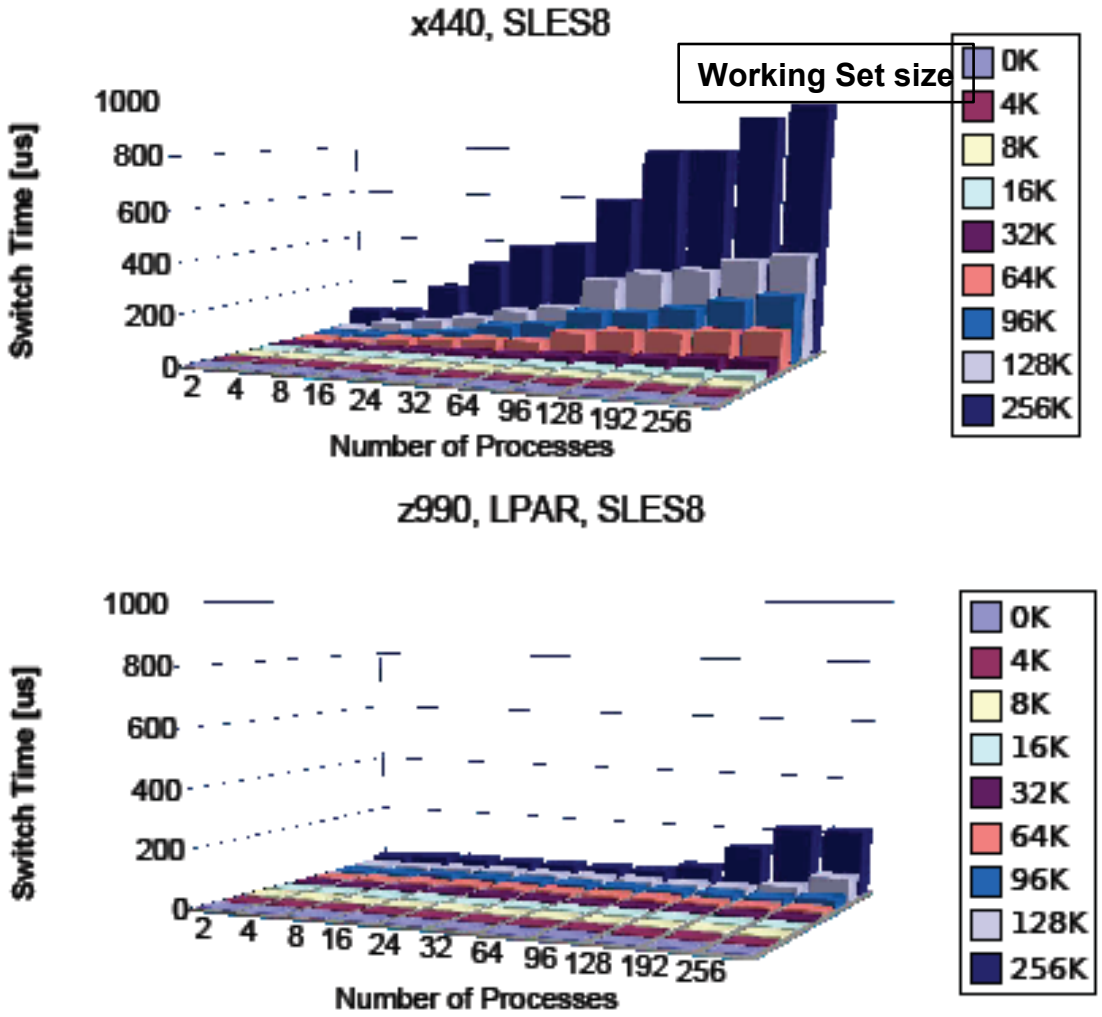
- Carichi di lavoro misti e contemporanei stressano l'uso della cache, richiedendo un numero alto di context switches
- Working sets potrebbero essere troppo grandi per le dimensioni della cache
- La velocità di un processore "Fast" non è pienamente sfruttata a causa dei continui "cache misses"

Consolidated Server Chip Design



- La cache System Z è in grado di contenere più working sets e di dimensioni maggiori
- La velocità del Processore è ottimizzata in relazione alle dimensioni e all'uso della cache

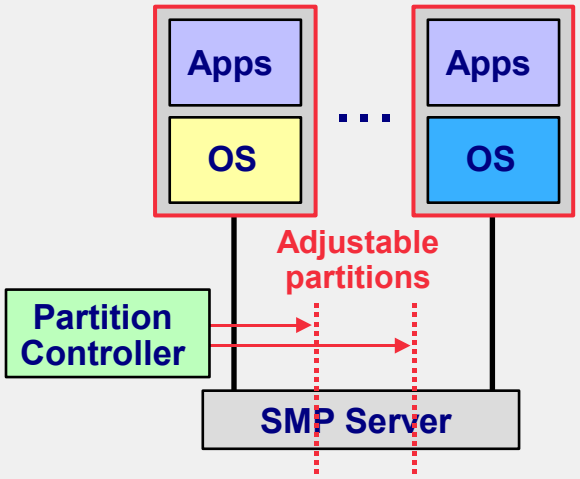
Considerazioni sulla Scalabilità: Context Switching



- Virtualizzazione, - per definizione, comprende il “context switching”
- Il Tempo richiesto per eseguire un “context switching” , è un indicatore del memory time

Approcci alla Server Virtualization

Partizionamento Fisico

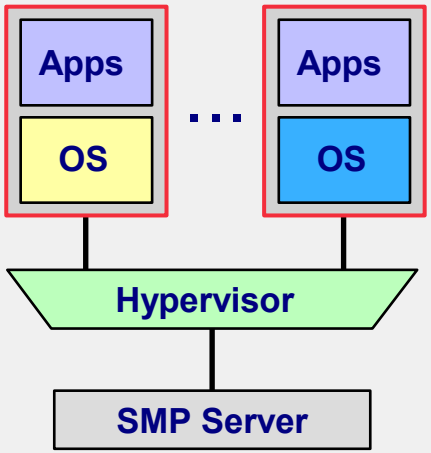


Server suddiviso in frazioni, in ognuna delle quali gira un OS

Physical partitioning
S/370 SI-to-PP and PP-to-SI, Sun Domains, HP nPars

Logical partitioning
pSeries LPAR, HP (PA) vPars

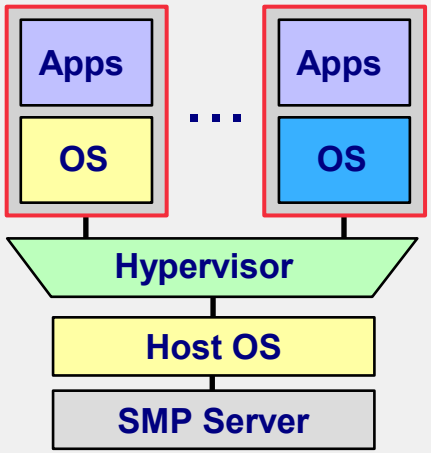
Firmware su "metallo"



Firmware fornisce la condivisione a "grana fine" di tutte le risorse

System z PR/SM and z/VM
POWER Hypervisor
VMware ESX Server
Xen Hypervisor

Firmware su SW



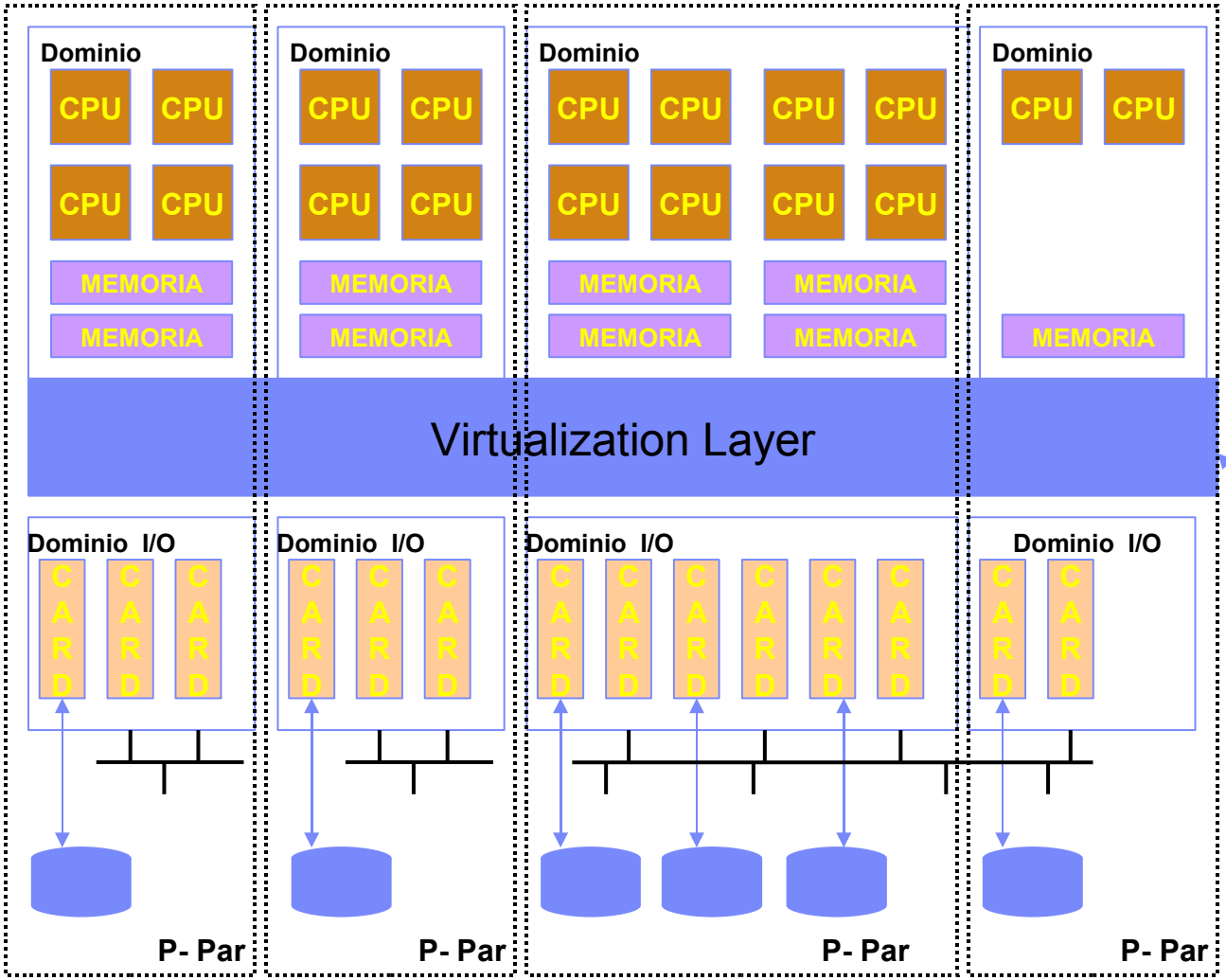
Firmware usa I servizi del OS per la condivisione delle risorse

VMware GSX
Microsoft Virtual Server
HP Integrity VM
User Mode Linux

Outlook:

- "Firmware su Metallo" diventerà la soluzione dominante per i servers (alta efficenza e disponibilità
- "Firmware su SW" utilizzato per workstation per clients,
- "Hardware partitioning" è ormai un approccio superato

Partizionamento Fisico - PPAR



Partizionamento fisico PPAR

Si tratta di una **tecnica** usata da HP , SUN ed IBM solo nei Sistemi di Architettura **INTEL**

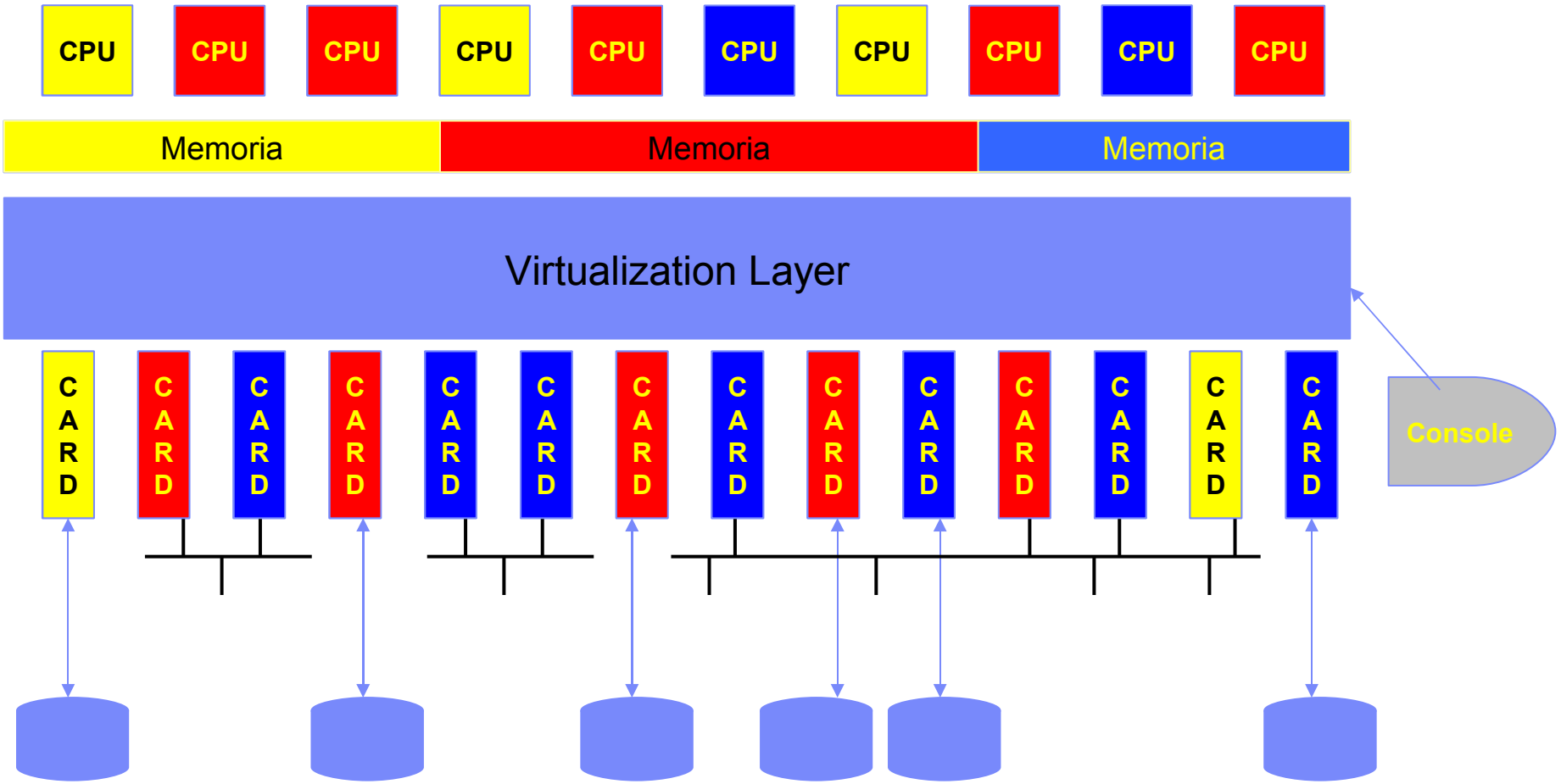
.

Le Partizioni Fisiche sono in genere correlate con la struttura fisica della macchina e poco modificabili.

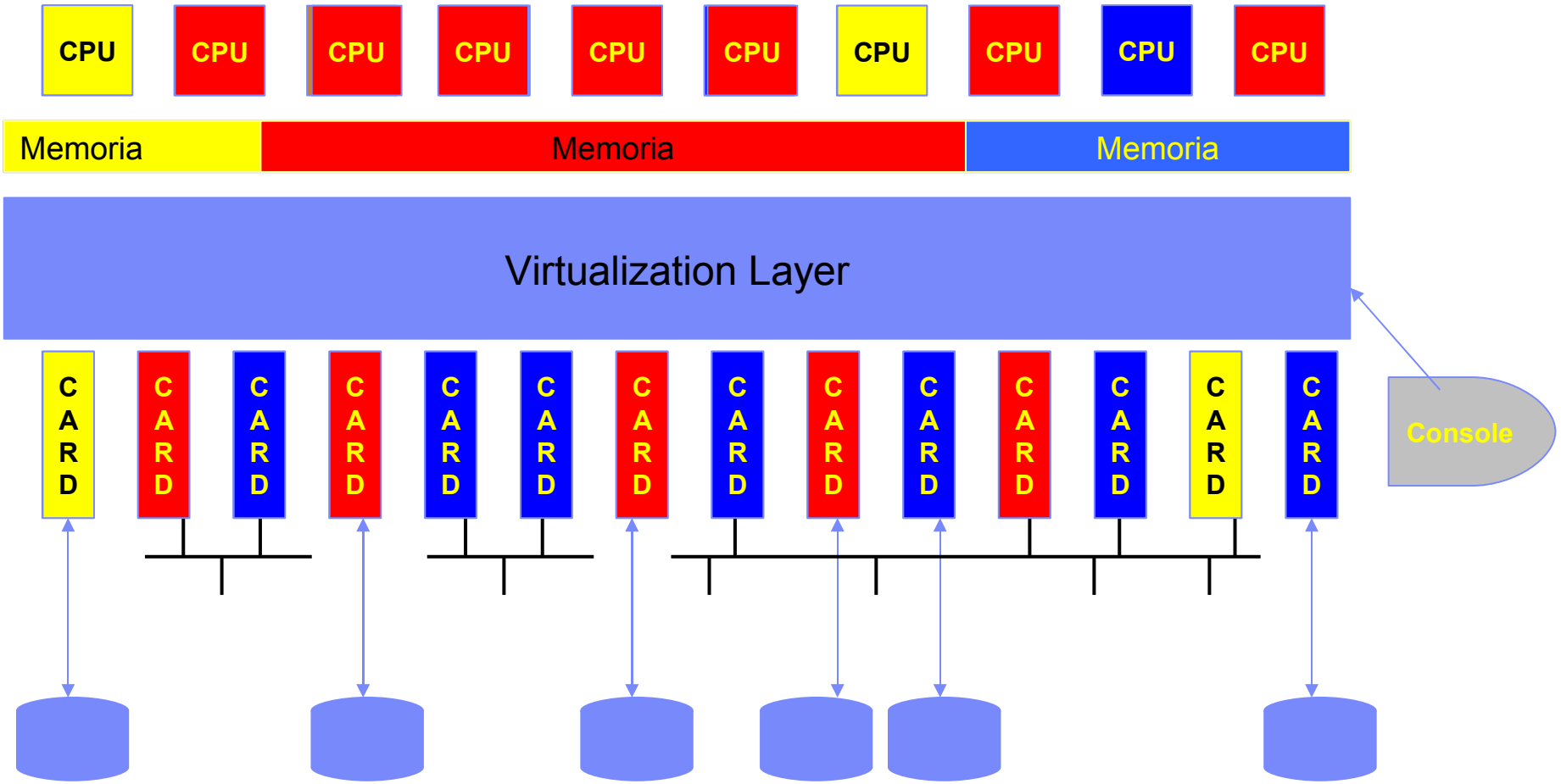
Le risorse non sono condivise e/o virtualizzate

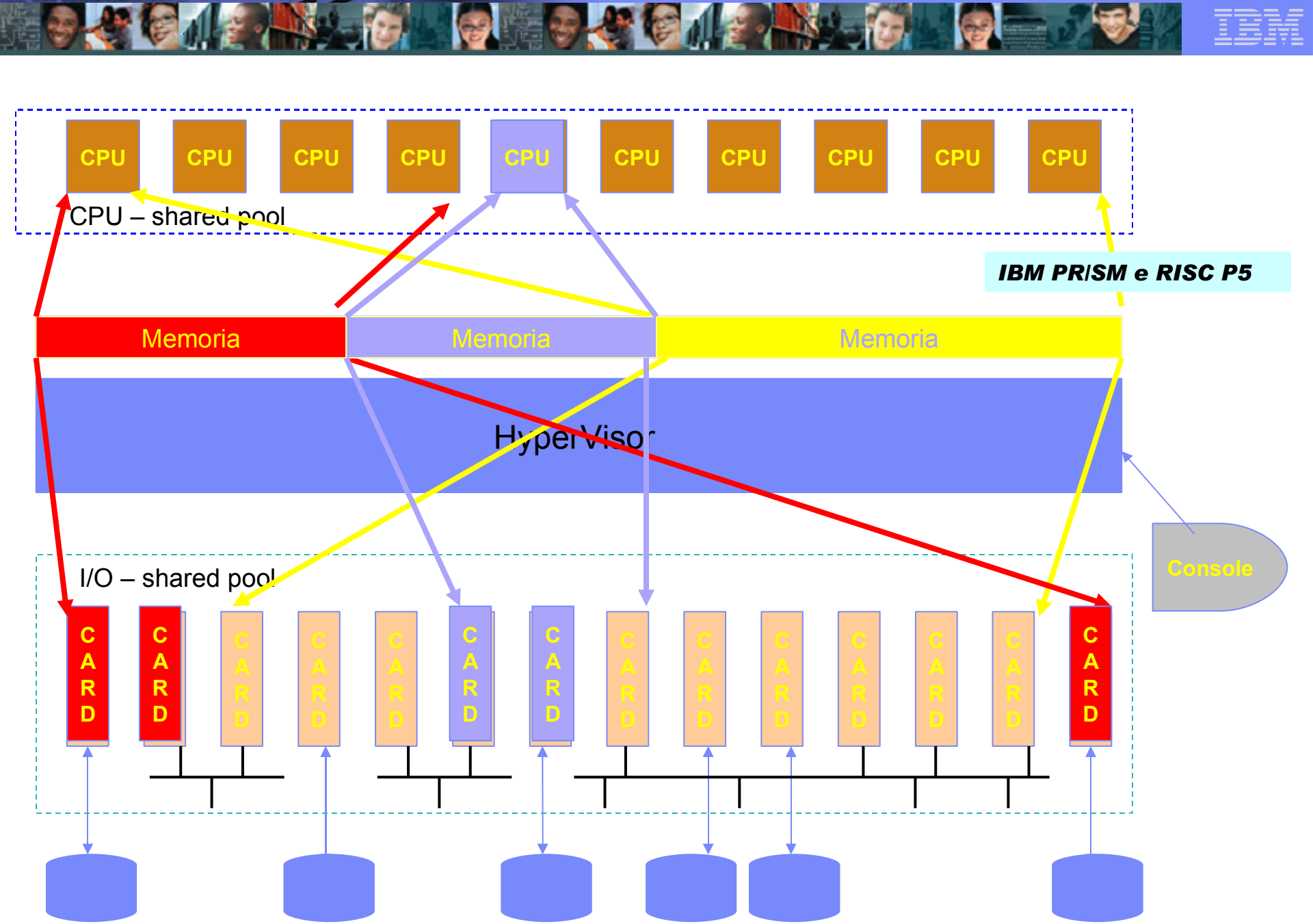
Partizionamento Logico a risorse dedicate- LPAR

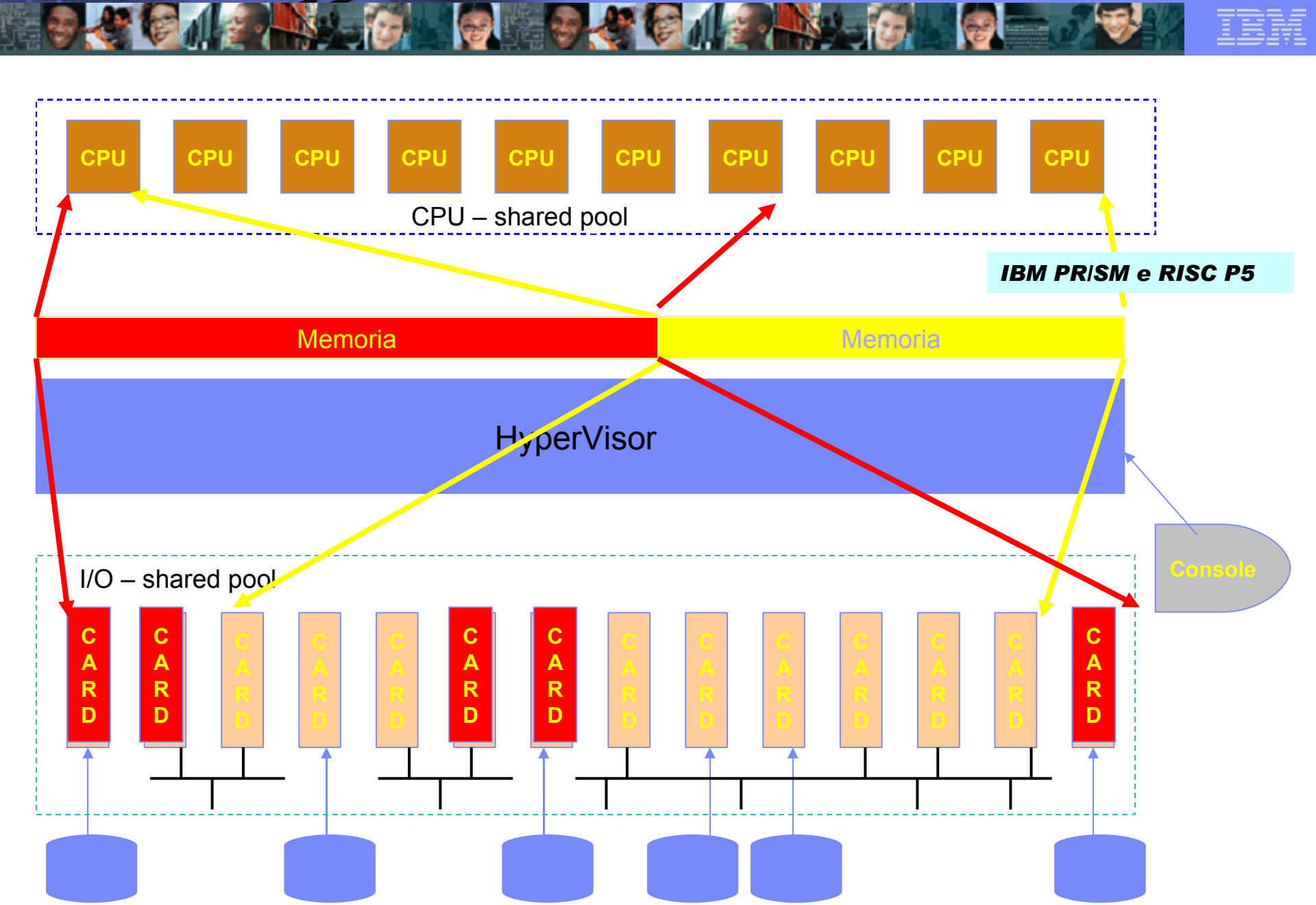
Si tratta di una tecnica usata daIBM nei Sistemi RISC con P4



Partizionamento Logico Dinamico – Dyn LPAR







Caratteristiche del PR/SM

- Livelli Multipli di Virtualizzazione
- Completa Condivisione di Tutte le risorse
- Supporta Diversi Sistemi Operativi
- Partizioni Logiche con differenti TOD
- Allocazione Dinamica delle Risorse
- Eventuali Risorse Dedicate
- Capping
- Intelligent Resource Director

Numero di LPAR :

3.	Z900	= 15
4.	Z800	= 15
5.	Z890-110	= 15
6.	Z890	= 30
7.	Z990	= 30
8.	Z9-109	= 60

Un Sistema con un solo processore su **z9-109** puo' avere **60 LPAR** , quindi la granularita' minima di una LPAR e' di **un sessantesimo di processore.**

Vocabolario

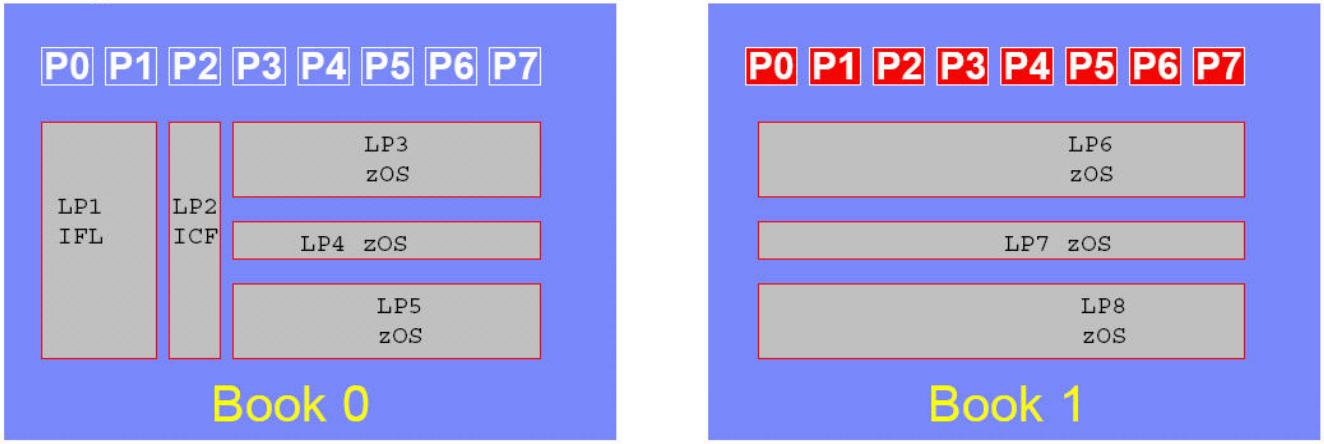
- **LP – Logical processor**
  - Processore logico , che verrà dispacciato su un processore fisico
- **PP – Physical Processor**
  - Processore fisico, uno dei processori presenti in macchina sul quale verranno eseguiti i LP
- **LPAR**
  - Partizione Logica
- **Weight**
  - Peso assegnato alla partizione logica
- **Share**
  - Percentuale di allocazione della macchina fisica

Definizioni

- **Processori Logici Dedicati**
  - Assegnati ad un Processore Fisico (PP)
  - Ad uso esclusivo di una LPAR
  - LPAR con PP dedicati sottraggono elementi dal Pool dei processori disponibili
  - LPAR con PP dedicati INACTIVE dona i suoi PP al Pool
  - LPAR con PP dedicati configurati OFFLINE dona i suoi PP al Pool
- **Processori Logici Shared (LP)**
  - Assegnati ad un Pool di Processori Fisici
  - Share% = Peso (Weight) del LPAR diviso il Peso delle LPAR “Active”
  - E’ possibile superare lo Share% se le altre LPAR non usano la loro allocazione
  - Hard Capping non permette di superare lo Share%
  - Soft Capping (WLC) non permette di superare la Capacità Definita (MSU)

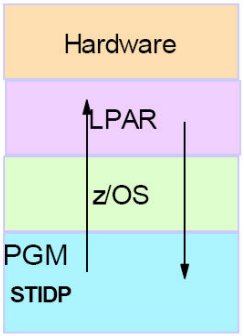
Dispatching – PR/SM System z9

- **La struttura Multi-Book dei System z9 e dei z990 Trex, ha caratteristiche diverse per l'accesso alla Memoria**
  - Accesso alla Cache L2 da parte di PP di Books differenti
    - 2 x Book, 32MB x Book, Processor Cache
  - Accesso

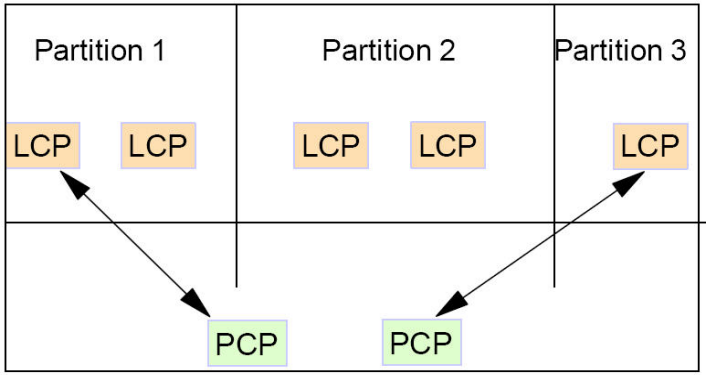


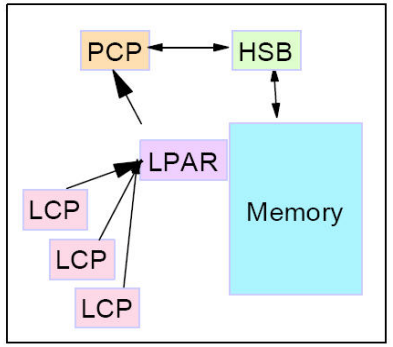
- **Il PR/SM dei System z9 è concepito per garantire la migliore Performance ad un LPAR …**
  - Assicurando la migliore allocazione possibile della memoria e dei processori al momento dell'attivazione
    - Tentativo di allocare tutta la memoria di un LPAR sullo stesso Book
    - Per le LPAR con PP dedicati, tentativo di allocare i processori sullo stesso Book della memoria
  - Assicurando il "miglior Dispatching" possibile di un LP sul "miglior" PP
    - Ricerca del PP sullo stesso Book dell'ultimo dispatch
    - Ricerca del PP sullo stesso Book di dove risiede la memoria della LPAR (Primary Affinity Mask)

Le fonti di OVERHEAD



- **L'overhead è costituito da 3 componenti principali:**
  - il tempo speso dal PR/SM per una specifica LPAR (LPAR Mgmt)
    - setup di un LP della LPAR per il dispatching su un PP
    - emulazioni chieste dal Sistema Operativo
    - cambio del TOD
  - il tempo speso dal PR/SM per tutte le LPAR (*Physcal*)
    - ricerca di un LP da dispacciare
  - Cache L1 miss (allungamento del TCB time)
    - le caches L1 servono più di un LP
      - I System z9 hanno migliorato ulteriormente tale effetto

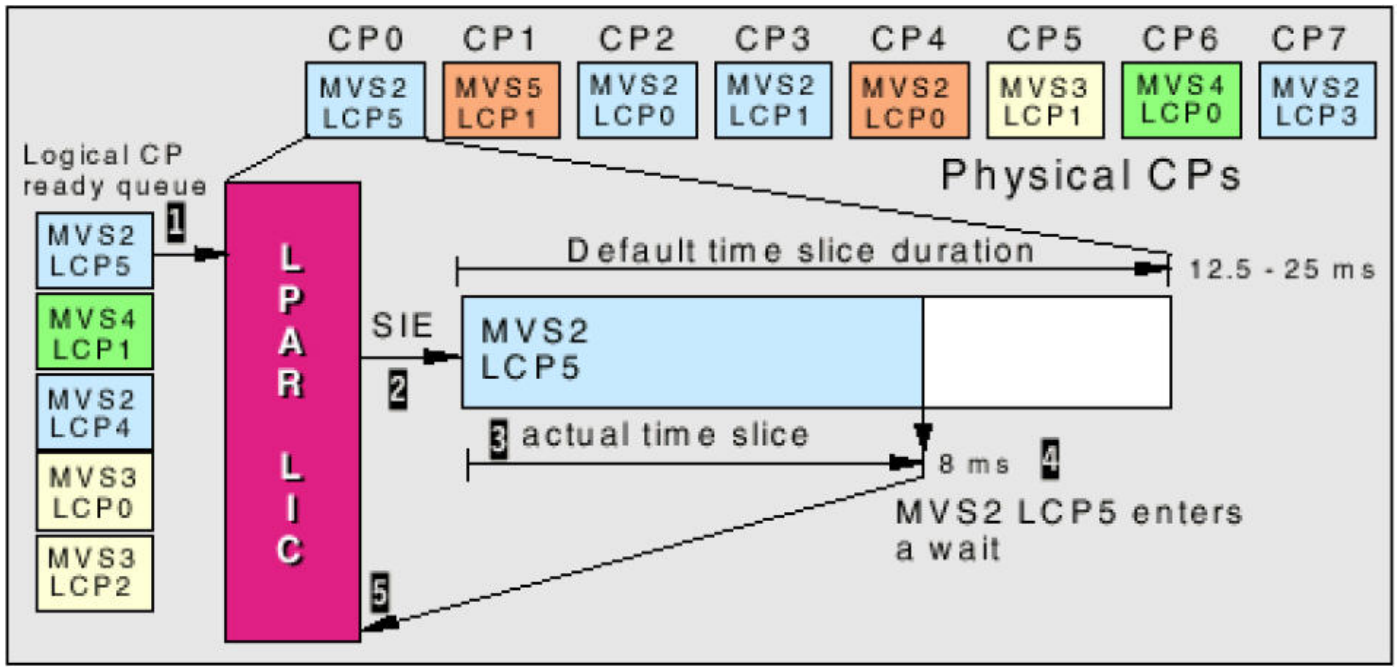




Wait Complete

Event Driven WC=NO

- Se un LP è in Wait State non aspetta la fine del Time Slice per cedere il PP ad un altro LP



Time Driven WC=YES

- Se un LP è in Wait State, attende la fine del Time Slice per cedere il PP ad un altro LP

Dispatching – L'influenza del LPAR Weight) - Priorità…2/2

Formule usate dal LIC

- Calcolo delle % fisica relativa a quella LPAR

$$WEIGHT(LPARx)\% = 100 * \frac{WEIGHT(LPARx)}{\sum WEIGHT\_LPAR\_Actives}$$

- Calcolo del numero di PP relativi a quella LPAR

$$TARGET(LPARx) = WEIGHT(LPARx) * (\#Non\_DED\_PP)$$

- Calcolo della % di PP per ogni LP

$$TARGET(LPx)\% = \frac{TARGET(LPARx)}{\#LP\_dans\_LPARx}$$

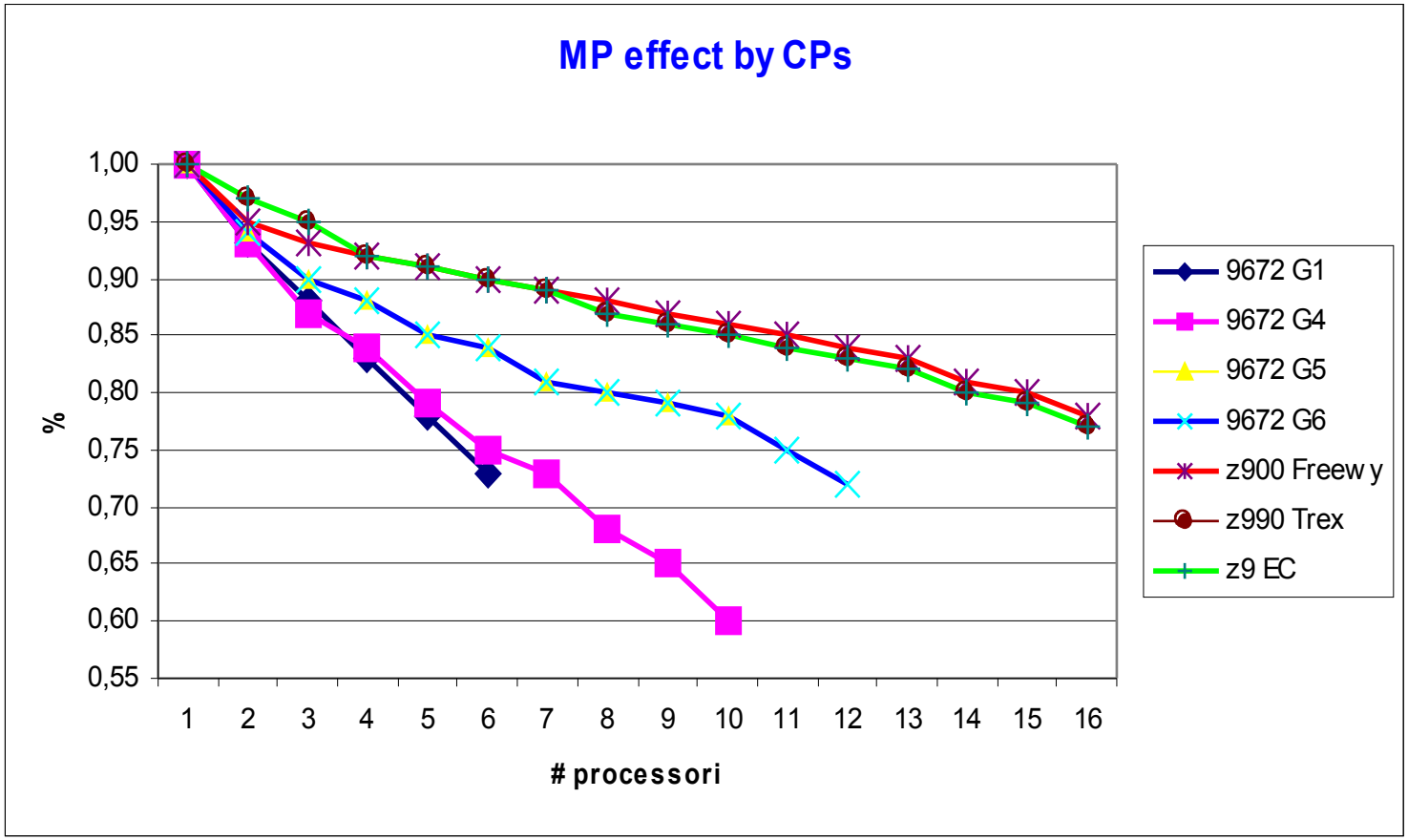
Se Target(LPx)% < Current

- La LP utilizza più risorse di quello che le è stato garantito e la Priorità Diminuisce all'interno della Ready_Q

Se Target(LPx)% > Current

- La LP utilizza meno risorse "garantite" e la Priorità Aumenta all'interno della coda

Effetto MP



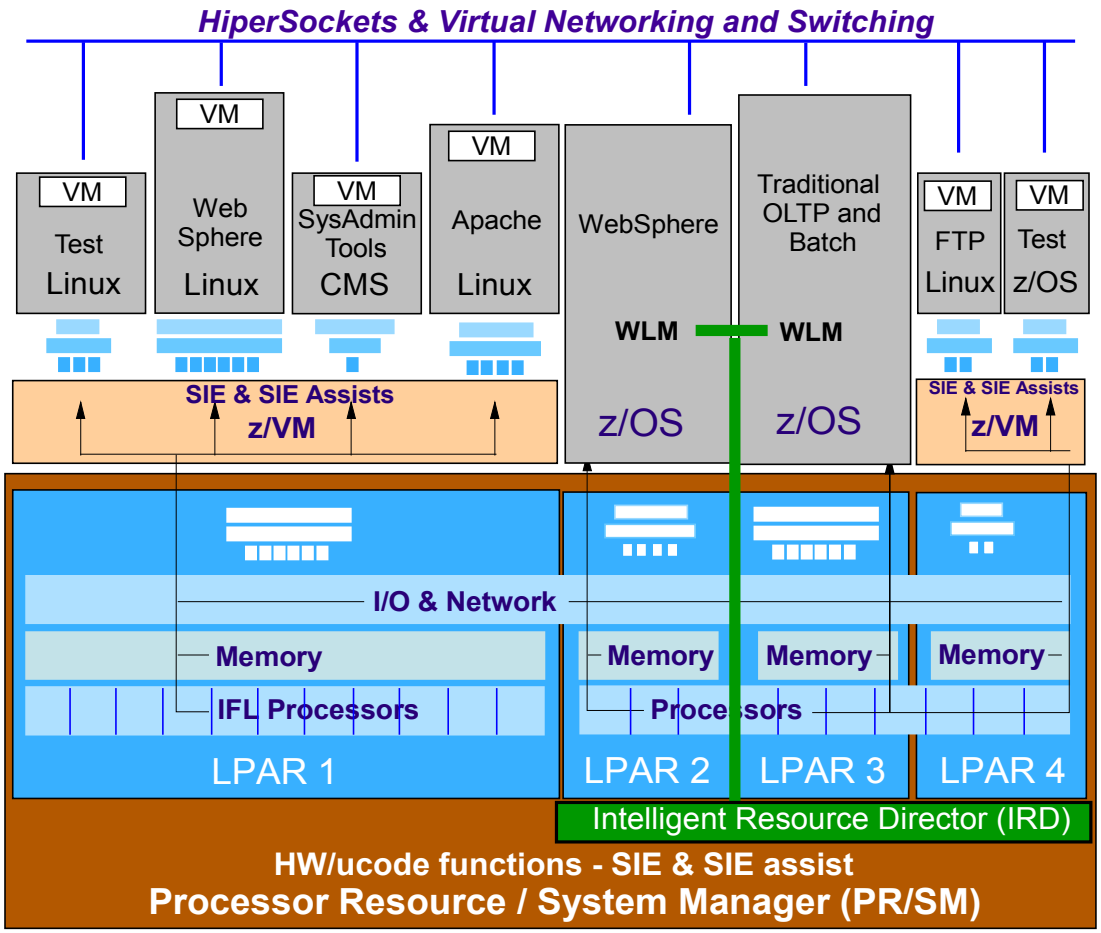
Fonte : Charyl Watson Research

Cos'è "Effetto MP" ?

Tempo speso nel "hardware" per le operazioni di coordinamento tra i vari Processori, azioni largamente "invisibili" al sistema operativo. Questo fenomeno accade ogni qualvolta un Processore accede o potrà accedere , nell'immediato (*instruction pipeline look-ahead*) qualsiasi risorsa hardware , la quale potrebbe essere, simultaneamente, aggiornata da un altro processo in corso.

Esempi di tale risorse sono:
> bytes in real storage,
> entries in a translation look-aside buffer,
> data in high-speed buffer (HSB or L1) cache,
> instruzioni in high-speed cache,
> instruzioni in pipelines di altri processori

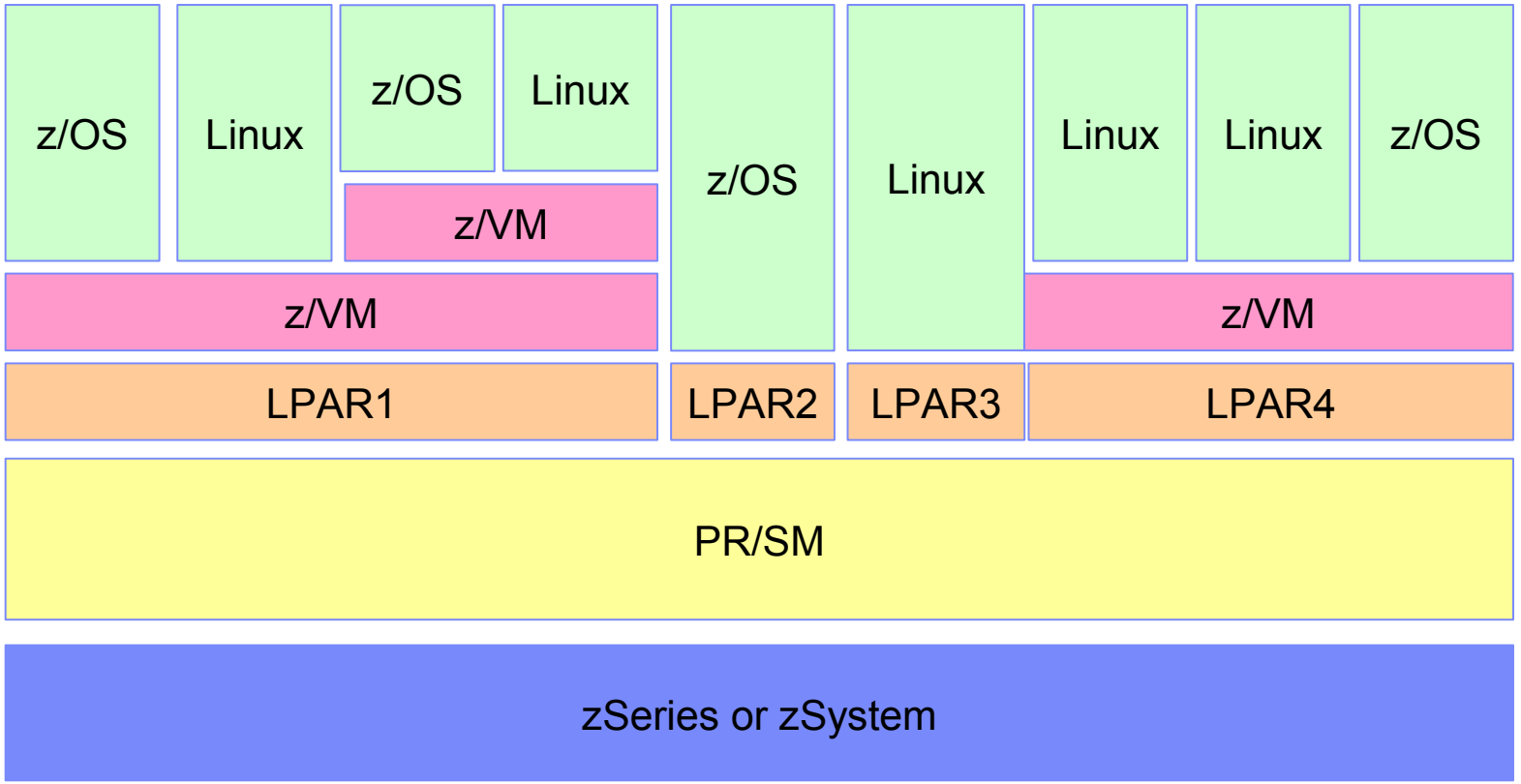
IBM System z CPU Virtualization



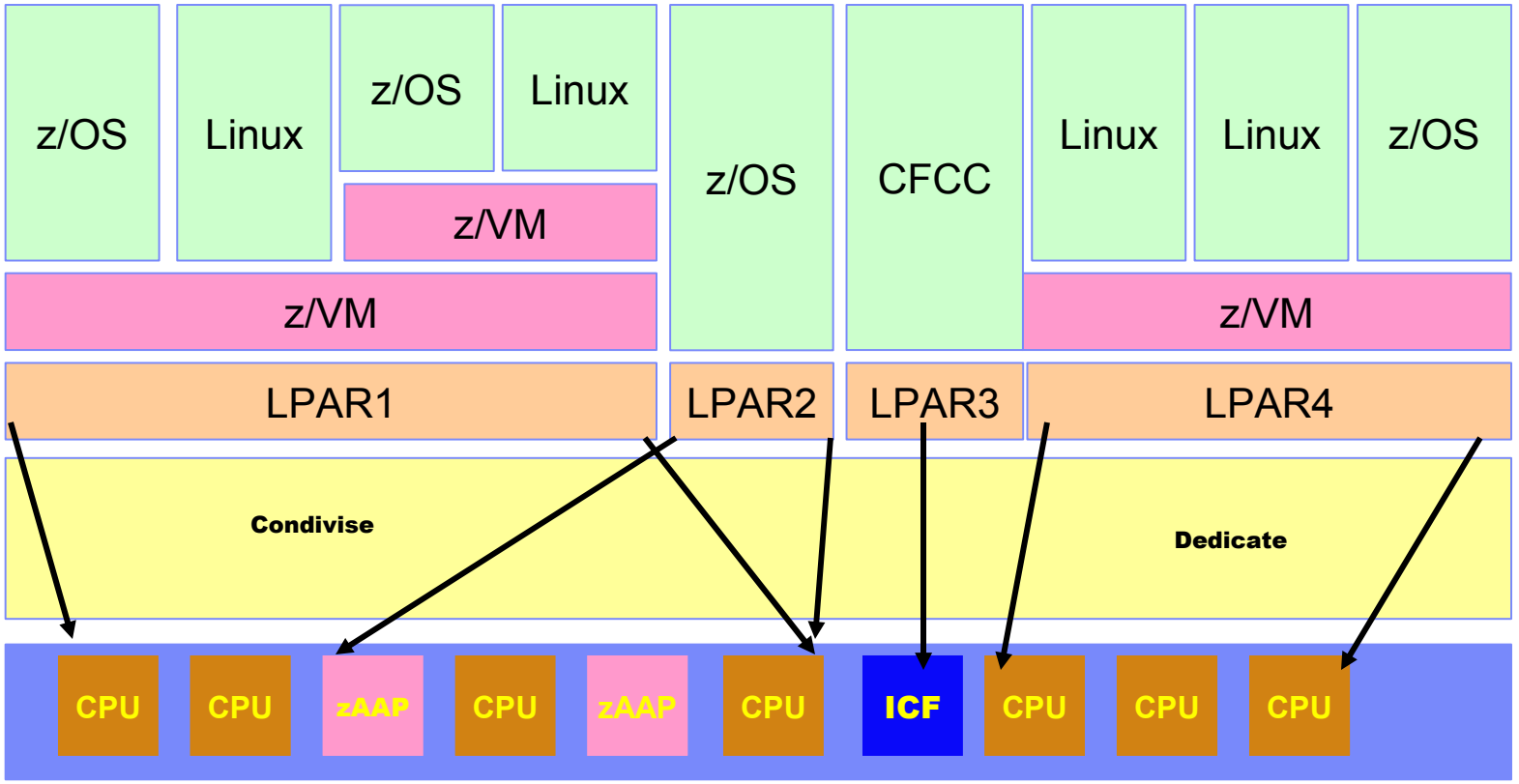
Virtualizzazione Multi-dimensionale

- HW: ***PR/SM (LPAR's)*** & SW: ***zVM (VM's)***
- **Shared or Dedicated pools di "CPU's"**
- **Shared Pools:**
- **Any Logical CPU runs on any physical CPU**
- **Dispatching dei logical CPU's è ottimizzato sulla base della locazione di memoria della LPAR & dispatch history**
- **Garanzia della capacità di LPAR = Weight**
- **Capped and uncapped LPAR's**
- **Time and Event Driven dispatching…..**
- **Event driven dispatching aumenta la reattività (metodo preferito)**
- **Time-slices – nell'ordine dei 5-10ms – per ridurre I costi di gestione**
- **Cicli di CPU non utilizzati da un LPAR sono "donati" ad altreLPAR's, e…**
- **Logical CP's sono ordinati per priorità sulla base del rapporto "Relative Weights/Logical CP"**
- **Low-priority LPAR's danno la precedenza alle LPAR con alta priorità**
- **zVM ha algoritmi di schedulazione per ottimizzare la reattività ed il throughput**
- **Le risorse Fisiche e Virtuali (CPU, I/O e memoria) possono essere "regolate" dinamicamente sia all'interno sia tra le LPAR's**

Livelli Multipli di Virtualizzazione HW + SW

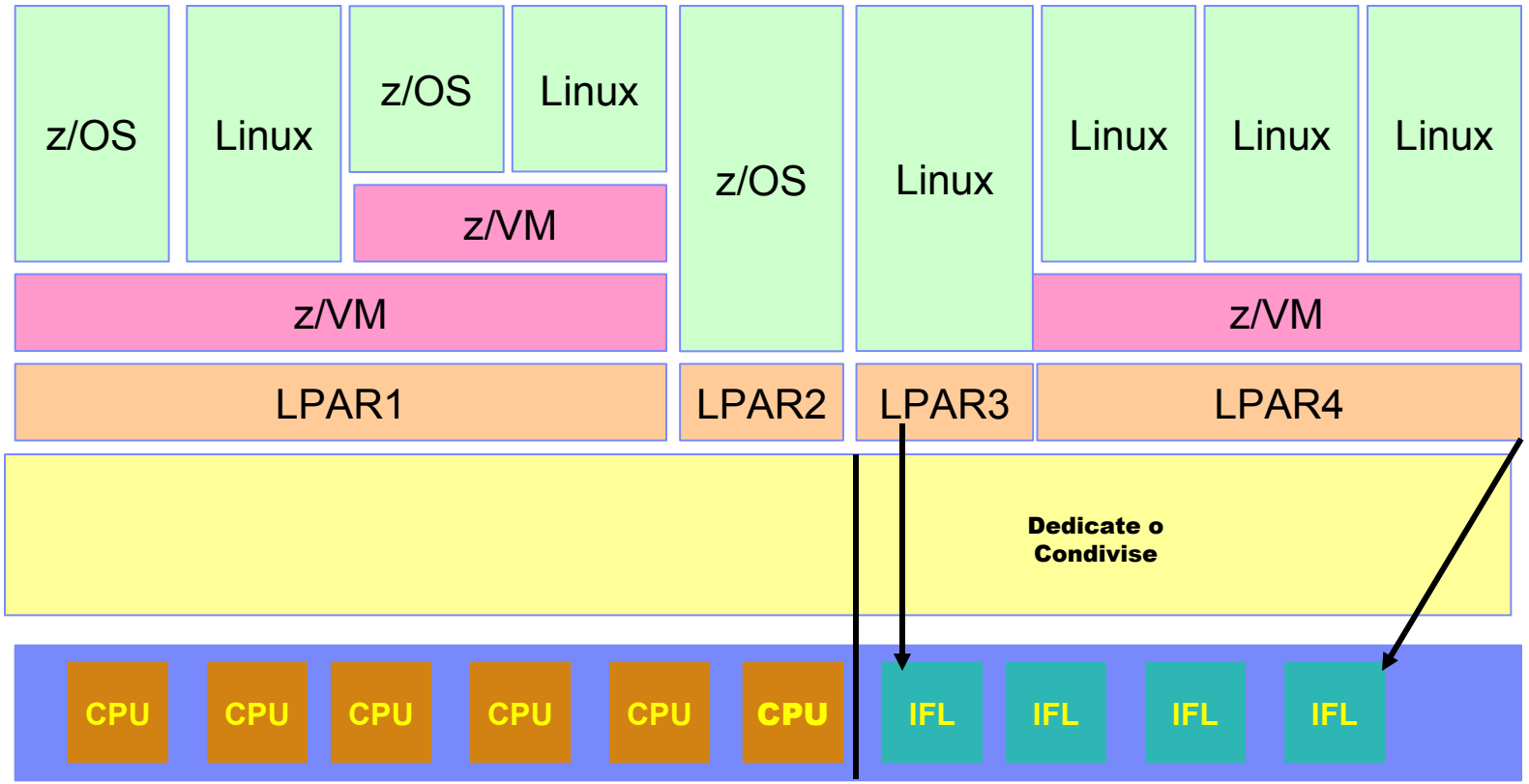


Completa Condivisione delle Risorse – CPU/ICF/zAAP



Completa Condivisione delle Risorse – CPU / IFL

I Processori di tipo IFL possono essere usati solo per LINUX o per LINUX Virtualizzato da z/VM



Gestione delle CPU Condivise – Dispatching e Peso e Capping

A Ciascuna **Partizione Logica caratterizzata** da un nome possono essere assegnate **CPU dedicate o Condivise.**

Se le CPU sono **dedicate** ad una partizione Logica la **potenza di calcolo** della partizione e' data dalla potenza equivalente di un Sistema zSeries di quel dato modello con quel dato numero di CPU in tutto.

Nel caso in cui vengano assegnate **CPU condivise** esse prendono il nome di **CP Logici**.

Il Numero di **CP Logici** assegnati ad una partizione ne determina **l'importanza e la priorita**' di esecuzione in caso di contesa di risorse.

Per calcolare la **priorita' di esecuzione** e la capacita' elaborativa assegnate ad una partizione con CPU condivise occorre tenere conto:

- **Del Numero di CP Logici Assegnati**
- Di un parametro utente detto **Peso** (weight)
- Della presenza di **Capping**

Consideriamo un semplice Esempio:

Dato un Sistema z/Architecture con 6 Cpu Fisiche definiamo in esso TRE partizioni logiche con CPU condivise.

Siano le partizioni logiche denominate:

VSEESA
ZOSTEST
ZVM

Supponiamo di definire per esse i seguenti parametri:

LP Name	Logical CPs	Weight
VSEESA	1	300
ZOSTEST	6	100
ZVM	2	900

Gestione delle CPU Condivise – Dispatching e Peso

Calcoliamo il 100% della capacita del sistema sommando i pesi dichiarati :

300 + 100 + 900 = 1300

Esprimiamo ora **in percentuale** il valore di ogni LPAR in base ai **pesi** dichiarati

Tale numero rappresenta **la percentuale della potenza globale, SHARE%** assegnata ad ogni LPAR in assoluto

Per calcolare quale e' la **frazione di singolo processore** che sara' assegnata ad ogni LPAR in caso di contesa di risorse occorre dividere **le percentuali assolute per il numero di CP Logici**

Su tali valori e' ammessa una tolleranza del 3.4% circa

LP Name	Logical CPs	Weight
VSEESA	1	300
ZOSTEST	6	100
ZVM	2	900

VSEESA	300/1300 = 23.1%
ZOSTEST	100/1300 = 7.7%
ZVM	900/1300 = 69.2%

VSEESA	23.1/1 CP = 23.1%
ZOSTEST	7.7/6 CPs = 1.3%
ZVM	69.2/2 CPs = 34.6%

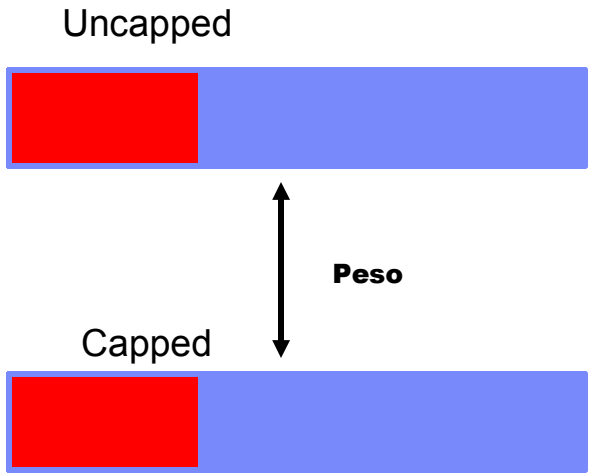
Gestione delle CPU Condivise – Capping

Pur avendo definito dei pesi specifici per ogni LPAR nel caso in cui non vi sia contesa di risorse, **ogni LPAR verra' servita mediante la risorsa CPU al meglio possibile** : In tal modo tutta la potenza disponibile sara' usata indipendentemente dai Pesi definiti da qualsiasi LPAR ne faccia richiesta.

Nel Caso in cui si presentino **contese** il PR/SM si occupera' di gestirle mantenendo i pesi specifici dichiarati e garantendo ad ogni LPAR la potenza calcolata in base ai pesi con uno scarto del 3.4%

Si definisce **Capping la possibilita' di forzare il rispetto dei Pesi su una LPAR anche in presenza di risorse disponibili in eccesso**

In caso di Capping tutte le LPAR cosi' dichiarate avranno la potenza determinata dal loro peso.



Caratteristiche delle Partizioni Logiche

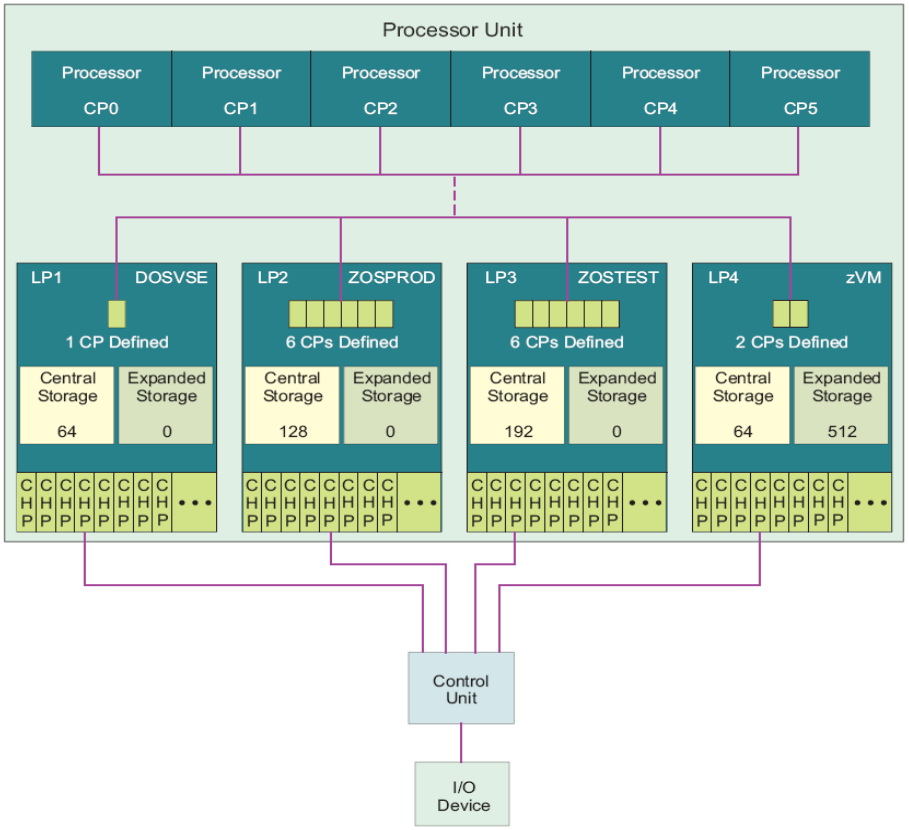


Figure 1-1. Characteristics of Logical Partitions

PR/SM Gestione dell' I/O – Multi Image Facility (MIF)

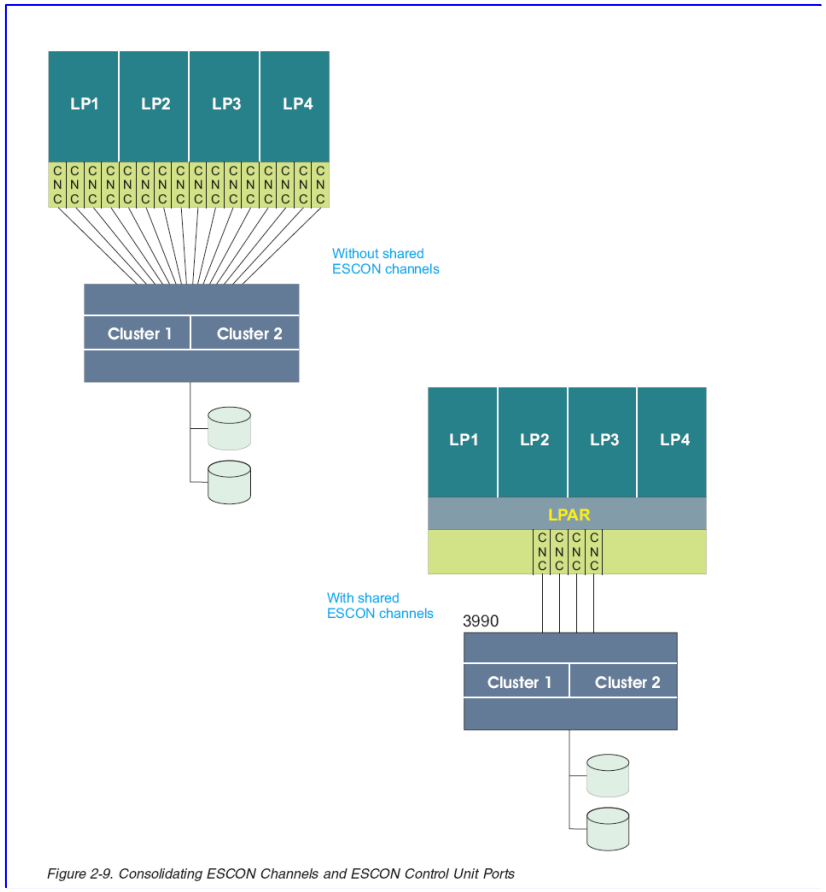


Figure 2-9. Consolidating ESCON Channels and ESCON Control Unit Ports

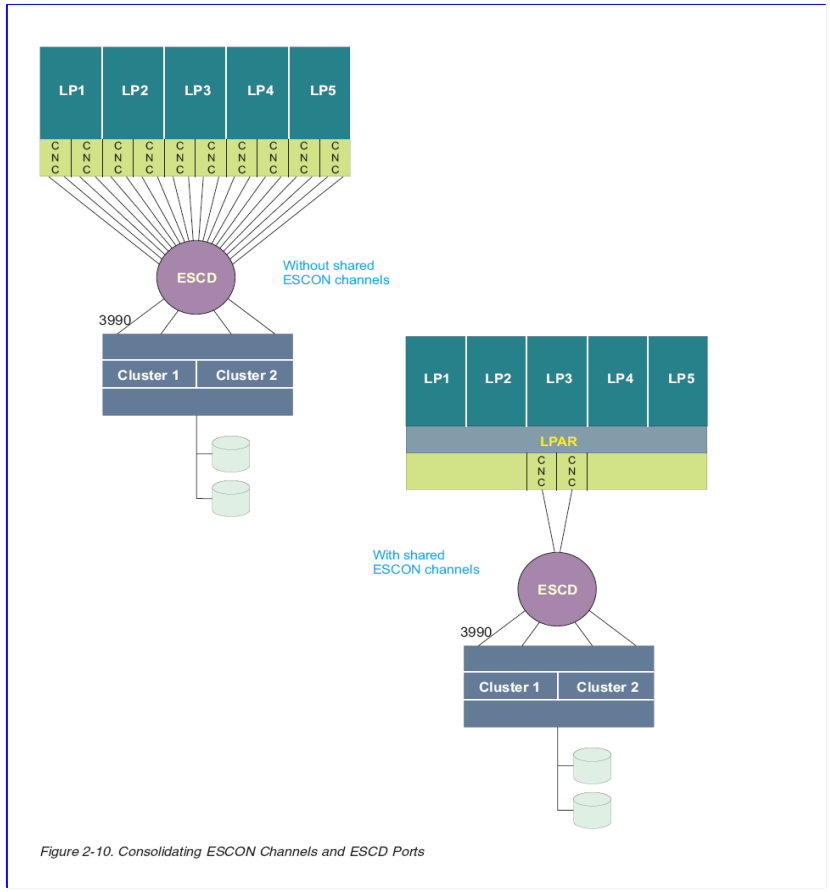
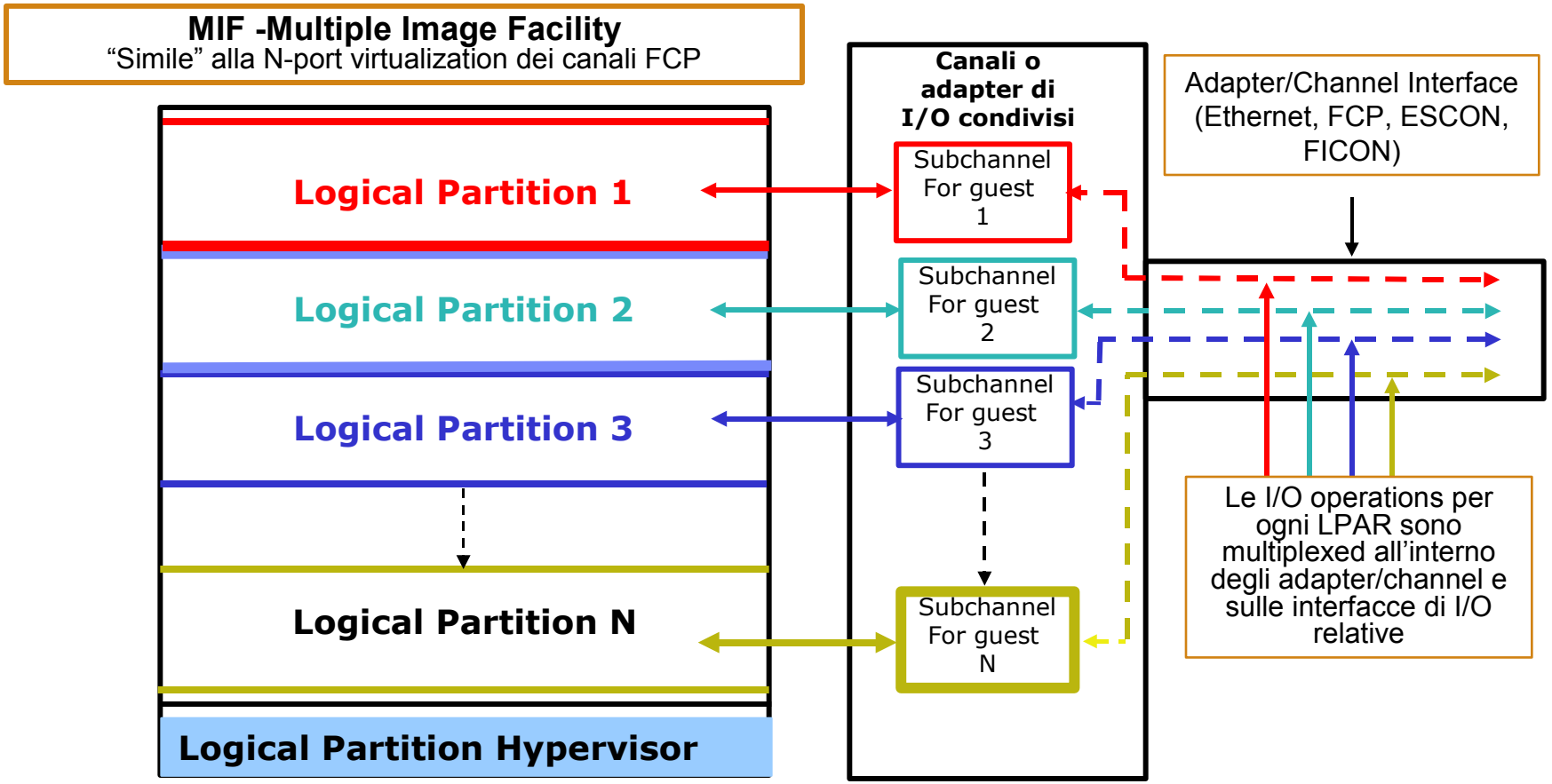


Figure 2-10. Consolidating ESCON Channels and ESCD Ports

System Z - MIF I/O Virtualization

- *Ogni I/O device ha un "subchannel"*
- *Il subchannel è il target di una I/O instruction e sovrintende al controllo del device ed al suo accesso*

- ***Ogni LPAR interagisce DIRETTAMENTE con gli I/O devices***



Gestione dell' I/O – Shared Devices & Logical Path

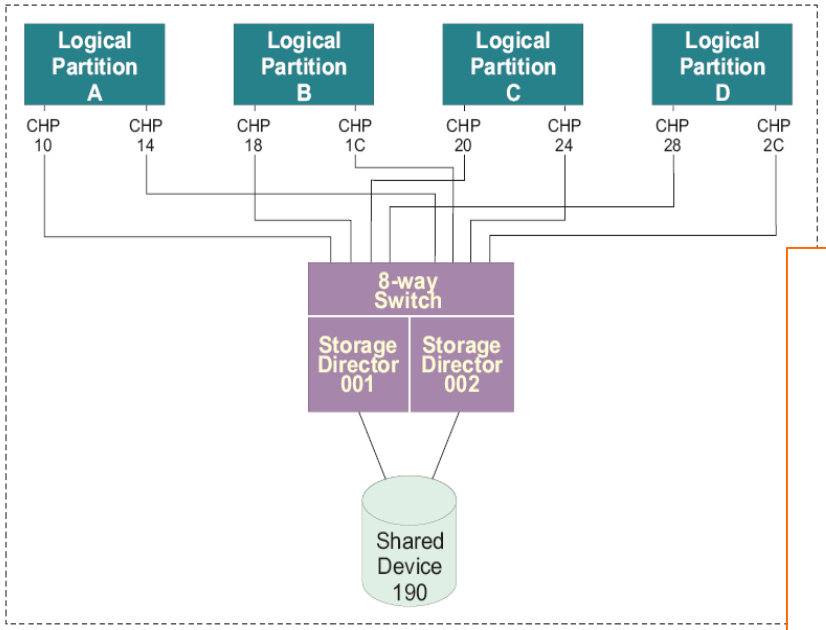


Figure 2-13. Physical Connectivity of Shared Device 190

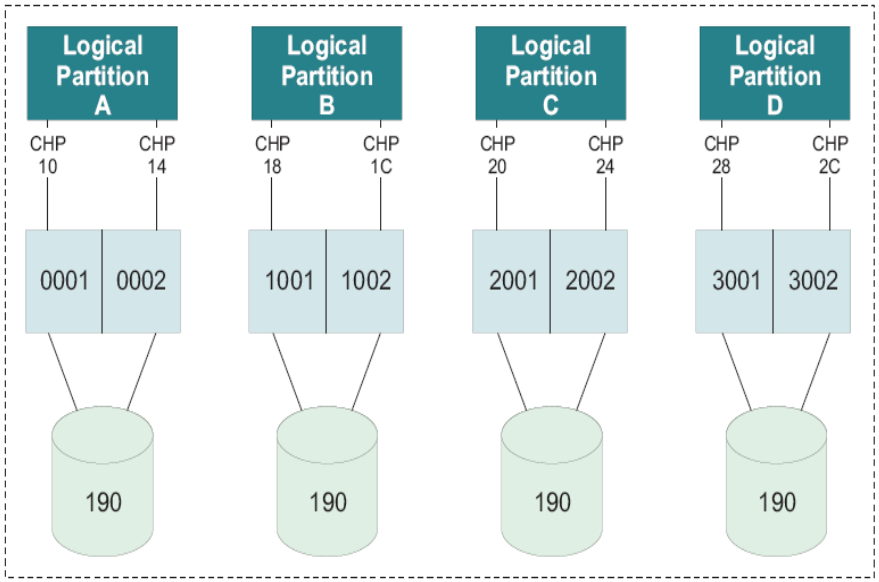
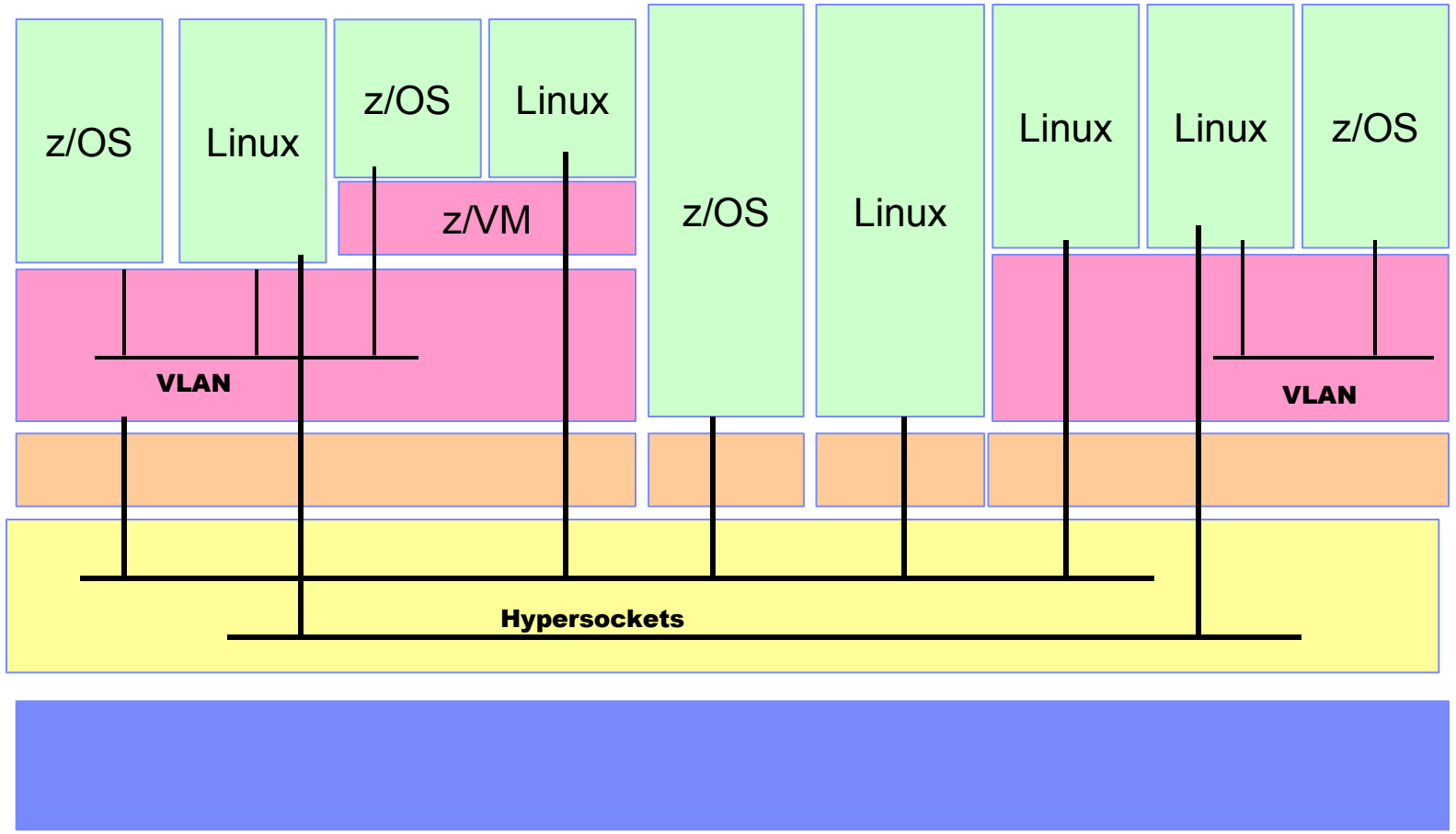
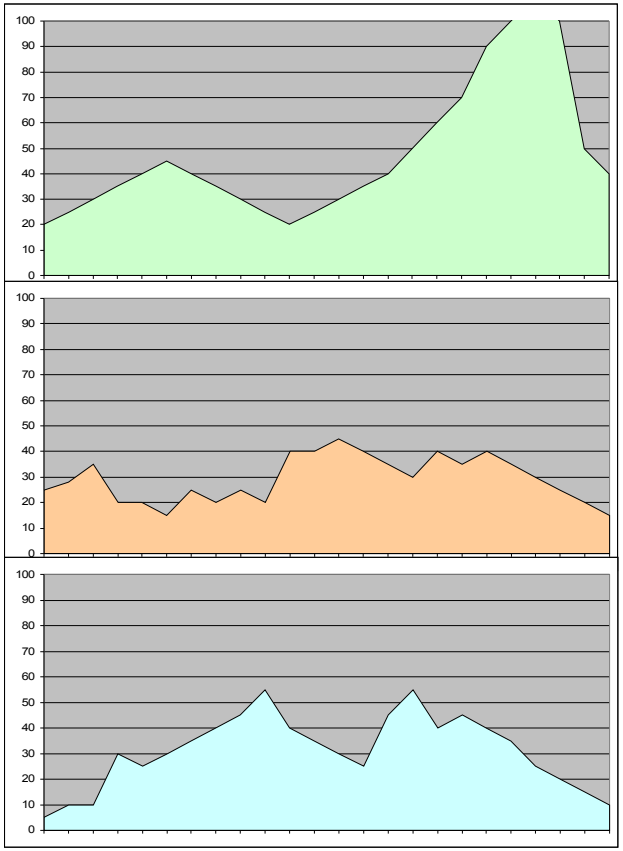


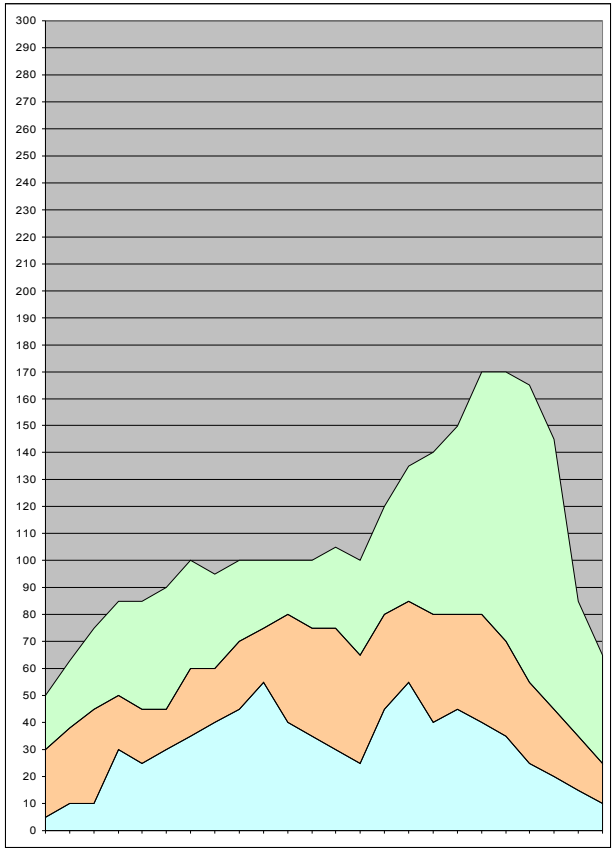
Figure 2-14. Logical View of Shared Device 190

Virtual I/O – Hypersockets e Virtual LAN



Consolidamento di Sistemi con Basso utilizzo





Se uno o piu' sistemi consolidati presentano un basso utilizzo il risultato sara' un risparmio complessivo della potenza necessaria

Ma fino a quanto questo recupero di potenza potra' essere realizzato ?

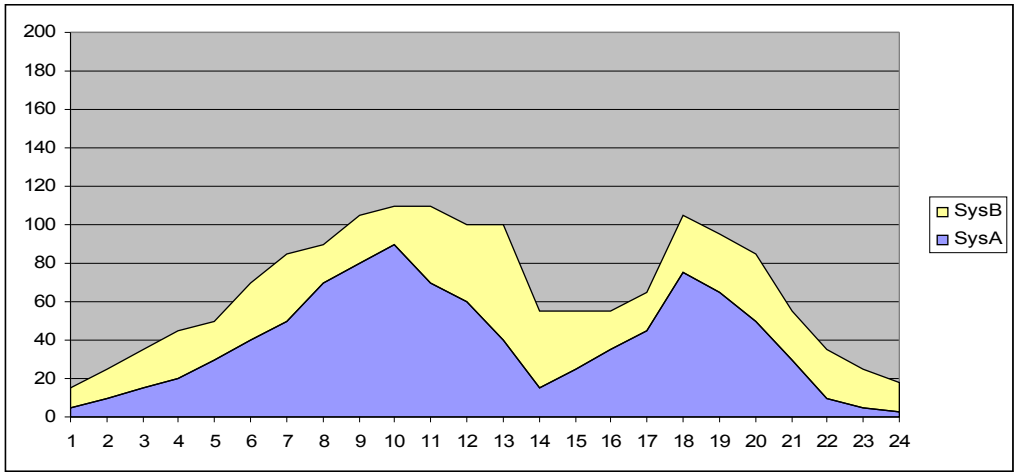
Minima Potenza Gestibile – Analizziamo un Caso di Studio

Studio del comportamento del consolidamento di due Sistemi detti SysA e SysB dei quali uno fortemente utilizzato ed uno scarsamente utilizzato.

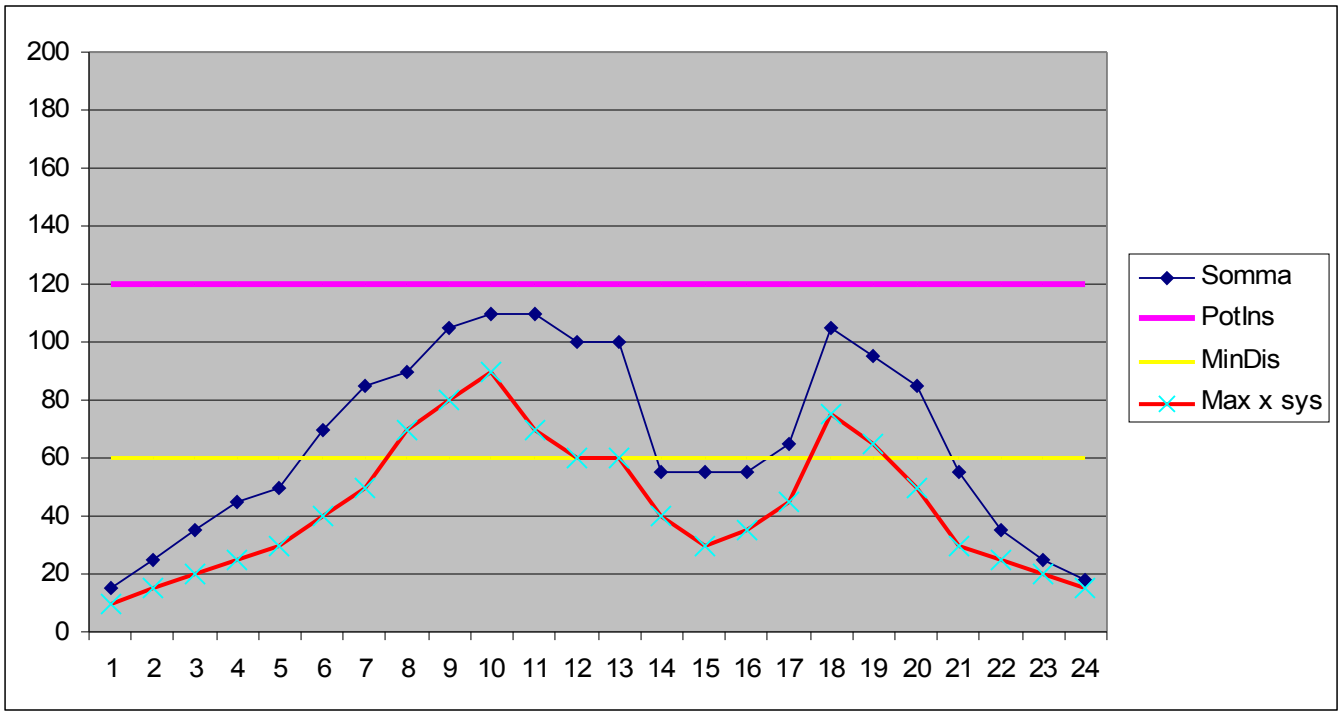
Ipotizziamo due tipi di Partizionatori logici:

A = in grado di suddividere il processore in due parti. (Minima Potenza ½ Processore)

B = in grado di suddividere il processore in quattro parti. (Minima Potenza ¼ di Processore)

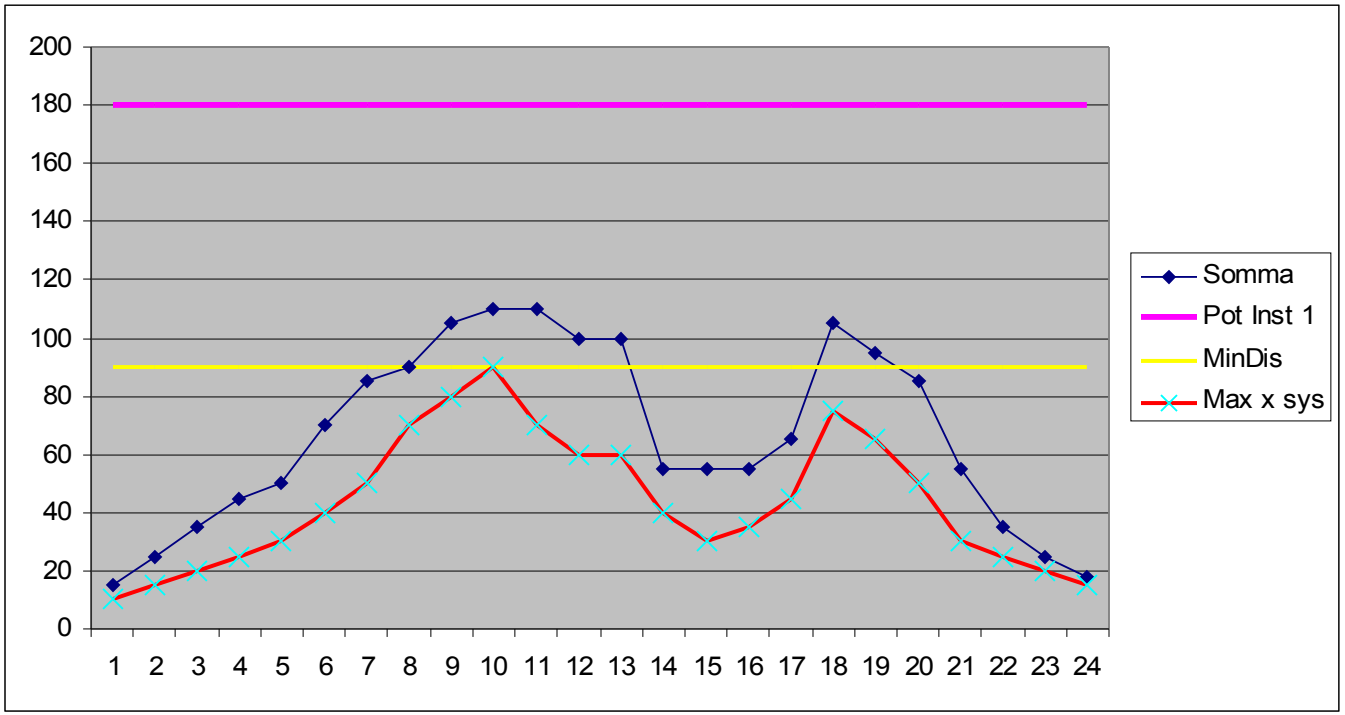


Minima potenza Gestibile – Partizionatore A (1/2 CPU)



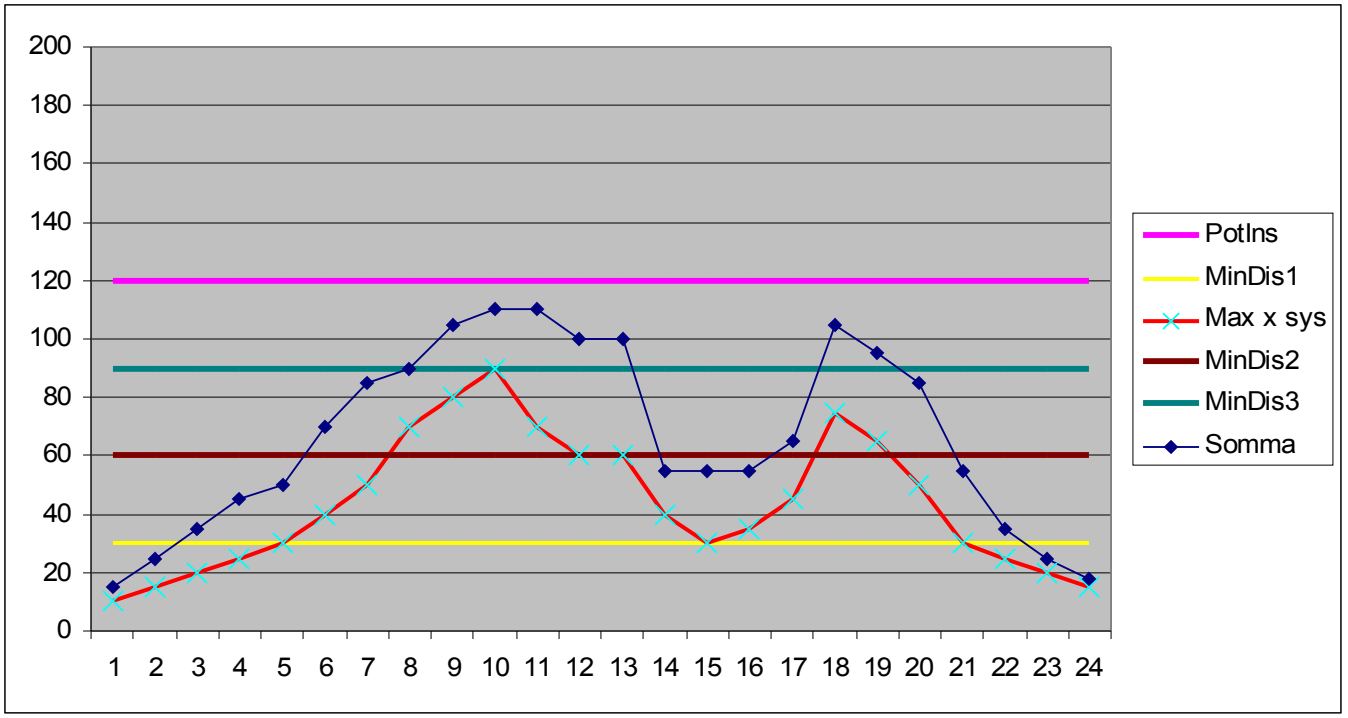
Benche' la somma dei due carichi non superi la potenza di 120 , il partizionatore di tipo A capace di suddividere il processore in due parti non sara' in grado di soddisfare in ogni istante le richieste concorrenti delle due partizioni logiche.

Minima potenza Gestibile – Partizionatore A (1/2 CPU)



Per potere gestire i due carichi concorrenti col partizionatore A occorrera' disporre almeno di una potenza di 180 a fronte di una somma che non supera 120 .

Minima potenza Gestibile – Partizionatore B (1/4 CPU)



Il Partizionatore B capace di suddividere il processore in quattro parti invece riuscira' a gestire perfettamente il carico con una potenza di 120 ,

La potenza minima gestibile e' molto importante al fine di ottimizzare carichi di diverso ordine di grandezza : Col PR/SM su z9-109 la potenza minima gestibile e' di un sessantesimo di CPU