IBM System P虚拟化技术

IBM PowerVM™on System p™

IBM中国技术支持中心
8008101818 - 5151
010-84981188 - 5151

安排:

- **IBM p系列服务器分区、虚拟化技术概念与功能概述**
  - 虚拟化概念
  - IBM Power5虚拟化技术
  - IBM Power6虚拟化技术
- **虚拟I/O服务器**
- **Work Load Partition技术**
- **Live Partition Mobility**
- **Integrated Virtual Ethernet技术**

IBM p系列服务器分区、虚拟化技术概念与功能概述

IBM PowerVM™on Power Systems ™

IBM中国技术支持中心
8008101818 - 5151
010-84981188 - 5151

IBM虚拟化发展的历史

A 40 year tradition culminates with PowerVM

1967	1973	1987	1999	2004	2007	2008
IBM develops **hypervisor** that would become VM on the mainframe	IBM announces first machines to do **physical partitioning**	IBM announces **LPAR on the mainframe**	IBM announces **LPAR** on POWER™	IBM announces POWER Hypervisor™ for System p™ and System i™	IBM announces POWER6™, **the first UNIX® servers with Live Partition Mobility**	**IBM announces PowerVM**

"[PowerVM] ensures that we are making the best possible use of hardware resources across our entire environment."

T N Rangarajan, VP of IT, Brakes India
August 2007

client quote source: Brakes India case study published at http://www.ibm.com/software/success/cssdb.nsf/CS/STRD-76DHWE?OpenDocument&Site=eserverpseries&cty=en_us

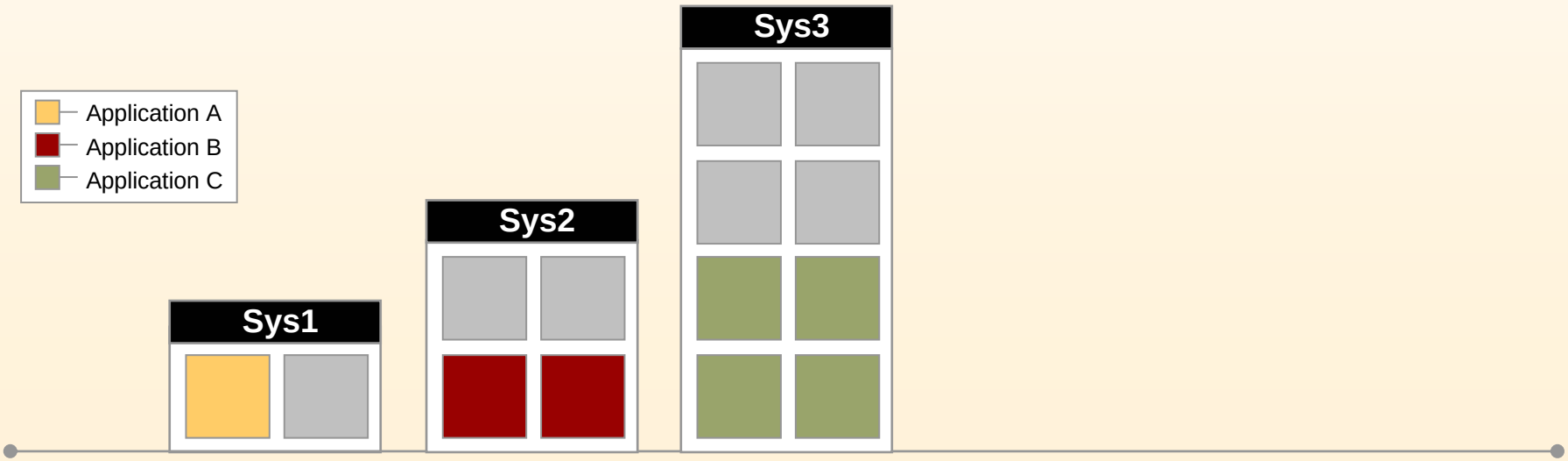
当前客户环境的现状是使用一台或多台独立的服务器运行一个应用

购买的多余的硬件资源造成了资源的闲置

more to manage

more costs

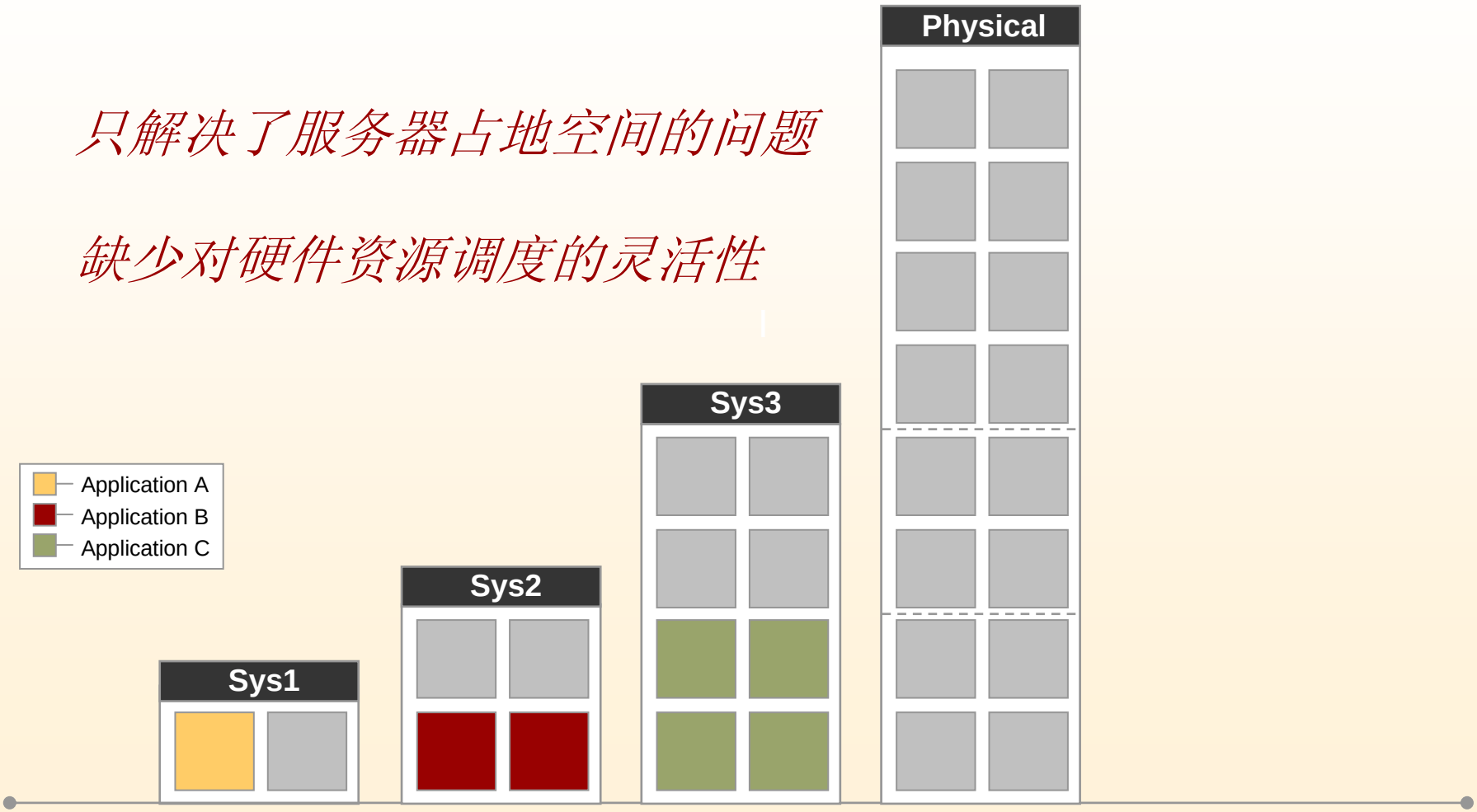
more headaches



一些服务器厂家提出了physical partitioning…技术

只解决了服务器占地空间的问题

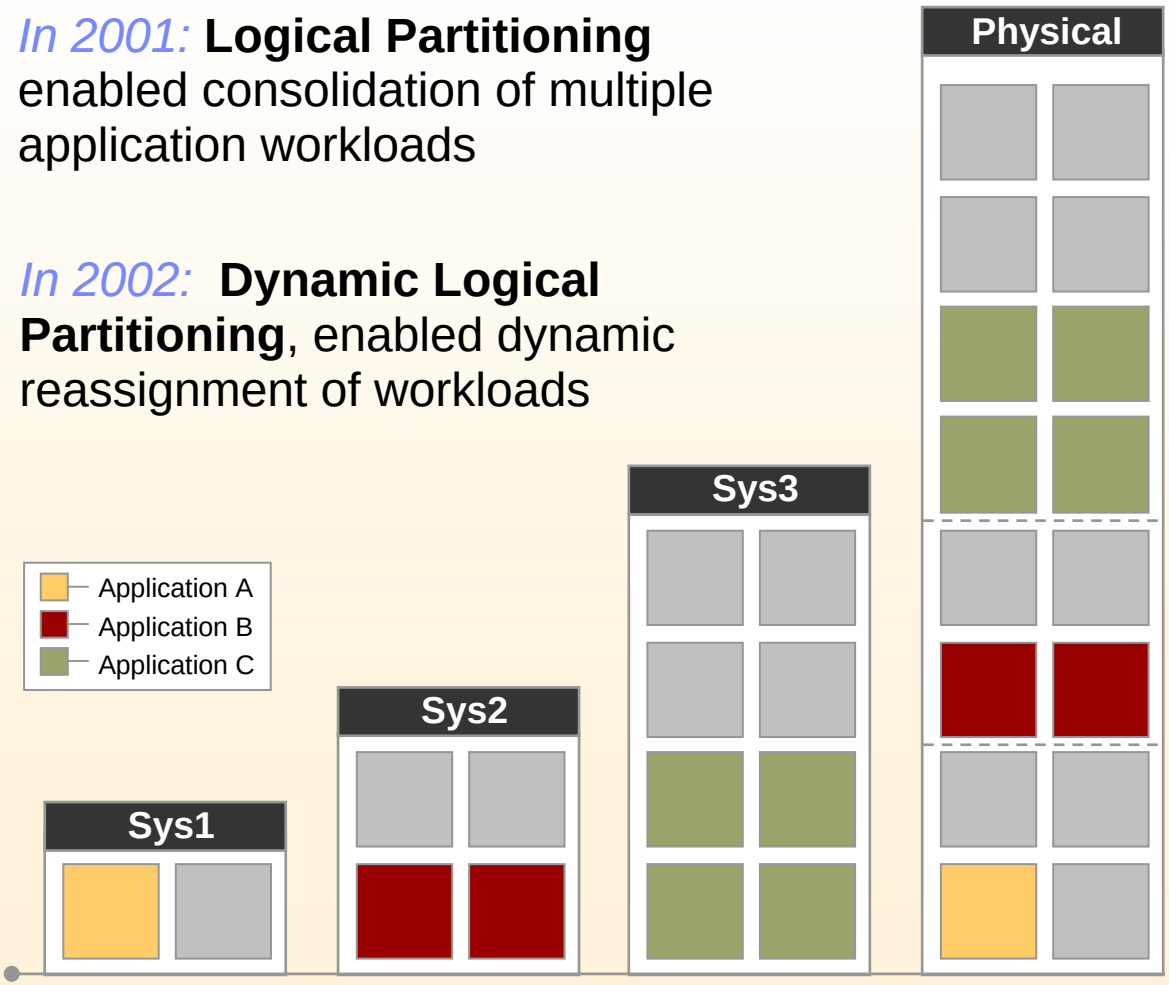
缺少对硬件资源调度的灵活性



IBM对分区技术进行了进一步的创新

In 2001: **Logical Partitioning** enabled consolidation of multiple application workloads

In 2002: **Dynamic Logical Partitioning**, enabled dynamic reassignment of workloads



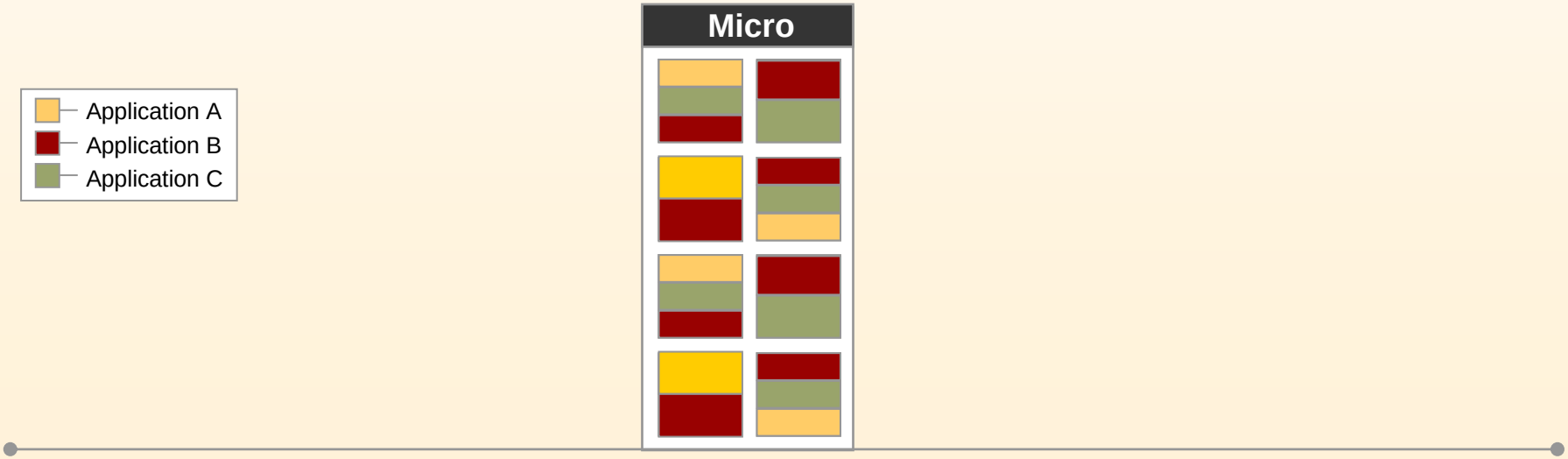
IBM Micro-Partitioning 技术给服务器注入了新的力量

相比逻辑分区技术，微分区技术可以在一台服务器上同时运行的应用更多，对CPU资源划分的更细。

automatically

for less

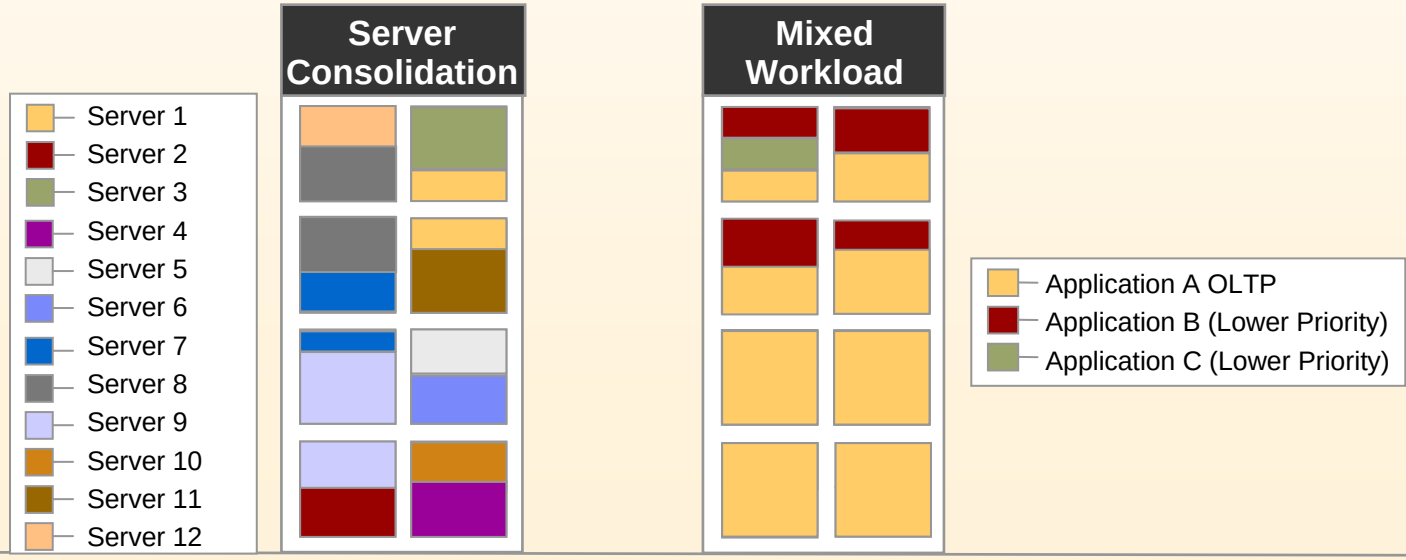
and, with fewer headaches



微分区也提供了更大的灵活性

Designed to support both server consolidation and a mixed workload

- *simplify your environment*
- *rapidly respond to changing needs*
- *drive higher system utilization*

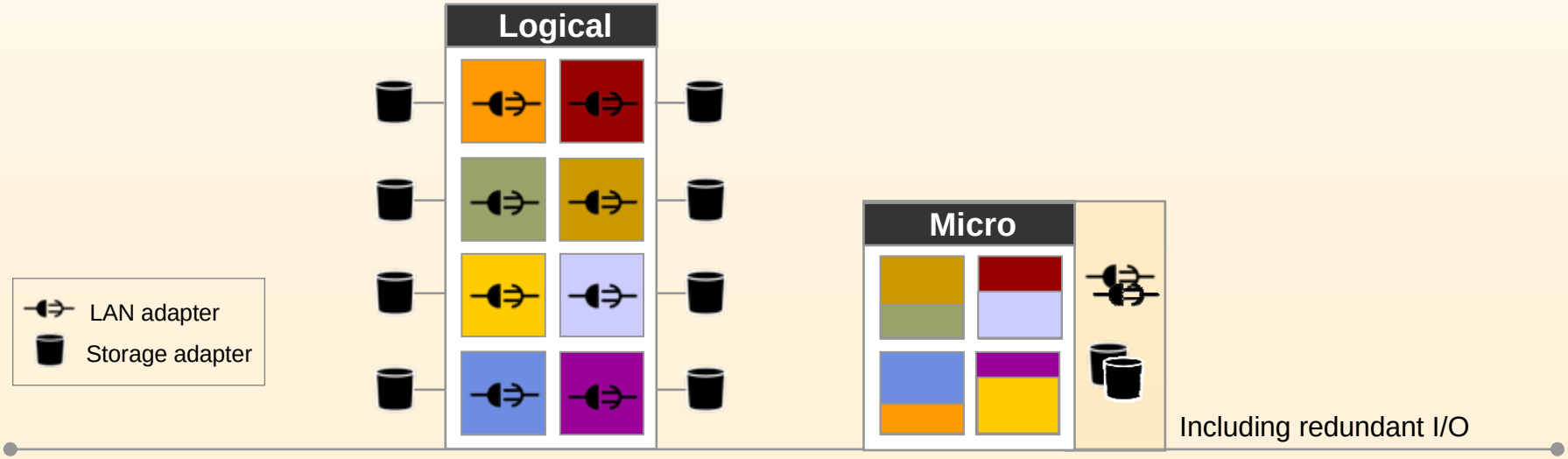


IBM 提供了共享资源的虚拟化

可以实现存储和网络设备的资源共享

— *fewer resources to purchase, configure and maintain*

— *simple and quick adjustments as business demands change*



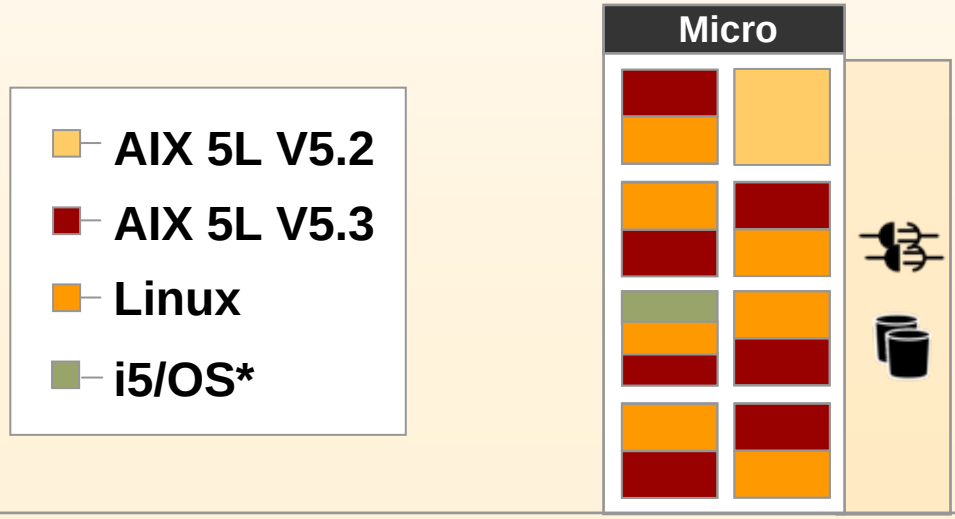
P5服务器的分区上可以安装 AIX 5L, Linux and i5/OS

Designed so you can use the right OS for your business

— *on the same system*

— *on the same processor*

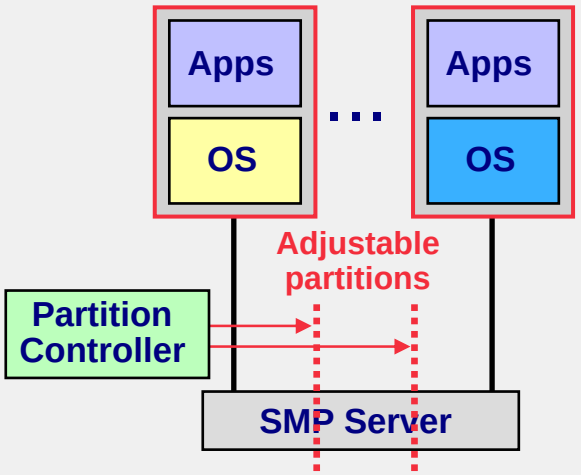
— *at the same time*



*Planned for select p5-570 models

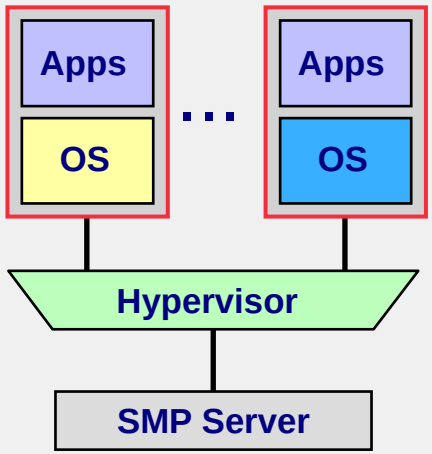
解析当前服务器虚拟化采用的三种方式

Hardware Partitioning



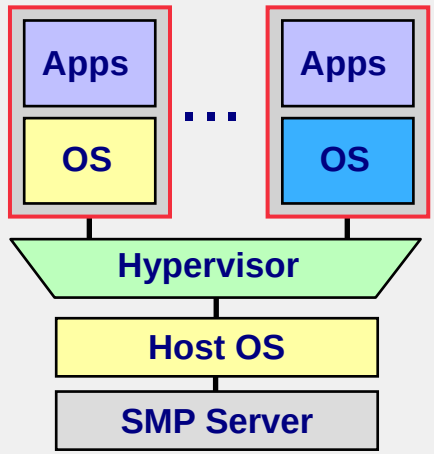
Physical partitioning
- S/370 SI->PP & PP->SI,
- Sun Domains
- HP nPartitions

Hypervisor: Type 1



Hypervisor software/firmware runs directly on server
- System z™ PR/SM and zVM™
- PowerVM Editions
- HP vPartitions
- VMware ESX Server
- Sun Logical Domains

Hypervisor: Type 2



Hypervisor software runs on a host operating system
- HP Integrity VM
- Solaris Containers
- AIX® V6.1 Workload Partitions

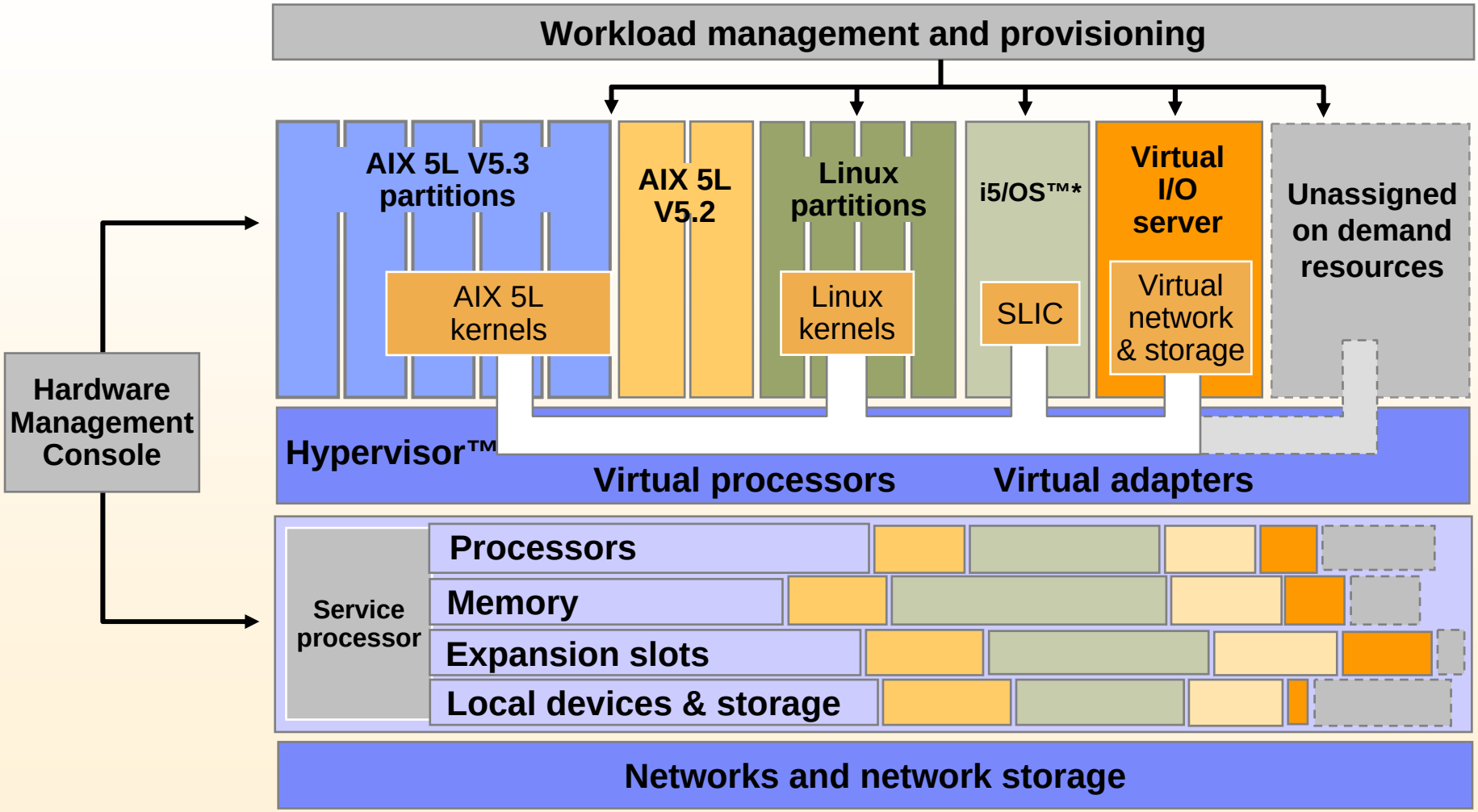
PowerVM

IBM System p

IBM power 5服务器虚拟化技术

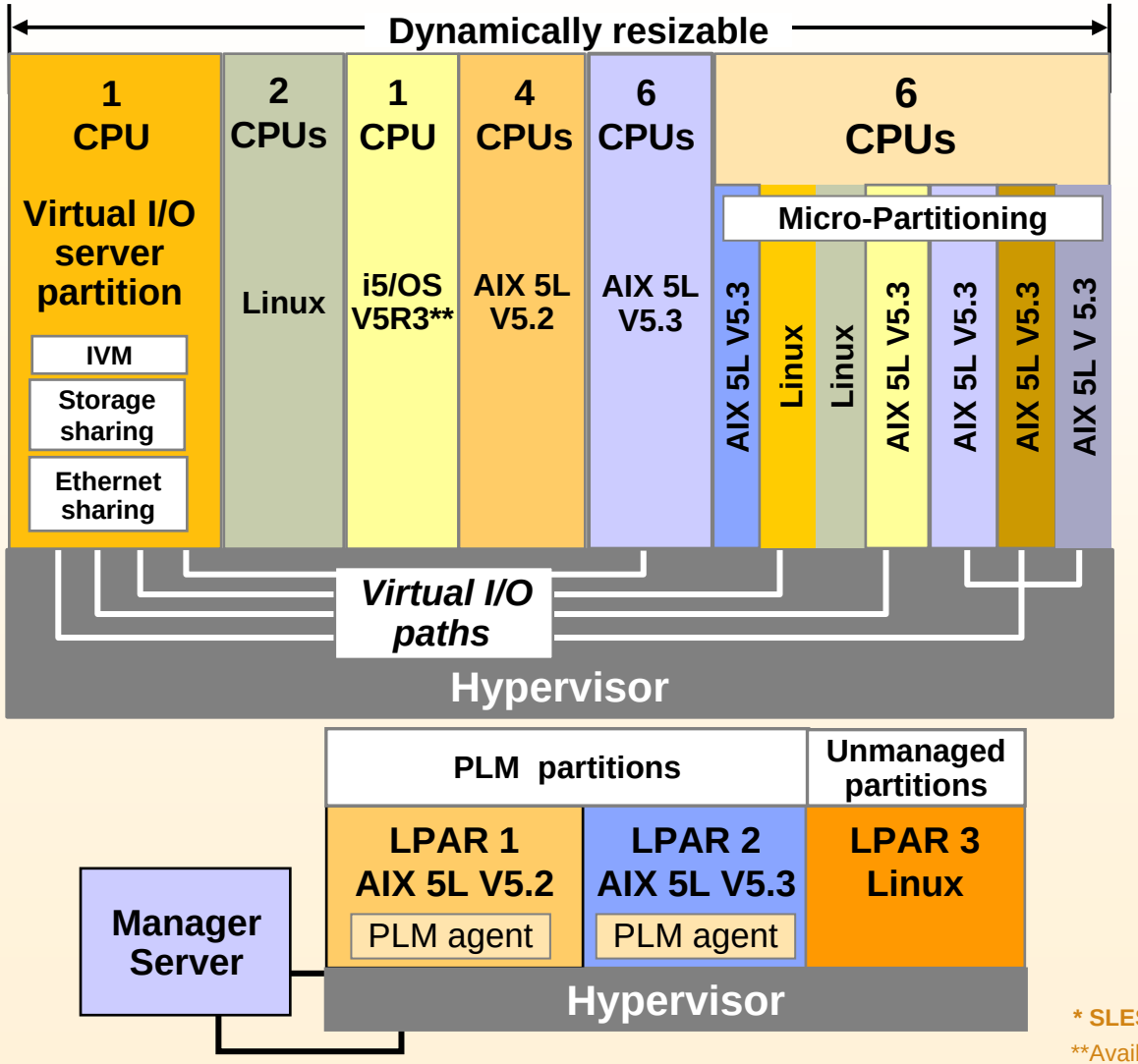
IBM PowerVM™on System p™

@server p5 系统虚拟化架构



*Available on 1.65 GHz POWER5 p5-570, p5-590 and p5-595 models

Advanced POWER Virtualization 选项



Virtual I/O Server

- Shared Ethernet
- Shared SCSI and Fibre Channel-attached disk subsystems
- Supports AIX 5L V5.3 and Linux* partitions

Micro-Partitioning

- Share processors across multiple partitions
- Minimum partition 1/10th processor
- AIX 5L V5.3, Linux*, or i5/OS**

Partition Load Manager

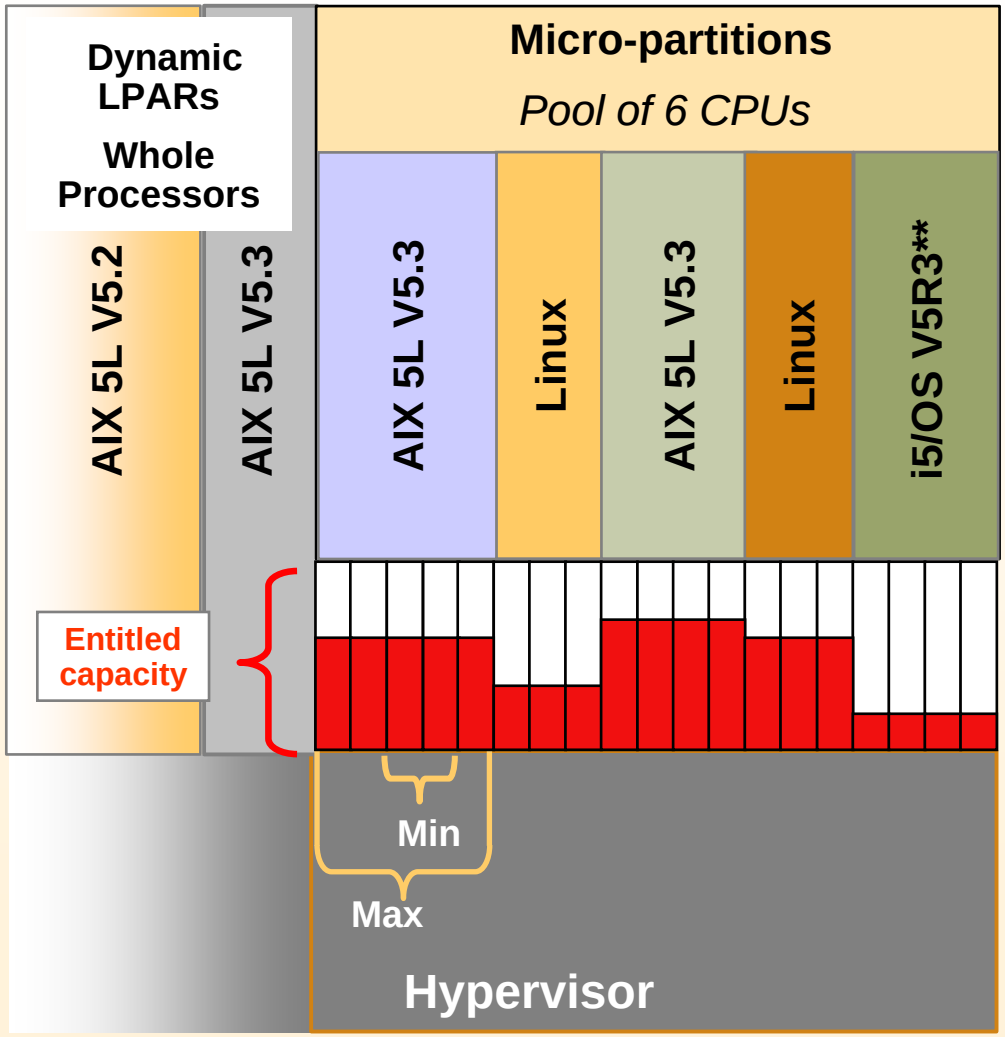
- Both AIX 5L V5.2 and AIX 5L V5.3 supported
- Balances processor and memory request

Managed via HMC or IVM

* SLES 9 or RHEL AS 3 and above

**Available on 1.65 GHz p5-570, p5-590 and p5-595 models

微分区技术



Note: Micro-partitions are optional.

Micro-Partitioning technology allows each processor to be subdivided into as many as 10 "virtual servers", helping to consolidate UNIX® and Linux applications.

Partitioning options

- Micro-partitions: Up to 254*
- Dynamic LPARs: Up to 32*
- Combination of both

Configured via the HMC

Number of logical processors

- Minimum/maximum

Entitled capacity

- In units of 1/100 of a CPU
- Minimum 1/10 of a CPU

Variable weight

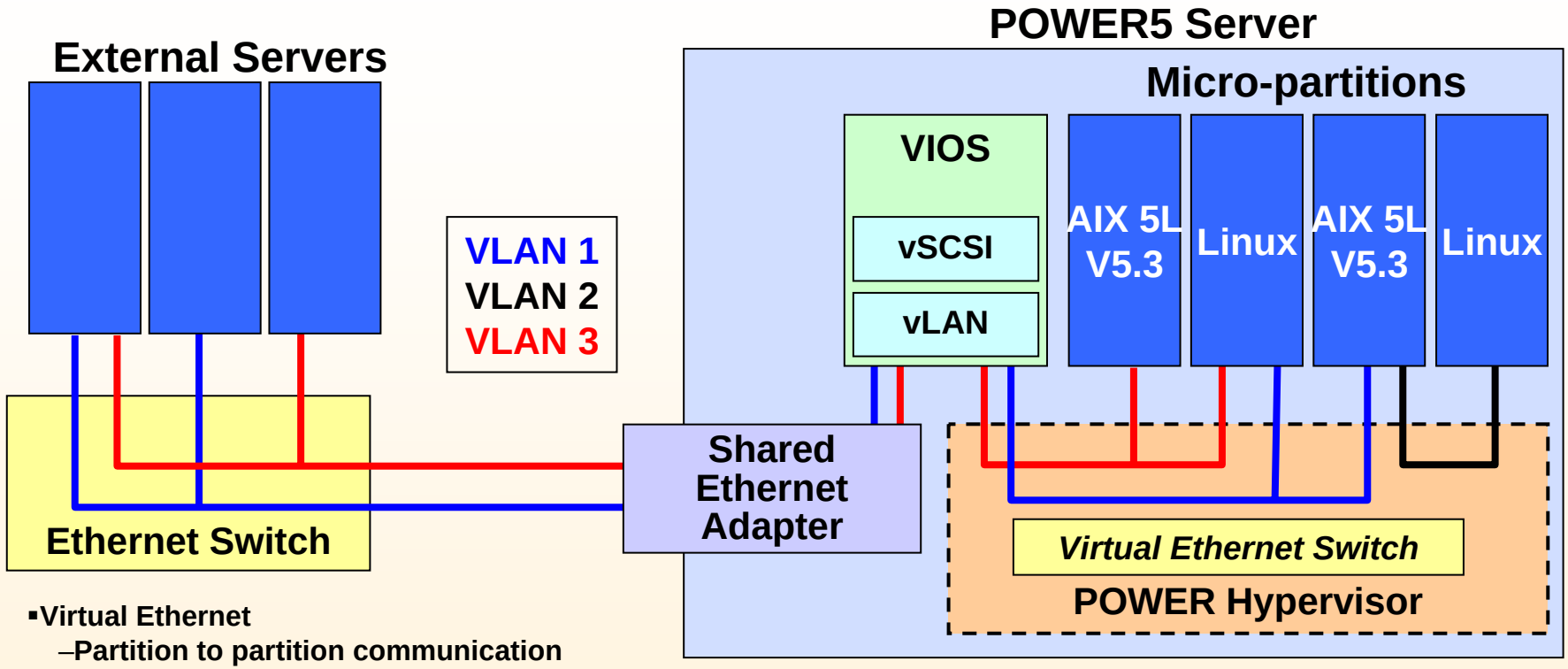
- % share (priority) of surplus capacity

Capped or uncapped partitions

*on p5-590 and p5-595
** on p5-570, p5-590, and p5-595

虚拟以太网络

Virtual Ethernet helps reduce hardware costs by sharing LAN adapters



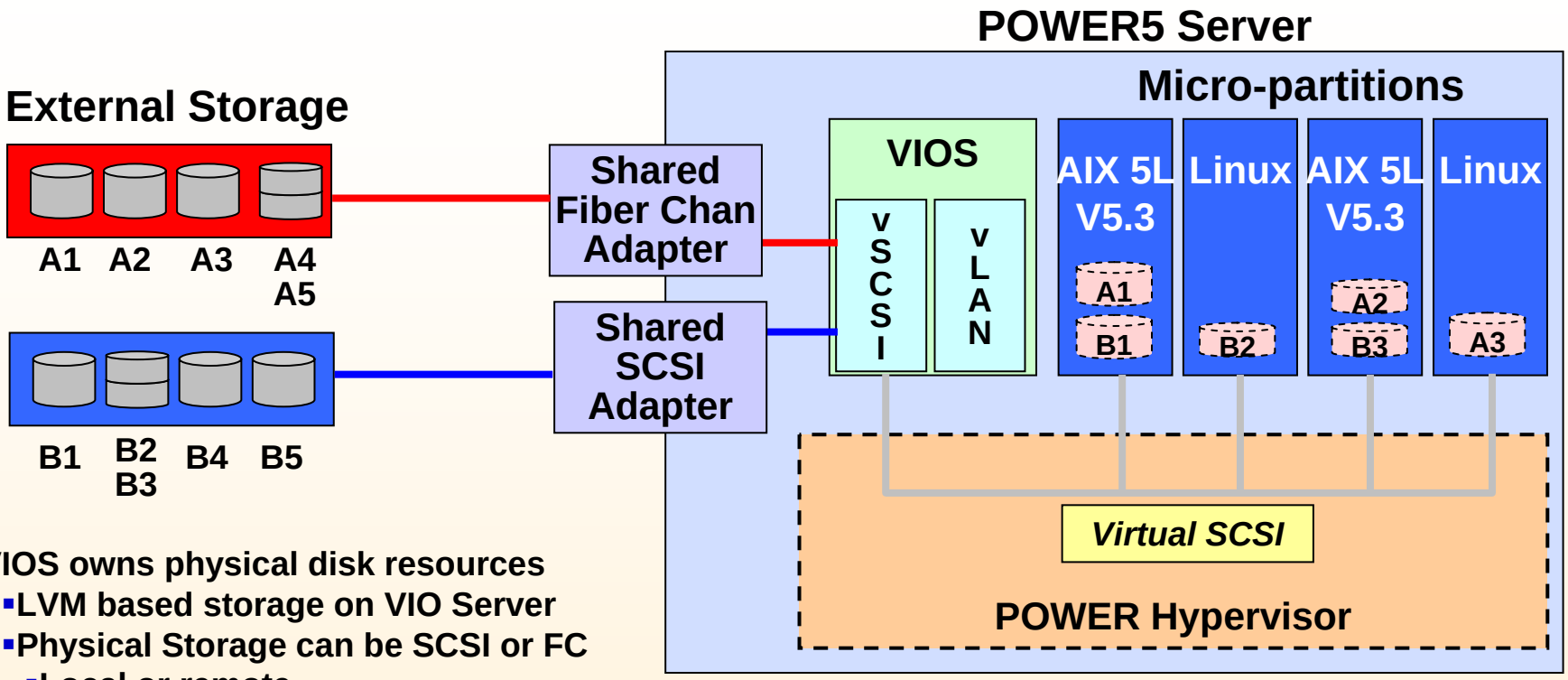
- Virtual Ethernet
  - Partition to partition communication
  - Requires AIX 5L V5.3 and POWER5
- Shared Ethernet Adapter
  - Provides access to outside world
  - Uses Physical Adapter in the Virtual I/O Server
- VLAN – Virtual LAN
  - Provide ability for one adapter to be on multiple subnets
  - Provide isolation of communication to VLAN members
  - Allows a single adapter to support multiple subnets

- IEEE VLANS
  - Up to 4096 VLANS
  - Up to 65533 vENET adapters
  - 21 VLANS per vENET adapter

Available via optional Advance POWER Virtualization or POWER Hypervisor and VIOS features.

虚拟存储

Virtual I/O helps reduce hardware costs by sharing disk drives



VIOS owns physical disk resources
- **LVM based storage on VIO Server**
- **Physical Storage can be SCSI or FC**
  - **Local or remote**

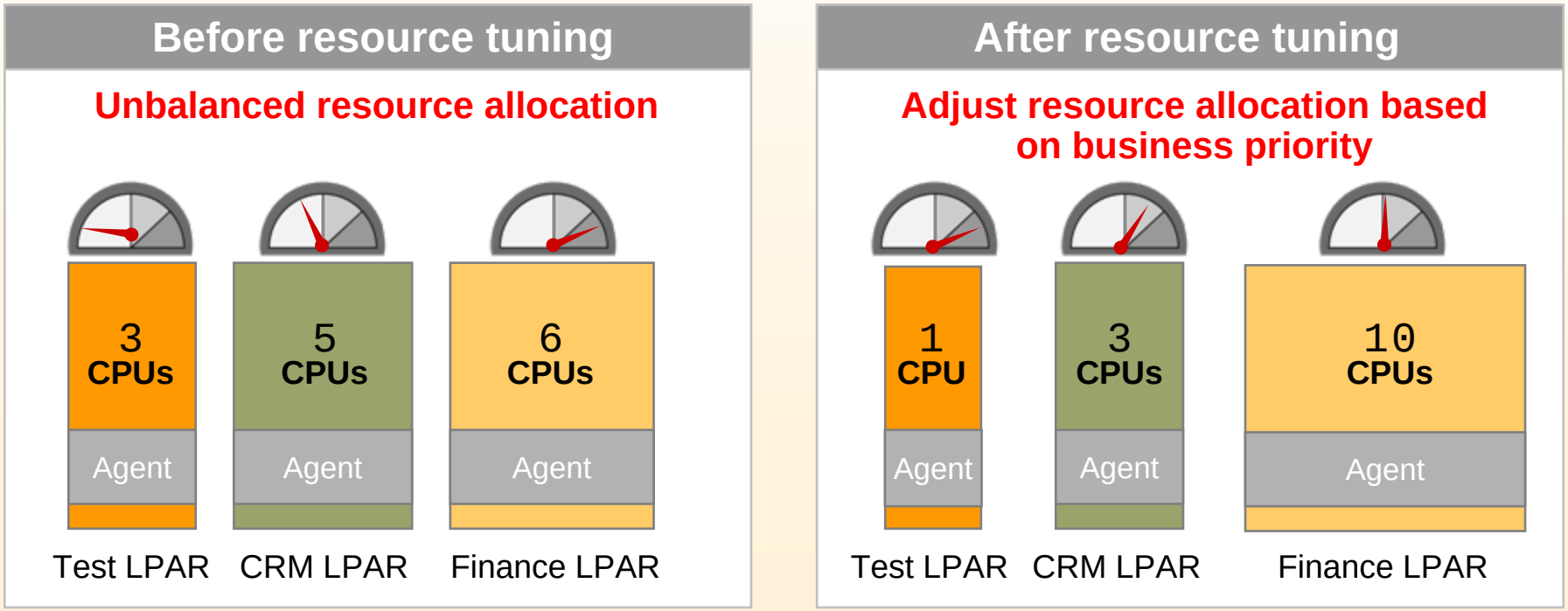
Micro-partition sees disks as vSCSI (Virtual SCSI) devices
- **Virtual SCSI devices added to partition via HMC**
- **LUNs on VIOS accessed as vSCSI disk**
- **VIOS must be active for client to boot**

Multiple LPARs can use same or different physical disk
- **Configure as logical volume on VIOS**
- **Appear a hdisk on the micro-partition**
- **Can assign entire hdisk to a single client**

Available via optional Advance POWER Virtualization or POWER Hypervisor and VIOS features.

Partition Load Manager for AIX 5L

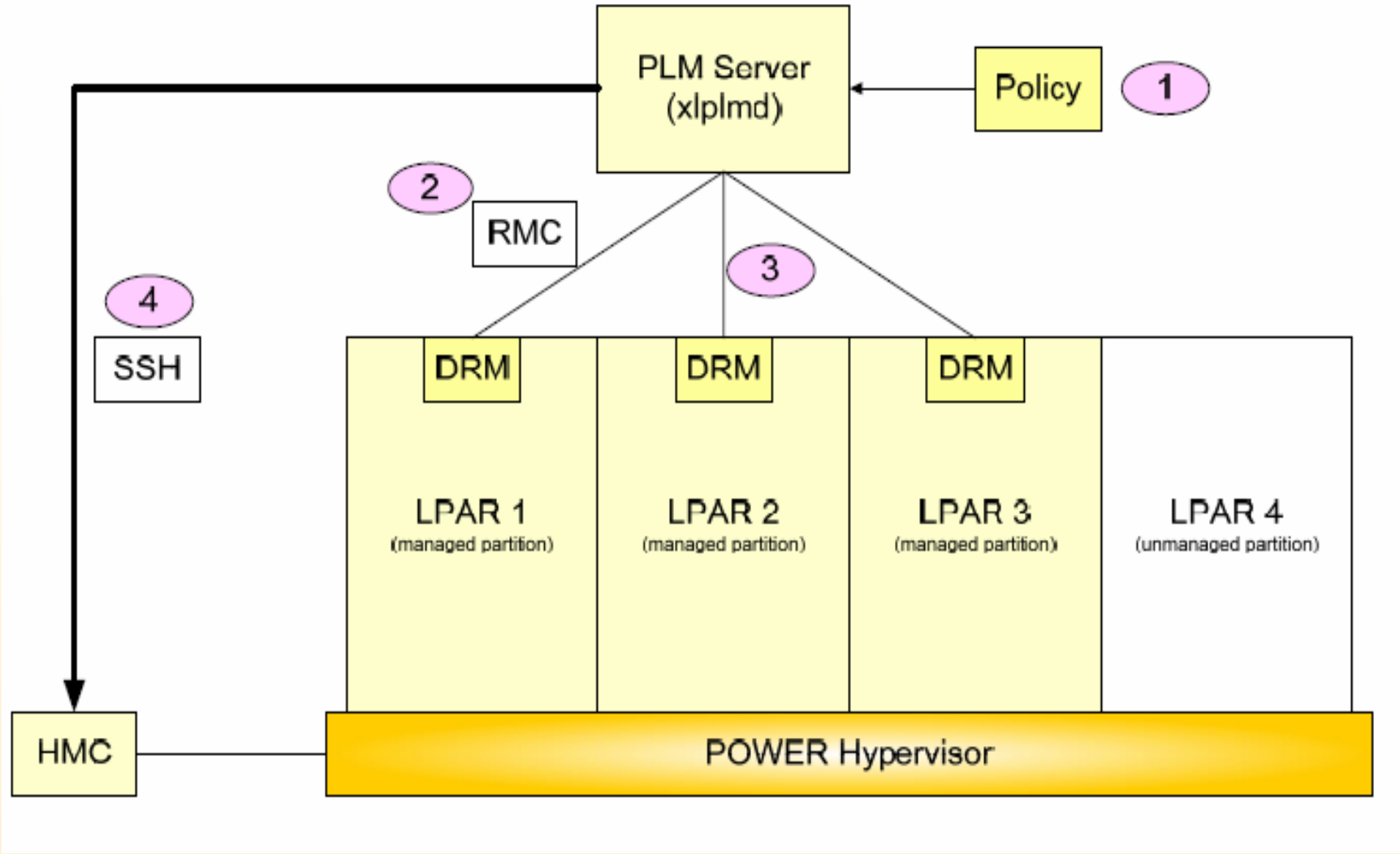
- 基于策略的，自动的对资源进行调整
- 动态地对CPU，内存进行调整

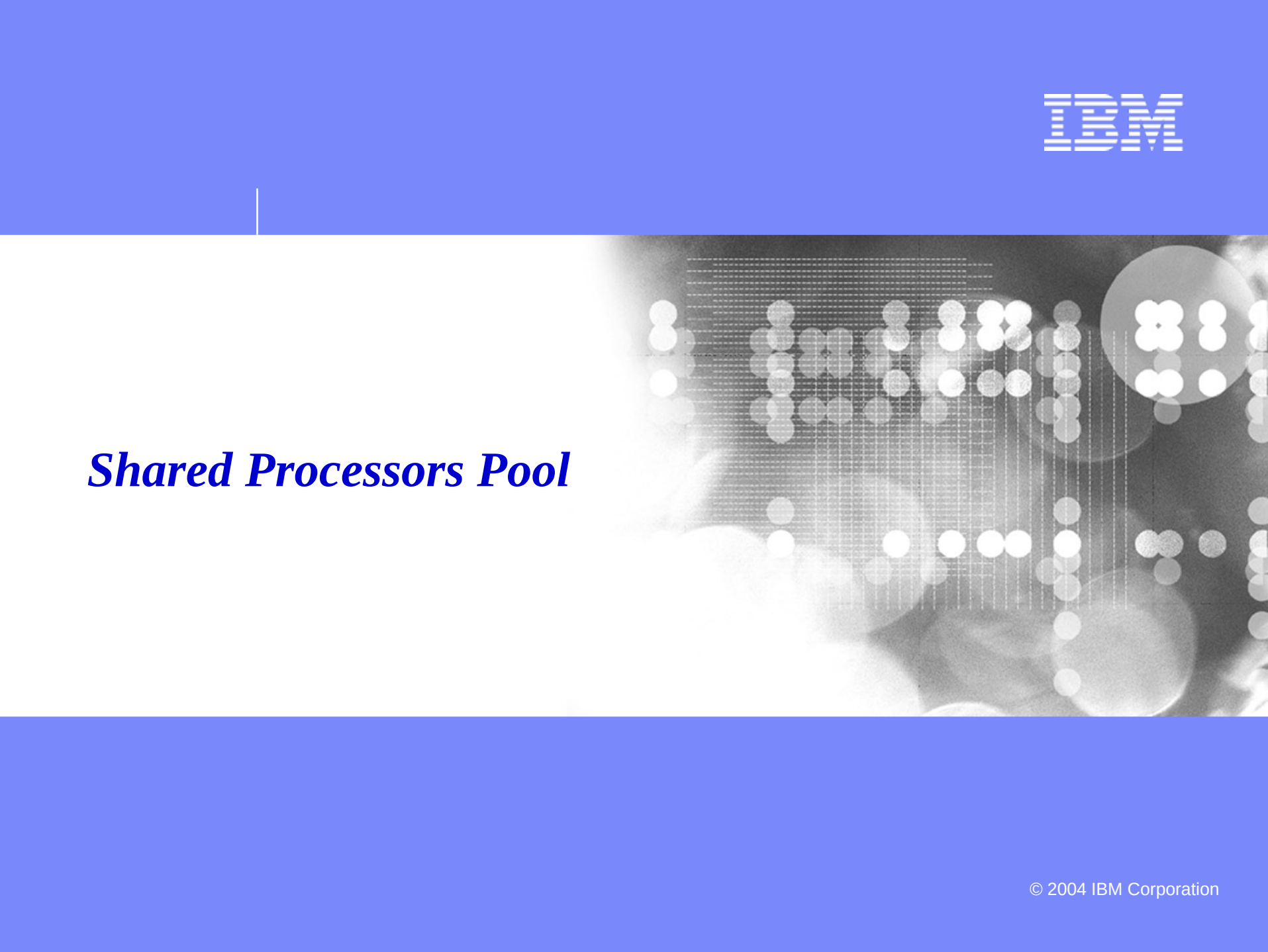


关于PLM 服务器

- The PLM server runs under AIX.
- The PLM server may reside either in a partition on the same server as the partitions being managed or on a different machine. When the PLM server runs in a partition, PLM is capable of managing its own partition.
- Multiple Partition Load Manager servers may be run on one AIX 5L system.
  (each has its own policy file)
- A partition can have, at most, one PLM manager.
- Management of all partitions in a system is not a requirement.
- One Partition Load Manager server can manage partitions within only one physical server.
- PLM cannot manage Linux partitions.

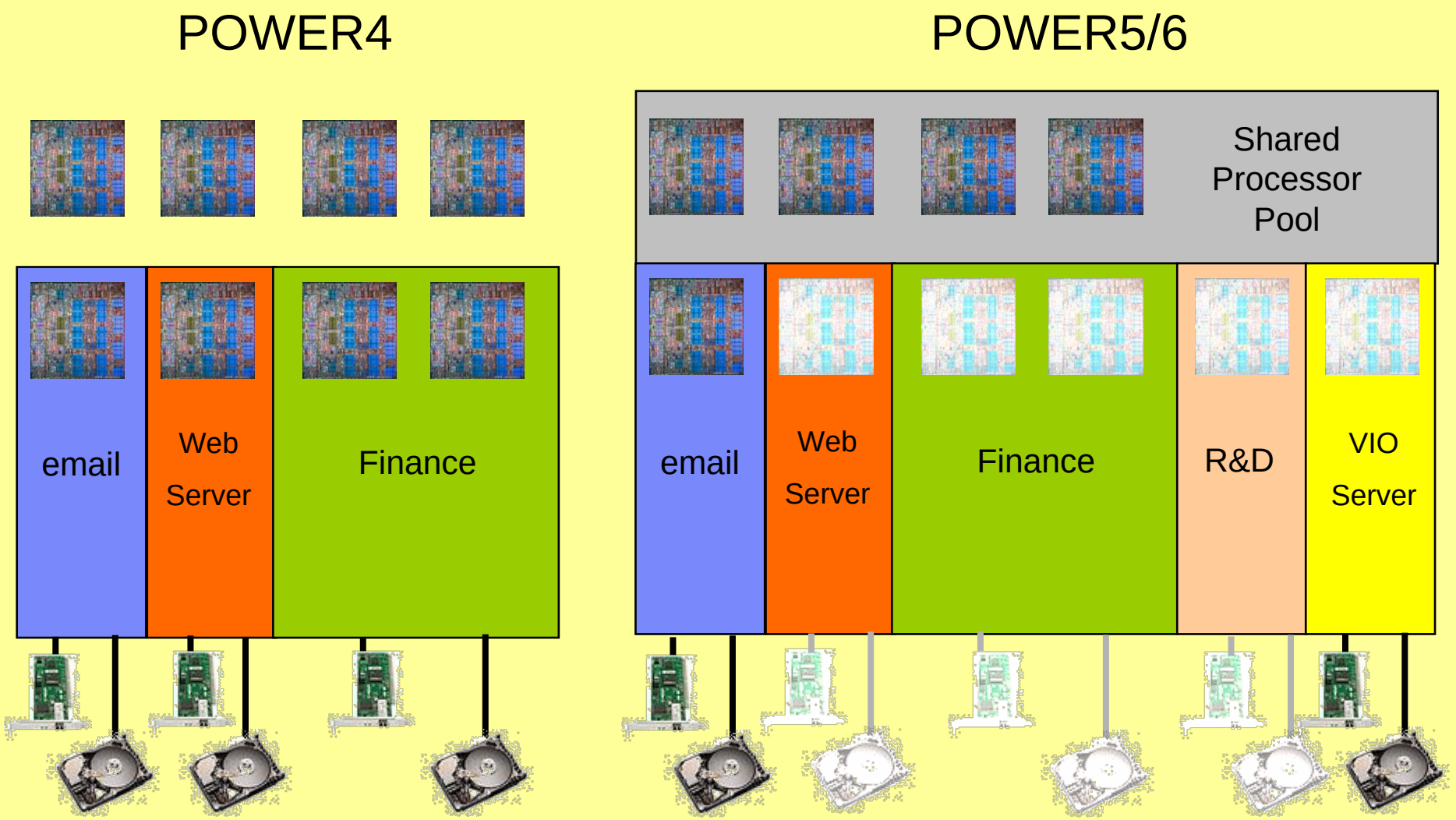
PLM 架构



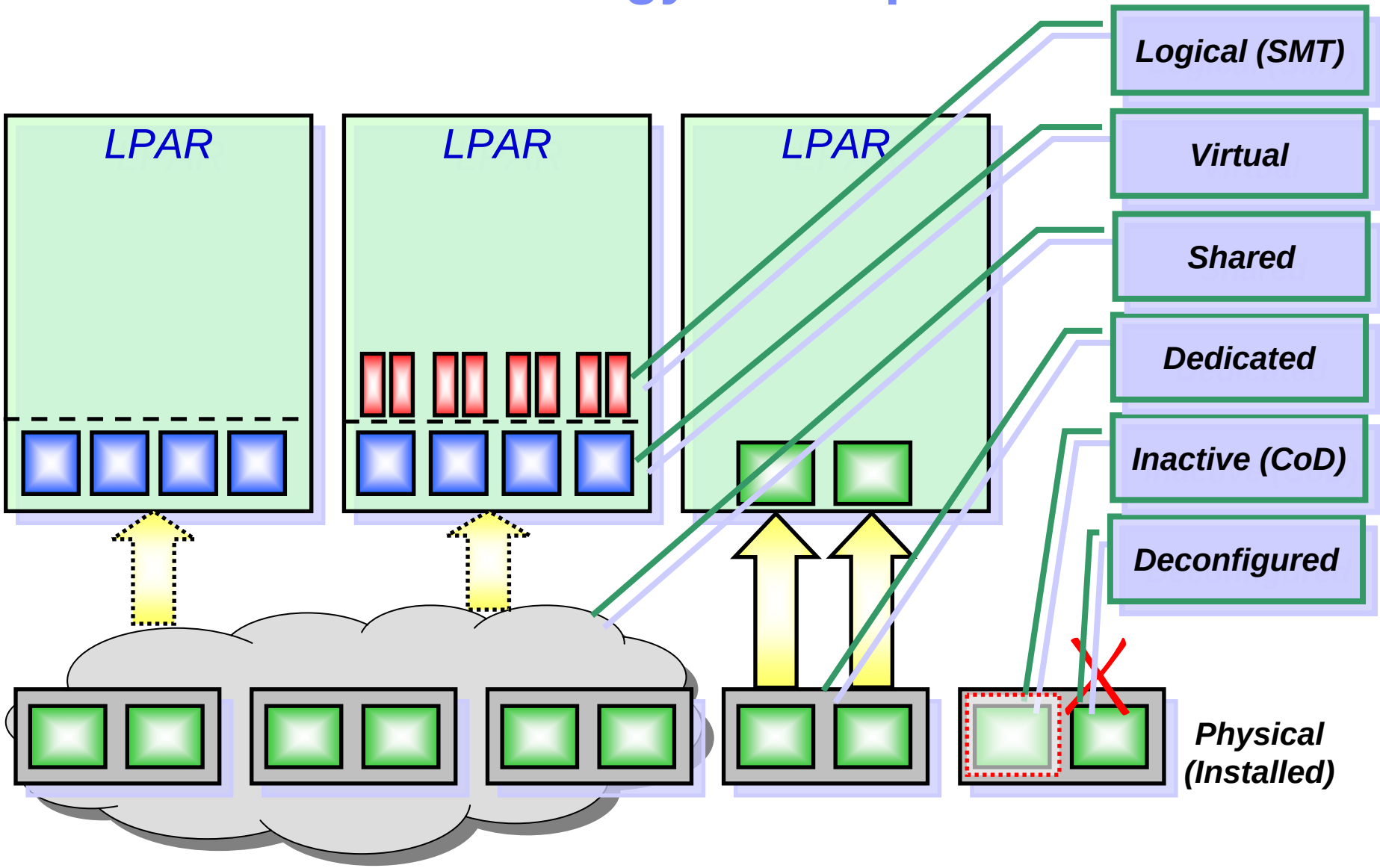


Shared Processors Pool

POWER4 vs POWER5/6 LPARs



Processor Terminology Concepts



Shared processor pool overview– Micro-partitions

- **LPARs are defined to be dedicated or shared**
  - Dedicated partitions use whole number of CPUs
  - Shared partitions use whole or fractions of CPUs from Shared Pool
- **Shared processor pool - subset of physical CPUs in a system**
  - all CPUs that are not in dedicated LPARs
- **Entitled capacity expressed in the form of number of 0.01 CPU units**
  - Desired: Size of partition at activation, between minimum and desired
  - Minimum: Partition won't start if Minimum capacity not available
  - Maximum: CPU that can't be exceeded in DLPAR operation
  - Divided among the virtual processors in the LPAR
- **Capped vs uncapped**
  - Capped: CPU Capacity limited to 'desired' entitlement
  - Uncapped: CPU Capacity limited by unused capacity in 'pool' and 'desired' virtual processors
- **SPLPARs run in 'virtual' processors**
  - Virtual cpus in an LPAR are dispatched on physical cpus in the p5 Server
- **LPAR weighting to determine preference when pool cycles are constrained**
  - Some LPARs more favored (up to weight 255), some less favored (down to weight 1)

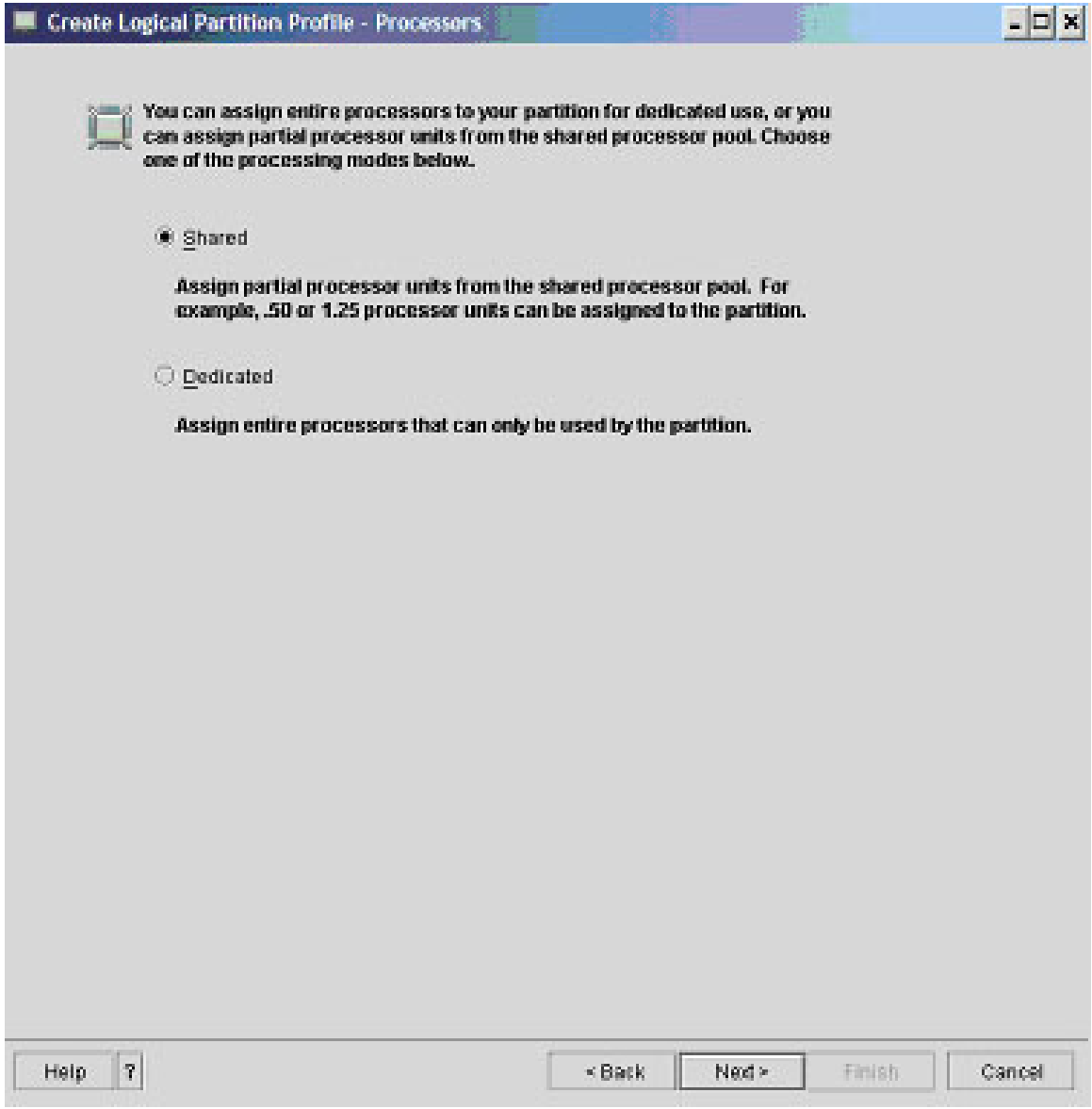
Virtual Processors in Shared Processor Pool

- Virtual processors represent concurrent operating system operations
- Entitled Capacity (physical cpu) is spread across these virtual processors
- Optimal number of virtual processors depends on the workload
  - Number of threads
  - What threads are doing
- Number of virtual processors (Minimum and Desired) is obtained by:
  - Rounding entitled capacity to next whole number
  - Example
    - Minimum = 0.50 (entitlement) -> 1 virtual processor minimum
    - Desired = 2.25 (entitlement) -> 3 virtual processors desired
- Maximum number of virtual processors is 10x entitlement
  - Do you want maximum 0.8 entitled over 8 virtual processors?
  - Some art, experimentation warranted
  - Some workloads need more concurrence, some need fewer and more powerful virtual processors

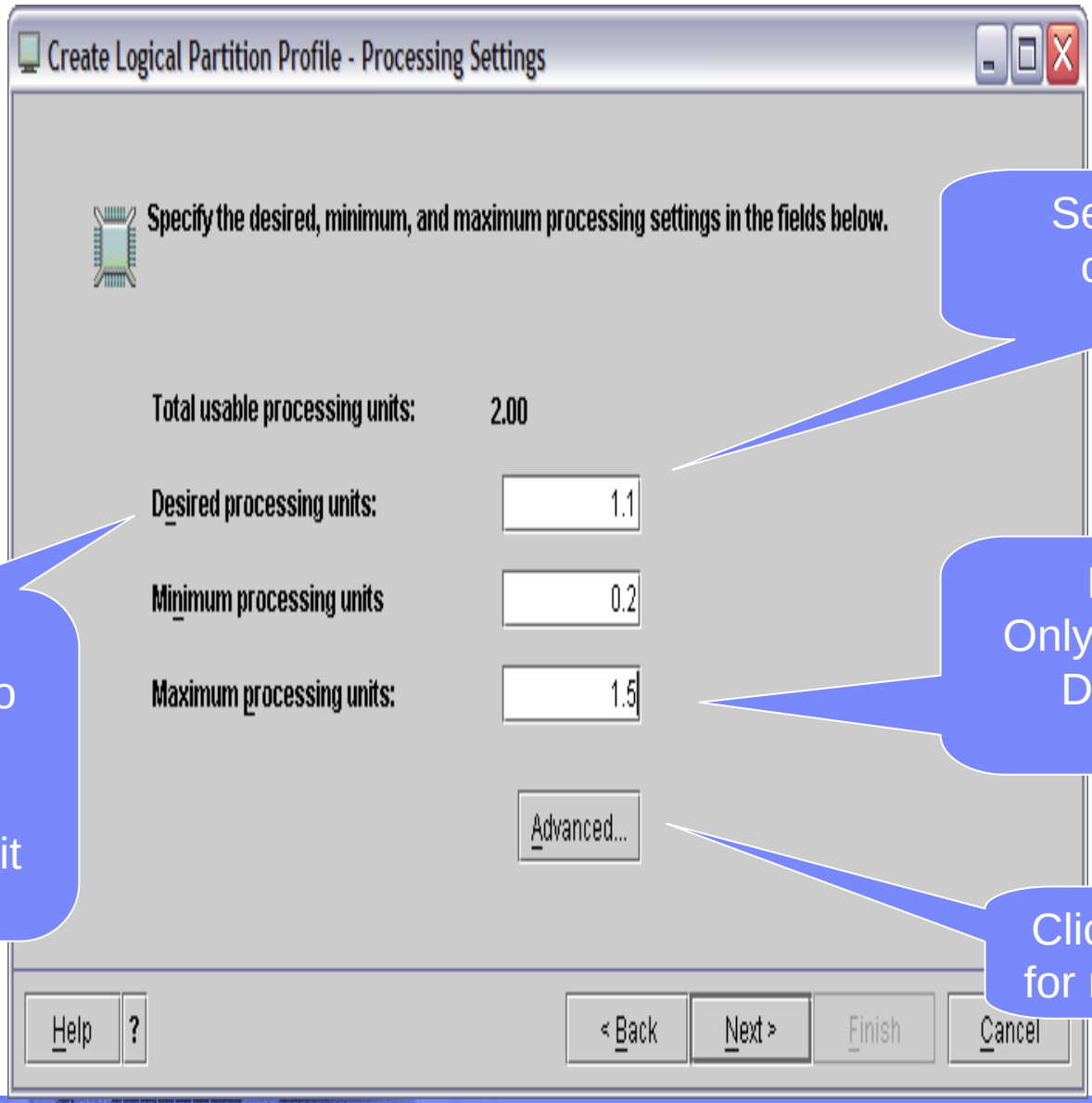
Virtual Processor Recommendations

- If you have more virtual processors than CPUs in the pool
  - ✓ You may have greater context switching
  - ✓ You may have greater overhead.
- If you have too few virtual processors too small
  - ✓ You limit the processing capacity of an uncapped partition
  - ✓ Cannot use more than 1.0 physical cpu cycles thru 1 virtual processor
- Example partition with desired 0.5 CPU units and excess capacity
  - ✓ With 1 virtual processor the partition cannot exceed 1.00 physical cpus
  - ✓ With 2 virtual processors the partition could use up to 2.00 physical cpus
  - ✓ With 5 virtual processors (max) physical capacity can reach 5.00
    - ❑ Maximum %ent is 1000% (0.1 entitled capacity, on 1 virtual processor, uncapped)

Shared Processor Configuration



Shared Processor Configuration



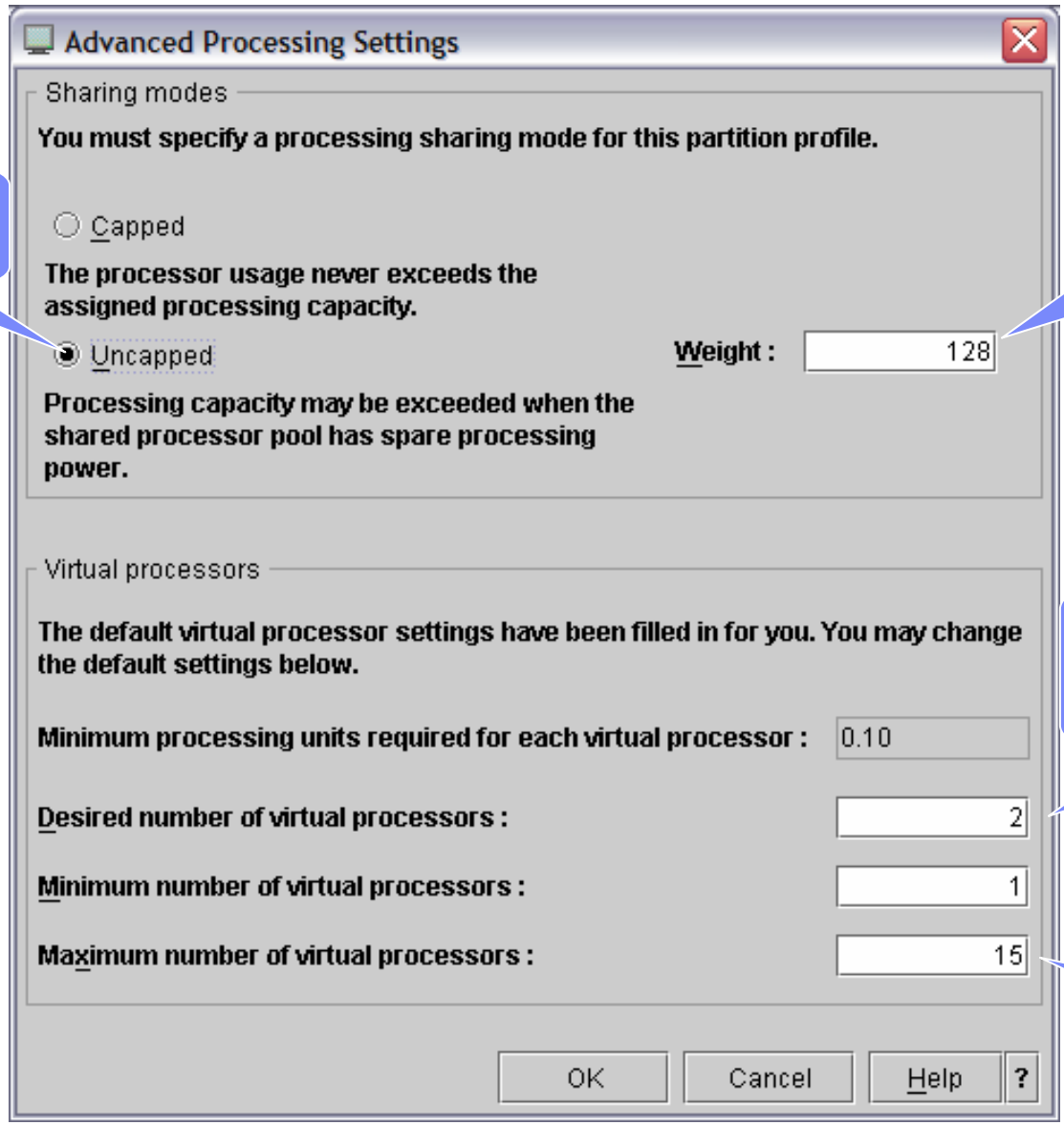
Set entitled capacity

Min: Required to activate Used for DLPAR limit

Max: Only used for DLPAR

Click Advanced for more options

Shared Processor Configuration



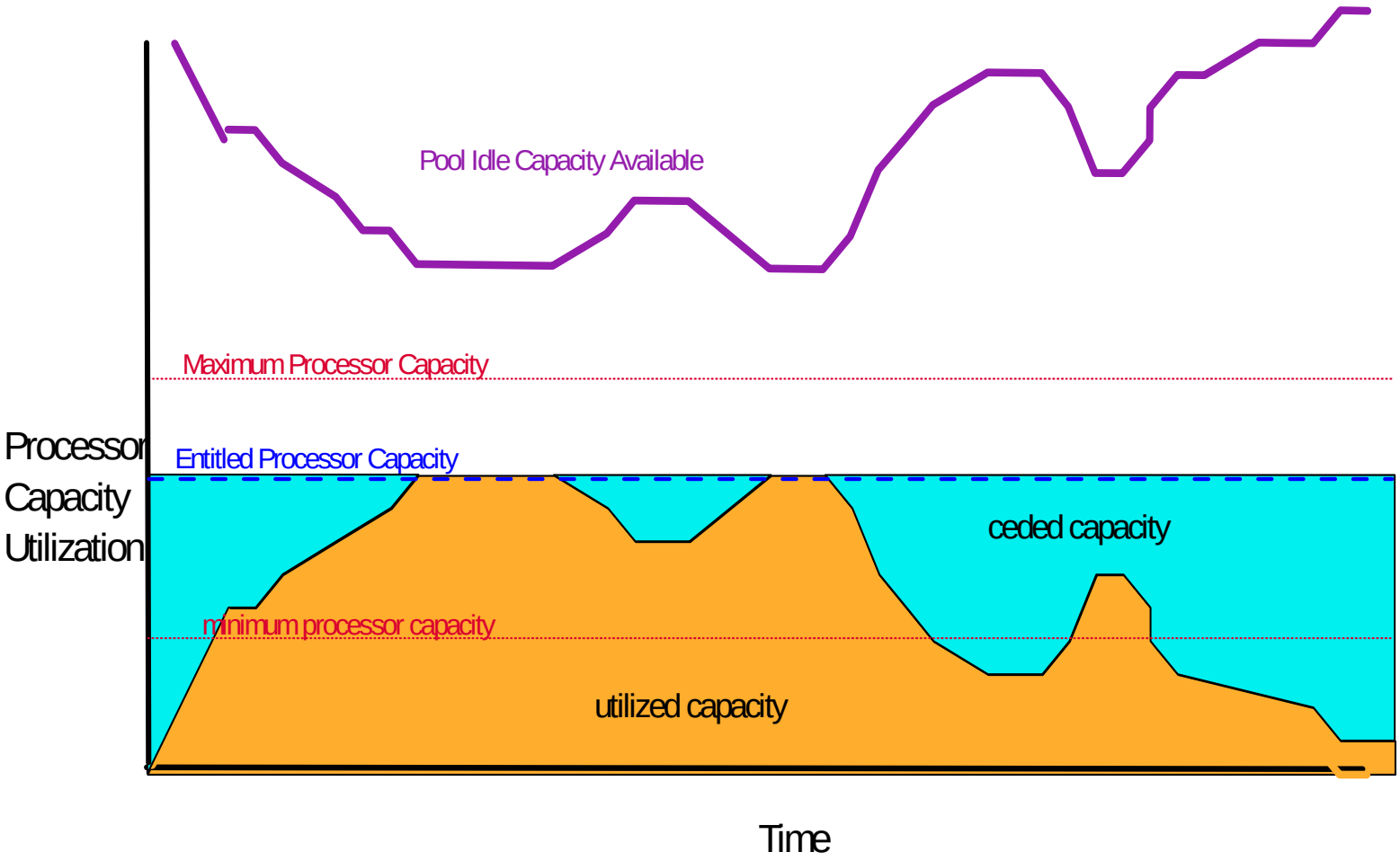
uncapped

Priority weight

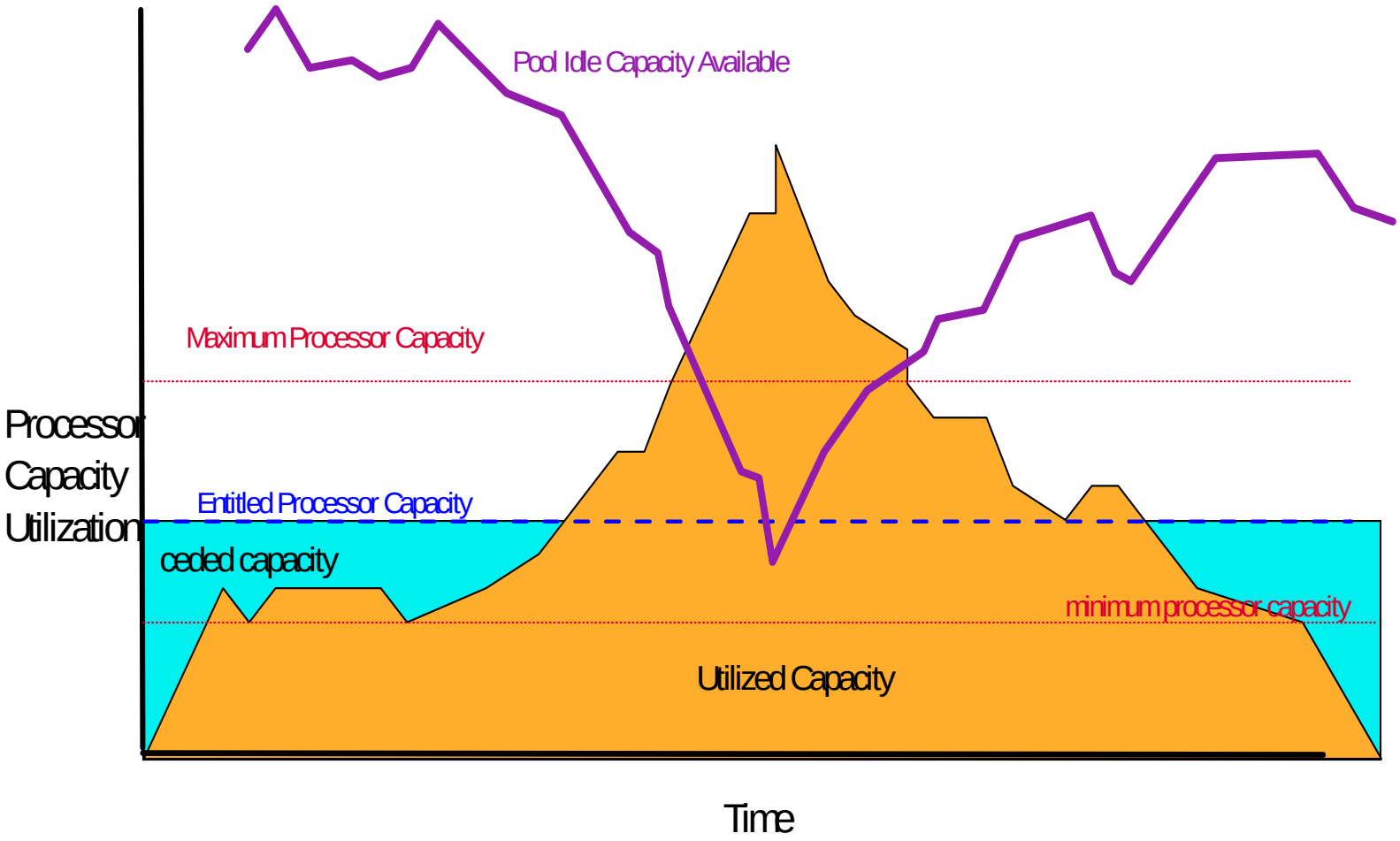
1.1 Rounded up

1.5 times 10

Capped Shared Processor LPAR



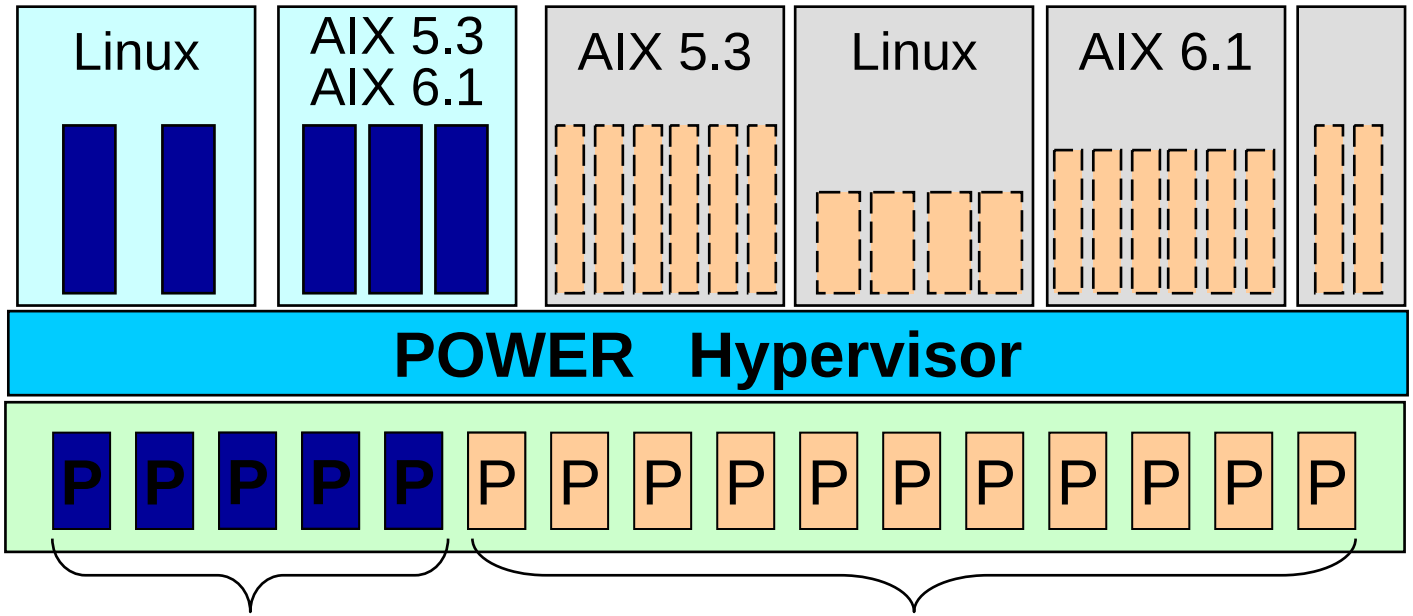
Uncapped Shared Processor LPAR



Shared Dedicated Capacity

- Shared Dedicated Capacity allows for the “donation” of spare cycles for dedicated processor partitions to be utilized by the shared pool
- User Benefits
  - Dedicated partition gets absolute priority for these excess cycles. Sharing will only occur when the dedicated partition has not consumed all resources and the uncapped partitions that have consumed all of their entitled capacity
- Cycles of a dedicated processor are “borrowed” for a short time to run an uncapped partition and then returned to the dedicated partition
- Shared Dedicated Capacity is supported on POWER6

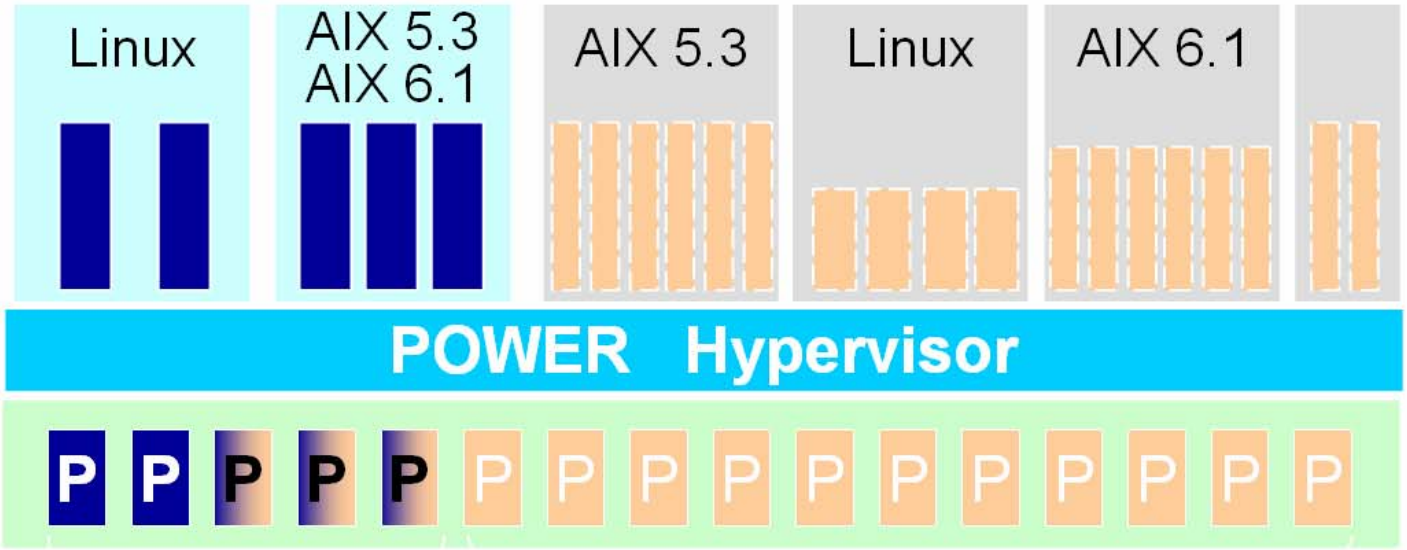
Planned Dedicated Shared Processors



Excess Dedicated Capacity Utilization

- Unused capacity in dedicated processor partitions can be “Donated” to shared processor pool
- Excess cycles will only be utilized by uncapped partitions that have consumed all of their entitled capacity.
- POWER6 Servers

Planned Dedicated Shared Processors



Dedicated / Shared Processors	Shared (Non-Dedicated) Processors
Excess Dedicated Capacity Utilization	▪ Unused capacity in dedicated processor partitions can be "Donated" to shared processor pool ▪ Excess cycles will only be utilized by uncapped partitions that have consumed all of their entitled capacity. ▪ POWER6 Servers

IBM System p

IBM power6系列服务器虚拟化技术

IBM PowerVM™on System p™

什么是PowerVM

Hardware and software that delivers industry-leading virtualization on IBM POWER processor-based servers for UNIX, Linux and i5/OS clients

Micro-Partitioning™

Subsystems

Workload Partitions

Live Application Mobility

PowerVM Editions features

Micro-Partitioning

Virtual I/O Server

Integrated Virtualization Manager

Live Partition Mobility

Lx86 (formerly System p AVE)

Logical Partitioning

IBM Power 6 虚拟化技术

- IBM Power 6服务器在继承了Power 4/5服务器虚拟化功能的同时,在虚拟化功能上又有了创新和发展.
- **PowerVM Editions**
  - PowerVM Express Edition
  - PowerVM Standard Edition
  - PowerVM Enterprise Edition
- **AIX 6**
  - PowerVM Live Application Mobility
  - PowerVM Workload Partitions
  - PowerVM Workload Partitions Manager
- **Components of PowerVM Editions** (Examples)
  - Live Partition Mobility (a feature of PowerVM Enterprise Edition)
  - Multiple Shared Processor Pools (a feature of PowerVM Standard and Enterprise Editions with POWER6)
  - Shared Dedicated Capacity (a feature of PowerVM Editions on POWER6 based processors)
  - Lx86 (a feature of PowerVM Editions)

PowerVM三种版本比较

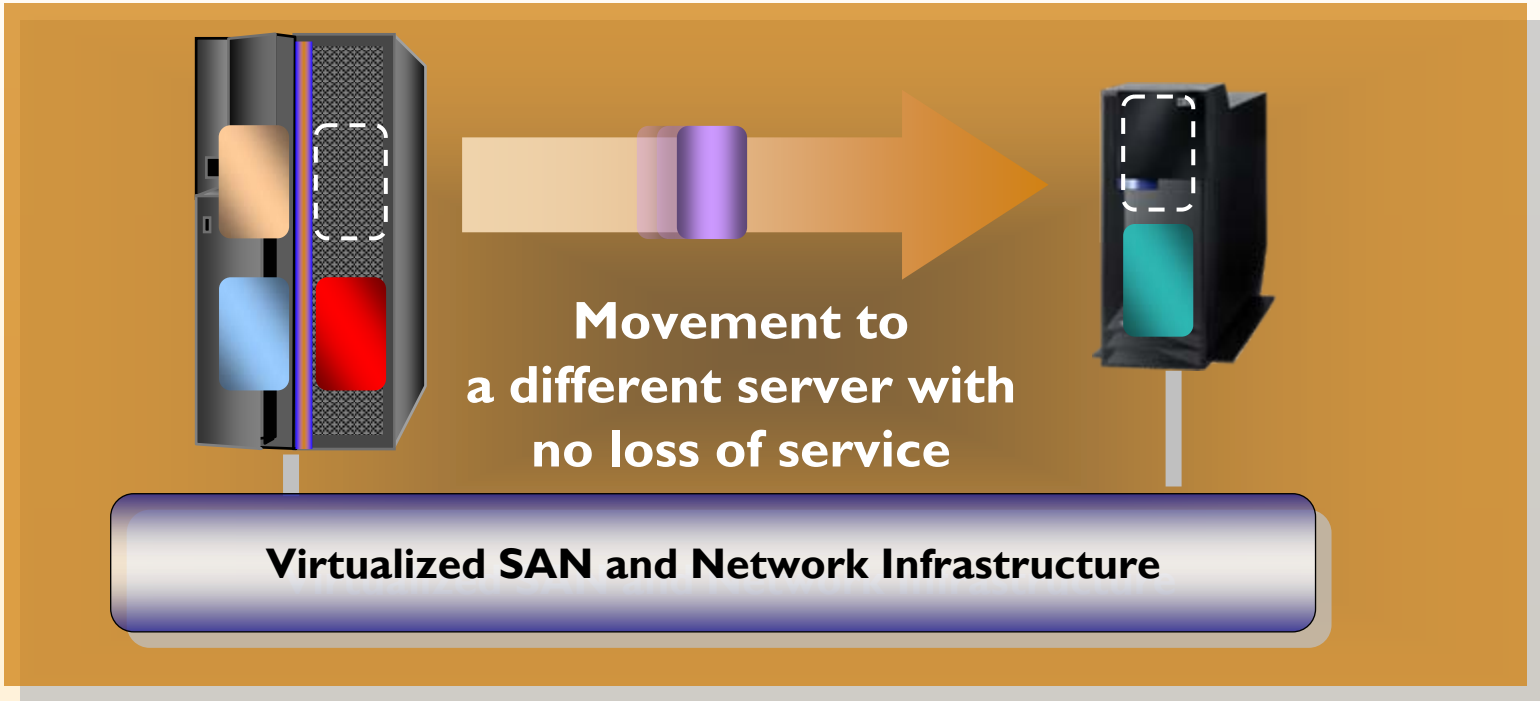
	Express Edition	Standard Edition	Enterprise Edition
Servers Supported	p520 / p550	p5, p6 and JS2X	p6 and JS22
Max LPARs	3 / Server	10 / Core	10 / Core
Management	IVM	IVM & HMC	IVM & HMC
VIOS	Yes	Yes	Yes
Live Partition Mobility	No	No	Yes
Shared Processor Pools	No	Yes **(HMC)**	Yes **(HMC)**
Shared Dedicated Capacity	Yes	Yes	Yes
PLM	No	No	No
Operating Systems	AIX & Linux	AIX & Linux	AIX & Linux
PowerVM Lx86	Yes	Yes	Yes

Upgrade from edition to another: Electronic Key

POWER6* Live Partition Mobility

允许在不中断**LPAR**上应用的情况下,将**LPAR**动态迁移到另外一台物理的服务器上.

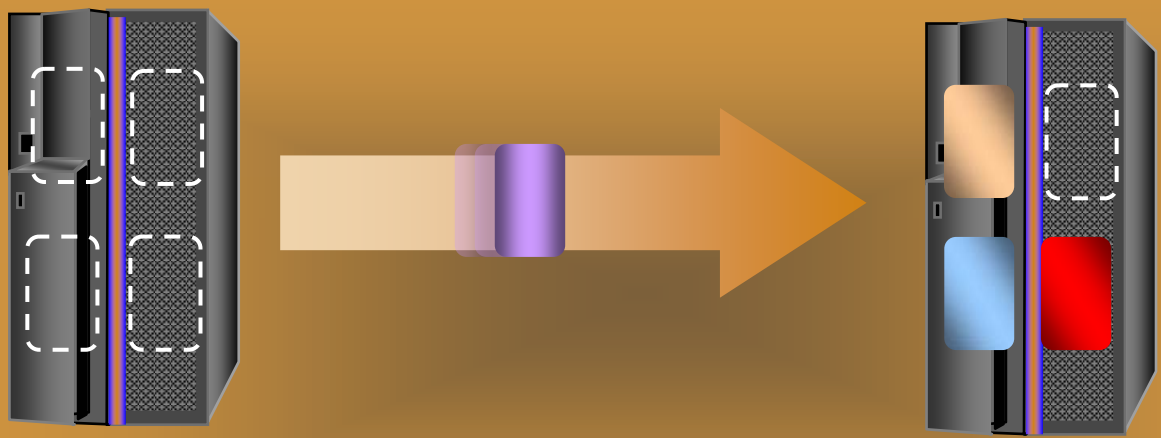
- ✓ **Reduce impact of planned outages**
- ✓ **Relocate workloads to enable growth**
- ✓ **Provision new technology with no disruption to service**
- ✓ **Save energy by moving workloads off underutilized servers**



* All statements regarding IBM future directions and intent are subject to change or withdrawal without notice and represent goals and objectives only. Any reliance on these Statements of General Direction is at the relying party's sole risk and will not create liability or obligation for IBM.

提高应用的可用性

- ***With Live Partition Mobility, planned outages for hardware and firmware maintenance and upgrades can be a thing of the past***

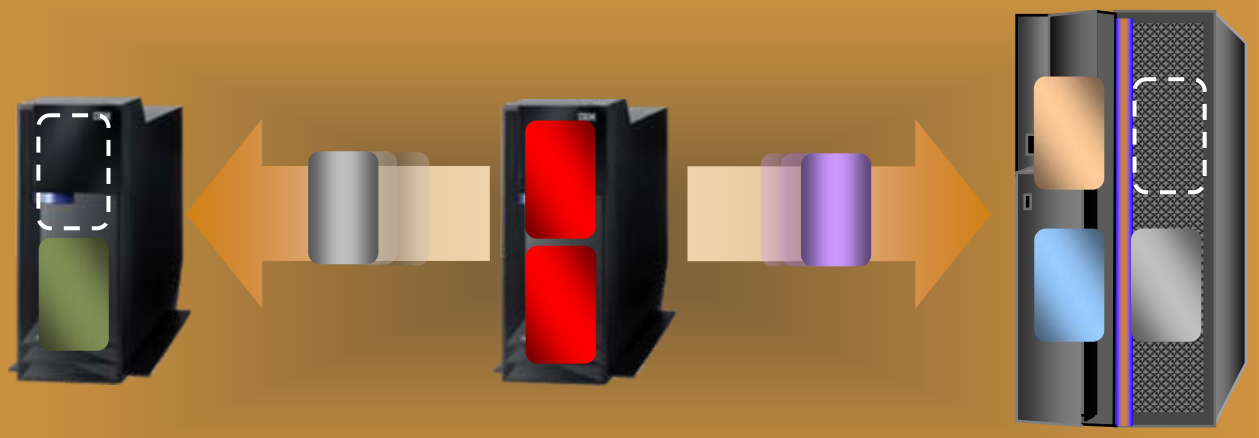


Relocate all partitions from one server to another when performing maintenance. Move the partitions back when maintenance is complete

* All statements regarding IBM future directions and intent are subject to change or withdrawal without notice and represent goals and objectives only. Any reliance on these Statements of General Direction is at the relying party's sole risk and will not create liability or obligation for IBM.

将工作负载动态的分配到多个服务器上

- ***As computing needs spike, redistribute workloads onto multiple physical servers without service interruption***



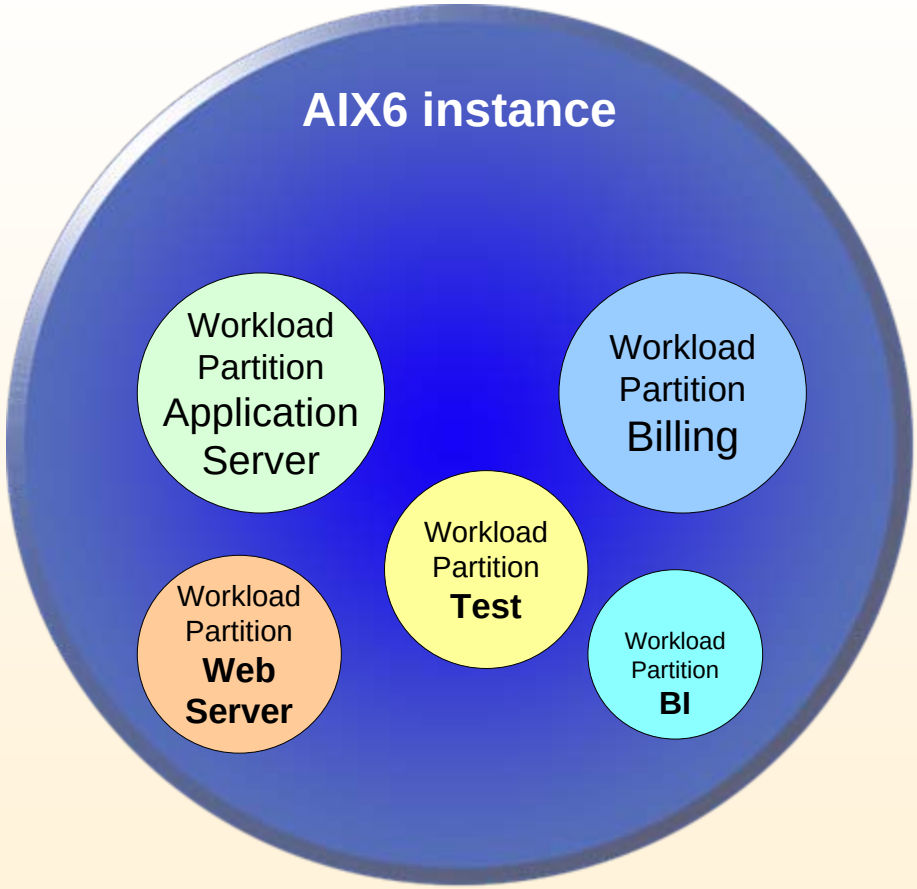
As one server gets overtaxed from a spike in demand, relocate partitions to other servers

* All statements regarding IBM future directions and intent are subject to change or withdrawal without notice and represent goals and objectives only. Any reliance on these Statements of General Direction is at the relying party's sole risk and will not create liability or obligation for IBM.

Workload Partitions 使AIX系统更加灵活

- Improved administrative efficiency by reducing the number of AIX images to maintain

- Comes as part of standard AIX 6
- AIX is partitioned in software
- Each Workload Partition (WPAR)
  - Obtains a regulated share of system resources
  - May have unique network & filesystems
  - A separate administrative & security domain
- Two types of Workload Partitions
  - System Partitions → comprehensive
  - Application Partitions → 1 app, quick & simple
- WPARs share the global system resources
  - Operating System, I/O, Processor, Memory



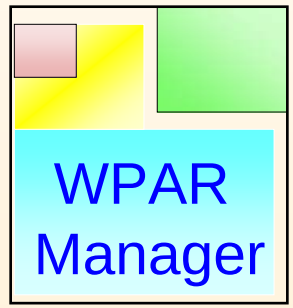
Pre-requisites → AIX 6 → POWER4, 5 or 6

两种WPAR AIX Offerings in 2007

- **AIX V6.1 (and AIX OpenBeta 6.0)**
  - Workload Partitions (WPAR) included in base AIX
  - In the default install image
  - Many WPARs on each machine or LPAR

- **Workload Partitions Manager***
  - Enablement for Live Application Mobility
  - Central Manager and agent on each machine
  - Cross System Management for Workload Partitions
  - Automated, Policy-based Application Mobility

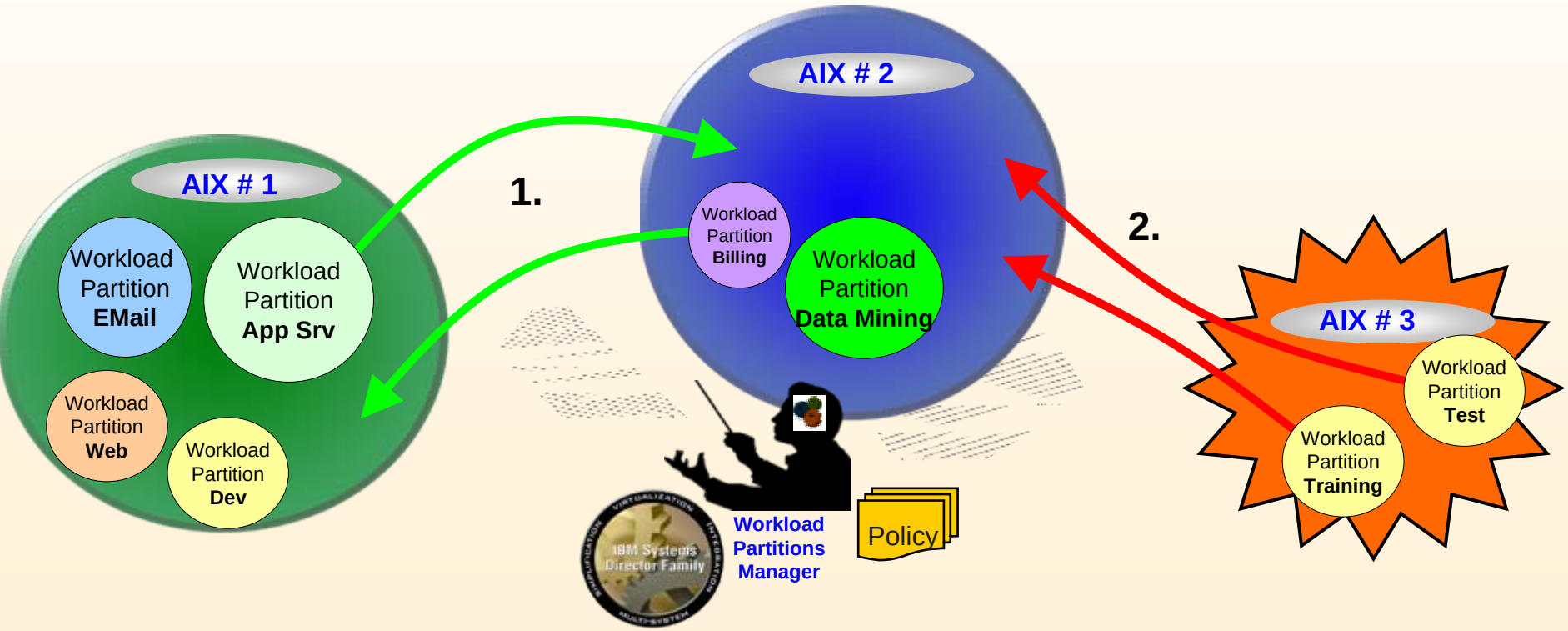
  * Part of the IBM System Director Family



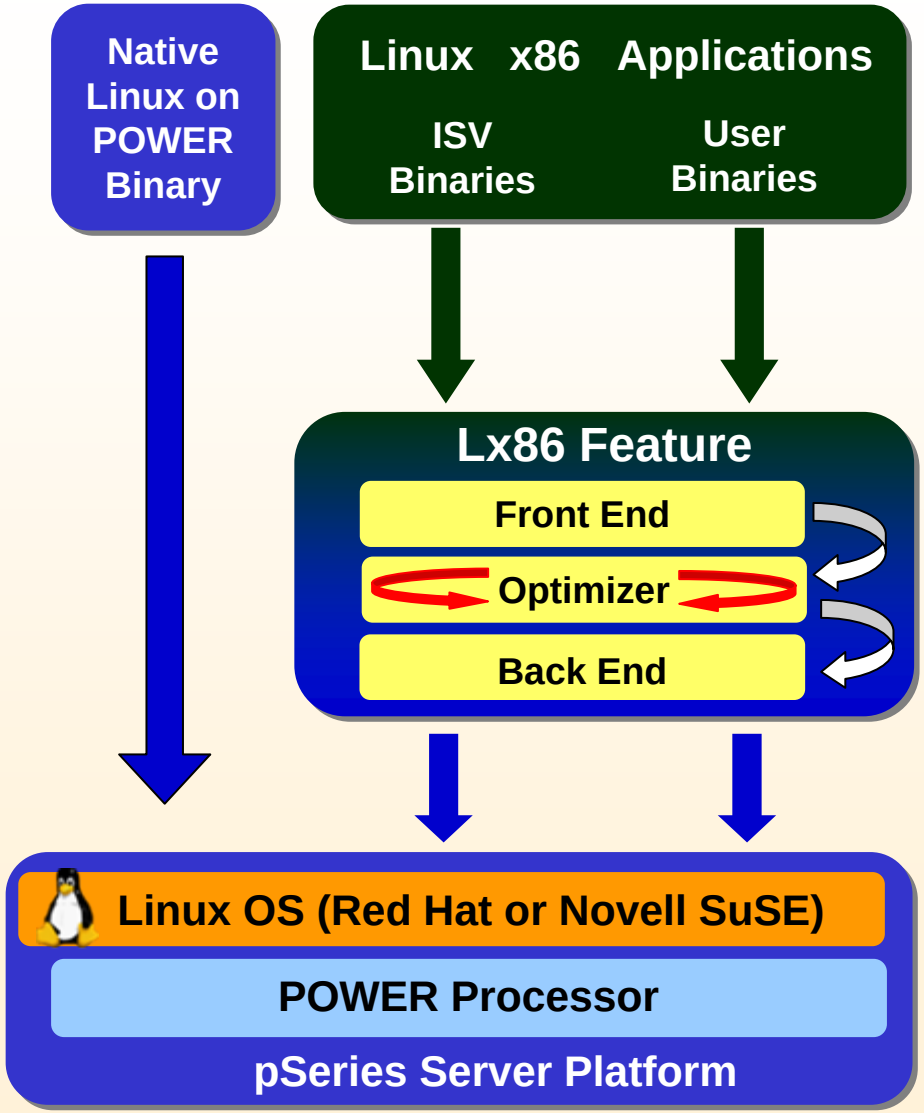
Live Application Mobility

Moving a running Workload Partition to another machine or LPAR

1. Multi-system workload balancing
   - IT optimization to maximize CPU, memory, I/O effectiveness
2. Empty a machine for application outage avoidance
   - e.g. Upgrade machine, AIX or firmware or for machine Repair

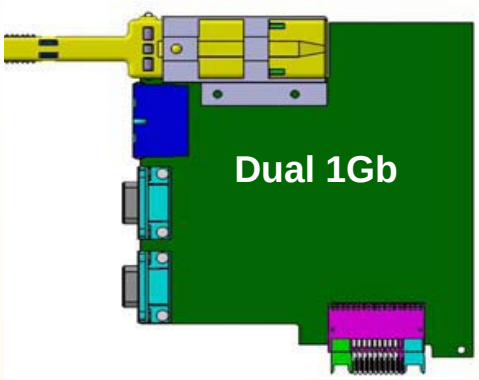


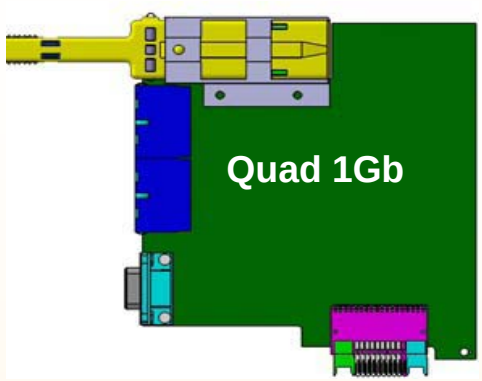
什么是PowerVM Lx86???

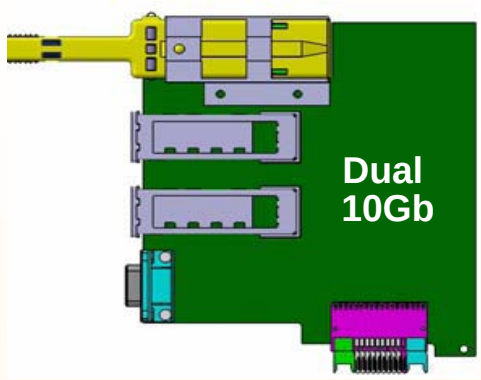


- **Dynamically translates and maps x86 Linux instructions to POWER**
- **Run any 32-bit Linux/x86 user space application binary**
  - –Requires no re-compilation
  - –Direct H/W access or apps using kernel mod's not supported
- **Native & translated applications interoperate**
- **Transparent to users**
  - –"It just runs" like on Linux/x86
- **Performance: 60-80% of native**
  - –"Good Enough" for many apps
  - –Translates blocks of code into intermediate representation
  - –Performs optimizations
  - –Stores optimized, frequently used blocks of code in cache
  - –Handles Linux OS call mapping
  - –Encodes binary for target POWER processor platform

集成的虚拟以太网卡(IVE)







Address Sharing:

- Dual 1GB: 16 MAC Addresses / pair Total: 16
- Quad 1GB 16 MAC Addresses / pair Total: 32
- Dual 10GB: 16 MAC Addresses / port Total: 32

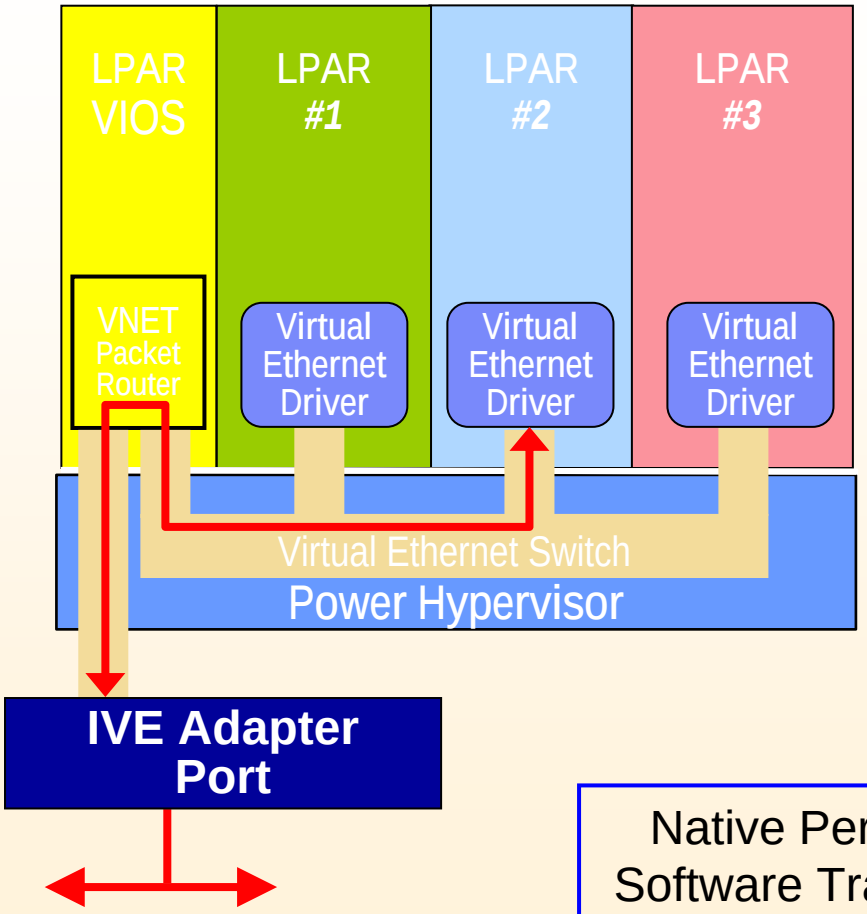
Non VIOS Partition: Address Sharing (MAC Addresses)

- Time Slicing “Physical” Ethernet adapter resources

VIOS Partitions: IVE logical/physical port is dedicated

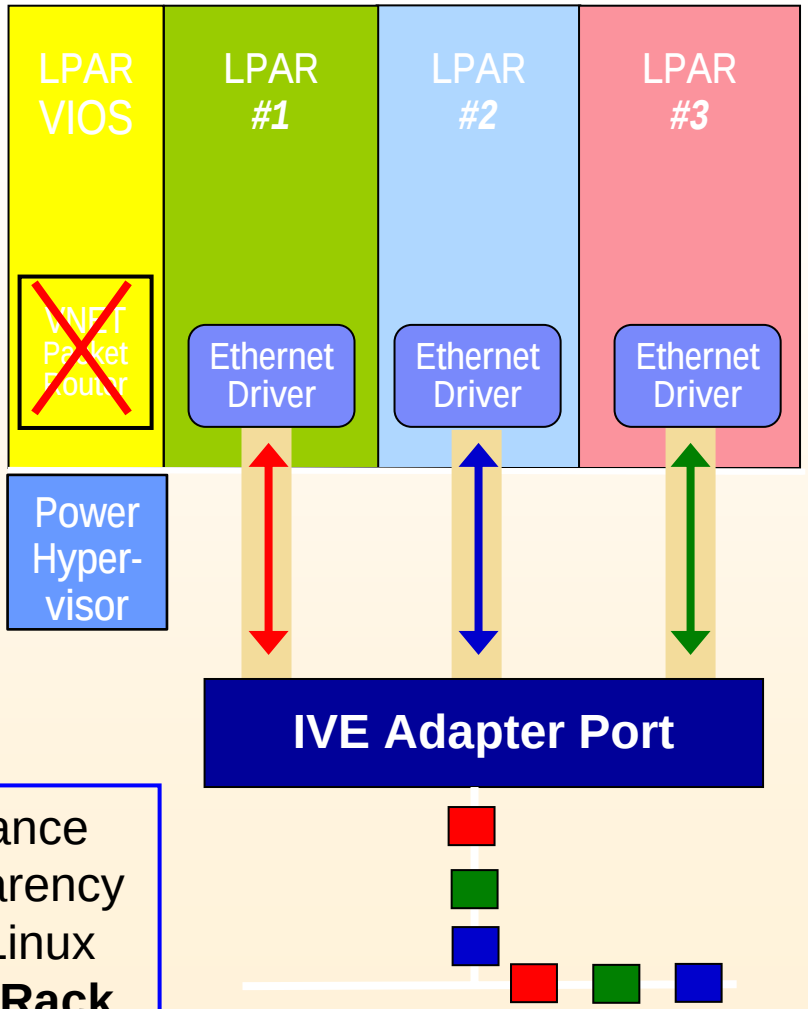
Integrated Virtual Ethernet How it works……

Option 1



Or

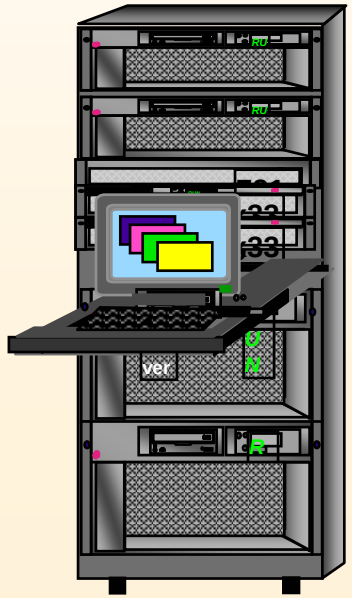
Option 2



Native Performance
Software Transparency
AIX 5.3 / 5.4 & Linux
P6 Midrange & Rack

硬件管理控制台 (HMC)

Ethernet support	▪ POWER5/6 provides Ethernet support of the HMC ▪ POWER4 system still requires RS232 connections
Management system POWER4 and POWER5 systems POWER5 systems	▪ Required for LPAR, CoD and clustering environments
Support for: p5-520, p5-550, p5-570, p5-590, and p5-595	▪ 7310-CR4/C06 is only for Power5 ▪ 7042-CR4/C06 for Power5/6 ▪ Support for up to 32 @server p5 systems with up to 254 partitions
New software load to support @server p6 systems	▪ POWER5 servers and POWER6 servers can be managed from the same HMC

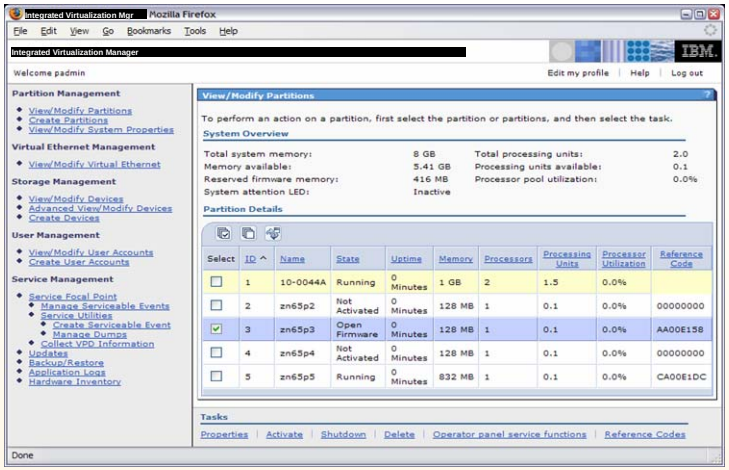


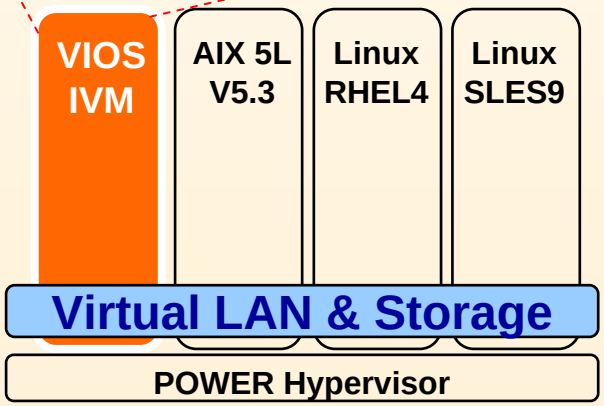
POWER6 HMC 增强

- **Support for POWER5 and POWER6 on same HMC**
- **Updated HMC hardware (Intel® technology refresh)**
- **No change: Hardware scaling support**
  - New Models: 7042-CR4 & 7042-C06
  - 32 physical systems
  - Up to 254 LPARS
- **Native browser access; WebSM no longer required**
  - Firefox 1.5.0.6 or later.
  - Microsoft® Internet Explorer 6.0 or later
- **Support for modified CSM on HMC**
- **Upgrade support for POWER5 HMC to POWER6**
  - 7310 will support POWER6 environment
  - New model type for POWER6: 7042
- **Internal modem support for rack models**
  - Available with CR4 models
  - Support will vary by geography

Integrated Virtualization Manager (IVM)

"HMC within a partition"





- Provides LPAR/virtualization support ***<u>without a physical HMC</u>***
  - Lower $$ entry point
- **Web-based**, intuitive/**user-friendly** interface
- Shipped with the Virtual I/O Server (VIOS)
- Supports **Creating/Management of I/O and LPARs** within a single physical server
- Subset of HMC Service functionality
- All I/O is virtualized – Virtual Console, Storage, Ethernet and Optical
- Available on these IBM systems:
  - System p5 505, 520, and 550 / 550Q
  - eServer p5 510, 520, and 550
  - OpenPower 710 and 720
  - Power6 520/550