

HP Integrity Essentials Global Workload Manager: Workload Management for HP Integrity Virtual Machines



Introduction	2
Why Use gWLM with Integrity Virtual Machines	2
Intelligent, Automatic, and Diverse Resource Management Capabilities Mapped to Your Business Priorities	3
Intelligent and Automatic Activation/Deactivation and Distribution of iCAP or TiCAP Resources Based on Business Priorities	3
The Net Gain that gWLM Provides	4
How gWLM Manages Integrity VM Resources	4
Notes Regarding gWLM Management of Virtual Machine Workloads	5
gWLM Policies You Can Establish for Virtual Machine Workloads	6
How Virtual Machine Entitlements Map to gWLM Policies	7
Overview: Configuring and Managing Virtual Machines as VSE Workloads	8
Configuring VM Host Virtual Machines as VSE Workloads and Applying gWLM Policies	8
Adding a Virtual Machine to the VM Host's SRD	9
Viewing the Behavior of the Virtual Machine Workloads Under gWLM Control	9
Practical Use Scenarios	11
Scenario 1: Putting gWLM in Control of Virtual Machines	11
Scenario 2: Taking Advantage of Instant Capacity Resources	12
Scenario 3: Taking Advantage of Temporary Instant Capacity Resources	14
Scenario 4: gWLM Maintains Virtual Machine Minimum vCPU When Resources Are Needed Elsewhere ..	15
Summary	17
Related information	18

Introduction

Virtualization is a key trend in the IT industry today, making it possible to run multiple “virtual machines” inside a physical machine for higher resource utilization and increased flexibility. The HP Virtual Server Environment (VSE), an integrated virtualization solution for HP Integrity server platforms, enables you to achieve a greater return on your IT investments by optimizing server resource utilization in real time according to business priorities. Through tight integration with partitioning, high availability, and utility pricing, HP VSE allows you to maintain service levels in the event of downtime and to pay for spare capacity on an as-needed basis.

HP VSE provides many virtualization technologies that pool and share resources to meet demands automatically and most effectively. One such technology is HP Integrity Virtual Machines (VM); another is HP Integrity Essentials Global Workload Manager (gWLM). This paper is for Integrity VM users who would like to reap greater benefits by using gWLM to centrally manage the resource utilization and service levels of virtual machines created using Integrity VM. The versions of the products addressed in this paper are Integrity VM Version 2.0 and gWLM Version A.02.50.

gWLM provides centralized, intelligent, policy-based resource management, allowing you to establish resource-sharing policies that can be used across multiple partitioned systems. Any resource managed by gWLM is associated with a collection of processes that is assigned to a specific workload. For example, the collection of processes that run within a particular virtual machine can be managed as one workload, while the processes that run within another virtual machine configured on the same Integrity VM can be managed as a separate workload. gWLM monitors workloads and automatically allocates resources to those workloads across partitions to increase server utilization while satisfying workload service level objectives. gWLM can manage the real-time resource allocation of hard partitions (HP hardware partitions or nPartitions), soft partitions (HP-UX Virtual Partitions, also called vPars, and virtual machines), resource partitions, and nested partitions. (Resource partitions consist of whole-core processor sets, known as PSETs, or sub-core Fair Share Scheduler groups, known as FSS groups. A core is the actual, physical data-processing engine within a processor. A single processor might have multiple cores.)

You can use gWLM with the Integrity VM product in several ways. For example, you might use gWLM to manage the VM Host as a single workload. In this case, your gWLM policies would apply to the VM Host while Integrity VM entitlements continue to determine resource allocation for the virtual machines. A more practical and effective use of gWLM is to use it to manage the virtual machines as individual workloads. In this way, you can have the flexible and powerful gWLM policies determine resource allocation for the virtual machines and thereby provide a tighter coupling between the behavior of the virtual machines and your business goals and needs. This paper focuses on the latter application of gWLM.

Why Use gWLM with Integrity Virtual Machines

HP Integrity Virtual Machines provides better use of your Integrity server capacity, allowing you to run multiple virtual machines on a single server and easily share computing resources. Using gWLM to manage your virtual machines extends and enhances their effectiveness. The most important benefits of using gWLM to manage your virtual machines are:

- Automatic and extensive resource management capabilities that you can align closely with your business priorities
- Automatic activation and deactivation of Instant Capacity (iCAP) and Temporary Instant Capacity (TiCAP) resources, distributing them as needed according to your business priorities
- Simplified, centralized management across multiple servers through a web-based central management system integrated with HP Systems Insight Manager (SIM); this facility provides a unified, consistent interface for management of multiple partition types

As a result, you can easily make more efficient and effective use of your Integrity VM resources. While maintaining your targeted business priorities and service levels, you can consolidate more workloads onto fewer physical systems, reducing the physical hardware required and thereby significantly reducing costs.

Intelligent, Automatic, and Diverse Resource Management Capabilities Mapped to Your Business Priorities

Both HP Integrity VM and gWLM provide mechanisms for dynamic resource allocation. The HP Integrity VM includes a fixed, built-in policy (entitlement) that determines a minimum percentage of physical CPU (cores) to be allocated for a virtual machine, based on the number of virtual CPUs (vCPUs) configured. (Note that HP Integrity VM documentation generally uses the term “physical CPU” instead of “core” to help distinguish between physical and virtual components.) If a virtual machine is busy and sufficient processing power is available on the VM Host system, the virtual machine can receive more than its entitlement. When there is contention for processing power among the virtual machines, each virtual machine is limited to its entitlement. As needed, you can manually change these entitlements dynamically (while the Integrity VM software and the applications or processes it serves are running) to address needs for reallocating resources. You cannot set a single entitlement to be applied to all the virtual machines. The Integrity VM software is not capable of activating or deactivating Instant Capacity resources as needed.

In contrast, gWLM includes intelligent, automatic resource management capabilities and provides you more extensive and varied resource allocation options so that you can control the behavior of your virtual machines to align more closely to your business needs. For example, if you want to consolidate multiple virtual machines that run workloads having different business priorities, gWLM can automatically change CPU resource allocations to individual virtual machines based on these priorities. gWLM can also maintain targeted utilization goals and service levels for each virtual machine.

gWLM manages resources for workloads based on policies that you define and apply to those workloads. You can establish multiple policies (individualized for particular virtual machine workloads) or you can establish a single policy for all the virtual machine workloads to minimize the number of policies. To establish a gWLM policy, you can choose from a wide variety of rules for allocating and sharing resources. For example, besides being able to assign the allocation owned by a virtual machine workload, where the owned allocation is similar to an Integrity VM entitlement, you can also assign a minimum allocation that can be greater than the Integrity VM configured resource minimum, and you can assign a maximum allocation that can be lower than the virtual machine’s configured size. Thus, you can set a limit for the amount of resources a virtual machine can use without having to change its configuration. In addition, gWLM allows you to assign a priority level and weight to each virtual machine workload, helping ensure that critical applications get the resources they need. As business priorities and conditions change, you can easily modify your policies to reflect the changing needs. gWLM then automatically reassigns resources to your workloads to improve resource utilization and maintain continuous service levels. For more information on gWLM policies relevant for managing Integrity VM, see “gWLM Policies You Can Establish for Virtual Machine Workloads” on page 6.

Intelligent and Automatic Activation/Deactivation and Distribution of iCAP or TiCAP Resources Based on Business Priorities

For Integrity VMs deployed on cellular, partitionable servers, gWLM can automatically activate or deactivate Instant Capacity (iCAP) and Temporary Instant Capacity (TiCAP) resources available to the VM Host; you pay for these resources only as needed. gWLM can instantly activate these resources to increase capacity to accommodate increased demands of virtual machine workloads or to replace failed physical CPU resources on the VM Host. gWLM distributes the activated resources according to your business priorities. They can be activated without changes to the virtual machine configurations; the virtual machines immediately take advantage of the additional processing power. Note that

activation of HP Instant Capacity resources when policies are applied to individual virtual machines is supported in gWLM V2.5 or later.

In addition, gWLM can synchronize resource management policies for virtual machine workloads that are defined as Serviceguard packages in an HP Serviceguard cluster, providing cost-effective, high-availability solutions. For more information about using gWLM with Serviceguard, see the gWLM documentation, available from:

<http://docs.hp.com/en/vse.html>

The Net Gain that gWLM Provides

Thus, using gWLM to manage Integrity VM enables you to consolidate more workloads onto fewer physical systems and to use fewer CPU resources. You reduce costs significantly while maintaining expected service levels. Moreover, to uphold your business priorities more effectively, gWLM policies allow you to assign priority levels to the virtual machine workloads and thereby ensure that mission-critical workloads always get the resources they require. In short, gWLM capitalizes on and extends the features of Integrity VM that you most value:

- Maximizing server utilization and scalability
- Allowing for quick and easy addition or reallocation of system resources
- Isolating operating environments
- Improving system availability
- Consolidating enterprise-class servers
- Rapidly deploying new environments
- Improving cost of ownership

How gWLM Manages Integrity VM Resources

Assuming that Integrity VM is already installed on your host system, you can enable gWLM to manage your virtual machines by installing the gWLM agent on the VM Host and perform the necessary gWLM daemon configuration tasks described in the VSE Management Software Installation and Update Guide. This document is available from:

<http://docs.hp.com/en/vse.html>

You use the following management software integrated with gWLM:

- HP Systems Insight Manager (SIM)
- HP Virtual Server Environment Management CMS—this is the system that will be your central management server (CMS)

Once you have performed the required tasks to set up your initial environment, the quickest and easiest way to start using gWLM to manage your virtual machines is to use the Manage Systems and Workloads wizard, which guides you through all the basic steps. For information on using the wizard, see the gWLM documentation, available from:

<http://docs.hp.com/en/vse.html>

The wizard allows you to configure each virtual machine as a VSE/gWLM workload, choose a policy for each workload, and configure the Shared Resource Domain (SRD) (which consists of the collection of workloads that are to share resources). An SRD can be a server or several nPartitions of a complex, with one or more of the nPartitions hosting virtual machines (that is, being a VM Host). The wizard allows you to set the resource allocation interval for the SRD, which establishes the frequency at which gWLM checks current resource consumption and re-allocates resources in accord with your policies. gWLM manages CPU resources in terms of cores, as observed from the gWLM agent.

When a virtual machine is controlled as a workload by gWLM and is running in a deployed SRD, CPU resource allocations for the virtual machine are determined by the gWLM policy applied to that

workload. At any given time, the virtual machine is automatically allocated resources based on that policy. The virtual machine's entitlement value as reported by `hpvmstatus -r` is actually the most recent allocation set by gWLM. (The entitlement set for the virtual machine in its configuration file for the Integrity VM software is not changed while the SRD is deployed.) gWLM prevents the virtual machine from using more than the gWLM-set allocation. When gWLM stops managing an SRD, it sets the entitlements of the running virtual machines to their Integrity VM enforced minimums (5%) so that all other stopped virtual machines have enough resources to be restarted.

You can create an SRD for a set of virtual machines that are already running. To start an additional virtual machine in a deployed SRD, you must establish a gWLM workload and policy for that virtual machine. It will be prevented from starting until a policy is established.

While managing Integrity VM resources, gWLM prevents you from changing a virtual machine's running (dynamic) entitlement (using the Integrity VM Manager or the Integrity VM command line interface (CLI)); however, you can use the Integrity VM Manager or the CLI to configure and manage other elements. (For more information on what you can and cannot change, see "Notes Regarding gWLM Management of Virtual Machine Workloads" on page 5.)

From a central location, gWLM continuously monitors the activity of the virtual machine workloads in the SRD to determine how resources should be allocated. This information is consolidated and stored on the central management server (CMS). If your VM Host is deployed in a hard partition with Instant Capacity resources, gWLM can meet the needs of your virtual machine workloads by automatically increasing CPU resources as demand rises or transferring CPU resources to other hard partitions where demand is greater. gWLM can automatically activate or deactivate Temporary Instant Capacity resources available to a VM Host based on the total resource demands of all the virtual machine workloads. In either case, gWLM distributes processor cycles based on your policies, and it attempts to maintain the minimum number of CPU resources to meet the demands of all the virtual machines. Taking advantage of Instant Capacity or Temporary Instant Capacity resources, gWLM can ensure that enough cores remain available on the VM Host to satisfy the virtual machine workload having the greatest number of configured vCPUs (each virtual machine must have a core available for every configured vCPU); gWLM maintains the required amount of cores even when one or more cores associated with that virtual machine become unavailable.

Notes Regarding gWLM Management of Virtual Machine Workloads

When using gWLM to manage virtual machine workloads, note the following. For more information, see the "Getting the Most Out of gWLM" topic in online help (access online help in SIM by selecting Tools à VSE Management, followed by the tab Shared Resource Domain, and then the question mark [?] in the top right corner).

- gWLM operates in one of two modes: managed mode or advisory mode. Advisory mode allows you to see what CPU requests gWLM would make for a workload—without actually affecting resource allocation. For virtual machines (as well as PSETs and FSS groups), gWLM can operate only in managed mode.
- When an SRD is deployed, only those virtual machines managed as workloads in the SRD are allowed to start. This ensures that gWLM has complete control of the VM Host's pool of resources. gWLM allocates resources only to those virtual machines managed in the SRD and that are started. Any virtual machine that is not managed in the SRD is stopped and will not have a gWLM policy associated with it; therefore, that virtual machine will not be allocated resources needed to run.
- When managing a VM Host as a workload (managing the nPartition that is serving as a VM Host), you should set the policies in the SRD so that the VM Host always has enough cores to meet at least the minimum required for its configuration of virtual machines (typically, this minimum is the greatest number of vCPUs assigned to any virtual machine). For information on this VM Host minimum, see the HP Integrity Virtual Machines documentation, available from:
<http://docs.hp.com/en/vse.html>

- When managing virtual machine workloads, if Instant Capacity is available and gWLM needs to deactivate cores on the VM Host to make more resources available elsewhere, gWLM always maintains the minimum number of cores required for the VM Host's virtual machines. For example, if a virtual machine has three vCPUs, gWLM will ensure that the minimum of three cores is maintained for that virtual machine. A virtual machine must have a core available for each configured vCPU. For an example, see "Scenario 4: gWLM Maintains Virtual Machine Minimum vCPU When Resources Are Needed Elsewhere" on page 15.
- When using gWLM with VSE to create an SRD, a workload named by default hostname.OTHER will be displayed, where hostname is the name of the system serving as the VM Host. This is the workload where gWLM places unassigned resources. You cannot remove this workload, although you can change its name from the default. This workload consists of all CPU resources consumed by processes running on the VM Host. For information, see "How Virtual Machine Entitlements Map to gWLM Policies" on page 7.
- You can determine whether a virtual machine is being controlled by gWLM by using either of the following methods:
 - Under VSE, on the System tab, select the "VM Host" link. In the resulting screen, see the "External Manager" entry.
 - Run the command `hpvmstatus -S` on the VM Host
- When controlling HP VM resources, gWLM allows you to use Integrity VM Manager and standard Integrity VM commands. However, while the virtual machine is running in a deployed SRD, gWLM prevents you from modifying its dynamically determined entitlement.
- While a virtual machine is running in a deployed SRD, do not change the number of configured vCPUs. To have gWLM recognize a vCPU configuration change, first undeploy and delete the existing SRD, then change the virtual machine's configuration and restart the virtual machine. Once the virtual machine has started, create a new SRD to contain that virtual machine workload.
- You can power a virtual machine on or off from the EFI console prompt while it is associated with a workload in a deployed SRD.
- If you want to stop managing a virtual machine with gWLM (removing the associated workload from the SRD), you must stop the virtual machine first. When gWLM stops managing a virtual machine, it sets the dynamic entitlement of the running virtual machine to 5% (by default) to ensure that it can be started when the SRD is undeployed. Once a virtual machine workload is removed from the SRD, the virtual machine is prevented from starting and from consuming VM Host resources; it will be allowed to start and consume VM Host resources once the SRD is undeployed.

gWLM Policies You Can Establish for Virtual Machine Workloads

gWLM provides several types of policies. This section describes the OwnBorrow policy. The OwnBorrow policy allows you to set the following values, which are expressed as percentages:

- The owned amount of CPU resources—the virtual machine is provided the owned amount when needed.
- The minimum amount of CPU resources a virtual machine should always have after lending unneeded resources—when the virtual machine is not busy, it can lend up to a designated amount to other machines; if the machine becomes busy again, it can reacquire the lent-out resources immediately up to its owned amount.

- The maximum amount of CPU resources a virtual machine should have upon borrowing—when the amount of resources owned is insufficient to handle the current workload, it can borrow up to the designated amount if resources are available.

In addition, the OwnBorrow policy allows you to set the following:

- Priority—Priority levels are assigned to workloads; gWLM addresses priority levels from highest to lowest, allocating resources to all requests at a given priority level before considering lower priority requests. The value of 1 is the highest priority.
- Weight— If, at some priority level, all requests cannot be satisfied, the remaining resources are distributed so that the total resource allocation for each workload is as near the proportion of its weight relative to the sum of all the weights as possible.

If gWLM has satisfied all resource requests at all priorities and resources are still available for allocation, it will distribute the remaining resources by weight. Again, this ensures that the total resource allocation for each workload is as near the proportion of its weight relative to the sum of all the weights as possible.

How Virtual Machine Entitlements Map to gWLM Policies

This section describes how Integrity VM entitlements map to gWLM policies. When a virtual machine is brought under control of gWLM, you must select a pre-defined policy or define a new one for that workload; there is no default policy that is automatically enforced for the workload. The mapping discussed in this section describes the gWLM policy that, when enforced, would create allocation behavior similar to the corresponding virtual machine entitlement. gWLM gives you the opportunity to set policies that improve the allocation behavior in alignment with your business priorities. You can then use gWLM real-time and historical reports to assess the performance of the workloads under those policies and tweak policy settings accordingly.

A virtual machine entitlement maps to a gWLM OwnBorrow policy, as follows:

- The gWLM policy minimum is the minimum allocation of CPU resources (as enforced by Integrity VM) that a virtual machine needs to keep running. The Integrity VM enforced minimum is 5% of CPU resources and applies to all virtual machines. The gWLM policy minimum is then calculated as 5% of the number of vCPUs, expressed as a decimal fraction. For example, the policy minimum for a virtual machine with two vCPUs is:

$$2 \text{ vCPUs} * 0.05 = 0.10 \text{ cores}$$
- The gWLM policy owned amount is the virtual machine entitlement amount based on the number of vCPUs; for example, a virtual machine with 2 vCPUs and a 25% entitlement would have a gWLM owned amount of:

$$2 \text{ vCPUs} * 0.25 = 0.50 \text{ cores}$$
- The gWLM policy maximum is based on the CPU resources required by, and reserved for, the VM Host, which is 1% of every physical CPU. Therefore, the gWLM policy maximum is calculated by subtracting 1% for every configured vCPU from the total number of configured vCPUs. For example, the gWLM policy maximum allocation for a virtual machine with 1 vCPU is:

$$1 \text{ vCPU} - (0.01 * 1 \text{ vCPU}) = 0.99 \text{ cores}$$
 For a virtual machine with 3 vCPUs, it would be:

$$3 \text{ vCPUs} - (0.01 * 3 \text{ vCPUs}) = 2.97 \text{ cores}$$

When viewing Integrity VM resources from gWLM management facilities, the displayed minimum, size, and maximum values are calculated in the same manner as the minimum, owned, and maximum amounts. (For an example, see the historical utilization graph in “Viewing the Behavior of the Virtual Machine Workloads Under gWLM Control” on page 9.)

Table 1 shows how various virtual machine configurations could be defined as gWLM OwnBorrow policy configurations based on the calculations used above. The first two columns indicate the vCPUs

and entitlement configured for a virtual machine, and the third column indicates the gWLM policy that would provide behavior similar to the given vCPU/entitlement configuration.

Table 1 Correspondences Between Virtual Machine Configurations and gWLM Policies

Configured vCPUs	Entitlement	gWLM OwnBorrow Policy	
1	25%	Minimum	0.05
		Owned	0.25
		Maximum	0.99
1	99%	Minimum	0.05
		Owned	0.99
		Maximum	0.99
2	25%	Minimum	0.10
		Owned	0.50
		Maximum	1.98
3	25%	Minimum	0.15
		Owned	0.75
		Maximum	2.97
4	25%	Minimum	0.20
		Owned	1.00
		Maximum	3.96

Note that the default workload host.OTHER “holds” the resources reserved for the host (mentioned in the discussion of the gWLM policy maximum, previously). Note that because Integrity VM software does not assign an entitlement for the VSE/gWLM workload host.OTHER, there is no mapping from Integrity VM entitlement to gWLM policy for this workload.

Overview: Configuring and Managing Virtual Machines as VSE Workloads

This section provides a brief overview of how you can configure and manage your Integrity VM virtual machines as VSE workloads. It describes how to:

- Bring your virtual machines under the control of gWLM (see “Configuring VM Host Virtual Machines as VSE Workloads and Applying gWLM Policies” on page 8)
- Bring an additional, subsequently configured virtual machine under the control of gWLM (see “Adding a Virtual Machine to the VM Host’s SRD” on page 9)
- Use gWLM to monitor the behavior of the various virtual machines under its control, in real time as well as historically (see “Viewing the Behavior of the Virtual Machine Workloads Under gWLM Control” on page 9)

Configuring VM Host Virtual Machines as VSE Workloads and Applying gWLM Policies

To configure your virtual machines as VSE workloads to be managed by gWLM and assign gWLM policies to these workloads, access the CMS from a web browser and use the Manage Systems and Workloads wizard available through the VSE Management menu bar. The wizard takes you through four steps to set up your virtual machine workloads under the control of gWLM:

1. Specify a host name for each of your Integrity VM virtual machines—this establishes your virtual machines as VSE/gWLM workloads.

2. Set SRD properties such as the SRD name, mode, state, and resource allocation interval, and whether to use Temporary Instant Capacity (the latter will be used in Scenario 3). To manage virtual machines, you must set the mode to managed; the defaults for other options are acceptable for virtual machines. (The other mode setting is advisory mode, which is not available for SRDs containing virtual machines, PSETs, or FSS groups.)
3. Specify workload and policy settings. Workload settings include the compartment type, which you would select as "HPVM". (A gWLM compartment is an nPartition, virtual partition, virtual machine, PSET, or FSS group. Multiple compartments are grouped to share resources within an SRD. gWLM manages each workload by adjusting the resource allocation for its compartment.)

For each workload, a drop-down menu allows you to select from a list of policies; or you can choose to create your own policy. When you define your own policy, the wizard prompts you for the various policy parameters. Create your policies by following the steps described in the gWLM documentation, available from:

<http://docs.hp.com/en/vse.html>

4. Review the SRD and workload settings to confirm.

Adding a Virtual Machine to the VM Host's SRD

Assume you have configured a new virtual machine on your Integrity VM Host and want to add it to the gWLM-managed collection. Use the wizard to add the new virtual machine as a VSE workload in the SRD by selecting the "Add VM..." button and then selecting the virtual machine to be added. Assign a policy to the virtual machine (step 3, above, select your custom-defined policy or create a new one specific for this workload). The new virtual machine must have a policy before you can start it (for more information, see "Notes Regarding gWLM Management of Virtual Machine Workloads" on page 5).

Viewing the Behavior of the Virtual Machine Workloads Under gWLM Control

gWLM provides real-time and historical reports showing resource utilization and resource allocation profiles. These reports are accessible through the VSE Management menu bar available in a web browser accessing the CMS. For example, to view gWLM allocations and virtual machine utilization over time, you can have gWLM generate a historical Workload Utilization graph.

To view gWLM reports:

1. Access the VSE Management menu bar and select the Shared Resources Domain tab.
2. If you want a real-time report, select the workload for which you want a report; otherwise, skip to the next step.
3. From the VSE management menu bar, select Report and then select the type of report you want.

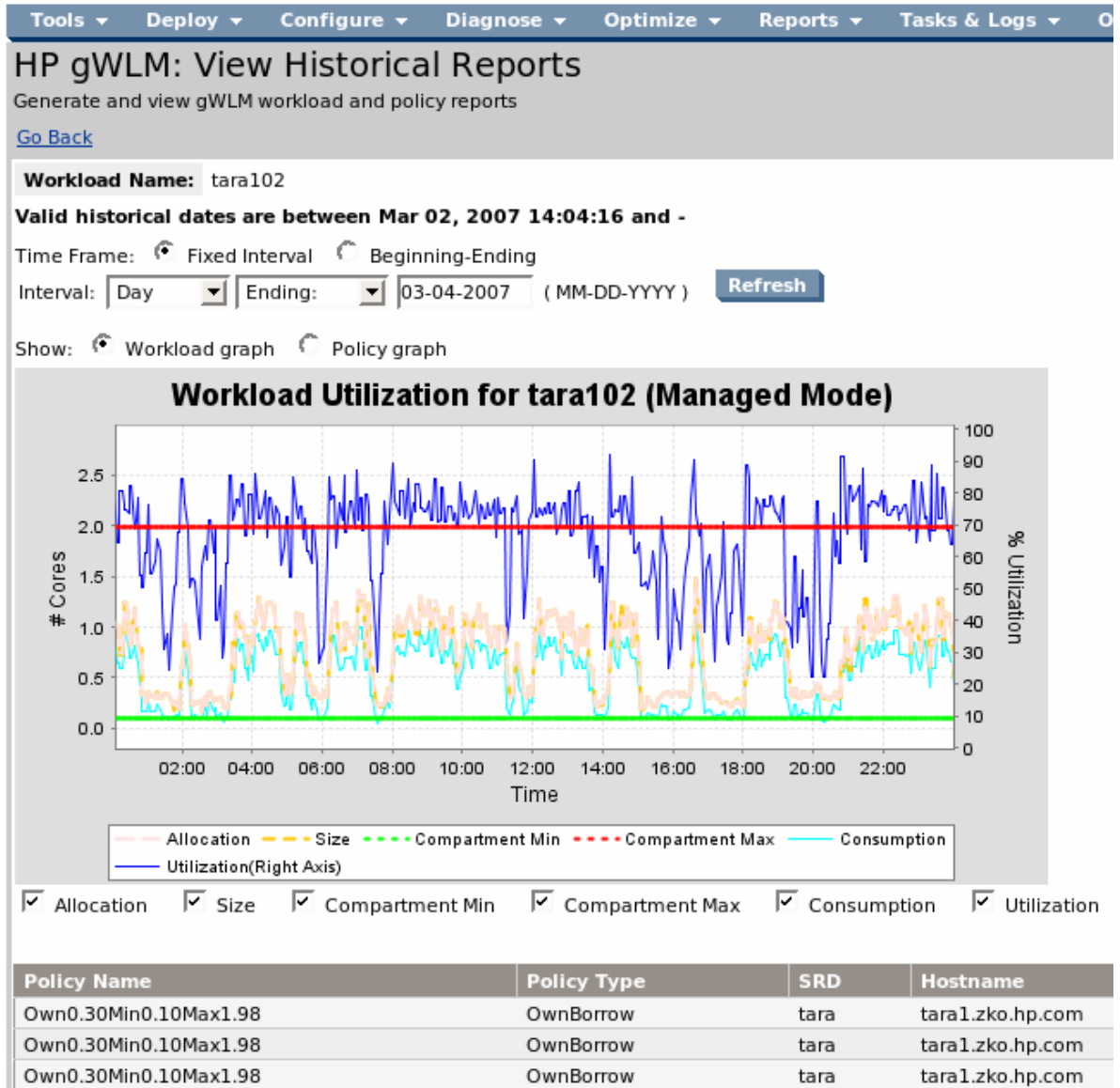
gWLM's historical reports are detailed and can be used for internal chargeback in a centralized-IT environment – for example, gWLM reports show which application instances borrow or lend excess resources and show that an application instance got its owned resources whenever it required them.

These reports can also help you understand better how gWLM policies work. Figure 1 shows the graph that is included in the historical report for a virtual machine workload named tara102, which has a policy named "Own0.3Min0.1Max1.98". The gWLM policy defines an owned value of 0.3, meaning that the workload is provided 0.3 cores when needed. The policy minimum is 0.1, which is the amount of cores owned by that workload even when the virtual machine is not busy; this value determines how much the workload can lend. The policy maximum of 1.98 is the limit upon borrowing; the workload can borrow as long as its consumption does not exceed 1.98 cores. The graph in Figure 1 shows the historical allocation over a two-day period, giving you a good look at how the workload's policy is faring. Not shown in this figure is other information reported with the graph, such as policy changes.

Note also that gWLM provides helpful monitoring tools from the gWLM CLI:

- `/opt/gwlm/bin/gwlm monitor`
Displays policy, workload, and SRD statistics.
- `/opt/gwlm/bin/gwlmreport`
Provides “topborrowers” and “resourceaudit” reports and more.

Figure 1 - Historical Data Report for Virtual Machine



The **horizontal red line** indicates the maximum amount of cores that this workload can consume, which corresponds to its virtual machine (compartment) configuration of 2 vCPUs. The **horizontal green line** indicates the minimum amount of cores that must be allocated at all times, which is about 0.1 for the workload or 0.05 cores for each vCPU configured for that workload. The **dark blue line** shows the percentage (% Utilization, measured on the vertical axis on the right of the graph) of the workload’s allocated cores being used at any given time. The **lighter blue line** tracks the actual amount of cores consumed. The **tan** and **orange** lines show, respectively, the amount of cores the workload actually has (size) at any given time and the amount gWLM allocated to it after arbitrating resource requests from the policies of all the workloads in the SRD. For compartments operating in managed mode (and therefore, all virtual machine workloads), these lines converge as shown in this

example. Note that size/allocation values “follow” the consumption; in other words, gWLM maintains an allocation and size level that meets current consumption. Thus, the current policy for workload tara102 is performing well. For more information on interpreting gWLM reports, see the online help.

Practical Use Scenarios

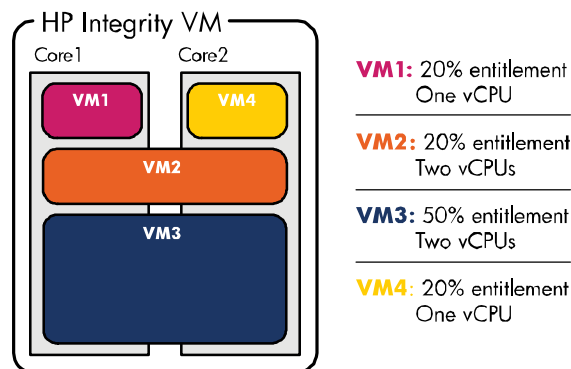
This section provides several scenarios demonstrating how you might use gWLM to manage Integrity VM virtual machines as workloads:

- Scenario 1: you have a VM Host with a set of virtual machines that you want managed by gWLM
- Scenario 2: you have a complex that includes a VM Host in one hard partition and a set of virtual partitions in another, and you want gWLM to make Instant Capacity (iCAP) resources available to the virtual machines as needed
- Scenario 3: you have a complex that includes a VM Host in one hard partition and a set of virtual partitions in another, and you want gWLM to make Temporary Instant Capacity (TiCAP) resources available to the virtual machines as needed
- Scenario 4: you have a complex that includes a VM Host with SMP virtual machines and want to make sure the SMP virtual machines maintain their minimum number of cores even when cores are needed elsewhere

Scenario 1: Putting gWLM in Control of Virtual Machines

Assume you have four virtual machines configured for your Integrity VM, as shown in Figure 2. Two of the virtual machines (VM1 and VM2) run mission-critical applications. The other two virtual machines (VM3 and VM4) run less critical applications. For normal operations, VM1 and VM2 typically require a smaller amount of cores than VM3, but they need more resources to meet peak demands. The entitlements and vCPU assignments are described on the right (where one vCPU is assigned for each core). The virtual machines shown in Figure 2 are running at their entitlement levels.

Figure 2 – Integrity VM Initial Configuration

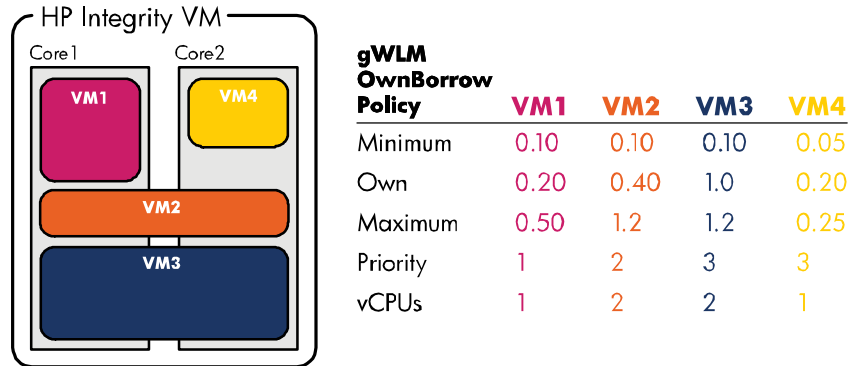


The Integrity VM entitlements ensure that the virtual machines receive, at minimum, the specified percentage of CPU resources. However, Integrity VM currently provides no mechanism to limit the resources consumed by a given virtual machine, other than allowing you to change the configured number of vCPUs. Thus, in this scenario, it is possible that a virtual machine with a less-critical application, such as VM3 or VM4, could consume more resources than its entitlement, preventing the more important applications running on VM1 and VM2 from receiving extra resources as needed. With gWLM, you can prioritize the virtual machine workloads to ensure that the higher-priority workloads always receive the extra resources they need prior to the lower-priority workloads. You can also assign a limit to how much certain workloads can consume.

Figure 3 shows the same virtual machines managed by gWLM. In this scenario, the highest-priority workload (VM1) has grown and requires more resources. VM1 must receive extra resources to meet its needs before VM2 is allocated extra resources. Assume you also want to make sure VM1 does not

use more than half of a core for its operations, and that VM2 and VM3 each do not use more than 60% of each core. In addition, VM3 typically does not need the resources it owns. To make this happen, you could set up the gWLM OwnBorrow policies shown in the table in Figure 3. These policies would ensure that VM1 (and VM2) can get more resources when needed by borrowing from VM3 and VM4; gWLM limits the resource allocation of the latter two less-critical workloads to make sure resources are always available for the higher-priority workloads. If VM1 and VM2 both need more resources, VM1 is ensured resources first (it has the highest priority, 1).

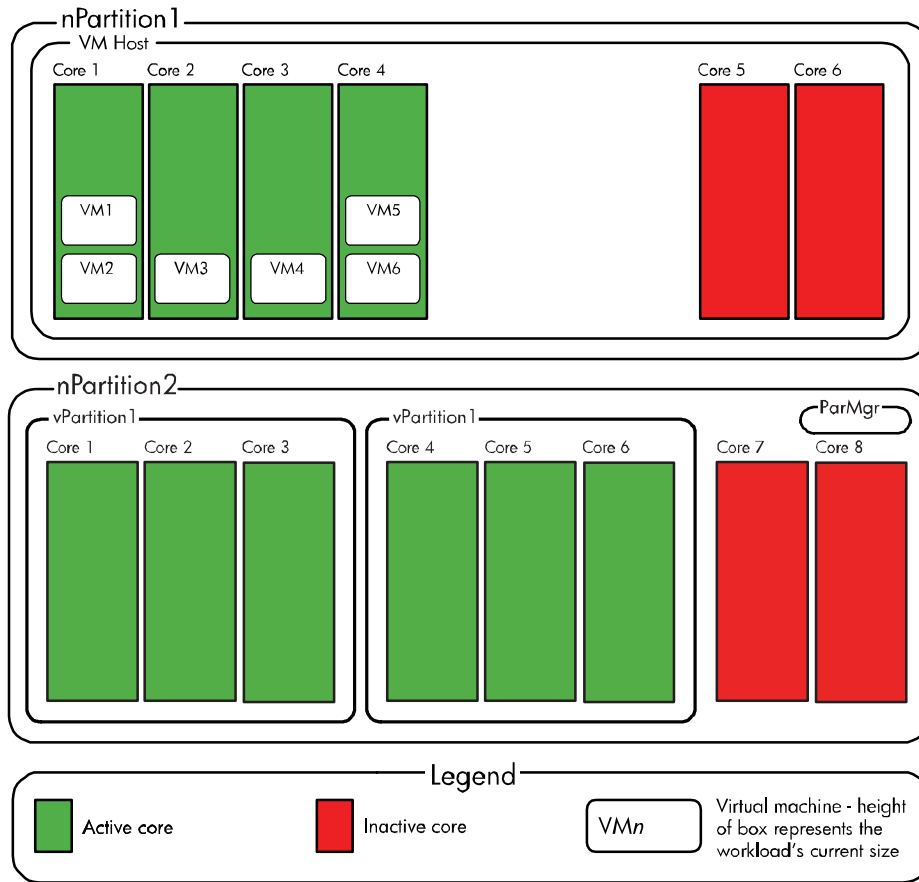
Figure 3 – Integrity VM Managed by gWLM



Scenario 2: Taking Advantage of Instant Capacity Resources

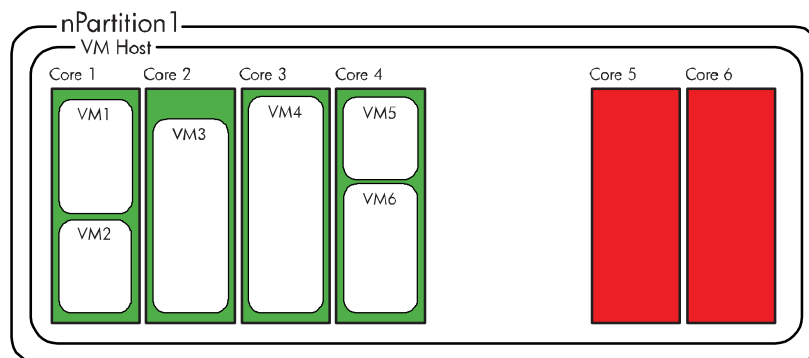
This scenario shows how you can use gWLM to activate (and deactivate) any number of iCAP cores to “migrate” usage rights to where they are needed. In this scenario, a cell-based complex consists of two hard partitions with a total of 14 cores. Of these cores, 10 have Instant Capacity usage rights while 4 are expected as inactive. As shown in Figure 4, the first partition (nPartition1) includes an Integrity VM Host that runs six lightly-loaded virtual machines sharing four cores. Two additional cores are inactive iCAP resources. Assume each virtual machine owns 25% of the resources of one core. For sake of simplicity, the virtual machines in this scenario each have one vCPU, and they are associated with one core only. A later figure shows an Integrity VM configuration where several virtual machines are associated with multiple cores (symmetric multiprocessing). The second partition (nPartition2) has two virtual partitions, each owning three cores, with two additional inactive iCAP cores. The nPartitions are monitored by the Partition Manager.

Figure 4 - Complex with Light Integrity VM Workloads (Uniprocessor) and Inactive Instant Capacity Resources



Now suppose several of the virtual machine workloads increase and the CPU utilization reaches an undesirable point, as in Figure 4A:

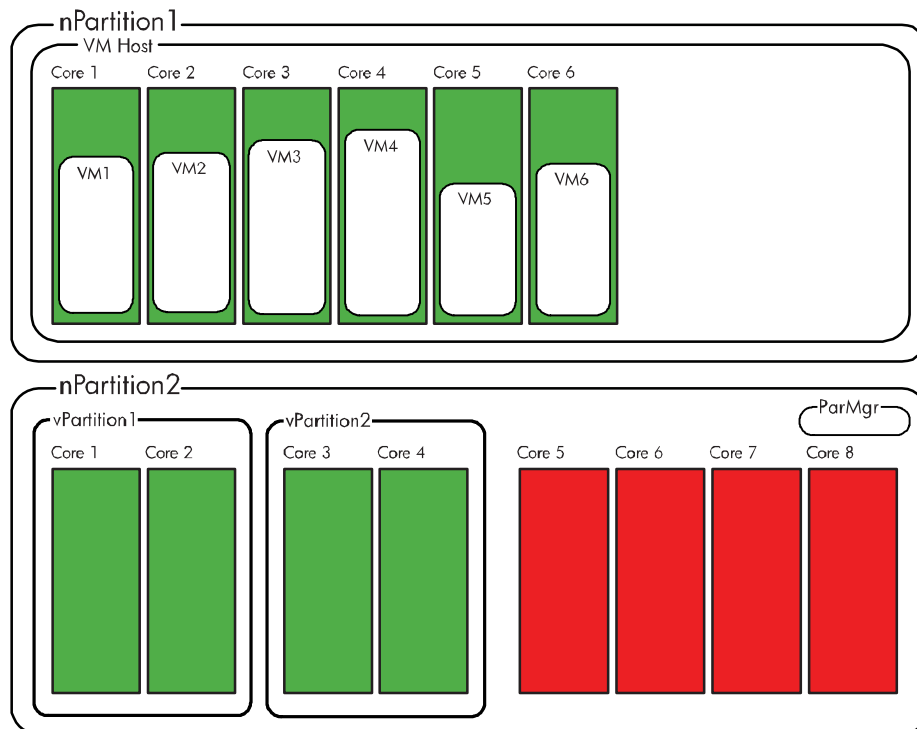
Figure 4A - Heavily Loaded Integrity VM Workloads (Uniprocessor)



gWLM allows you to form SRDs consisting of multiple types of compartments. The workloads in any compartments can then borrow resources from any of the other compartments in the same SRD. So, in this scenario, you could create an SRD that contains the six virtual machines in nPartition1 and the two virtual partitions in nPartition2. With Instant Capacity available in the complex, gWLM can automatically increase the number of cores available to the overloaded virtual machines by “moving” unneeded cores from the other hard partition. The cores are not actually moved physically from one hard partition to another: physical movement is simulated by deactivating cores where they are not needed (nPartition2) and activating an equivalent number of cores where they are needed (nPartition1), essentially shifting usage rights between machines. For example, in Figure 4B, gWLM

has deactivated one iCAP core in each vPar in nPartition2 (where there is less demand for CPU resources) and activated the equivalent iCAP resources (two cores) in nPartition1. The result is that the virtual machine workloads now have two additional cores to share their loads so that each virtual machine has all the processing power of a whole core to itself. You do not incur a charge for these extra (activated) cores. (The virtual machine workload sizes depicted in Figure 4B have changed slightly from the sizes in Figure 4A.) The vPars in nPartition2, whose workloads continue to be light, can easily meet their demands with one less core each. When the demands on the virtual machines return to normal again, gWLM can automatically deactivate the extra Instant Capacity cores on nPartition1 and shift usage rights back to nPartition2 as needed.

Figure 4B - Same Complex with Redistribution of iCAP Resources to Meet Demands



How You Enable gWLM to Use Instant Capacity: HP Instant Capacity must be installed on each nPartition. For more information, see the latest Instant Capacity User's Guide by going to the following location and selecting "Network and Systems Management" and then "Utility Pricing Solutions":

<http://docs.hp.com>

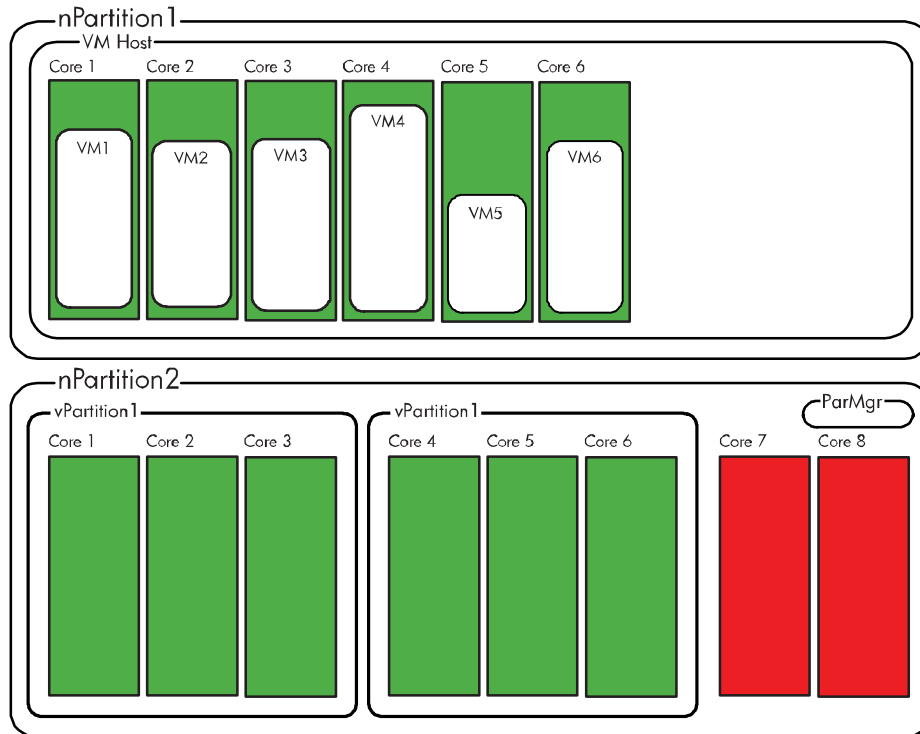
For information on enabling gWLM to take advantage of Instant Capacity resources, see the "Getting the Most Out of gWLM" topic in online help, especially the section "Setting Up npars to Maximize Resource Sharing (npar requirements)." Set up an OwnBorrow policy for each nPartition to make sure its number of owned CPU resources is the number you want the partition to have whenever needed.

Scenario 3: Taking Advantage of Temporary Instant Capacity Resources

This scenario shows how you can use gWLM to activate (and deactivate) any number of TiCAP cores to temporarily meet peak demands where needed. Assuming the original circumstances depicted in Figure 4A (with the overloaded virtual machines in nPartition1), if you have TiCAP rights, you can configure the SRD to support Temporary Instant Capacity. Then, when virtual machine workloads increase and CPU utilization reaches a certain point, gWLM can activate as many TiCAP resources as are needed until the demands on the virtual machine workloads diminish to a specified level or the prepaid amount of temporary capacity units expires. Additional cores can be activated for nPartition2

as needed, too. Figure 4C shows the net result when gWLM activates TiCAP resources for the virtual machines in Partition 1. In this case, the result for the virtual machines is similar to that of iCAP activation in the previous scenario; however, the three cores in each vPar in nPartition2 remain active. As a result, during the allotted time, more resources are active in the complex. At the same time, the virtual machines that originally had to share CPU resources (VM1 and VM2, VM5 and VM6) now have a whole core each.

Figure 4C - Same Complex with Redistribution of TiCAP Resources to Meet Demands



How You Enable gWLM to Use Temporary Instant Capacity: For information on how to configure and use TiCAP, see the gWLM documentation, available from:

<http://docs.hp.com/en/vse.html>

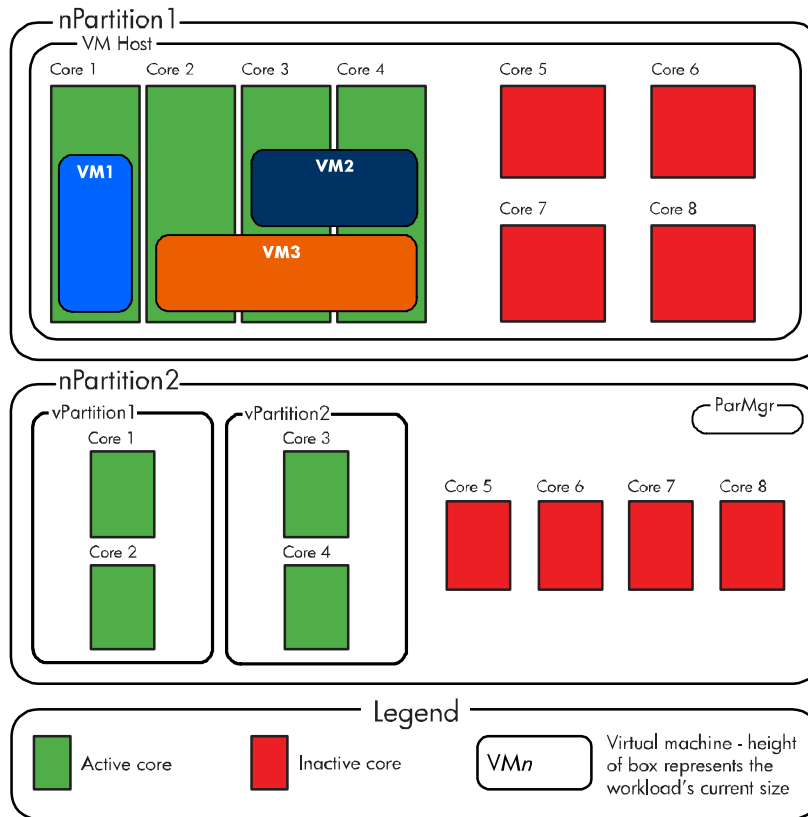
For more information on configuring Temporary Instant Capacity, see the latest HP Instant Capacity User's Guide available at the following location (select "Network and Systems Management" and then "Utility Pricing Solutions"):

<http://docs.hp.com>

Scenario 4: gWLM Maintains Virtual Machine Minimum vCPU When Resources Are Needed Elsewhere

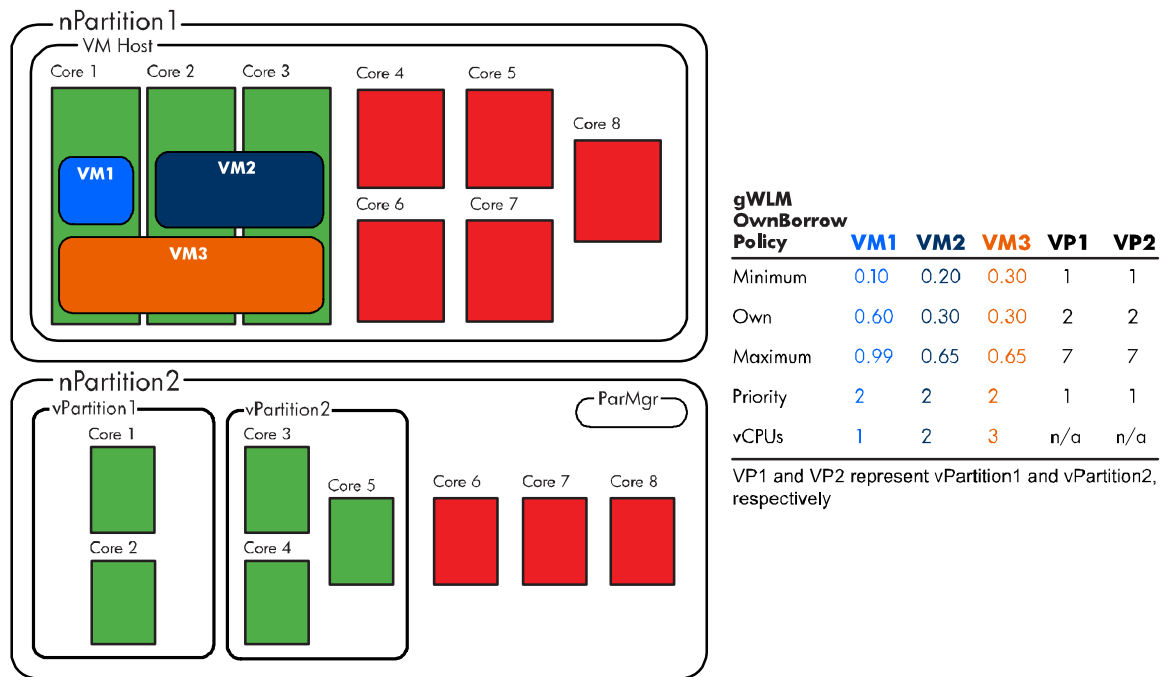
This scenario demonstrates how gWLM responds when managing SMP (multiprocessor) virtual machines whose cores are needed elsewhere. Given virtual machines VM2 and VM3 of nPartition1 in Figure 5, suppose the applications running in the virtual partitions in nPartition2 are higher priority than those running on the virtual machines in nPartition1. If vPartition2 in nPartition2 requires another core to meet a spike in demand, gWLM will borrow an active core from the VM Host if it can continue to maintain the minimum configuration of vCPUs required by each of the virtual machines.

Figure 5 - Multiprocessor Virtual Machines



Assume that 8 of the 16 cores in this complex have iCAP usage rights and that there is no available TiCAP balance. Figure 5A shows that gWLM can migrate an active core (Core4) from the VM Host to make it available for vPartition2. (If a TiCAP balance was available, Core5 in nPartition2 would be activated without reducing the active cores for the VM Host in nPartition1.) While doing so, it maintains the minimum number of vCPUs configured for VM2 and VM3. Virtual machines VM2 and VM3 are assigned to Core1 to maintain their required vCPU counts (VM2: 2 vCPUs; VM3: 3 vCPUs). The table in Figure 5A shows the gWLM policies you might apply to the various components in this scenario. The gWLM policy for VM1 ensures that it can lend the resources needed by VM2 and VM3. gWLM makes sure VM2 and VM3 always have their required minimum number of cores. Note that the number of cores (physical CPUs) on the Integrity VM Host must be at least the same as the highest vCPU count of any virtual machine it hosts.

Figure 5A - Multiprocessor Virtual Machines After Lending an Active Core



Summary

By managing your Integrity VM resources with gWLM, you can ensure that their behavior aligns more effectively with your business priorities. In addition, you can consolidate more workloads onto fewer physical systems to make better use of your IT investments and to reduce costs while maintaining targeted service levels. You can make all this happen easily, thanks to the simplified, web-based GUI of the CMS, a centralized management facility integrated with HP SIM to make management of multiple systems convenient. Through this facility, you can centrally define and manage gWLM policies across a large number of servers and a variety of partitions with ease.

gWLM automatically adjusts resource allocation based on real-time demand, allowing for resources to be shared when they are plentiful and to be dedicated to the highest-priority applications when demands spike. gWLM automatically expands or shrinks the amount of processing resources allocated to individual virtual machines as needed, and it can automatically activate or deactivate Instant Capacity or Temporary Instant Capacity resources available to the VM Host—in either case, you pay only for what you use. Thus, gWLM ensures that your mission-critical workloads always get the resources they need while you get the best return on your hardware investments.

Related information

The following references provide useful background information on related products and topics:

- HP Integrity Essentials global Workload Manager (gWLM)—<http://www.hp.com/go/gwlm>
 - For gWLM white papers and documentation, click the [Information Library](#) link on the right
 - For gWLM online help, access online help in SIM by selecting Tools à VSE Management, followed by the tab Shared Resource Domain, and then the question mark [?] in the top right corner
- HP Virtual Server Environment—<http://www.hp.com/go/vse>
- HP Adaptive Enterprise and virtualization—<http://www.hp.com/go/virtualization>
- HP Systems Insight Manager—<http://www.hp.com/go/hpsim>
- HP Instant Capacity—<http://www.hp.com/go/utility>

© 2007 Hewlett-Packard Development Company, L.P. The information contained herein is subject to change without notice. The only warranties for HP products and services are set forth in the express warranty statements accompanying such products and services. Nothing herein should be construed as constituting an additional warranty. HP shall not be liable for technical or editorial errors or omissions contained herein.

March 2007