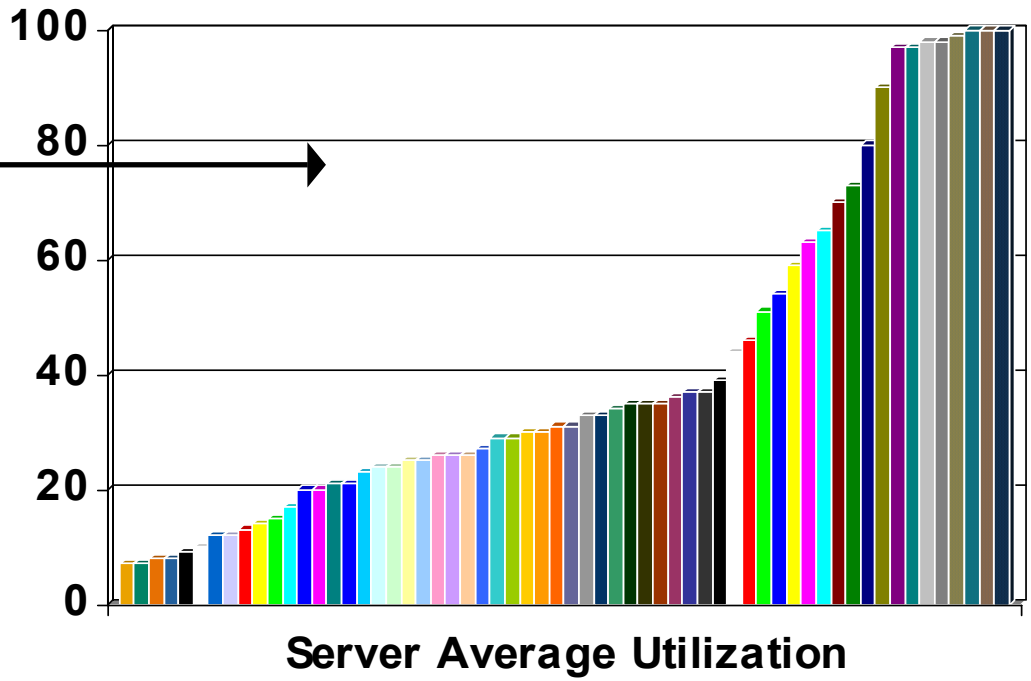
Architecting an Adaptive Infrastructure Solution Using Today's HP-UX Technologies

Dan Herington
WW Technical Program Manager, HP

Why You Need an Adaptive Infrastructure

Tremendous amount of unutilized capacity



Yet these systems are unable to handle the load

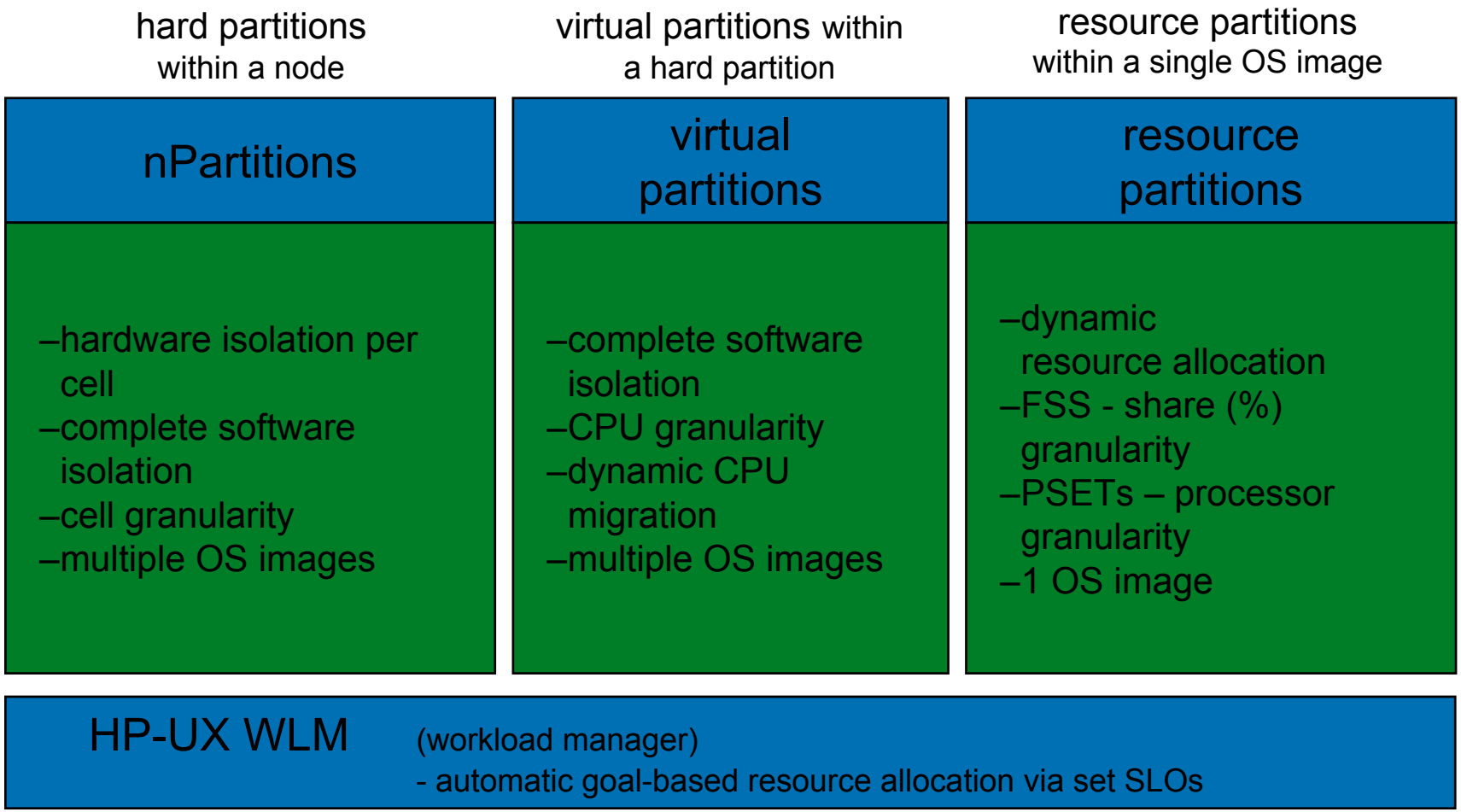
- Overall utilization less than 50%
- Some applications still not able to meet performance requirements

HP-UX Adaptive Infrastructure Technologies

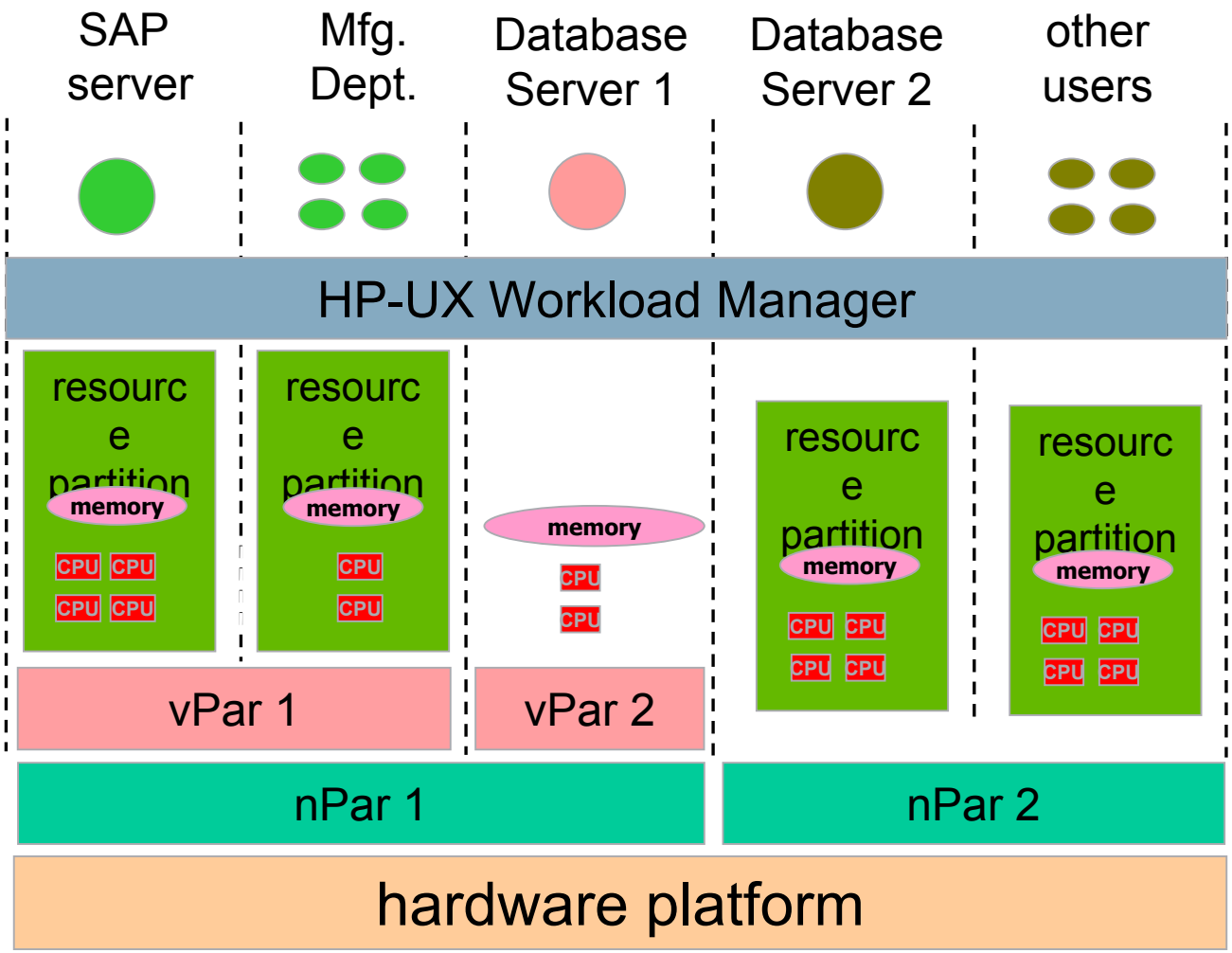
Definition of Partitioning

Partitions are physical or logical mechanisms for
isolating operational environments
within single or multiple
servers to offer the
flexibility of dynamic resizing
while ensuring that
applications can enjoy
protection from
unrelated events
that could otherwise cause disruption,
interruption, or performance degradation.

Adaptive infrastructure on HP-UX
HP's Partitioning Continuum



HP-UX = Broadest Partitioning Portfolio



application space

WLM to automatically resize partitions

Adaptive Resource Partitions inside of an OS image

vPars

nPars

hp 9000

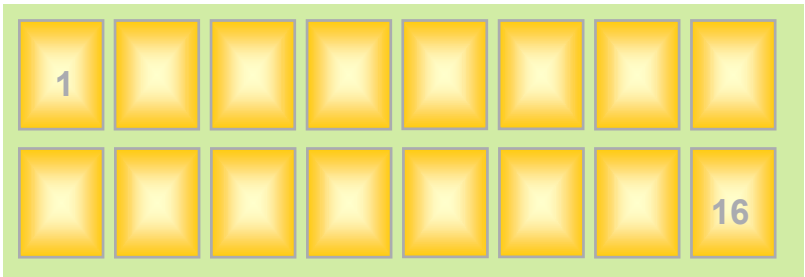
HP-UX On-Demand Technologies

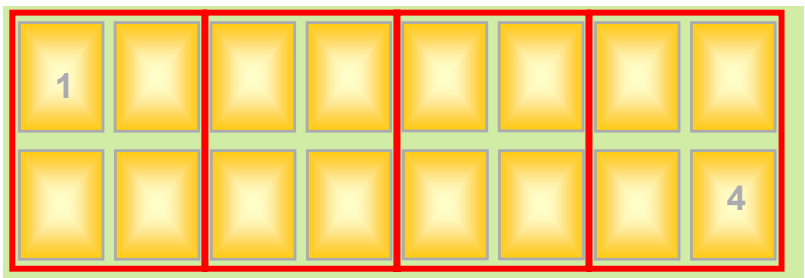
- Instant Capacity on Demand (iCOD)
  - Activate new permanent capacity when needed
- Instant Capacity on Demand – Temporary Capacity (TiCOD)
  - Activate/deactivate new temporary CPU capacity when needed
- Pay-per-Use Utility Computing – (PPU)
  - Lease systems based on CPU utilization

nPars

nPartitions

Multiple applications on the same server with full electrical isolation between partitions

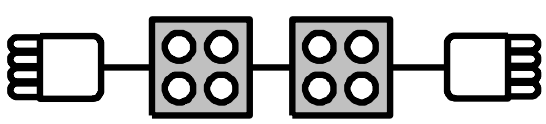




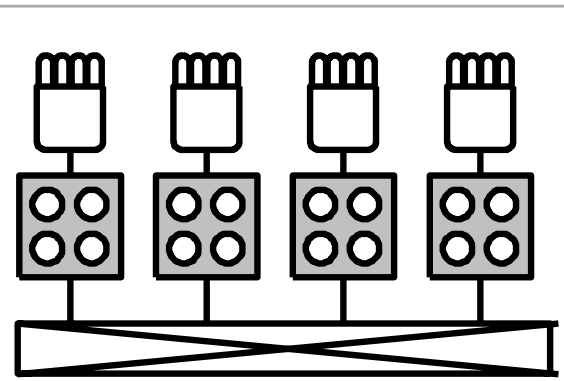
- **Increased system utilization**
  - partitioning Superdome into physical entities: up to16 nPartitions
- **Increased Flexibility: Multi OS**
  - Multi OS support: HP-UX, Linux (*), Windows (*)
  - Multi OS version support
  - Multiple patch level support
- **Increased Uptime**
  - hardware and software isolation across nPartitions
  - MC/ServiceGuard support (within Superdome or to another HP 9000 server)

hp's cellular architecture is very flexible

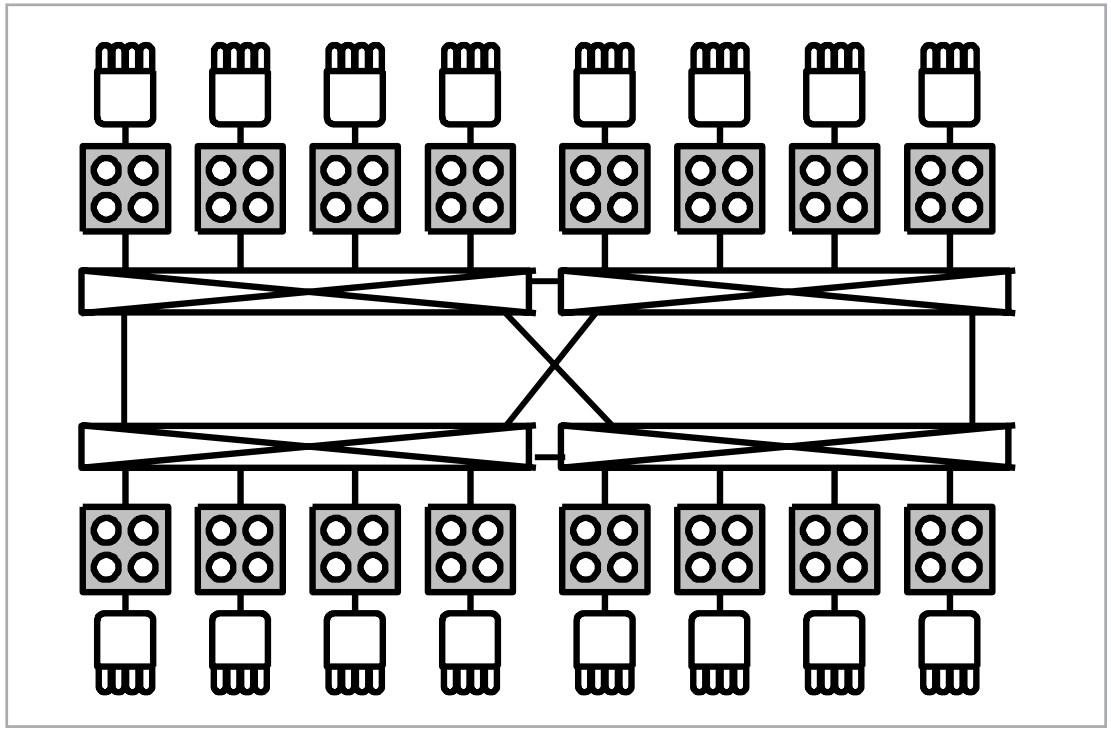
8 Socket system



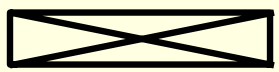
16 Socket system



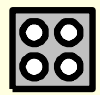
64 Socket system



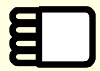
Legend



Two crossbar switches

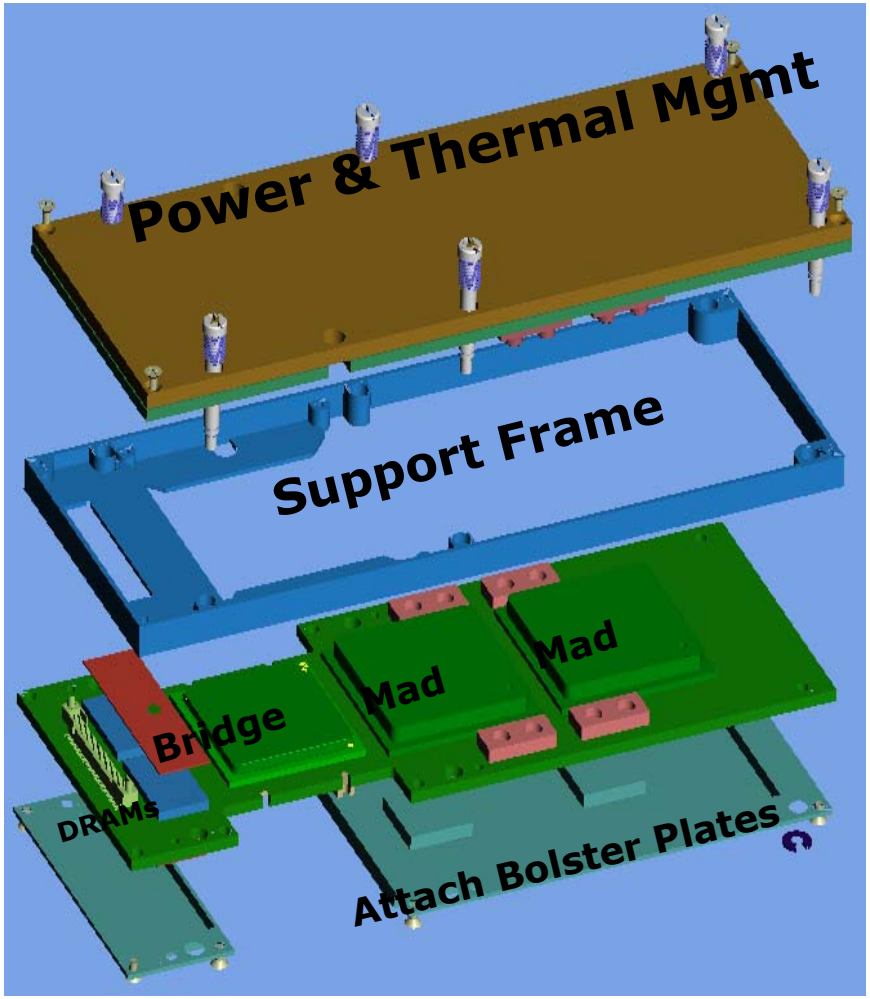


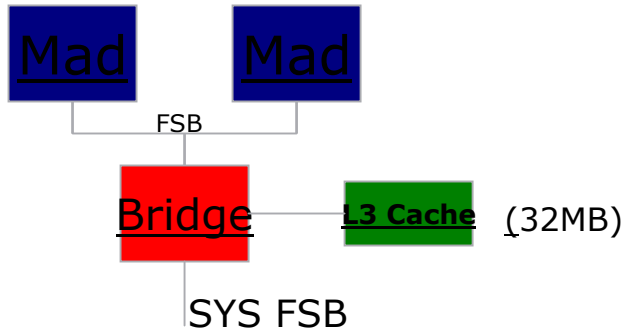
4 Socket Cell



I/O Backplane

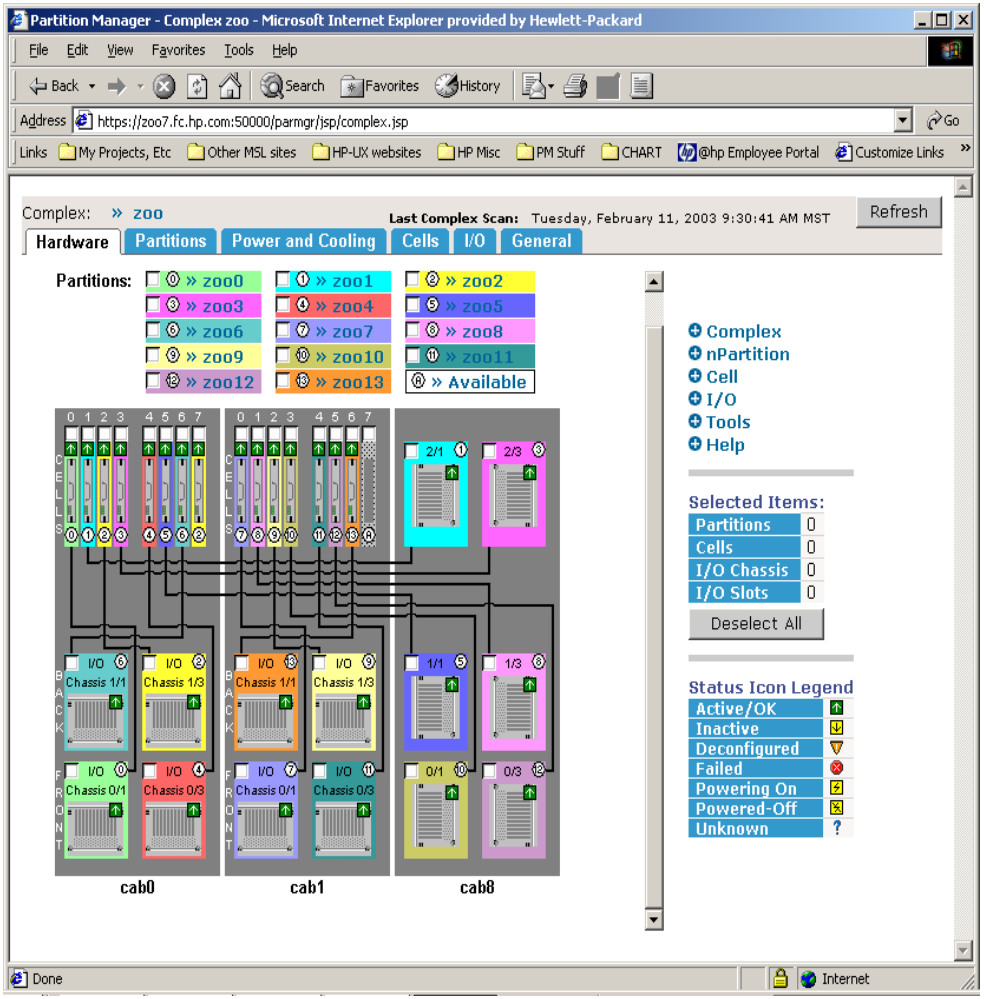
MX2 System Daughtercard





Partition Manager New Features

Significant Changes from ParManager on HP-UX 11i

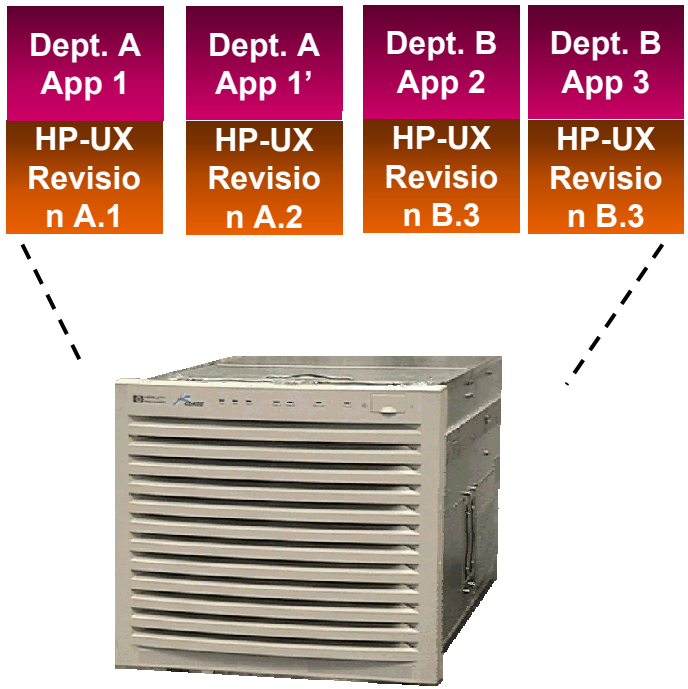


- ✓ New web interface
- ✓ Graphical “big picture” views of
  - nPars
  - Hardware in complex
- ✓ Supports new OS/HW features
  - Cell local memory for HP-UX 11i v.2 partitions
  - nPartition configuration privilege
- ✓ Remote admin of Superdome complex
- ✓ Compatible with iCOD/pay-per-use
- ✓ Native on Windows (2H 2004)

vPars

HP-UX Virtual Partitions

Multiple HP-UX instances running on the same system or in the same nPar



Increased system utilization

- partitioning a single physical server or hard partition into multiple virtual partitions for rp5405,rp5470, rp7400, Superdome, rp8400, rp7410, rp8420, rp7420

Increased Flexibility

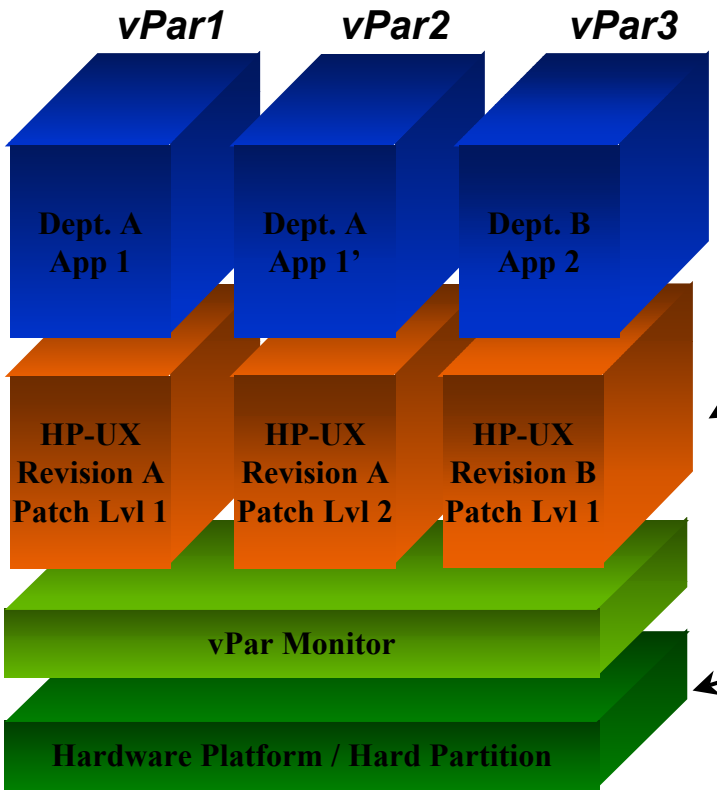
- multiple independent instances of HP-UX
- dynamic CPU migration across virtual partitions

Increased Isolation

- application isolation across virtual partitions
- OS isolation
- individual reconfiguration and reboot

vPars logical overview

- multiple applications or multiple instances or versions of the same application
- provides name space and resource isolation



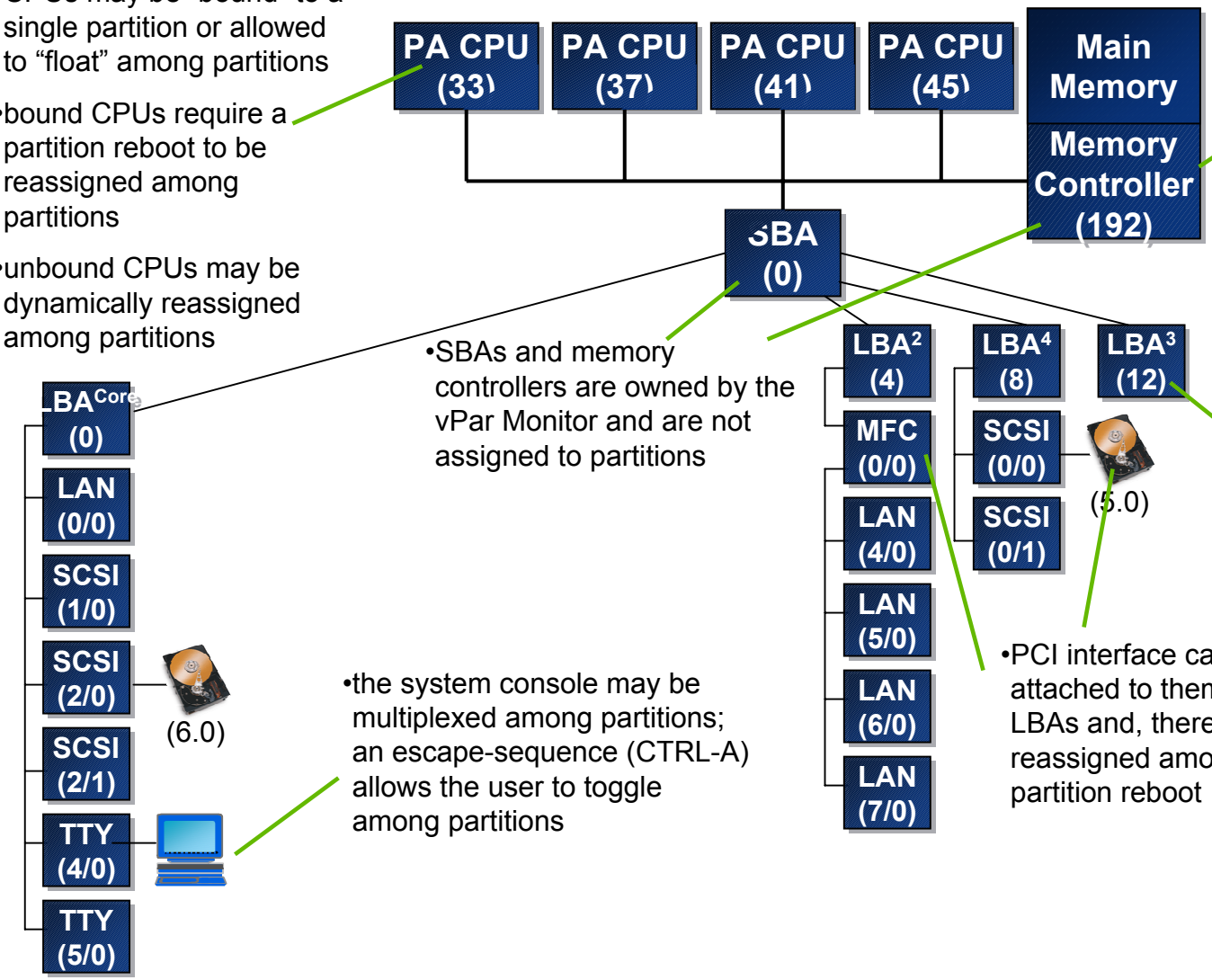
- each operating system instance tailored specifically for the application(s) it hosts
- operating systems instances are given a user-defined portion of the physical resources
- provides name space and resource isolation

- creates illusion of many separate hardware platforms
- manages shared physical resources
- monitors health of operating system instances

- supported on rp5470, rp7400, Superdome, rp8400, rp7410, rp8420, rp7420 systems
- no additional platform support required

Partitionable Resources

- CPUs may be "bound" to a single partition or allowed to "float" among partitions
- bound CPUs require a partition reboot to be reassigned among partitions
- unbound CPUs may be dynamically reassigned among partitions



- main memory is allocated to partitions in multiples of 64MB ranges
- adding or removing memory to or from a partition requires a partition reboot

- SBAs and memory controllers are owned by the vPar Monitor and are not assigned to partitions

- LBAs are bound to a single partition
- adding or removing LBAs to or from a partition requires a partition reboot

- the system console may be multiplexed among partitions; an escape-sequence (CTRL-A) allows the user to toggle among partitions

- PCI interface cards and the devices attached to them are connected through LBAs and, therefore, cannot be logically reassigned among partitions without a partition reboot

Resource Partitions

Resource Partitioning

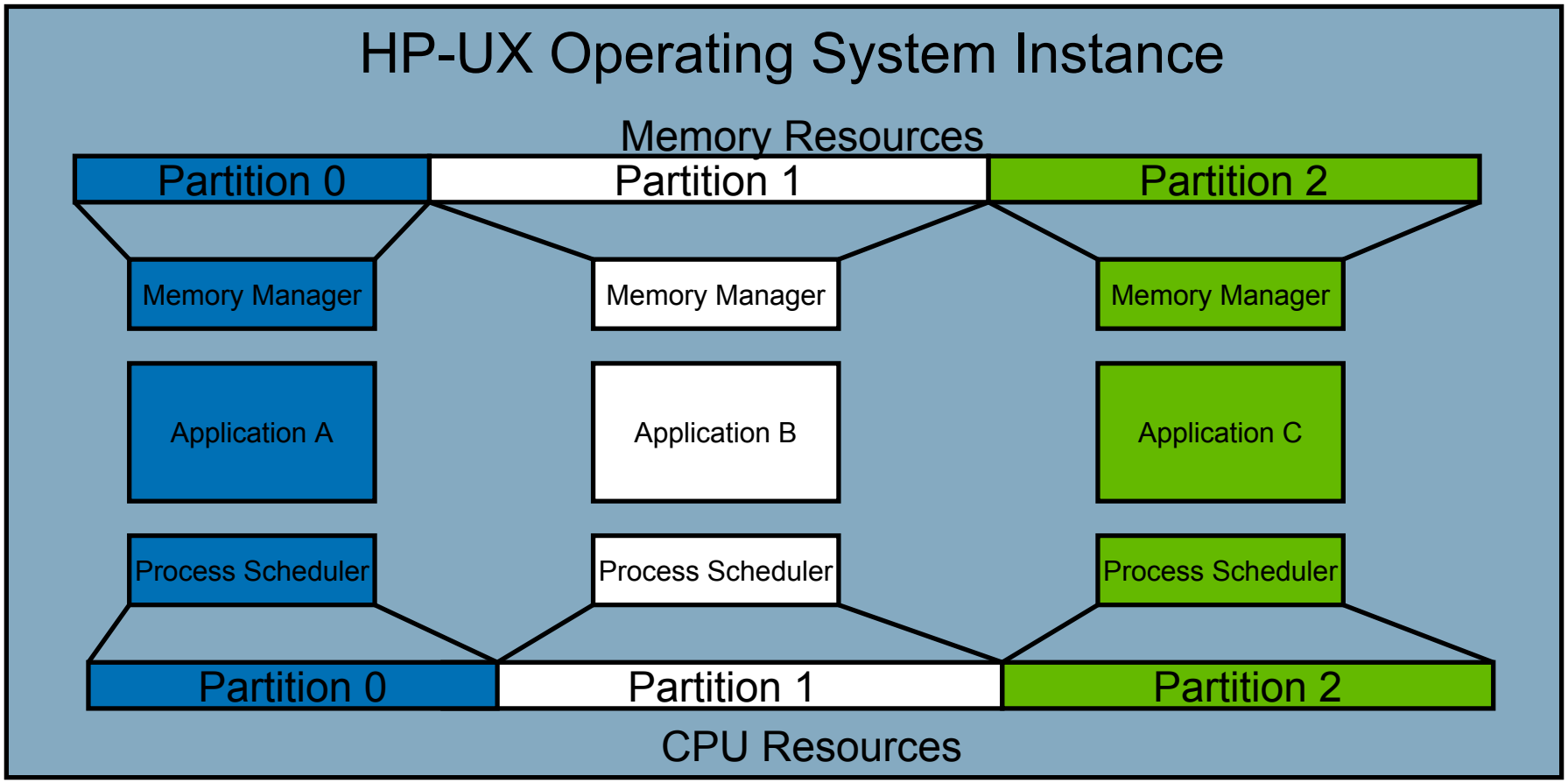
- The Problem:
  - Competition for resources on a consolidated server
- The Solution:
  - Resource Partitioning with Process Resource Manager (PRM)
- PRM is used to configure resource partitions and assign groups of processes to run in each partition

Resource Partitioning Features

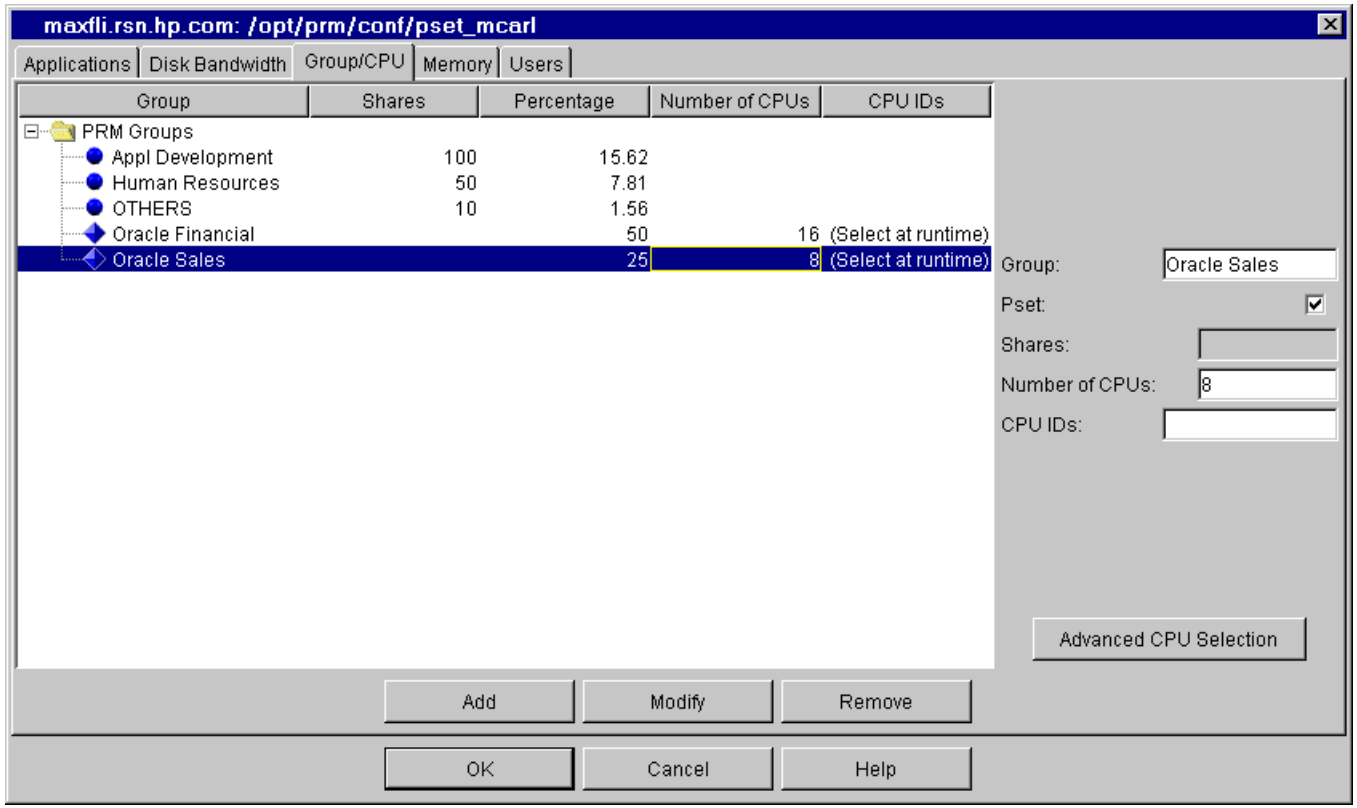
- Supports hierarchical partitions
- Resource controls:
  - CPU Controls
    - CPU allocation by percentage, shares, or whole processors
    - Optional capping in FSS partitions
    - Concurrent FSS and PSETs
  - Real memory controls
    - Each partition gets a separate memory manager in 11i
  - Disk bandwidth
    - Both LVM and Veritas VxVM Volume Groups
  - Automatic process assignment to partition
    - Users/Groups
    - Executable path/Process name
    - Children automatically run with parent by default

Resource Partitions

Apps are running in the same OS, but have separate process schedulers and separate memory managers



PRM GUI



This screenshot shows two PSET groups and 3 fair share groups configured using the PRM Java based GUI

On-Demand

Instant Capacity on Demand (iCOD)

- System acquired with inactive processors
- Processors are paid for when they are activated
  - Price paid is current price when activated
- CPUs can be activated on-line – no reboot required
- Excellent solution for expected growth

- iCOD is licensed for an entire complex
- CPUs can be deallocated in one nPar and activated in another

iCOD Temporary Capacity (TiCOD)

- Alternative purchasing model for iCOD processors
- Temporary Capacity is purchased in 30 Day increments
  - 30 CPU-Days = 43,200 CPU-Minutes
- Any number of iCOD CPUs can be activated
- Activating processors causes the iCOD software to deduct minutes from the “bank”
- Deactivating the iCOD processors stops the deductions

- Excellent solution for:
  - Short term peaks in application load
  - Activation of additional capacity upon failover of a large workload onto a failover server

Pay-per-Use Utility Computing (PPU)

- Type of lease
- Acquire a system with peak capacity required
- Monthly charge based on base payment plus a variable payment based on actual resource usage
- 2 utilization measurement models
  - Active CPU – CPUs are activated/deactivated and variable payment is based on how long CPUs were active
  - Percent Utilization – All CPUs are active and the system is monitored for CPU utilization – variable payment is based on average utilization of all CPUs
- Excellent solution for highly variable loads, especially revenue generating loads because costs are in line with revenues

Workload Manager

Target Problem

- Handling Peaks in Load on Mission Critical Applications

Traditional Approach

- Overprovisioning
  - Lots of dedicated Unix servers
  - Excess capacity on each
  - Gartner states that the average IT organization utilizes their infrastructure at approximately 35% of capacity
- Drawbacks
  - Cost of underutilized capacity
  - Difficult to manage many systems

New Solutions
The Adaptive Infrastructure

- Dynamically reconfigurable partitions
  - nPars with iCOD
  - Virtual Partitions
  - Resource Partitions
- Capacity on Demand
  - iCOD
  - iCOD Temporary Capacity (TiCOD)
  - Pay Per Use (PPU)
- Application Consolidation
  - Run multiple workloads on a single Unix system
- Spare Capacity Consolidation
  - Provide spare capacity for multiple apps on the same system or systems

HP Workload Manager

- HP WLM is a state-of-the-art dynamic workload manager for HP-UX servers
  - It automatically adapts the partition configuration based on the loads on the applications running in those partitions and your business priorities
  - Supports:
    - Resource partitions and vPars
    - Automatic activation/deactivation of iCOD and pay-per-use CPUs
    - Resource partition memory reallocation when workloads are activated/deactivated due to failover or batch job activation
- WLM helps you comfortably increase utilization while still ensuring that your mission critical applications maintain their performance requirements

WLM Service Level Objectives

SLO's use goals, constraints, and conditions.

An SLO consists of:

- **A workload (partition)**
- **Constraints (min, max cpu)**
- **A goal**
- **Priority**
- **Conditions (time of day, event, etc)**

Group A

Min CPU: 20%

Max CPU: 50%

Group A receives 3 shares for each additional user.

Policy applies 9am to 5pm AND

when ServiceGuard Package XYZ

WLM goal types

- Any of the following can be used to allocate resources to a workload:
  - resource utilization
    - CPU entitlement based on utilization of current entitlement
    - Easiest to configure – no data required
  - direct measurement of the performance of the workload
    - response time
    - throughput
  - measurement of load on application
    - number of users/processes
    - queue length

WLM 2.1
Major New Features

- Itanium Support
- Automatic PSET CPU Migration
- BEA Weblogic toolkit to collect load metrics from Weblogic
- Monitoring GUI – graphing of WLM allocation of resources and actual utilization by workloads
- Auditing (billing) utilities – utilities that accumulate the actual usage of resources by each workload over time, csv formatted for upload to your favorite billing package
- Advisory mode - to allow customers to monitor their workloads without turning on WLM controls
- Transient group support – Resource partitions are created when an application starts (eg. on failover, or batch job startup) – ensures resources are not allocated to workloads that are not running

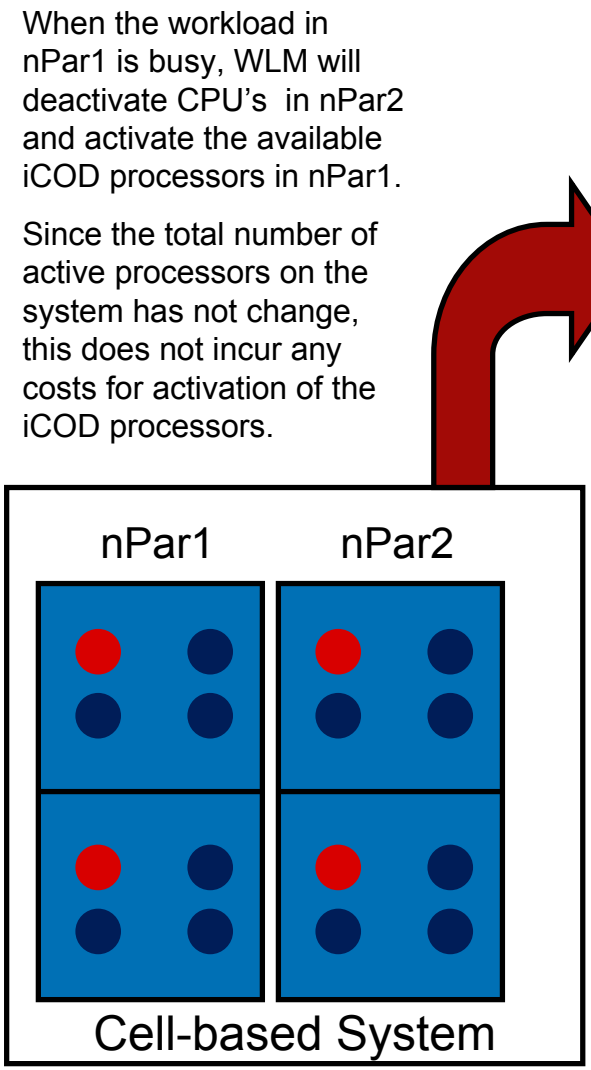
New Features in WLM 2.2

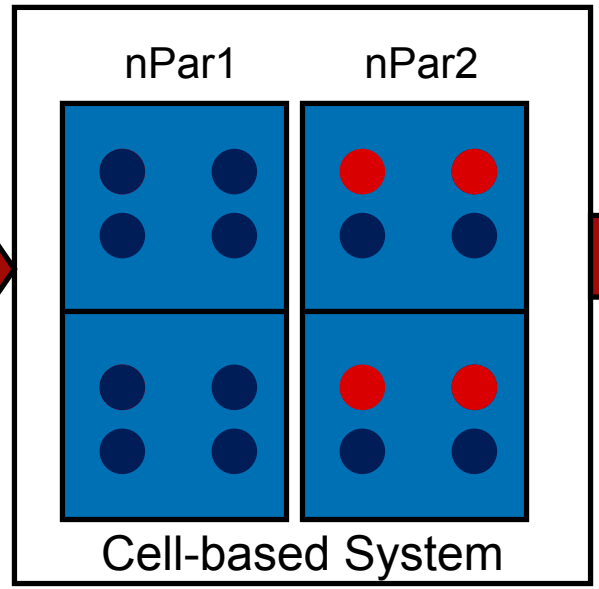
- Support for goal-based CPU allocation across nPars using available iCOD processors
- Support for Temporary iCOD activation/deactivation
- Remote Monitoring GUI
- Remote Configuration GUI
- Support for PSET based transient groups
- Configuration Wizard enhancements

WLM support for Hard Partitions (nPars)

When the workload in nPar1 is busy, WLM will deactivate CPU's in nPar2 and activate the available iCOD processors in nPar1.

Since the total number of active processors on the system has not change, this does not incur any costs for activation of the iCOD processors.

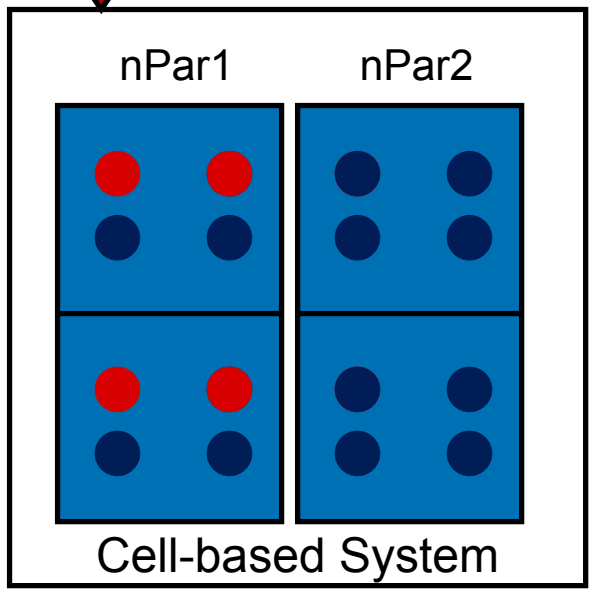




This is a 4-cell 16-CPU server with 2 nPars – each with 2 cells.

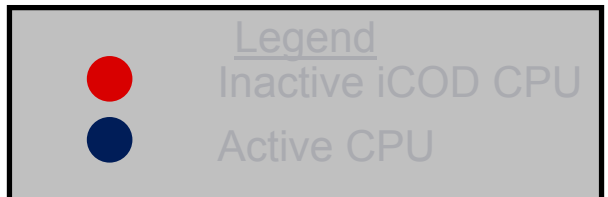
12 active processors on the system and 4 available iCOD processors

WLM will use goal-based Service Level Objectives to determine which nPar the 12 active processors should be running in.

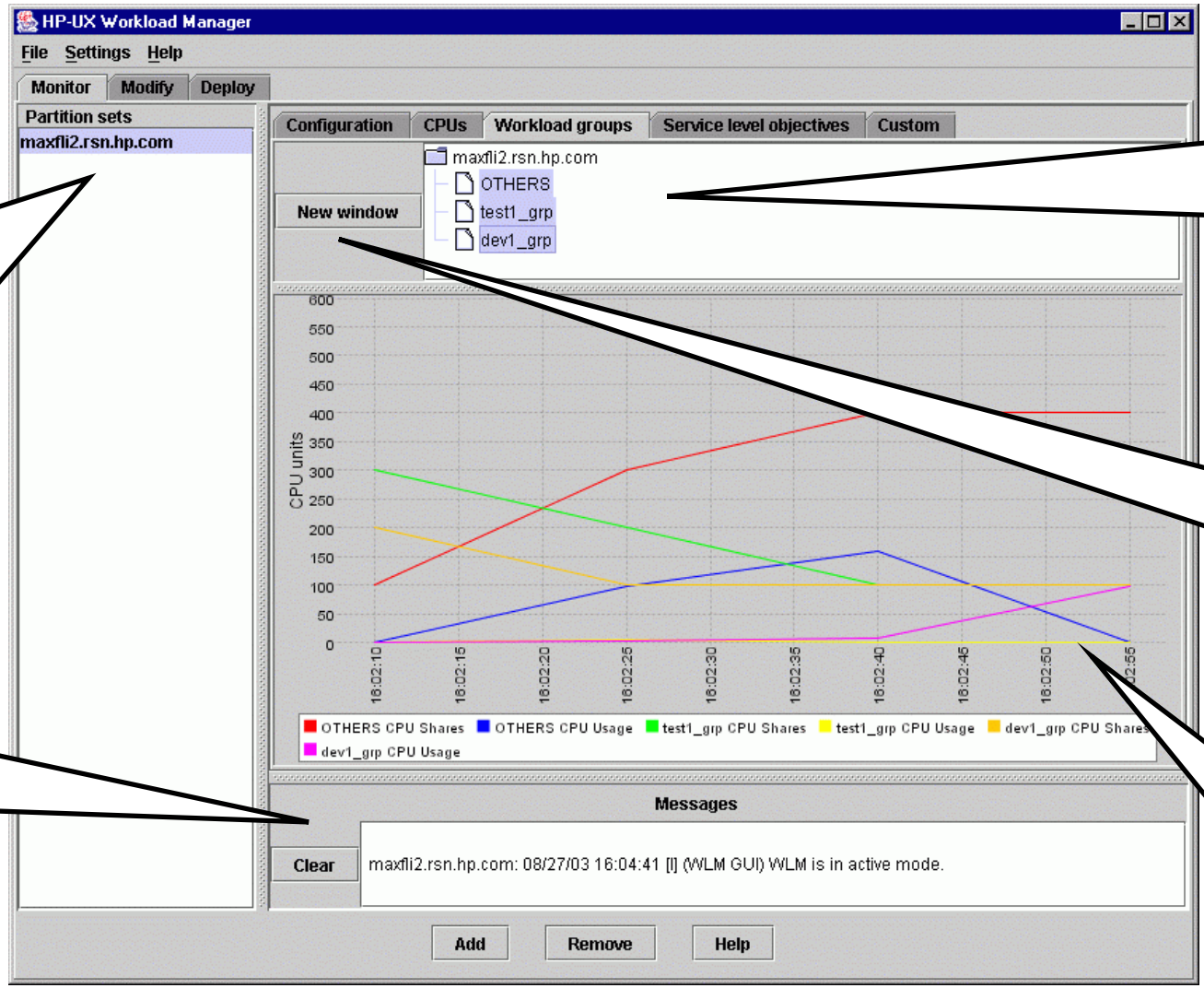


When the workloads in nPar2 get busy, WLM can deactivate CPUs in nPar1 and activate them in nPar2.

This allows each nPar to scale from 4 to 8 CPUs depending on the status of the workloads running in each nPar.



Remote Monitoring GUI



User Defined Sets of HP-UX partitions running WLM

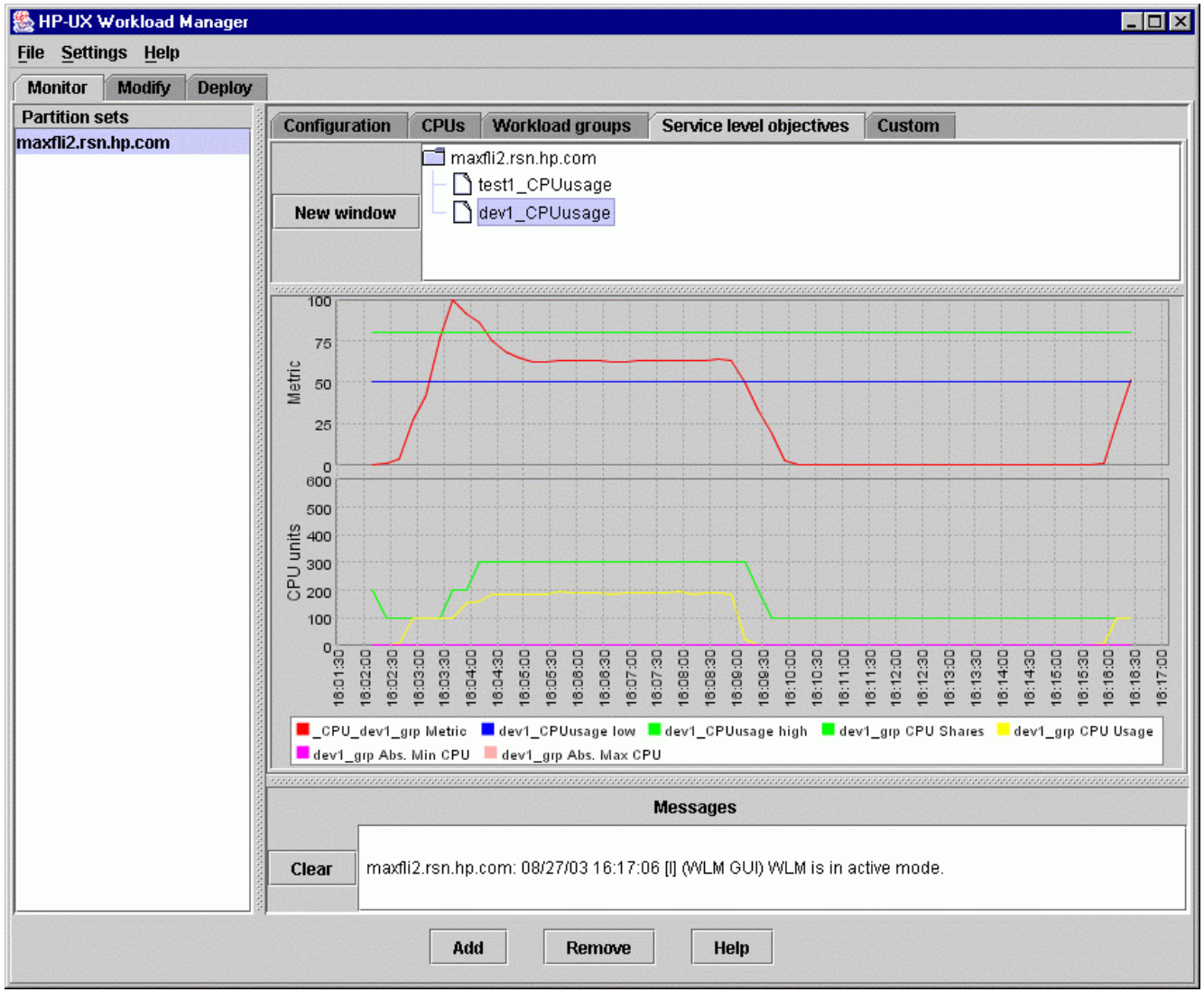
User selected partitions for each graph

Opens the graph in a separate window

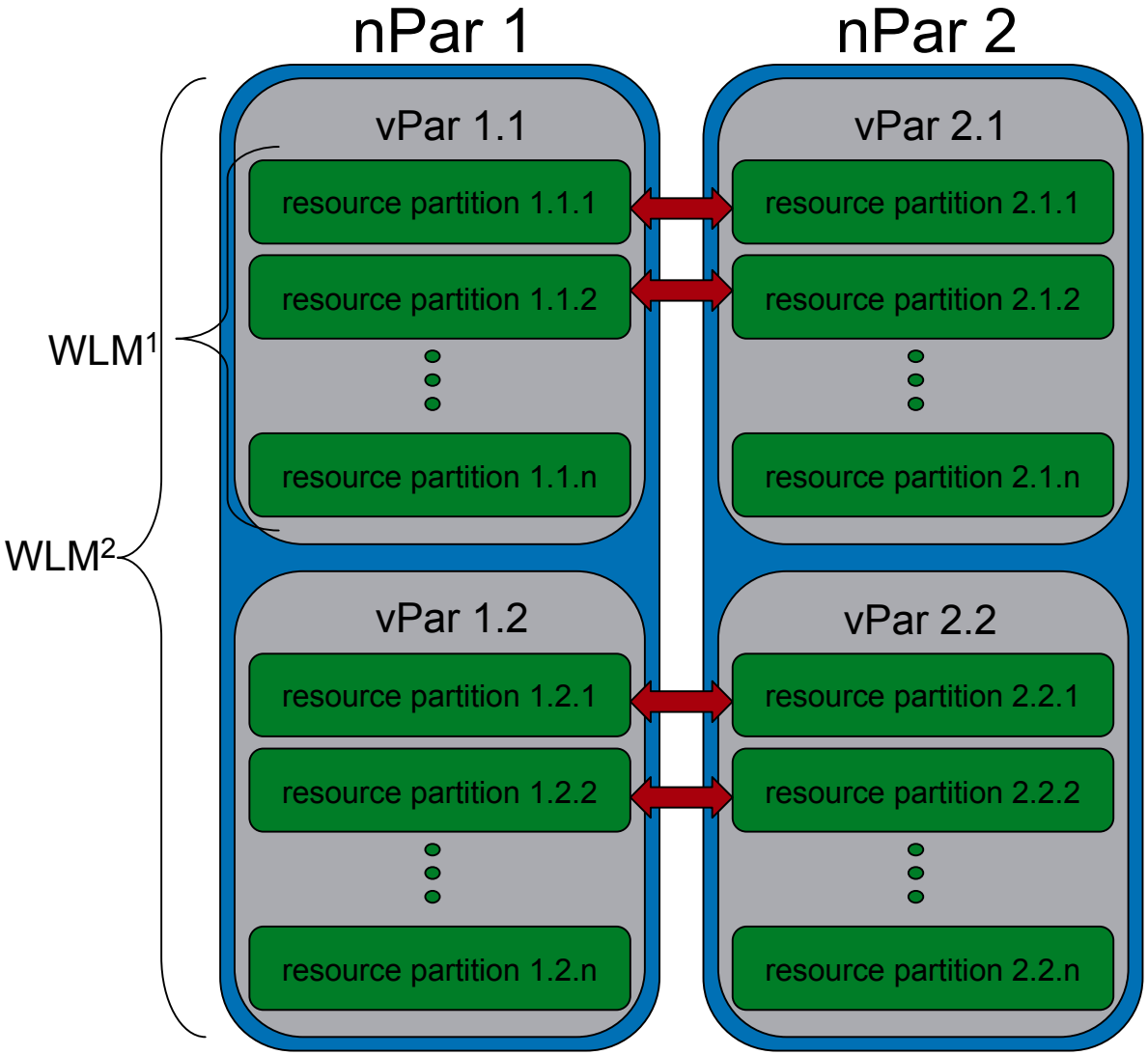
Remote WLM messages displayed here

Graph data displayed here

Remote Monitoring GUI



Resource management of your adaptive infrastructure



- 2 nPars provides
  - hardware fault isolation
- 2 vPars within each nPar provides
  - software fault isolation
  - OS version isolation
- Any number of resource partitions (one for each major application, or group of same priority minor applications) in each vPar provides:
  - resource isolation
- WLM[1] automatically allocates CPU resources as needed to resource partitions
- WLM[2] automatically allocates CPUs as needed to vPars
- Failover across nPar boundary (indicated by ⟷) provides
  - HA for both hardware and software faults
  - WLM will reallocate resources upon failover

Architecting A Solution

Common Benefits of all Partition Types supported at different levels

- maximize system utilization
- resource isolation
- os isolation
- support for full line of HP 9000 servers
- os version support
- ease of setup and management
- flexible CPU resources
- partition stacking
- iCOD, PPU support
- wlm support

Benefits/Strengths

Benefit	nPars	vPars	prm/psets	prm/fss
Maximize system utilization	Good	Better	Better	Best
Resource isolation	Best	Better	Better	Good
Os isolation	Best	Better	No	No
Support for all 9000 servers	sd,8400, 7410	l,n,sd,8400,7 410	All	All
Os version support	11i	11i	11i	10.20, 11.x
Ease of setup	Good	Better	Best	Best
Ease of management/TCO	Good	Better	Best	Best
CPU resource flexibility	Good	Better	Better	Best
iCOD/PPU support	Yes	iCOD/%PPU	Yes	Yes
WLM support	March 04	Yes	Yes	Yes

choosing between partitioning technologies

- nPars
- vPars
- PSET Resource Partitions
- Fair Share Scheduler Resource Partitions

nPars

- nPars is the only partition type that has:

Hardware Fault Isolation

Windows & Linux Support

  - A hardware fault in one partition will not effect the other partitions
  - You can also do hardware maintenance in one partition while the other partitions are running
- Single CPU resource migration is possible if iCOD CPUs are available on the system
  - WLM will automate this in 2.2 (March 04)

vPars

- Why choose vPars over nPars?
  - vPars provides:
    - Dynamic processor movement without rebooting the partition
    - Single cpu granularity without need for iCOD
    - Can run within an nPar
- Why choose vPars over resource partitions?
  - vPars provides:
    - Software fault isolation
    - Different versions of the OS
    - Application isolation

Resource Partitions

- Why choose resource partitions over nPars or vPars?
  - Allows shared I/O – no need to duplicate hardware for each partition
  - Much easier to implement
  - Much lower TCO - single os instance to manage
  - Can run within an nPar and/or a vPar
- PSETs provides:
  - Processor isolation – apps have sole access to processors in the group
  - Memory isolation on top of PSETs
- FSS provides:
  - More granular CPU allocation
  - More partitions

When to use On-Demand Technologies

- iCOD is useful for deferring cost of anticipated growth
  - Resources can be added very quickly
  - Resources can be added while the system is on-line
- TiCOD is useful for short-term spikes in load or for failover server
  - Costs can be managed/budgeted
- PPU is most useful for highly variable loads
  - Particularly for revenue generating workloads because costs vary in line with revenues

WLM

- WLM is NOT a partitioning technology, it provides automatic movement of CPU resources to workloads that need them to meet SLOs
- WLM provides:
  - Automatic CPU resource allocation across Resource Partitions, vPars, and nPars with iCOD
  - Truly maximizes CPU utilization
  - Automatic response to ServiceGuard failovers
  - Guaranteed consistent performance during varying loads on the application
  - iCOD/TiCOD integration
  - Minimizes utility(PPU) computing costs through automatic allocation/de-allocation of utility CPUs

Gotcha's/Incompatibilities

- iCOD/vPars/WLM incompatibility will be removed in March 04.
- vPars does not support Active CPU PPU – this will be resolved in the X.X release of vPars in MONTH of 04.
- PSETs/vPars – vPars CPU migration is NOT supported when PSETs are being used in an affected vPar – this will be resolved in the X.X release of vPars in MONTH of 04.
- WLM 2.2 (March 04) will allow nPar/iCOD migration OR vPar CPU migration OR Auto PPU/TiCOD activation/deactivation – no two will be supported in the same config. This will be resolved in the WLM 2.3 release in September 04.

Key Takeaways

- **All of these options provide the ability to consolidate applications or consolidate data centers and ensure that each app has a minimum amount of resources.**
- **If resource contention is the top issue, resource partitioning is the easiest to set up, the easiest to manage and provides the most flexibility.**
- **If HA is the top issue, nPars provides hardware fault isolation and vPars provides software fault isolation.**
- **If I/O chassis space is limited, resource partitions can be used without requiring duplication of I/O.**
- **If applications don't coexist well on the same OS image, nPars or vPars are the right solution.**
- **If the applications have varying loads and varying priorities, WLM can be used to ensure the resources get used to the best business advantage possible.**
- **Consider using On-Demand technologies (iCOD, TiCOD, PPU) where there are varying loads**


invent