Architecting a Solution with the HP-UX Partitioning Continuum and Workload Manager

The Adaptive Infrastructure for HP-UX

Mark Shaw

Sr. Systems Engineer

High Performance Systems Lab, HP

Author: Dan Herington

HP WORLD 2003

Solutions and Technology Conference & Expo

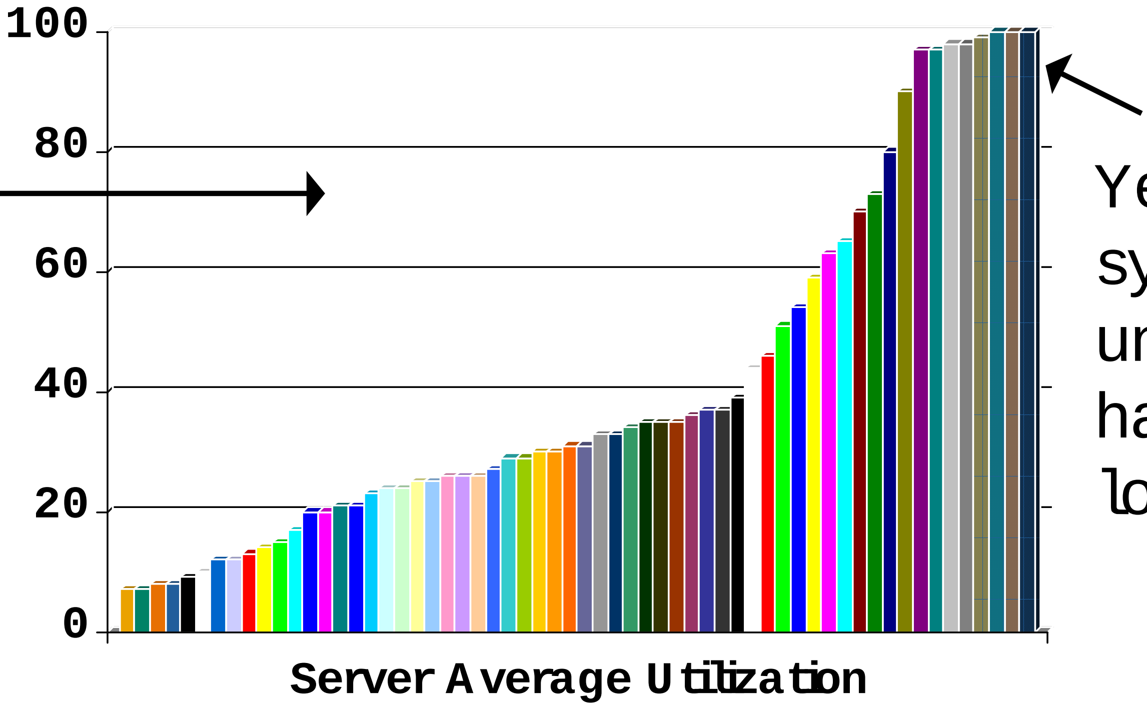
Partitioning Continuum Overview

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.

Why You Need an Adaptive Infrastructure

Tremendous amount of unutilized capacity

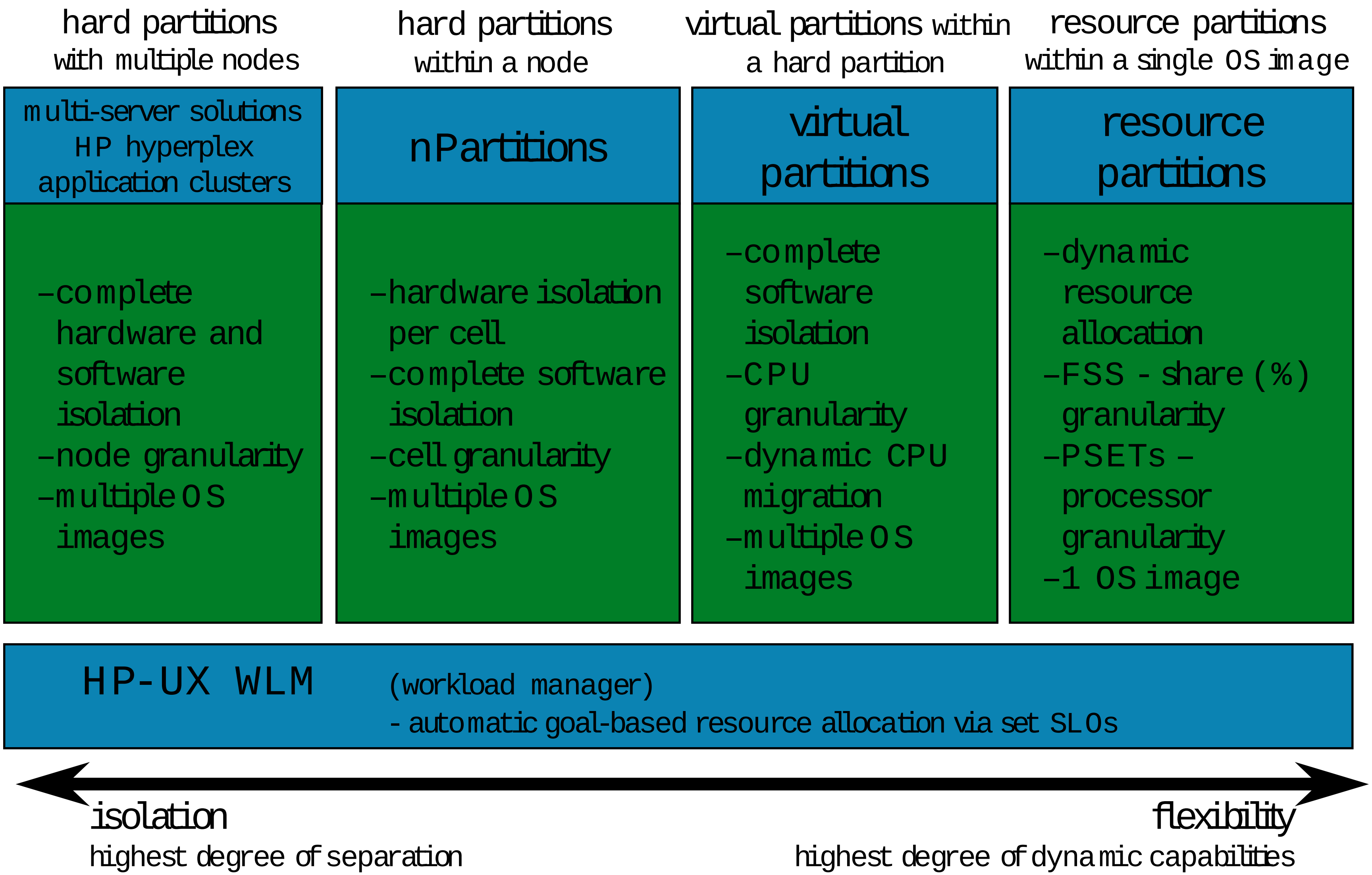


Yet these systems are unable to handle the load

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

Adaptive infrastructure on HP-UX

HP's Partitioning Continuum



hard partitions with multiple nodes

multi-server solutions
HP hyperplex
application clusters

- complete hardware and software isolation
- node granularity
- multiple OS images

hard partitions within a node

nPartitions

- hardware isolation per cell
- complete software isolation
- cell granularity
- multiple OS images

virtual partitions within a hard partition

virtual partitions

- complete software isolation
- CPU granularity
- dynamic CPU migration
- multiple OS images

resource partitions within a single OS image

resource partitions

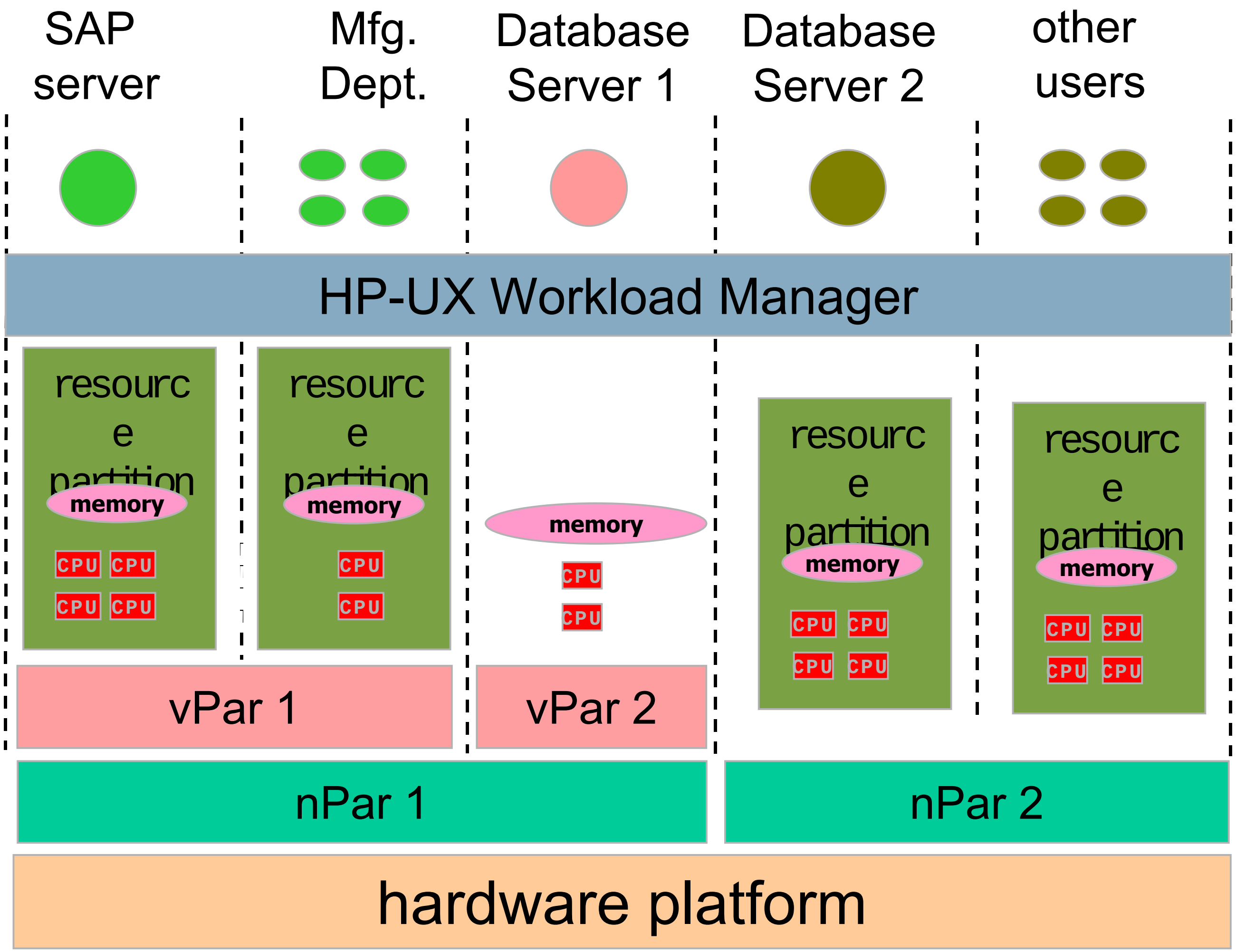
- dynamic resource allocation
- FSS - share (%) granularity
- PSETs – processor granularity
- 1 OS image

HP-UX WLM (workload manager)
- automatic goal-based resource allocation via set SLOs

isolation
highest degree of separation

flexibility
highest degree of dynamic capabilities

HP-UX = Broadest and Most Flexible Partitioning Portfolio



application space

WLM to automatically resize partitions

Adaptive Resource Partitions inside of an OS image

vPars

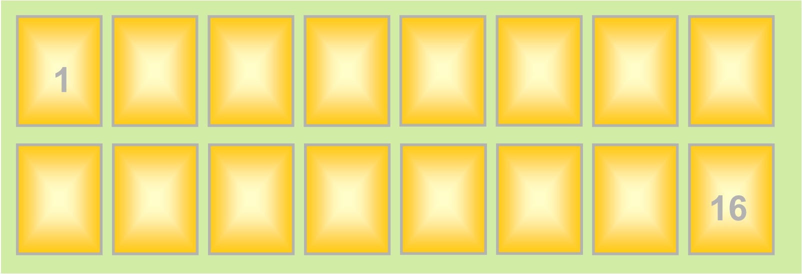
nPars

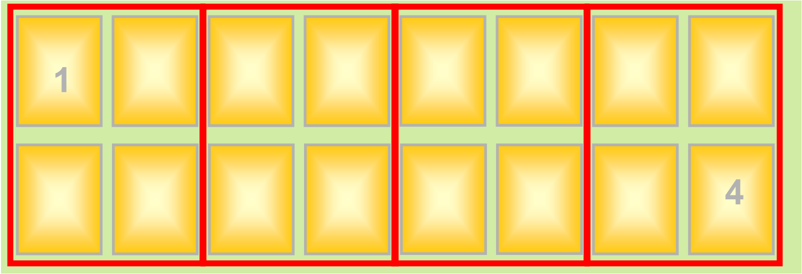
hp 9000

Hard Partitions - 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 Cellular Systems Family

SuperDome
(PA-RISC & IPF)

16 cells, 64 cpu's
512 DIMM slots
192 PCI slots

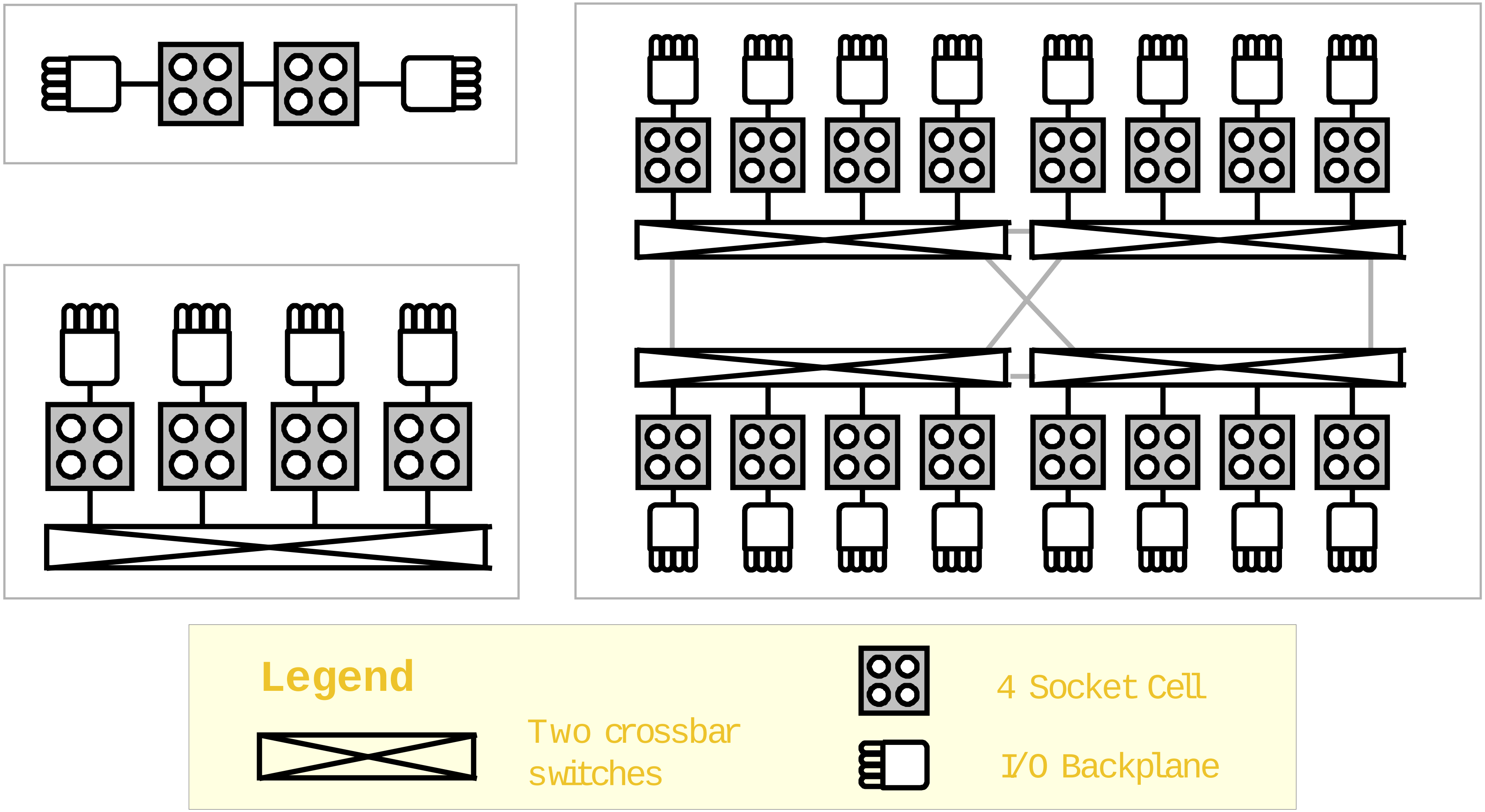
PA-RISC: rp8400
IPF: rx8620

4 cells, 16 cpu's
64 DIMM slots
32 PCI slots,
17U high, 19" rack

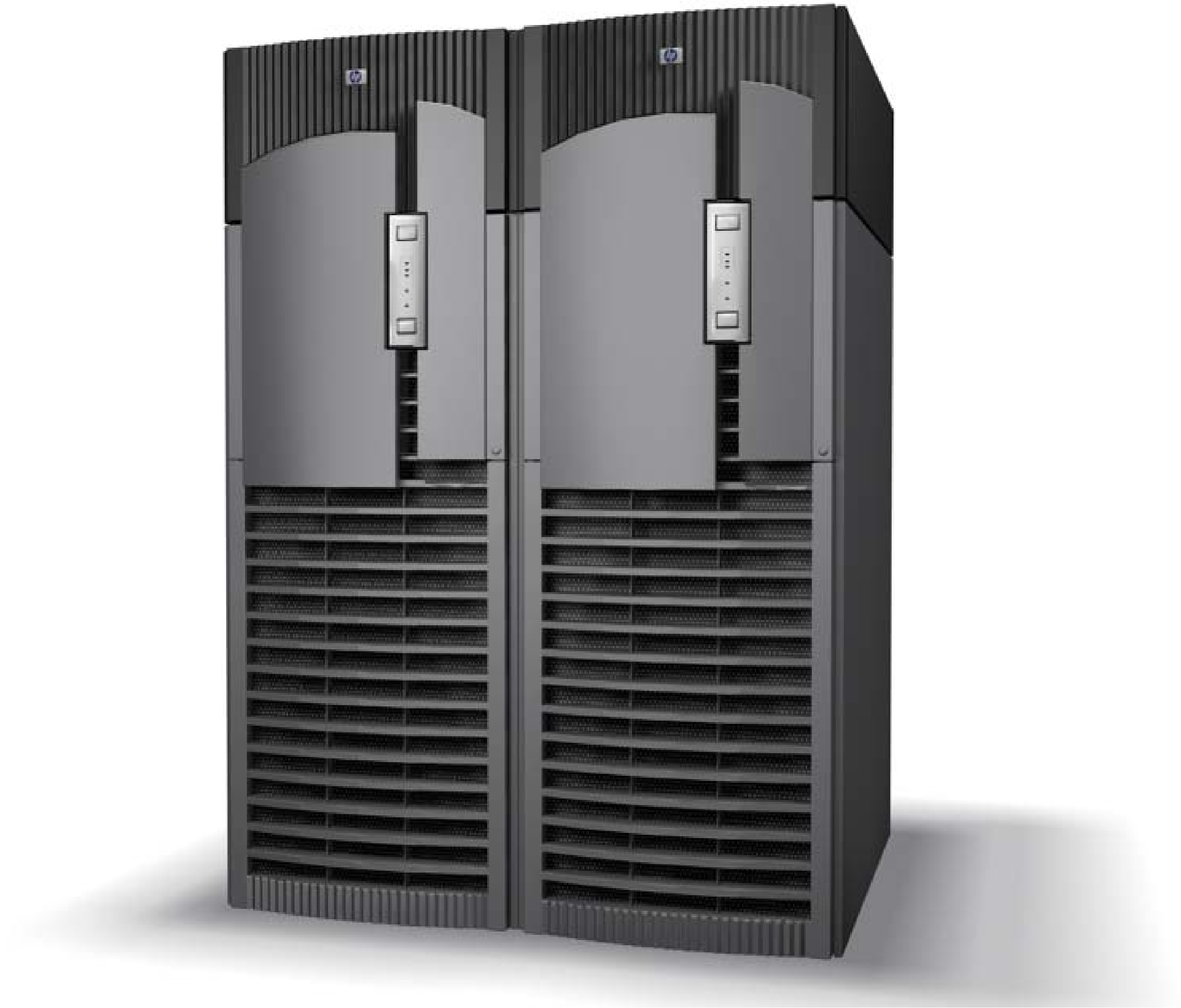
PA-RISC: rp7410
IPF: rx7620

2 cells, 8 cpu's
32 DIMM slots
16 PCI slots,
10U high, 19" rack

hp’s cellular architecture is very flexible



New Superdome industrial design



unprecedented high availability (across Superdome, rp8400, rp7410)

Keep it running

n+1 features (hot swappable)
1. cabinet blowers
2. i/o fans
3. dc power supplies
4. cell backplane dc power supplies

error correction
1. ecc on cpu cache
2. parity protected cpu & i/o links
3. single wire correction on fabric & i/o
4. ecc on all fabric and memory paths
5. Chip-kill memory

redundant ac input power (optional)

Fix it fast

diagnostic features:
1. test station
   - asic level scan tools
   - remotely accessible via lan
2. enhanced predictive support
3. high availability observatory
4. ems monitoring system
5. dynamic processor resilience
6. dynamic memory resilience

fault isolation technologies

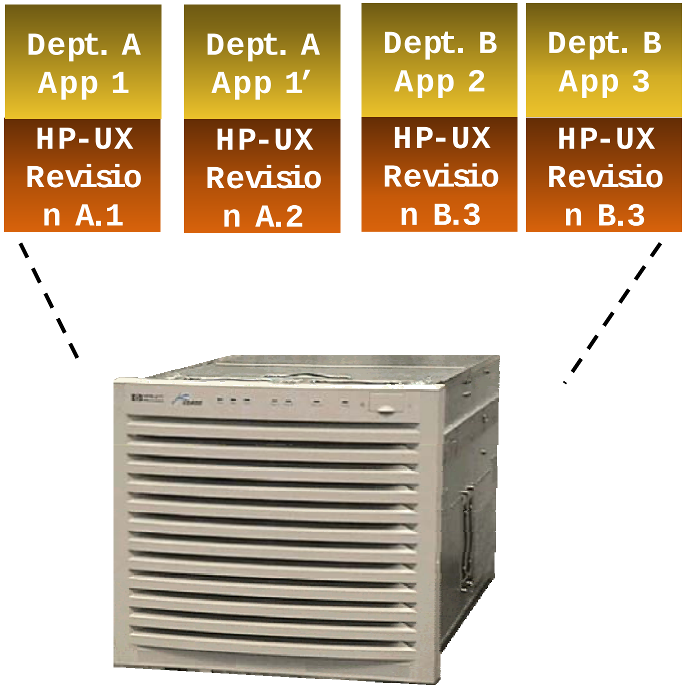
online removal, replacement:
1. cell assemblies*
2. i/o cards

* note: os version dependent

Virtual Partitions - 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 rp5470, rp7400, Superdome, rp8400, rp7410

Increased Flexibility

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

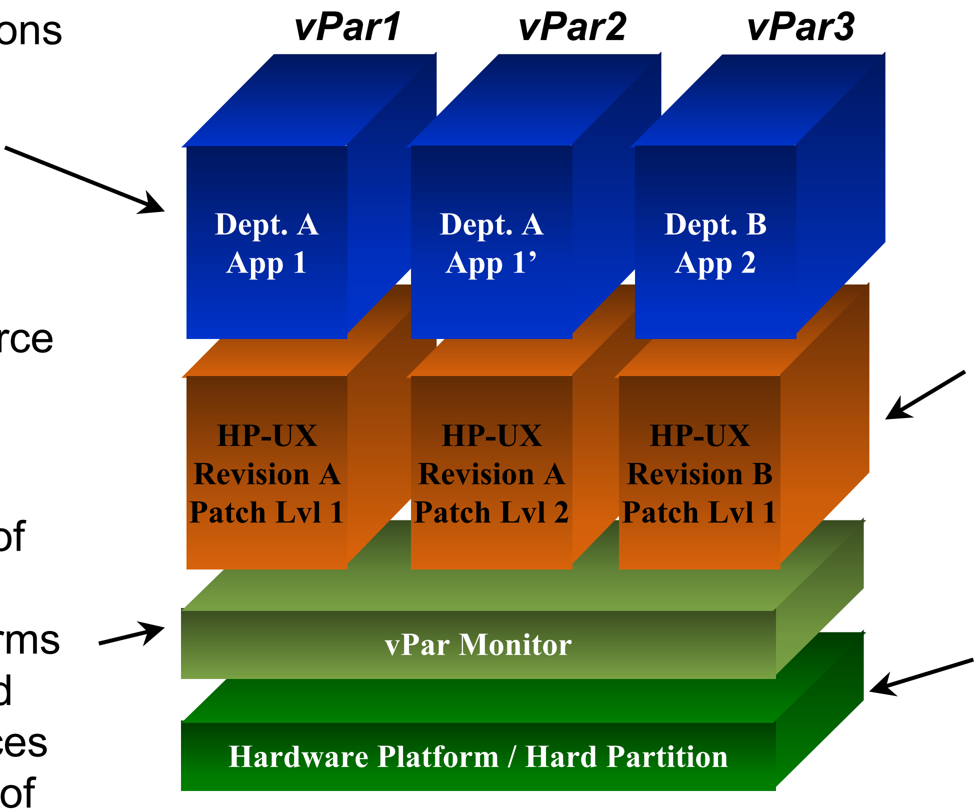
Increased Isolation

- 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

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



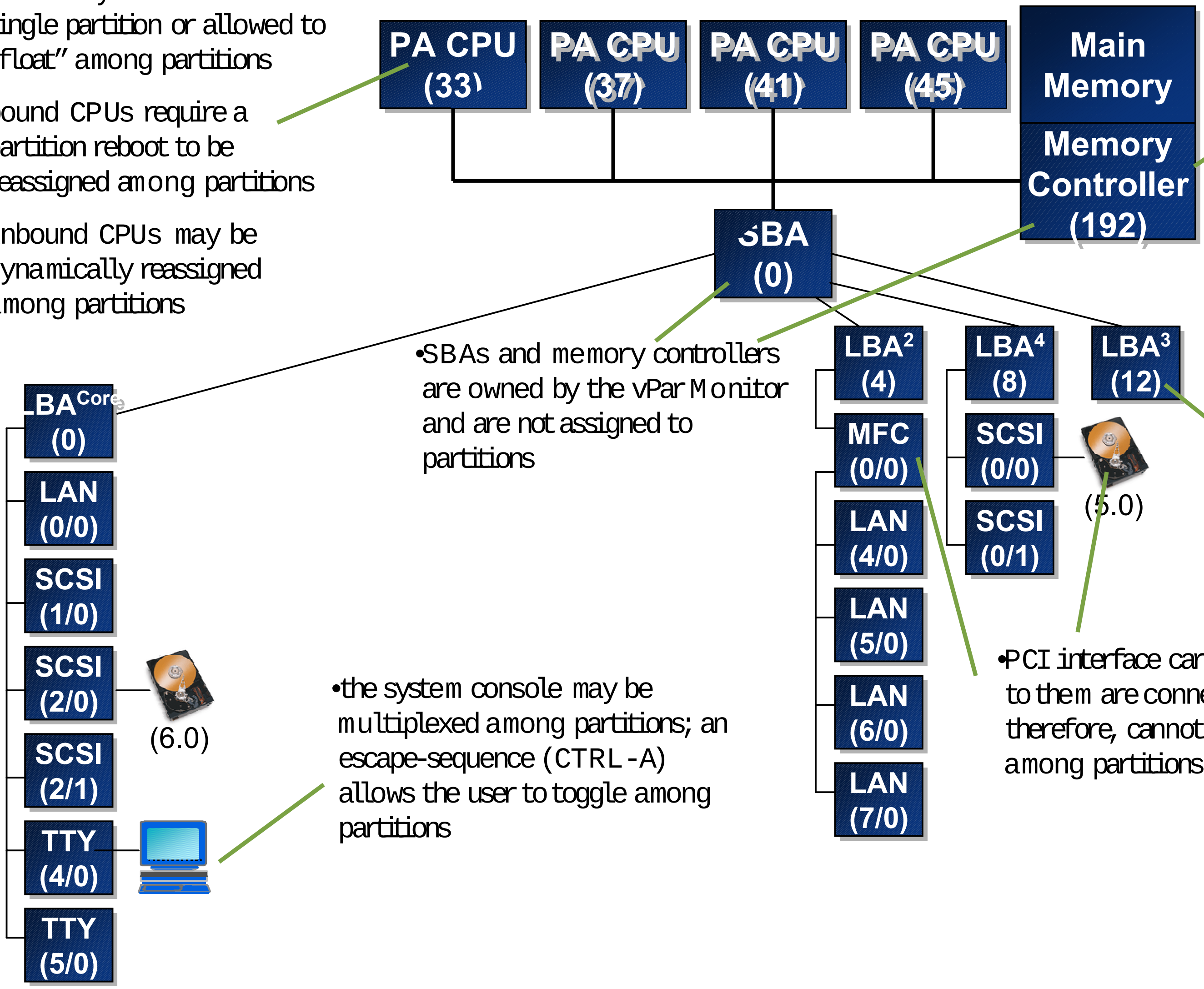
vPar1
vPar2
vPar3
Dept. A App 1
Dept. A App 1’
Dept. B App 2
HP-UX Revision A Patch Lvl 1
HP-UX Revision A Patch Lvl 2
HP-UX Revision B Patch Lvl 1
vPar Monitor
Hardware Platform / Hard Partition

- 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

- supported on rp5470, rp7400, Superdome, rp8400, rp7410 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 - PRM

Resource Partitioning

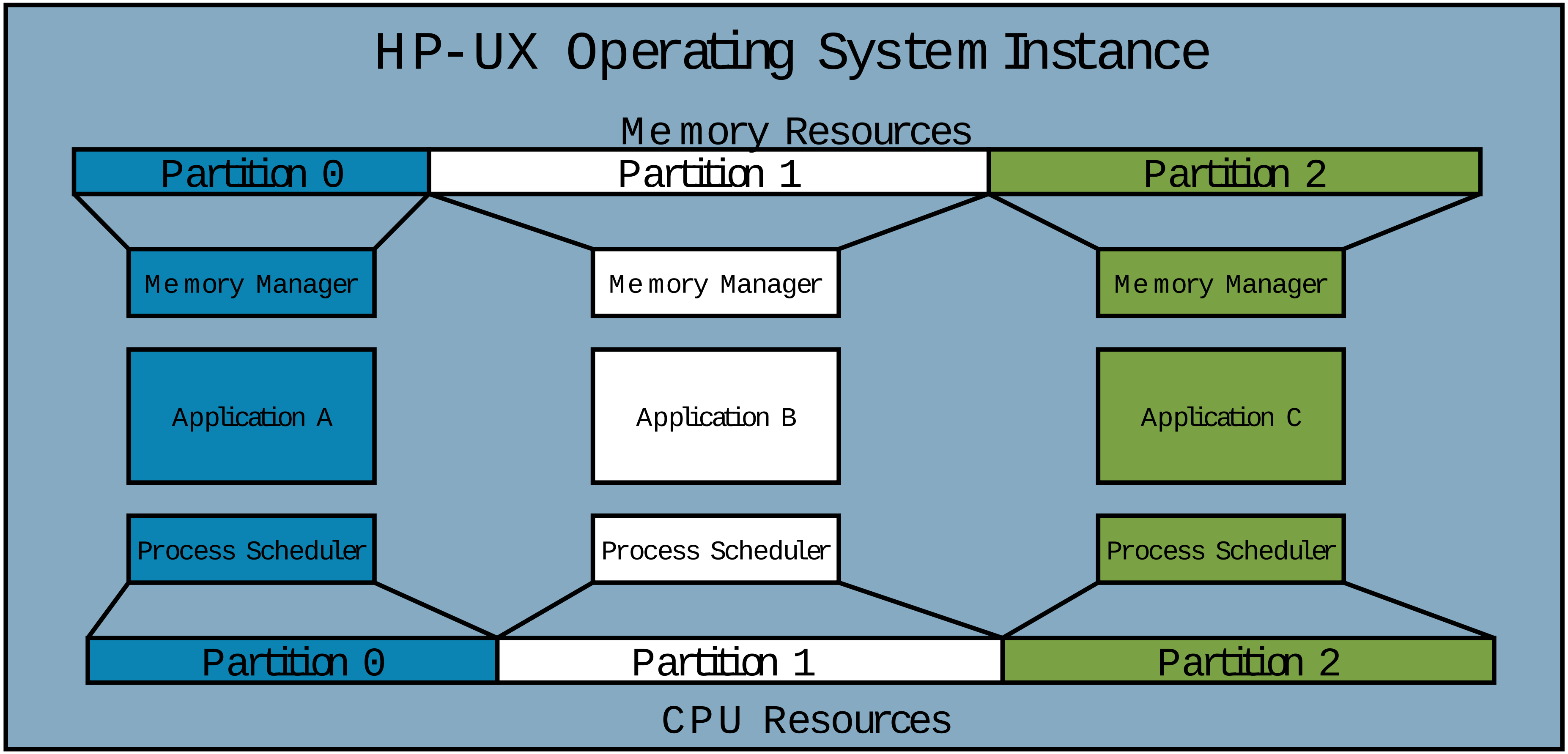
- The Problem:
  - Competition for resources on a consolidated server
- The Solution:
  - Resource Partitioning with Process Resource Manager (PRM)
- PRM is used to configure resource partitions and assign groups of processes to run in each partition

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



FSS CPU Management

- Fair-share scheduler sits on top of standard Unix scheduler in the kernel
- FSS allocates CPU ticks to partitions based on entitlements
- Shares are then allocated using standard scheduling
- Unused CPU cycles are available to other groups (when not capping).
- Supports “capping” mode

FSS CPU Management - Carousel Algorithm

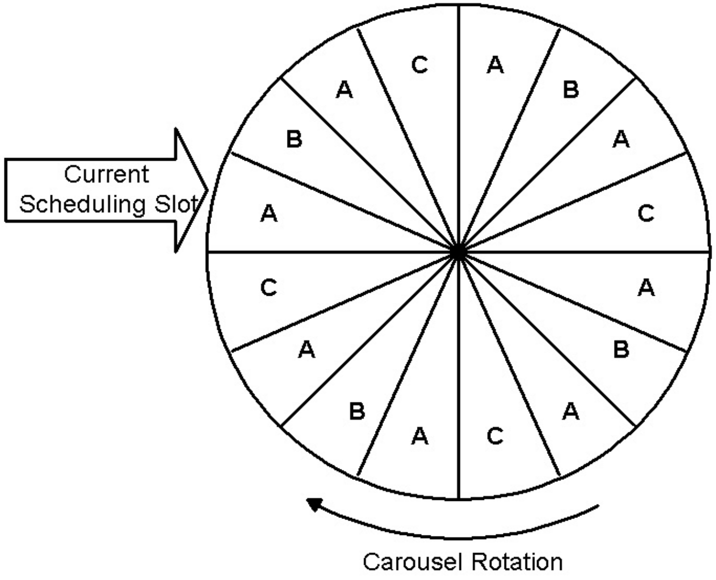
- Shares become slots in a carousel
- Each slot is 10 ms.
- Processes in group get first shot at CPU during their tick
- If capping is off carousel turns if no processes need CPU

Example:

Group A has 8 shares or 50%

Group B has 4 shares or 25%

Group C has 4 shares or 25%



Resource Partitioning with Processor Sets

- PSET is a PRM group type
- CPU is allocated on whole CPU boundaries
- A configuration can contain both FSS and PSET groups
  - The FSS groups run in the default PSET
- Standard Unix scheduler allocates CPU within a PSET
  - Separate process schedulers for each partition, as with FSS groups as well
- PRM uses PSET system calls to manipulate PSETs
- Configuration using configuration file or xprm

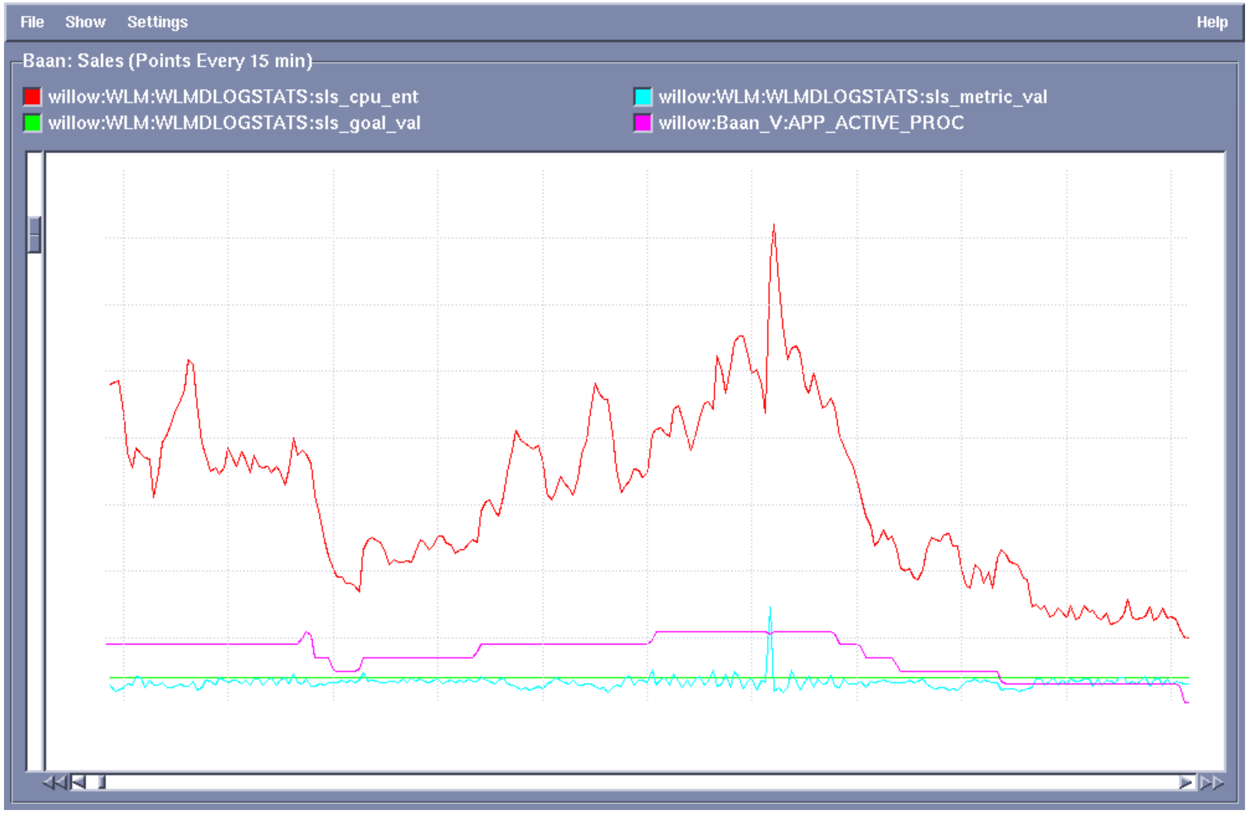
Resource Partition Memory Management in 11i

- Memory Resource Groups (MRGs) are implemented in the 11i kernel
- MRGs are mapped to PRM groups
- Each MRG contains its own copy of a memory subsystem
  - Processes in the group will page if they allocate more memory than their entitlement
- PRM is the only supported interface to MRGs
- Supported on HP-UX 11i and above

Workload Manager

Target Problem

- Handling Peaks in Load on Mission Critical Applications


File
Show
Settings
Help
Baan: Sales (Points Every 15 min)
🟥 willow:WLM:WLMDLOGSTATS:sls_cpu_ent
🟩 willow:WLM:WLMDLOGSTATS:sls_goal_val
🟦 willow:WLM:WLMDLOGSTATS:sls_metric_val
🟪 willow:Baan_V:APP_ACTIVE_PROC

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
  - Virtual Partitions
  - Resource Partitions
- 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
- Capacity on Demand
  - iCOD
  - Pay Per Use (PPU)

HP Workload Manager

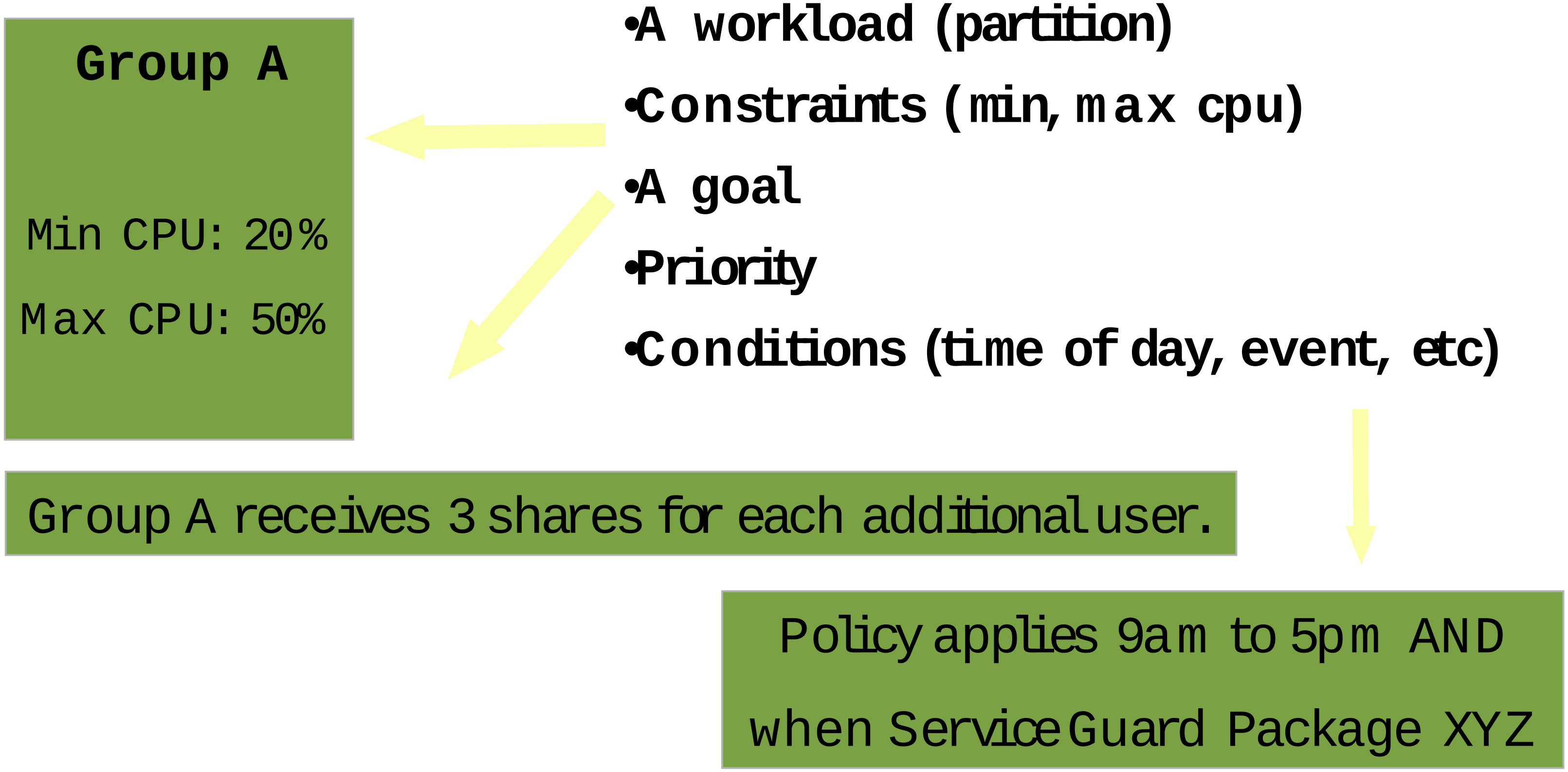
- HP WLM is a state-of-the-art dynamic workload manager for HP-UX servers
  - It automatically adapts the partition configuration based on the loads on the applications running in those partitions and your business priorities
  - Supports:
    - Resource partitions and vPars
    - Automatic activation/deactivation 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
  - direct measurement of the performance of the workload
    - response time
    - throughput
  - measurement of load on application
    - number of users/processes
    - queue length

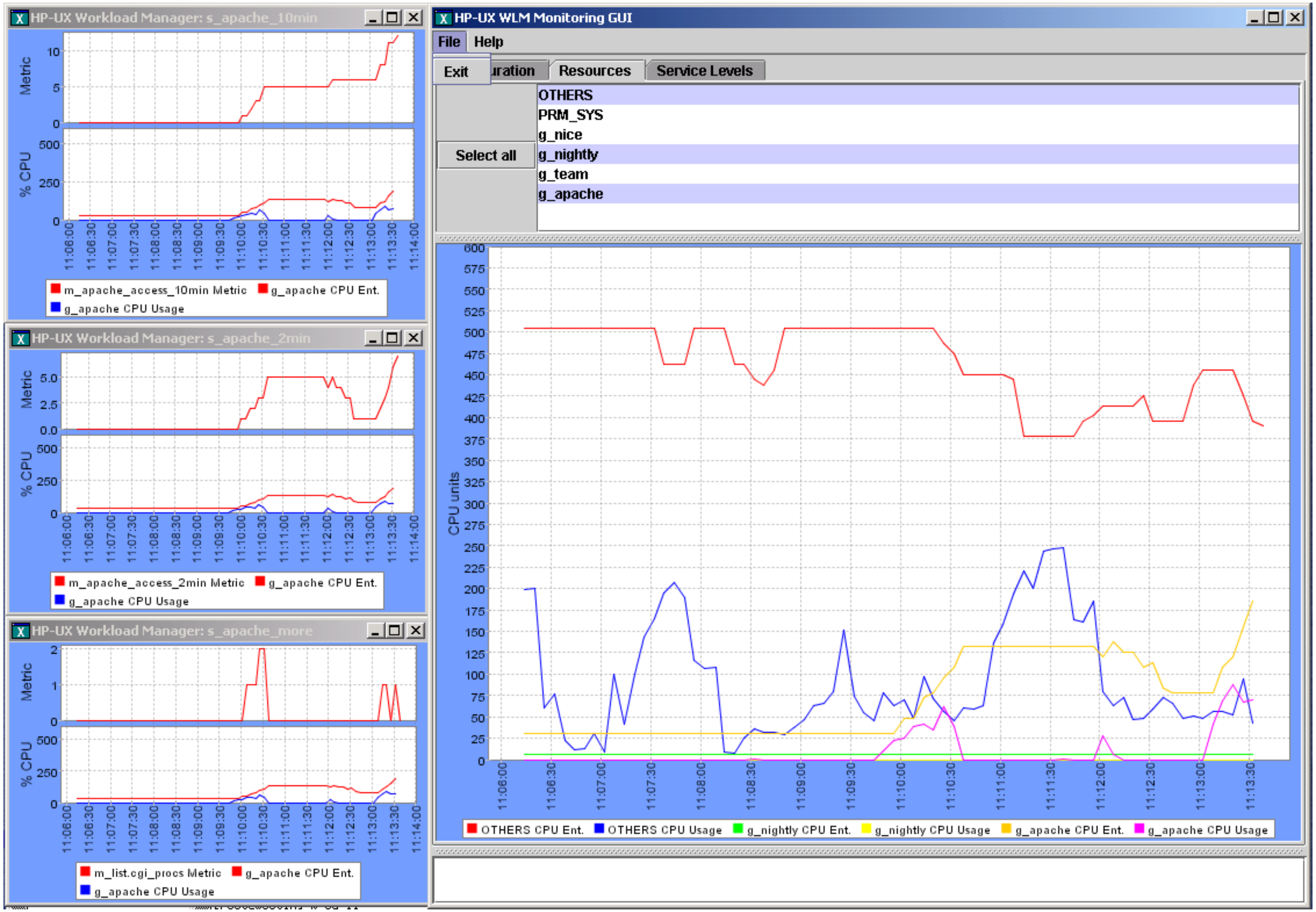
ISV Toolkits

- We are developing toolkits for databases and major applications
- These will allow customers to quickly implement policies most appropriate to each app right out of the box
- Current Toolkits:
  - Oracle Database
  - BEA WebLogic
  - Job Duration – SAS
  - Apache/Java/Oracle Apps
  - SNMP/Pay-Per-Use
  - SAP - Coming

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

Monitoring GUI Dashboard View



Architecting an Adaptive Infrastructure 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, 7410	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 only	Yes	Yes
WLM support	Coming	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

- A hardware fault in one partition will not effect the other partitions even during maintenance
- Consolidating Windows, Linux, and HP-UX on the same platform

vPars

- Why choose vPars over nPars?
  - vPars provides:
    - Dynamic processor movement without rebooting the partition
    - Single cpu granularity
    - 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

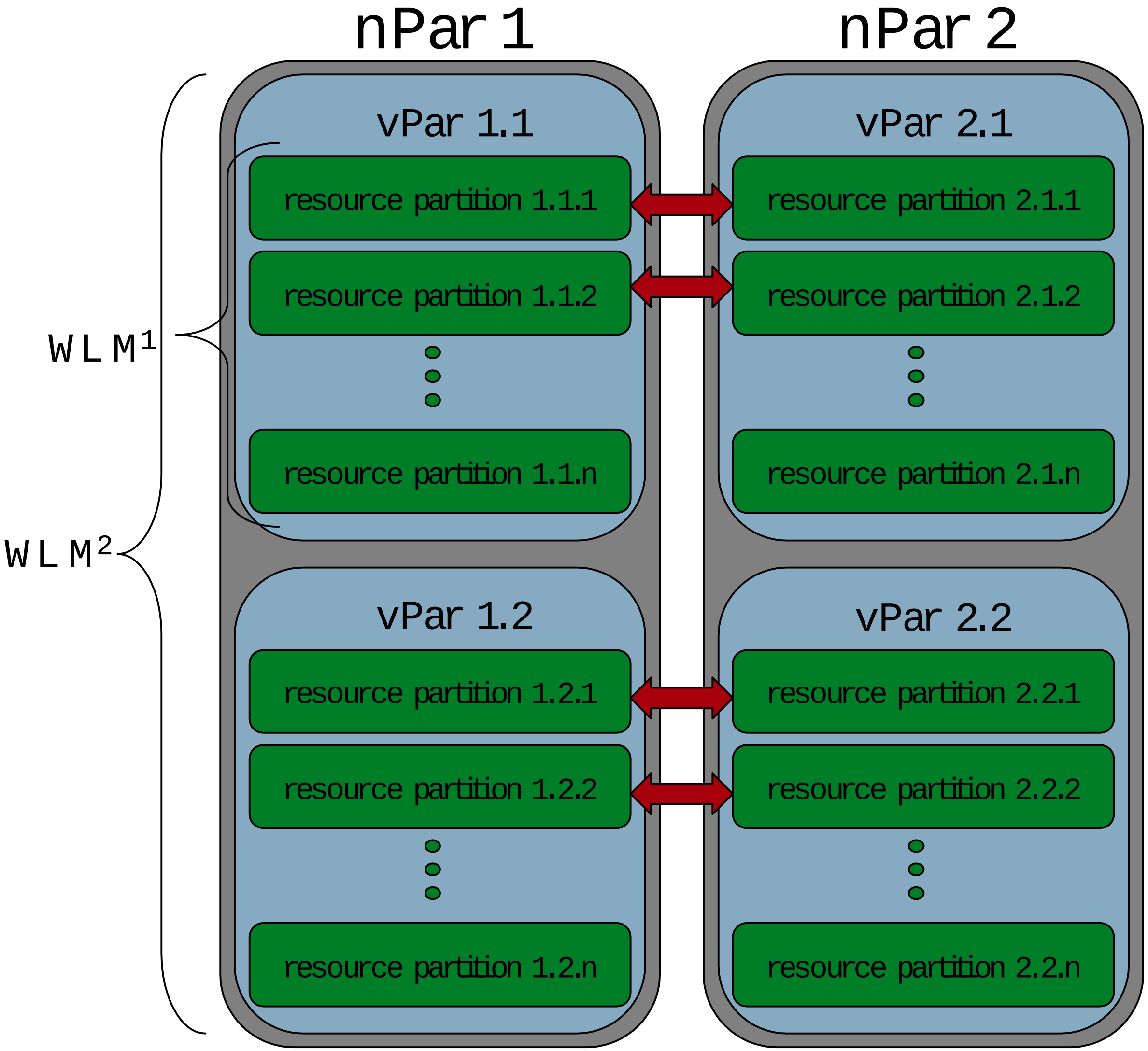
WLM

- WLM is NOT a partitioning technology, it provides automatic movement of CPU resources between partitions to meet SLOs
- WLM provides:
  - Automatic CPU resource allocation to meet SLOs
  - Truly maximizes CPU utilization
  - Automatic response to ServiceGuard failovers
  - Guaranteed consistent performance during varying loads on the application
  - iCOD integration
  - Minimizes utility(PPU) computing costs through automatic allocation/de-allocation of utility CPUs

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.
- If consolidating Windows, HP-UX, and Linux, use nPars.

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

WLM and Utility Computing

- WLM can minimize utility computing costs
  - Utility CPUs are turned on only when needed to meet service level objectives
  - Utility CPUs are turned off as soon as they are no longer needed

Case Studies

Wachovia Bank

WACHOVIA

“HP-UX Workload Manager helped us to maximize system utilization and reduce our capacity management costs in our consolidated Oracle Hotel environment.

In addition, the integration of HP-UX Workload Manager with MC/ServiceGuard helped us to prioritize production database instances on failover."

Tommy Simmons
UNIX Team Leader
Wachovia Bank

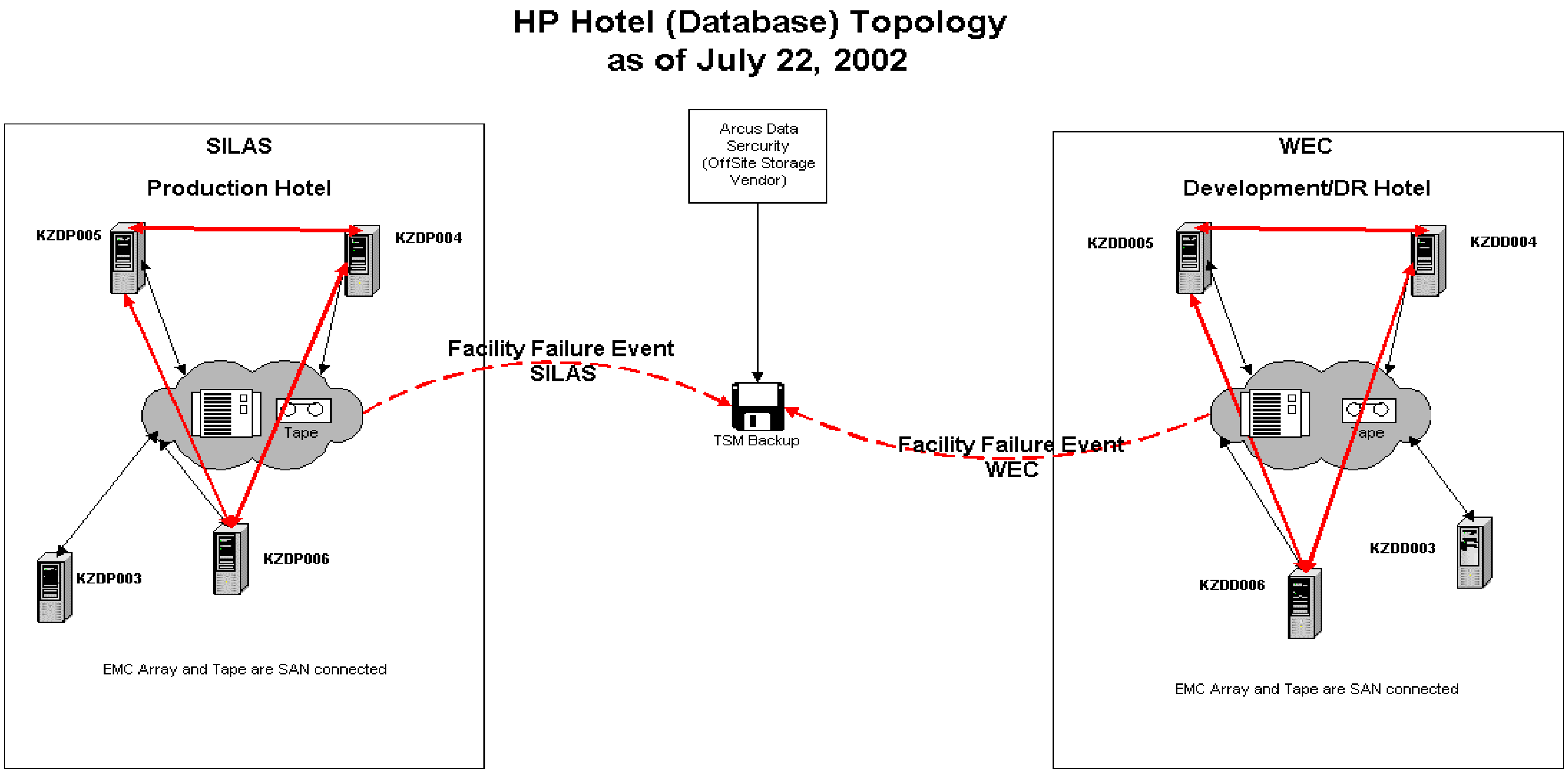
Wachovia: Why Consolidate?

- Decrease Time to Market
- Reduce Support Costs
  - Hardware
  - Software
  - Human Resources
- Standardization of Systems
- Expansion of System Unlimited
- Provides for High Availability and Disaster Recovery

Policies and Procedures

- Elements to Define
  - Standard Installation Components
  - Upgrade/Migration Policy
  - Desupported Software Policy
  - Standards for Backup/Recovery and DR
  - Resource Allocation
  - Penalty for Non-Compliance

Wachovia Architecture


HP Hotel (Database) Topology
as of July 22, 2002
SILAS
Production Hotel
KZDP005
KZDP004
Tape
KZDP003
KZDP006
EMC Array and Tape are SAN connected
Arcus Data Sercurity (OffSite Storage Vendor)
Facility Failure Event SILAS
TSM Backup
Facility Failure Event WEC
WEC
Development/DR Hotel
KZDD005
KZDD004
Tape
KZDD003
KZDD006
EMC Array and Tape are SAN connected

Hotel Room:

Disk: 18GB
SGA:250MB

Red Line denote hardware or package failover

Black lines to servers from EMC denote fibre connection

All Servers except: Hewlett-Packard N4000-750 6 (750mHz) processors with HP-UX 11i (64-bit) and 16.0gb RAM

kzdp003/kzdd003: Hewlett-Packard N4000 (750mHz) processors with HP-UX 11.0 (64-bit) and 16.0gb RAM. These will not be HA and will be used for obsolete versions of Oracle in which the customer can not upgrade.

Case study: large financial institution

- 29,000 employees in 40 countries
- 82 year history
- Earned nearly $1.8 billion in 2001

Customer requirements

- Simplify deployment of new WebLogic instances
- Minimize systems management overhead
- Increase utilization by providing dynamic allocation of resources as loads fluctuate on these web applications

Constraints on the solution

- Single JVM runs better with dedicated CPUs on a large SMP

Case study: large financial institution

Solution architecture

- Running multiple WebLogic instances in a single OS instance simplifies deployment of new instances while reducing the management overhead of multiple OS instances that would be required with vPars
- Resource partitions with PSETs provides isolation of a set of whole CPUs to individual WebLogic instances

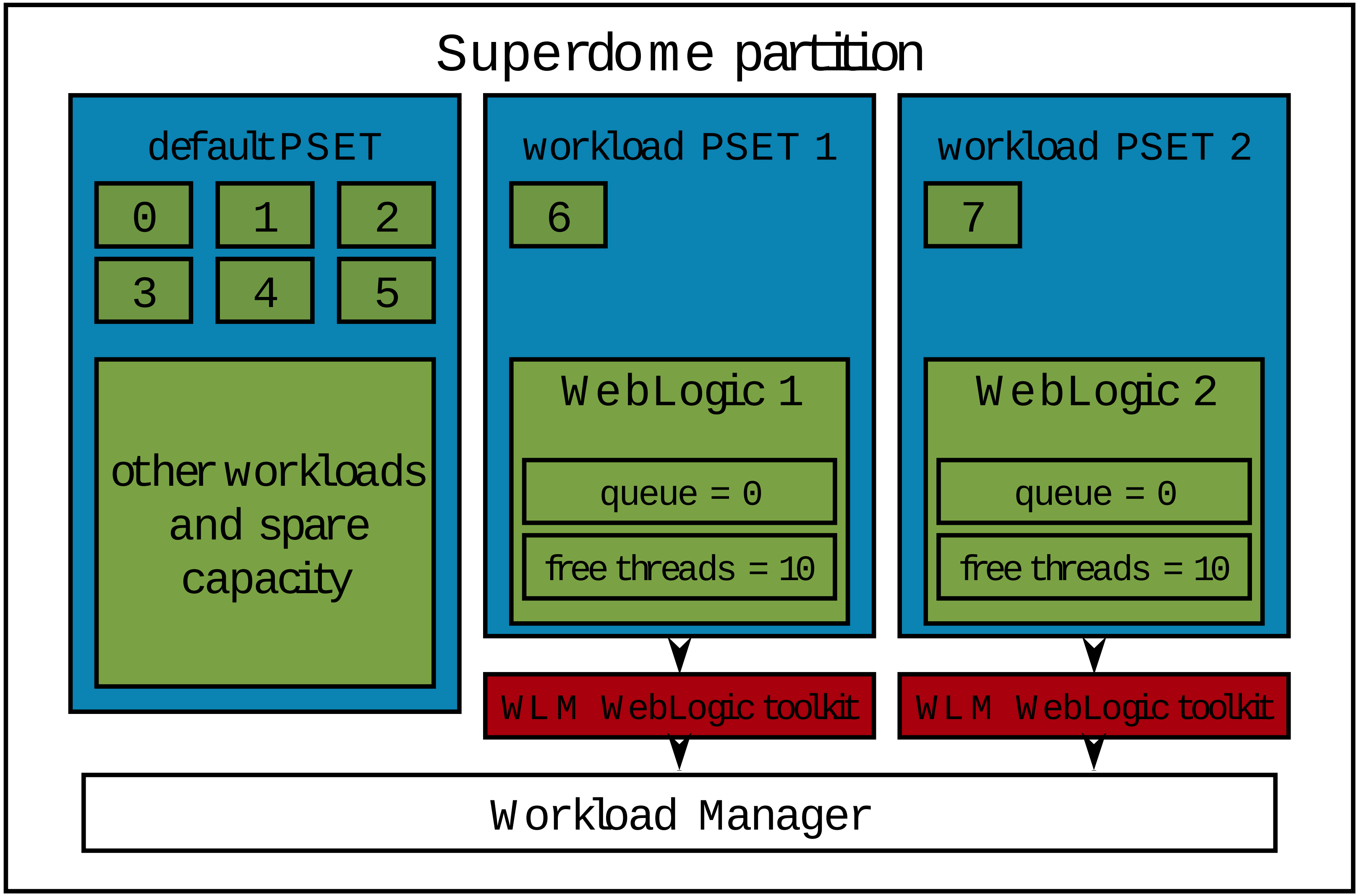
Challenges

- Customer required automatic allocation of CPUs to WebLogic instances, but WLM 2.0 didn’t support PSET CPU migration

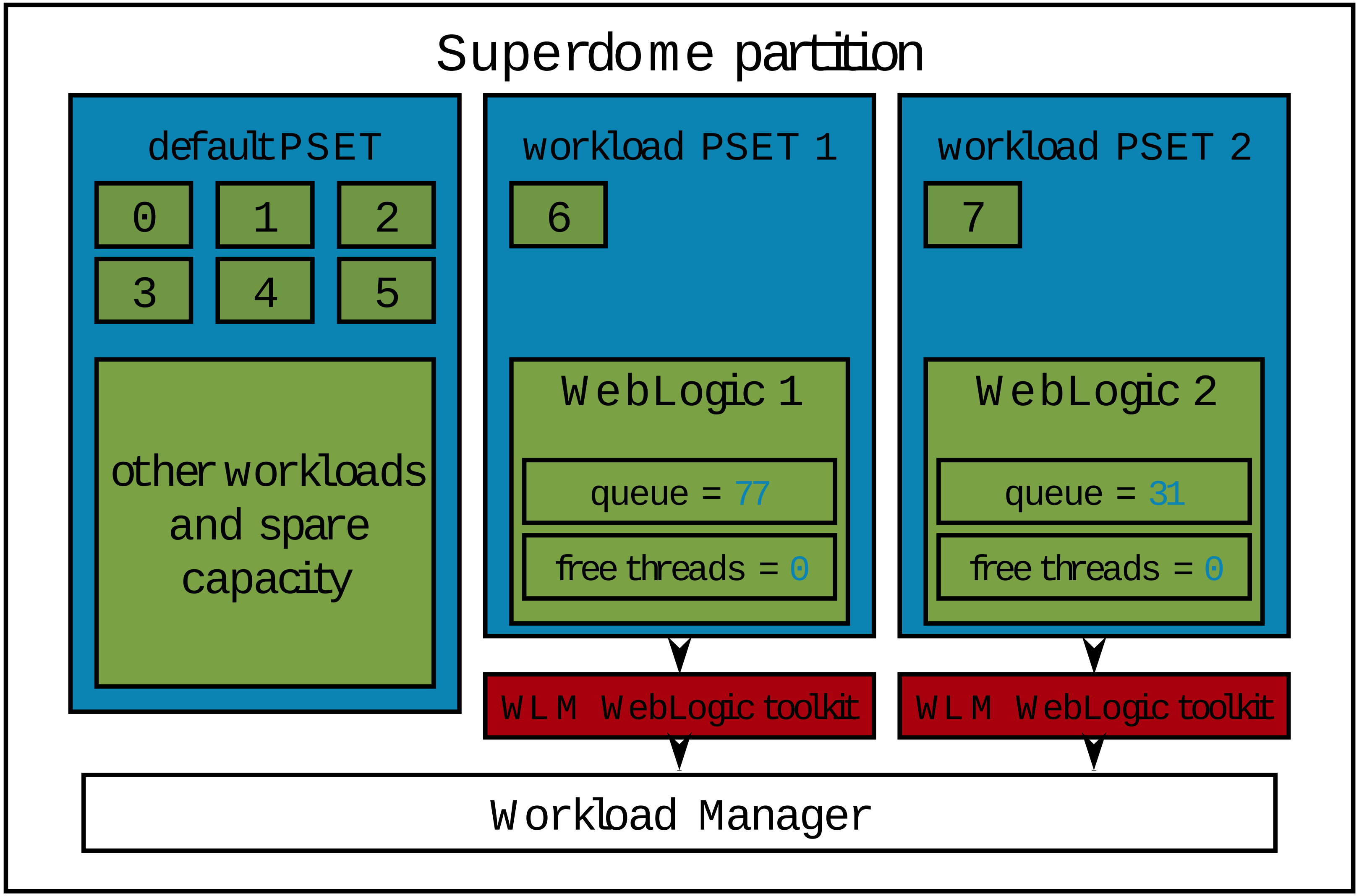
WLM 2.1 requirements

- In order to satisfy the requirements of this solution, the HP consulting organization contacted the lab and the following features were included in the 2.1 release:
  - support for automatic CPU migration between PSETs
    - allows each JVM to get a set of dedicated CPUs
  - BEA WebLogic toolkit
    - provides WebLogic queue length and free thread pool statistics to WLM in real time
- This combination provides for:
  - automatic allocation of the correct number of CPUs to each WebLogic application
  - the ability to prioritize each WebLogic instance and ensure that higher priority instances get preference in cases of simultaneous high load on multiple applications

WLM Optimization of Resources



Load is added to each WebLogic



Superdome partition

default PSET

0 1 2

3 4 5

other workloads and spare capacity

workload PSET 1

6

WebLogic 1

queue = 77

free threads = 0

workload PSET 2

7

WebLogic 2

queue = 31

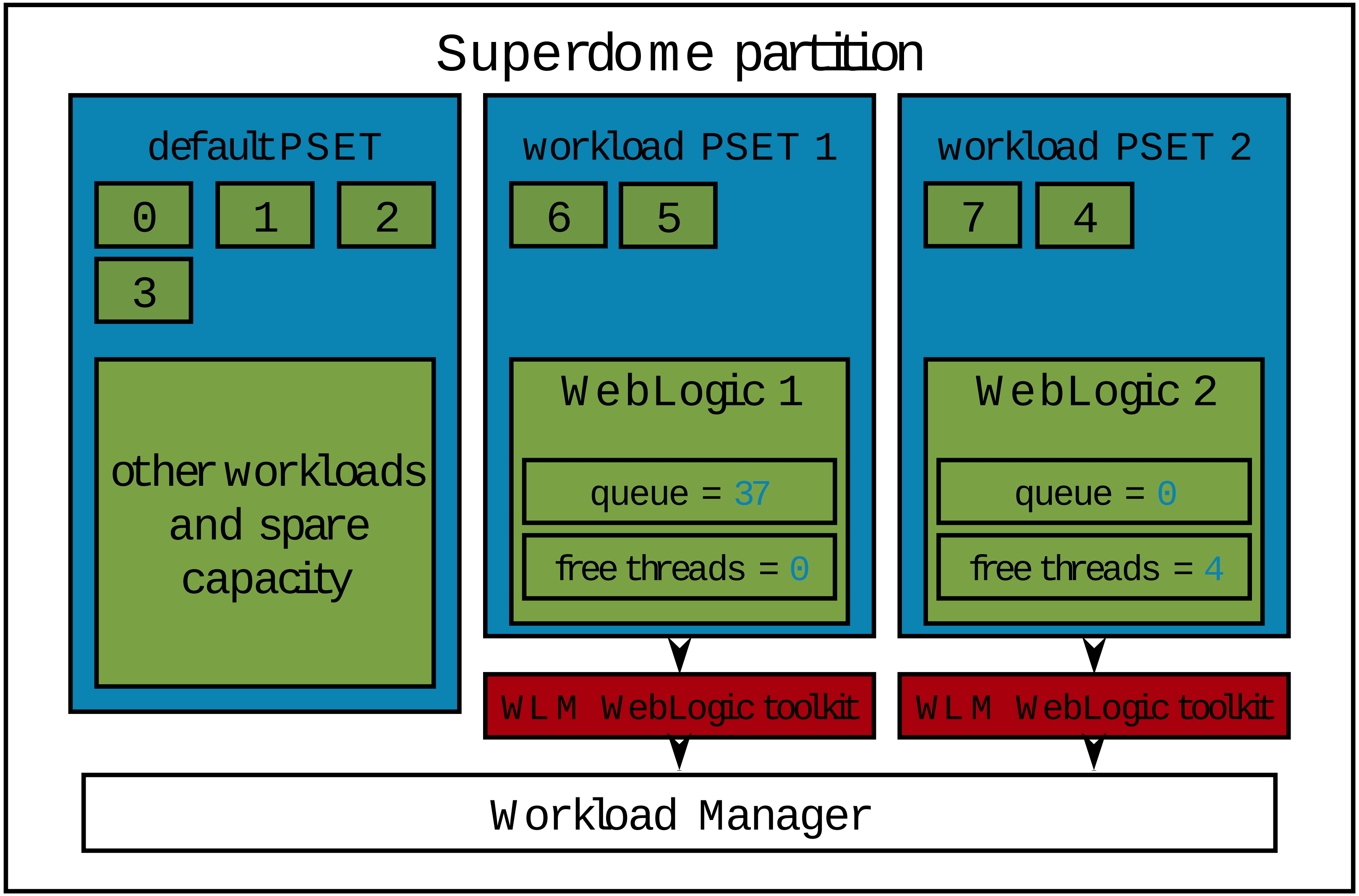
free threads = 0

WLM WebLogic toolkit

WLM WebLogic toolkit

Workload Manager

WLM allocates additional CPUs



Superdome partition

default PSET

0 1 2 3

other workloads and spare capacity

workload PSET 1

6 5

WebLogic 1

queue = 37

free threads = 0

workload PSET 2

7 4

WebLogic 2

queue = 0

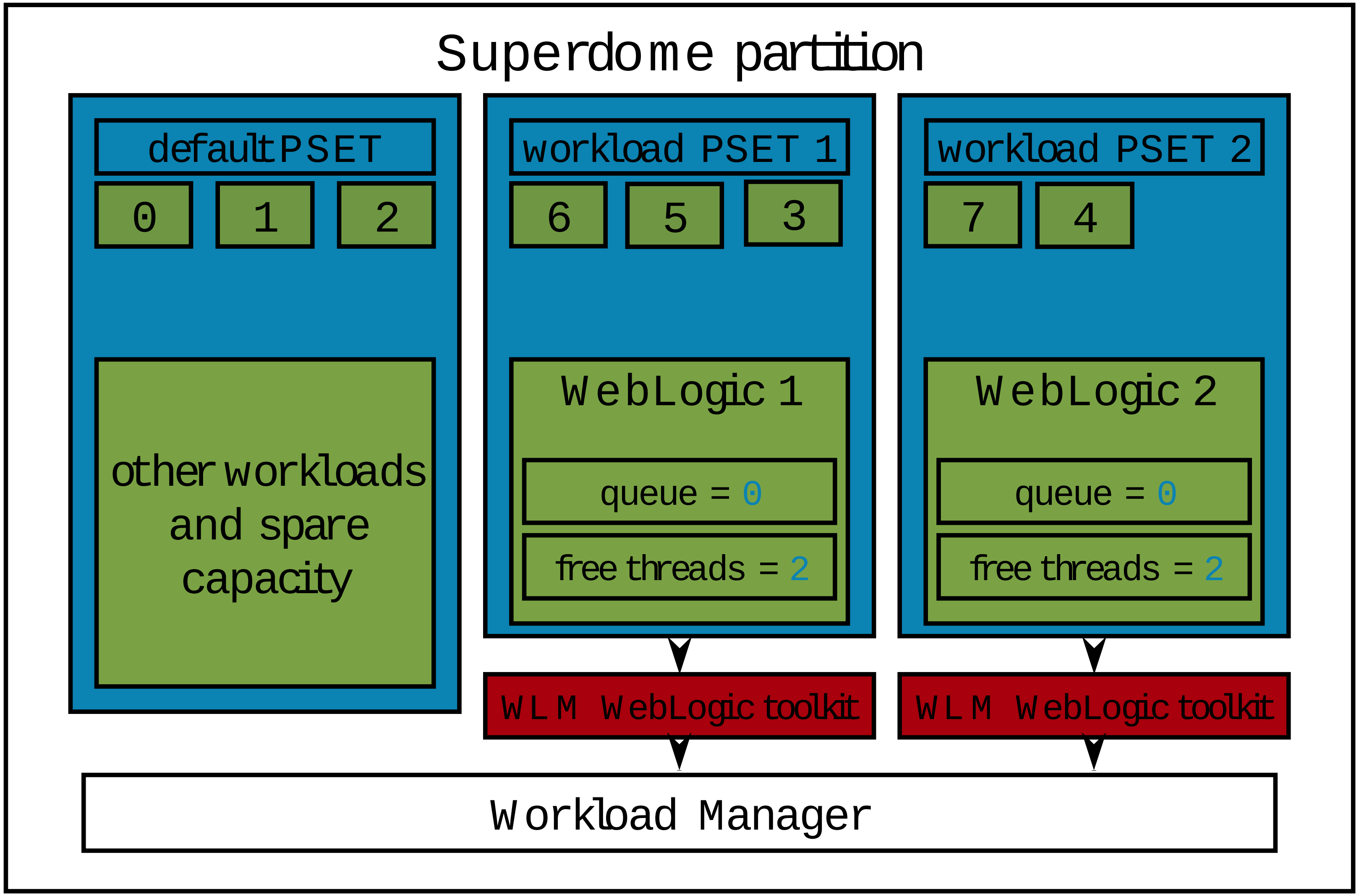
free threads = 4

WLM WebLogic toolkit

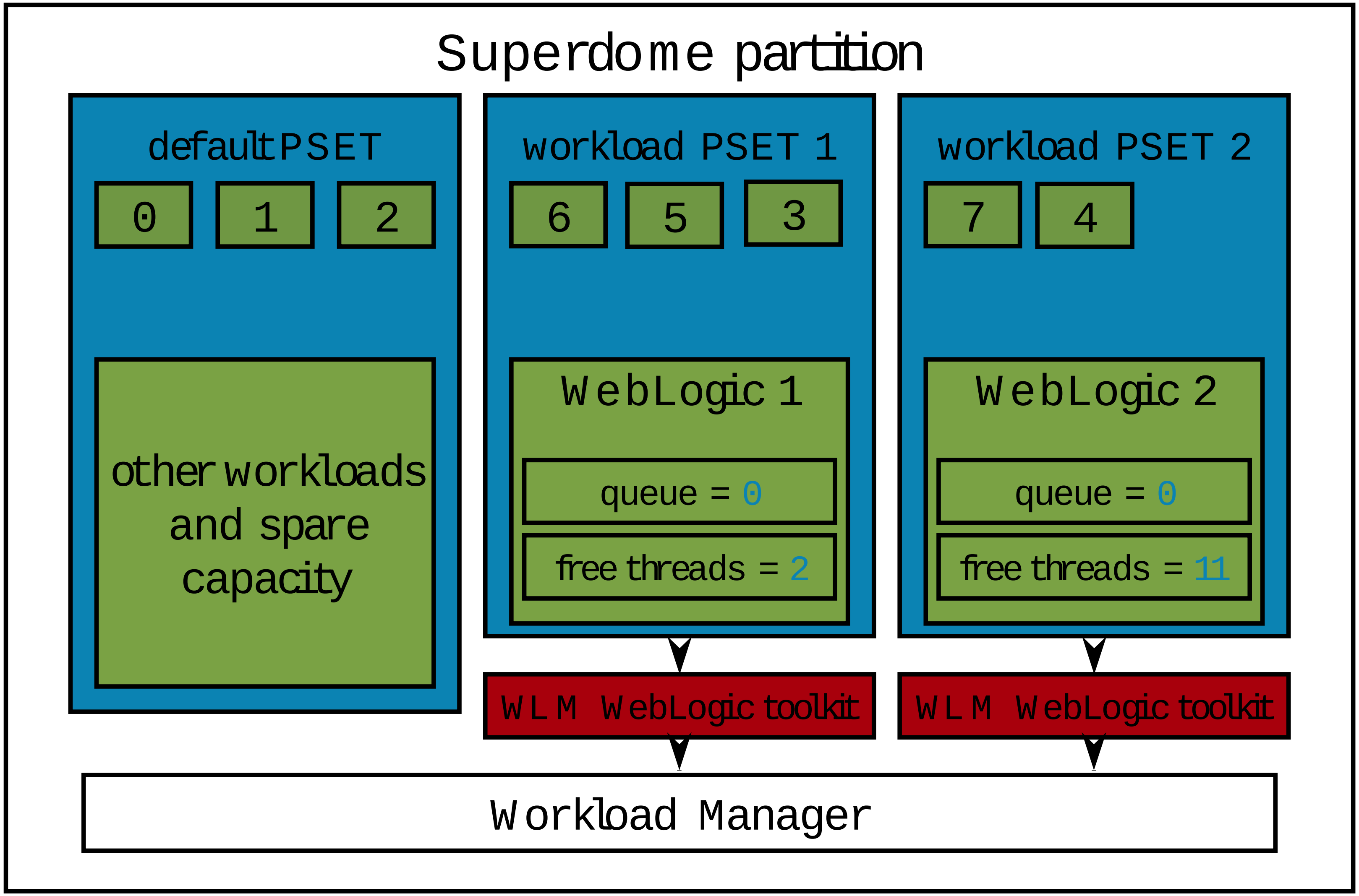
WLM WebLogic toolkit

Workload Manager

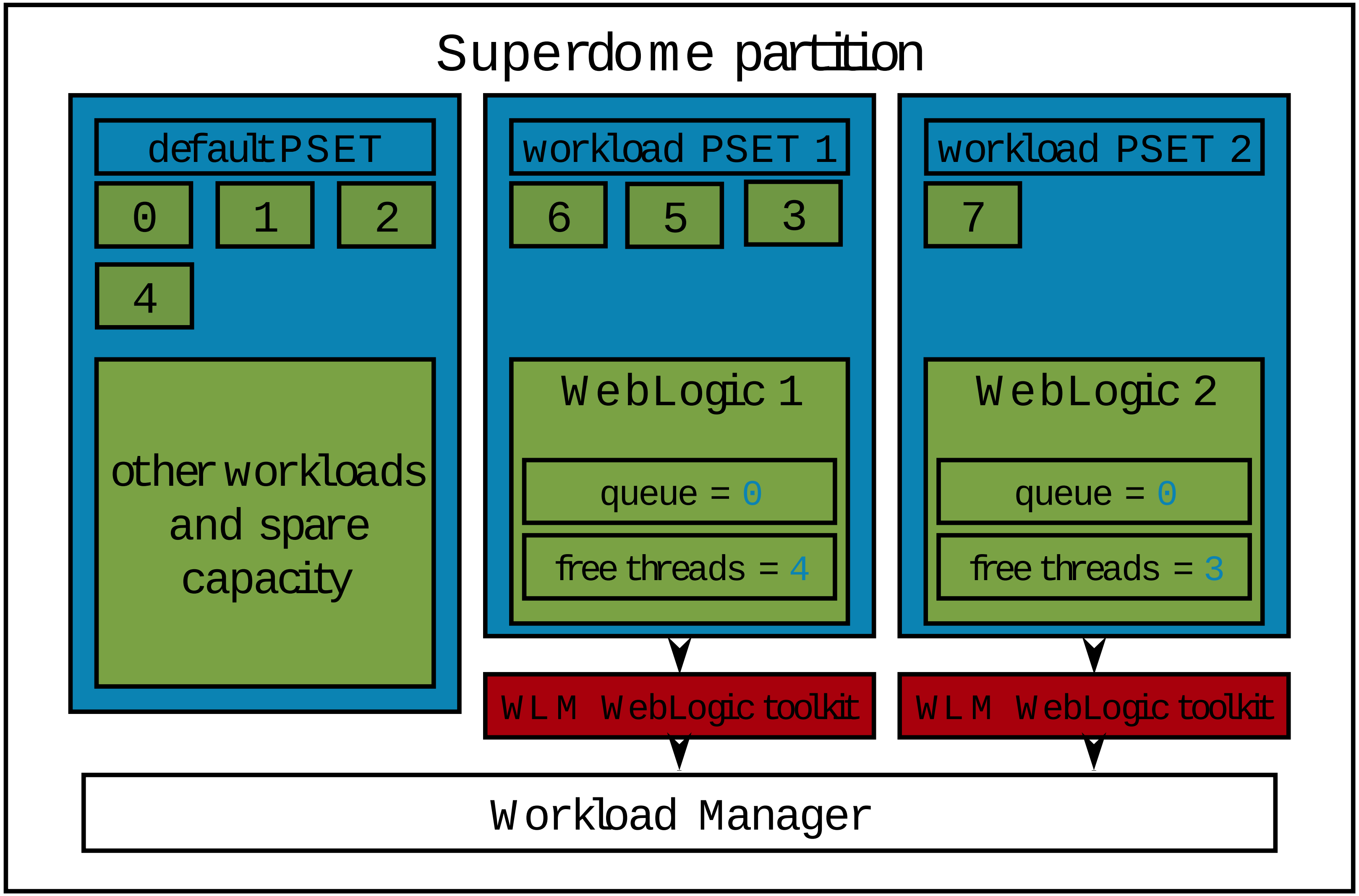
More is still needed for WebLogic 1



Load decreases on WebLogic 2



WLM moves a CPU back to the pool for other workloads



National Semiconductor
The Challenge

- 5 SAP and Oracle servers, two production and 3 development
- Production applications compute needs growing and the existing servers could not expand
- At the same time, the development servers were underutilized

National Semiconductor
The Solution

- Consolidate to 2 x rp8400s
- Install 3 vPars on each rp8400
- Use WLM to automatically move CPU resources between vPars based on CPU utilization, priorities and failover status
    - installation and configuration of WLM on all vPars took less than 2 hours
- MC/ServiceGuard for high availability

National Semiconductor
The results

- Production applications can borrow CPUs from development applications to ensure Service Level Objectives are met
- Production applications are now in a high availability environment that automatically adapts to failover conditions

Summary

- HP's adaptive infrastructure provides a more agile environment for running your many applications
- Your lower cost adaptive infrastructure:
  - fewer OS images to manage
  - faster deployment of new applications
  - higher utilization while still maintaining performance of your highest priority applications
  - better data for capacity planning

Related Presentations

- HP-UX Partitioning break-out session #2072
  - Mon 4:20
- HP-UX Partitioning hands-on lab #2473
  - Tues 8:00, 2:30, Fri 8:00
- HP-UX Workload Manager hands-on lab #2474
  - Tues 10:10, 4:40, Fri 10:10
- Itanium® 2-based Servers break-out session #2554
  - Thurs 9:30
- HP’s High Performance Systems Strategy #2594
  - Thurs 1:30
- How HP has Added Value to Itanium Servers – NDA track
  - Tues 4:50

HP WORLD 2003

Solutions and Technology Conference & Expo

Interex, Encompass and HP bring you a powerful new HP World.

interex
shared knowledge • shared power

encompass
AN HP USER GROUP

hp
invent