VMware® Distributed Power Management Concepts and Use

vmware®

Table of Contents

VMware ESX 4 and VMware vCenter Server 4

Consolidation of physical servers into virtual machines that share host physical resources can result in significant reductions in the costs associated with hardware maintenance and power consumption. VMware® Distributed Power Management (VMware DPM) provides additional power savings beyond this initial benefit by dynamically consolidating workloads even further during periods of low resource utilization. Virtual machines are migrated onto fewer hosts and the un-needed ESX hosts are powered off. When workload demands increase, ESX hosts are powered back on and virtual machines are redistributed to them. VMware DPM is an optional feature of VMware® Distributed Resource Scheduler (VMware DRS).

This information guide provides a technical overview of VMware DPM operation in VMware® ESX™ 4 and VMware vCenter™ Server 4. It is intended for VMware partners, resellers, and customers who want detailed information on VMware DPM functionality in that release.

The guide covers the following topics:

VMware vSphere™ and Cluster Services

VMware DPM Usage

VMware DPM Operation

VMware DPM Advanced Options

VMware DPM and Datacenter Monitoring Tools

VMware DPM Usage Scenario

Resources

VMware vSphere and Cluster Services

One of the key management constructs in VMware vSphere 4, which comprises VMware ESX 4 and VMware vCenter Server 4, is the cluster. Grouping multiple ESX hosts into a cluster enables you to manage them as a single compute resource. The cluster services that bring about this benefit include VMware® Fault Tolerance (VMware FT), VMware® High Availability (VMware HA), VMware DRS, and VMware DPM.

VMware FT and VMware HA handle host and virtual machine failures in a cluster of ESX hosts. It respects your settings for the desired policies and the associated resources to be set aside for use by virtual machines in the event of a failure. VMware FT and VMware HA implements mechanisms for detecting problems and restarting virtual machines. The comprehensive "VMware vSphere Availability Guide" (see Resources for a link) presents information on VMware FT and VMware HA operations. VMware FT and VMware HA failover resource constraints are respected by VMware DRS and VMware DPM.

VMware DRS manages the allocation of resources to a set of virtual machines running on a cluster of ESX hosts with the goal of fair and effective use of resources. VMware DRS makes virtual machine placement and migration recommendations that serve to enforce resource-based service level agreements, honor system- and user-specified constraints, and maintain load balance across the cluster even as workloads change. The best practices paper "Resource Management with VMware DRS" (see Resources for a link) provides material on VMware DRS usage and best practices.

VMware DPM saves power by dynamically right-sizing cluster capacity according to workload demands. VMware DPM recommends the evacuation and powering off of ESX hosts when both CPU and memory resources are lightly utilized. VMware DPM recommends powering ESX hosts back on when either CPU or memory resource utilization increases appropriately or additional host resources are needed to meet VMware HA or user-specified constraints. VMware DPM executes VMware DRS in a what-if mode to ensure its host power recommendations are consistent with the cluster constraints and objectives being managed by VMware DRS. VMware DPM is now a fully supported feature in VMware vSphere 4. However, please keep in mind that since VMware DRS does not automatically migrate VMware FT-enabled virtual machines, VMware DPM will not power off hosts running VMware FT virtual machines until they are manually migrated off using VMware® VMotion™. Consider placing VMware FT virtual machines on hosts that will not be considered for power off by VMware DPM.

VMware DPM Usage

The comprehensive "vSphere Resource Management Guide" (see Resources for a link) is the primary user guide for VMware DRS and VMware DPM. This section covers a subset of the contents of the guide to provide context for later sections that describe details of the VMware DPM algorithm.

Hosts Entering and Exiting Standby

Hosts powered off by VMware DPM are marked by vCenter Server as being in standby mode, indicating that they are available to be powered on whenever needed. VMware DPM operates on ESX 4 and even ESX 3.5 hosts that can be awakened from a powered-off (ACPI S5) state via either wake-on-LAN (WoL) packets or these out-of-band methods: Intelligent Platform Management Interface (IPMI) or HP Integrated Lights-Out (iLO) technology. WoL packets are sent over the VMware VMotion networking interface by another ESX 3.5 or 4 host in the cluster, so VMware DPM keeps at least one such host powered on at all times. IPMI and iLO require the Baseboard Management Controller (BMC) of the host to be properly configured. Before enabling VMware DPM on an ESX host, it is important to manually test the "exit standby" procedure for that host to ensure that it can be powered on successfully via one of the three protocols. This can be done using the vSphere Client. For wake-on-LAN compatibility requirements, see the VMware knowledge base article "Wake-On-LAN Compatibility as a Prerequisite for Distributed Power Management" (see Resources for a link).

The reason VMware DPM evacuates hosts and powers them down to the ACPI S5 state is that hosts typically use 60 percent or more of their peak power when totally idle, so the power savings possible with this approach are substantial. Vacating a host and placing it in a lighter sleep state than S5, such as the ACPI S3 "suspend-to-RAM" state, can consume an order of magnitude more power than ACPI S5 because of the need to keep the host's RAM powered on. For more information on the ACPI global power states, see Appendix: ACPI Global Power States or "Advanced Configuration and Power Interface Specification" (see Resources for a link).

After VMware DPM has determined the number of hosts needed to handle the load and to satisfy all relevant constraints and VMware DRS has distributed virtual machines across the hosts in keeping with resource allocation constraints and objectives, each individual powered-on host is free to handle power management of its hardware. For CPU power management, ESX 3.5 and 4 place idle CPUs in C1 halt state. ESX 4 also has support for host-level power-saving mechanisms through changing ACPI P-states; also known as dynamic voltage and frequency scaling (DVFS). DVFS runs CPUs at a lower speed and possibly at a lower voltage when there is sufficient excess capacity where the workload will not be affected. DVFS is "off" by default but can be turned on by setting the Power.CpuPolicy advanced option to "dynamic" for the hardware that supports it. Host-level power management is synergistic with VMware DPM. Even though it can provide additional power savings beyond VMware DPM, it cannot save as much power as VMware DPM does by powering hosts down completely.

Enabling and Disabling VMware DPM

VMware DPM leverages VMware DRS when it automatically migrates virtual machines away from ESX hosts that are to be powered off. Thus VMware DPM requires that you enable VMware DRS on clusters where it runs. To enable VMware DRS, right-click the cluster in vCenter Server and select **Edit Settings**. Check the box labeled **Enable VMware DRS** in the Cluster Settings dialog box.

You can then enable VMware DPM on the VMware DRS cluster. In the same Cluster Settings dialog box, select **Power Management** from the left pane as shown in Figure 1. VMware DPM is disabled (set to Off) by default. Enable VMware DPM by selecting either **Manual** or **Automatic**. In manual mode, execution of VMware DPM recommendations requires confirmation by the user. In automatic mode, VMware DPM recommendations are executed without user confirmation. Finally, you can set the **DPM Threshold** using the slider bar to be more conservative or more aggressive as described in the next section.
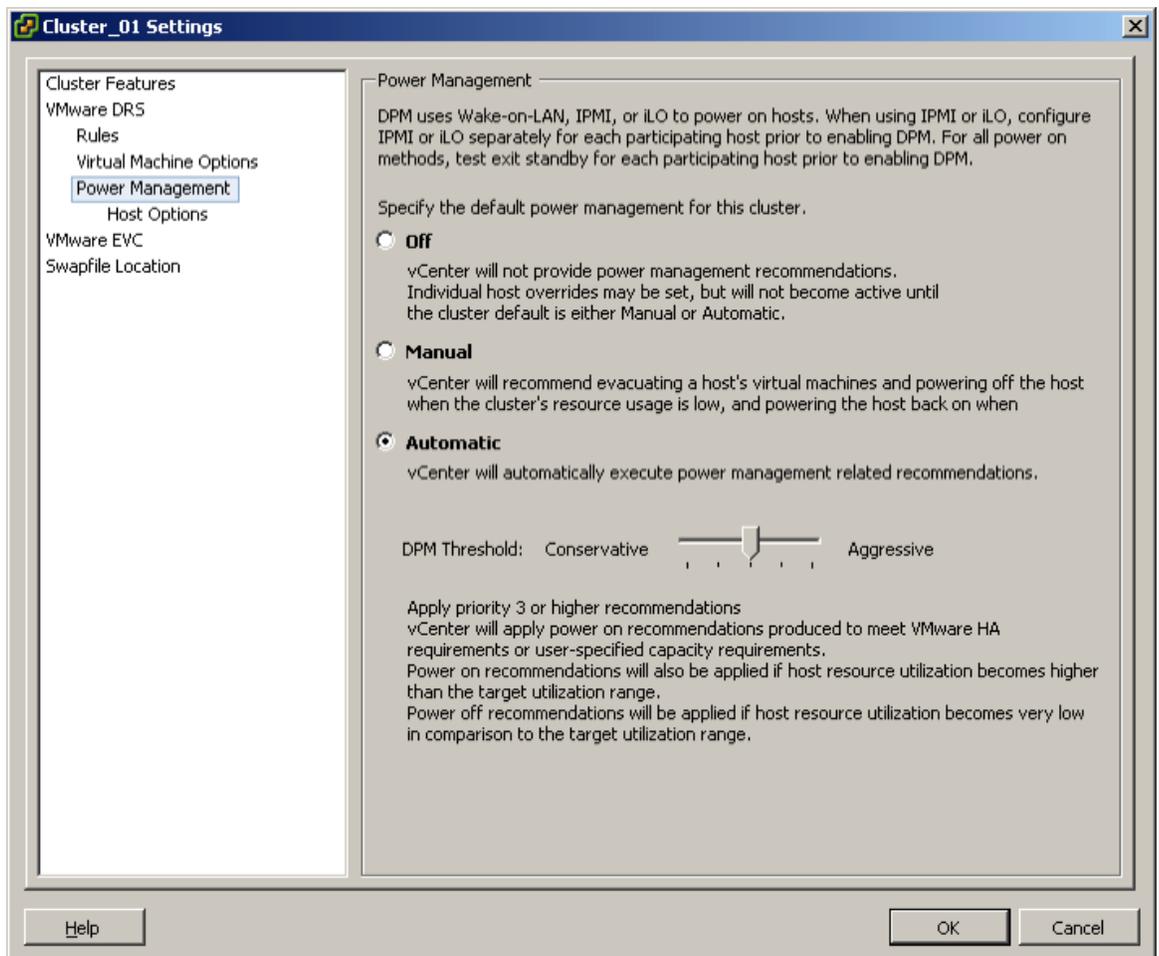

Figure 1. Enabling VMware DPM

VMware DPM Recommendation Rankings

VMware DPM power recommendations are assigned priority levels from priority 1 (most important) to priority 5 (least important). These levels signify the expected importance of particular recommendations given the current utilization of ESX hosts across the cluster and given the constraints, if any, on powered-on capacity.

Host power-off recommendations are ranked from priority 2 to priority 5. A high priority level (i.e., priority 2) for a power-off recommendation signifies a larger amount of underutilized, powered-on capacity in the cluster. Hence the recommendation with the higher priority level presents a more attractive opportunity for power savings.

Host power-on recommendations are rated as priority 1 to priority 3. Power-on recommendations generated to meet VMware HA or optional user-specified powered-on capacity requirements receive priority 1 rankings. Power-on recommendations produced to address high host utilization are rated as either priority 2 or priority 3, with the higher priority level indicating that overall host utilization is closer to saturation.

VMware DPM recommendation rankings are compared to the user-configured DPM Threshold as shown in Figure 1. VMware DPM, with **DPM Threshold** set at priority 1 (most conservative), would generate only the most important (i.e., priority 1) recommendations. Setting **DPM Threshold** at priority 5 (most aggressive) would generate all recommendations. The default **DPM Threshold** setting is priority 3 (middle level), so those VMware DPM recommendations below the threshold (i.e., priorities 4 and 5) are discarded. In manual mode, the VMware DPM power recommendations are displayed in the vSphere Client, allowing the VMware vSphere administrator to choose whether to execute them or not. In automatic mode, the recommendations are executed automatically. ESX host power-on recommendations produced by VMware DRS as prerequisites for VMware DRS migrations also receive priority levels. The VMware DRS priority levels do not correspond to VMware DPM priority levels and are governed by the setting for the VMware DRS migration threshold.

Setting Host Options

By default the VMware DPM automation setting applies to all hosts in the cluster. This is shown by selecting **Host Options** from the left pane of the Cluster Settings dialog box. **Use Cluster Default** is set for each host, but you can override this default setting on a per-host basis. For example, you should set any hosts in the cluster that cannot be powered on via WoL, IPMI, or iLO to **Disabled**. You should also set to Disabled any other hosts that you never want VMware DPM to power off. Other possible override settings are **Always Manual** and **Always Automatic**. These per-host override settings are meaningful only when VMware DPM is enabled for the cluster as a whole.

VMware DPM Operation

The goal of VMware DPM is to keep the utilization of ESX hosts in the cluster within a target range, subject to the constraints specified by the VMware DPM operating parameters and those associated with VMware HA and VMware DRS. VMware DPM evaluates recommending host power-on operations when there are hosts whose utilization is above this range and host power-off operations when there are hosts whose utilization is below it. Although this approach might seem relatively straightforward, there are key challenges that VMware DPM must overcome to be an effective power-saving solution. These include the following:

• Accurately assess workload resource demands. Overestimating can lead to less than ideal power savings. Underestimating can result in poor performance and violations of VMware DRS resource-level SLAs.

• Avoid powering servers on and off frequently, even if running workloads are highly variable. Powering servers on and off too often impairs performance because it requires superfluous VMotion operations.

• React rapidly to sudden increase in workload demands so that performance is not sacrificed when saving power.

- Select the appropriate hosts to power on or off. Powering off a larger host with numerous virtual machines might violate the target utilization range on one or more smaller hosts.

- Redistribute virtual machines intelligently after hosts are powered on and off by seamlessly leveraging VMware DRS.

The following subsections describe in detail the key components of the VMware DPM decision-making algorithm that address these challenges:

- Evaluating Utilization describes the method VMware DPM uses to evaluate host utilization.

- Host Power-off Recommendations and Host Power-off Cost-Benefit Analysis explain the procedure VMware DPM uses to determine when there is excess host capacity.

- Host Power-on Recommendations covers the process for ensuring that host capacity is powered on when needed.

- Host Sort for Power-on or Power-off Evaluation describes the order in which VMware DPM evaluates hosts to determine whether they should be powered on or powered off.

VMware DPM is run as part of the periodic VMware DRS invocation (every five minutes by default), immediately after the core VMware DRS cluster analysis and rebalancing step is complete. VMware DRS itself may recommend host power-on operations, if the additional capacity is needed as a prerequisite for migration recommendations to honor VMware HA or VMware DRS constraints, to handle user requests involving host evacuation (such as maintenance mode), or to place newly powered-on virtual machines.

Evaluating Utilization

VMware DPM evaluates the CPU and memory resource utilization of each ESX host and aims to keep the host's resource utilization within a target utilization range. VMware DPM may take appropriate action when the host's utilization falls outside the target range. The target utilization range is defined as:

```
Target resource utilization range = DemandCapacityRatioTarget ±
DemandCapacityRatioToleranceHost
```

By default, the utilization range is 45% to 81% (that is, 63% ±18%).

Each ESX host's resource utilization is calculated as demand/capacity for each resource (CPU and memory). In this calculation, demand is the total amount of the resource needed by the virtual machines currently running on the ESX host and capacity is the total amount of the resource currently available on the ESX host. A virtual machine's demand includes both its actual usage and an estimate of its unsatisfied demand, to account for cases in which the demand value is constrained by the ESX host's available resources. If an ESX host faces heavy contention for its resources, its demand can exceed 100 percent. VMware DPM computes actual memory usage using a statistical sampling estimate of the virtual machine's working set size. It also computes the estimate of unsatisfied demand for memory using a heuristic technique.

VMware DPM calculates an ESX host's resource demand as the aggregate demand over all the virtual machines running on that host. It calculates a virtual machine's demand as its average demand over a historical period of time plus two standard deviations (capped at the virtual machine's maximum demand observed over that period). Using a virtual machine's average demand over a period of time, rather than simply its current demand, is intended to ensure that the demand used in the calculation is not anomalous. This approach also smoothes out any intermediate demand spikes that might lead to powering hosts on and off too frequently. The default period of time VMware DPM evaluates when it calculates average demand that may lead to host power-on recommendations is the past 300 seconds (five minutes). When it calculates average demand for host power-off recommendations, the default period of time VMware DPM evaluates is the past 2400 seconds (40 minutes). The default time period for evaluating host power-on recommendations is shorter because rapid reactions to power on hosts are considered more important than rapid reactions to power off hosts. In other words, providing the necessary resources for workload demands has a higher priority than saving power.

If any host's CPU or memory resource utilization during the period evaluated for host power-on recommendations is above the target utilization range, VMware DPM evaluates powering hosts on. If any host's CPU and any host's memory resource utilization over the period evaluated for host power-off recommendations is below the target utilization range and there are no recommendations to power hosts on, VMware DPM evaluates powering hosts off.

In addition, when VMware DPM runs VMware DRS in what-if mode to evaluate the impact of host power-on and power-off, CPU and memory reservations are taken into account (as well as all other cluster constraints). VMware DRS will reject proposed host power-off recommendations that will violate reservations and VMware DRS will initiate host power-on operations to satisfy reservations.

Host Power-off Recommendations

If the host resource utilization evaluation described in Evaluating Utilization leads VMware DPM to evaluate recommending powering off a host to address low utilization, VMware DPM iterates through the powered on hosts to select one or more hosts that can be powered off to bring the hosts' utilization levels back up, ideally into the target utilization range. The order of iterating through the powered on hosts is described in Host Sort for Power-on or Power-off Evaluation.

For each powered on host, VMware DPM evaluates it as a candidate host and invokes VMware DRS on a theoretical scenario in which this candidate host is powered off in the cluster. To quantify the impact that powering off a candidate host would have in increasing the number of highly utilized hosts in the cluster or increasing host utilization levels, VMware DPM computes a low score value, which measures the amount of host underutilization, for each resource (CPU and memory). These are called cpuLowScore and memLowScore. Each value is computed as the sum of the weighted distances below the target utilization for only those hosts below that target.

```
{RESOURCE}LowScore = Sum across hosts below target utilization for RESOURCE of
(targetUtilization - hostUtilization)
```

In this calculation, RESOURCE is CPU or memory. The same formula is used for evaluating both resources.

VMware DPM compares the low score values for the cluster without the candidate host powered off to the low score values if the candidate host is powered off. If either the cpuLowScore value or the memLowScore value is improved for the cluster with the candidate host powered off and if the values of cpuHighScore and memHighScore (see Host Power-on Recommendations) for the resulting cluster are not worse than that with the host kept powered on, VMware DPM generates a recommendation to power off the host. This includes recommendations to first migrate virtual machines off that host. VMware DPM continues to iterate through the powered on candidate hosts for power-off evaluation as long as the cluster contains any hosts below the target utilization range for CPU and memory resources.

VMware DPM also evaluates three additional factors that affect placing a host in standby mode.

- VMware DPM does not recommend any host power-off operations (hence VMware DPM is effectively disabled) if the VMware DRS migration recommendation threshold is set to priority 1. With this setting, VMware DRS generates VMotion recommendations only to address constraint violations and not to rebalance virtual machines across hosts in the cluster, meaning that when VMware DPM runs VMware DRS in what-if mode to evaluate the impact of powering on a standby host, VMware DRS does not produce any non-mandatory recommendations to move virtual machines to those hosts.

- VMware DPM rejects powering down a host if it, by entering standby mode, would take the powered on capacity of the cluster below the specified minimum, MinPoweredOnCpuCapacity and MinPoweredOnMemCapacity (see Powered-on Capacity Options).

- VMware DPM does not power down a host if the conservatively projected benefit of placing that host into standby mode does not exceed by a specified multiplier the potential risk-adjusted cost of doing so, as described in Host Power-off Cost-Benefit Analysis.

Host Power-off Cost-Benefit Analysis

When VMware DPM powers off an ESX host, the operation has a number of potential associated costs such as the following:

- Cost of migrating any running virtual machines off of the associated host, such as the cost of CPU and memory required for VMotion.

- Loss of the host's resources during the power-down and subsequent power-on operations.

- Power consumed during the powering-down period.

- Loss of performance if the host's resources become needed to meet demand while the host is powered off.

- Power consumed during the powering-up period.

- Costs of migrating virtual machines back onto the host.

For each host evaluated for a power-off recommendation, VMware DPM compares these costs, taking into account an estimate of their associated risks, with a conservative projection of the power-savings benefit that can be obtained by powering off the host. This analysis step is called VMware DPM power-off cost-benefit.

VMware DPM compares the host power-off benefit to the host power-off cost, both measured as CPU resources and memory resources. VMware DPM cost-benefit analysis accepts a potential host power-off recommendation only if the benefit is greater than or equal to the cost multiplied by PowerPerformanceRatio (default of 40) for both resources, meaning only if there is significant benefit. (The term resource refers to both CPU and memory.) The power-off benefit and power-off cost is computed as follows.

The power-off benefit you achieve by powering off the candidate host is the resource savings during the estimated time when no additional hosts are required for power on. We call this time estimate StableOffTime. In other words, StableOffTime is the estimate of time while the candidate host is off before the overall utilization of the cluster is expected to significantly increase. The power-off benefit is actually computed as the aggregate of each resource over the StableOffTime. StableOffTime is computed as:

```
StableOffTime = ClusterStableTime - (HostEvacuationTime + HostPowerOffTime)
```

In other words, StableOffTime is the estimated time that the virtual machines' utilization will remain stable (ClusterStableTime), minus the time needed to power off the candidate host (HostEvacuationTime + HostPowerOffTime). The time to power off a host includes the time to evacuate its virtual machines as well as conduct an orderly shutdown. ClusterStableTime is the VMware DPM calculation of when the cluster will require one or more hosts to be powered on to satisfy increasing resource utilization. StableOffTime can be equal to or less than zero when ClusterStableTime is especially low. In that case, VMware DPM does not evaluate the candidate host for a power-off recommendation because no benefit would be realized.

The power-off cost is calculated as the summation of the following estimated resource costs:

- Migration of virtual machines off of the host before powering it off.

- Unsatisfied resource demand during host power on at the end of ClusterStableTime.

- Migration of virtual machines back onto hosts after the hosts are powered on.

To estimate the last two points above, VMware DPM computes the number of hosts that need to be powered on at the end of ClusterStableTime using a conservative projection that the demand of each virtual machine will rise to a high level. Specifically, this projection is the mean of each virtual machine's actual utilization over the previous 3600 seconds (60 minutes) plus three standard deviations.

Host Power-on Recommendations
If the evaluation of host resource utilization described in Evaluating Utilization leads VMware DPM to evaluate recommending host power-on operations to address high utilization, VMware DPM iterates through the standby hosts to select one or more hosts that can be powered on to bring the hosts' utilization levels back down, ideally into the target utilization range. The order of iterating through the standby hosts is described in Host Sort for Power-on or Power-off Evaluation.

For each standby host, VMware DPM evaluates it as a candidate host and invokes VMware DRS on a theoretical scenario in which this candidate host is powered on in the cluster. VMware DPM computes a high score value for each resource, called cpuHighScore and memHighScore, in the same fashion as it computed the low score values (see Host Power-off Recommendations). It computes each value as the sum of the weighted distance above the target utilization for only those hosts above that target. If either the cpuHighScore value or the memHighScore value is stably improved for the cluster with the candidate host powered on, VMware DPM generates a power-on recommendation for that candidate host.

In comparing the high score values, if the memory resource is overcommitted on hosts in the cluster, VMware DPM gives reduction in memory utilization higher importance than it gives impact on CPU resources. VMware DPM continues to iterate through the candidate hosts for power-on evaluation as long as there are any hosts in the cluster exceeding the target utilization range for either CPU or memory resources. For efficiency purposes, VMware DPM skips over any candidate host that it finds equivalent to another candidate host that was rejected in the same round for power-on evaluation. The current candidate host is equivalent to a prior candidate when they are both VMotion compatible with each other and when the current host has the same or fewer CPU and memory resources as the prior host.

VMware DPM then recommends powering on any additional hosts needed to reach a minimum amount of powered on CPU or memory resources, which is the maximum of:

•Any values specified by VMware HA.

•Values set by the user.

•Values defined by default.

The default minimum powered on CPU and memory resources (see Powered-on Capacity Options) are 1MHz and 1MB, respectively—that is, at least one host in the cluster is kept powered on. Hosts powered on solely to reach a specified minimum amount of CPU or memory resources are not needed to accommodate the virtual machines currently running in the cluster and may be idle.

Host Sort for Power-on or Power-off Evaluation
Typically, more than one ESX host is evaluated for VMware DPM power-on or power-off operations. This means that hosts must be sorted into a particular order for evaluation.

For both VMware DPM power-on and power-off operations, ESX hosts in VMware DPM automatic mode are evaluated before hosts in VMware DPM manual mode. Hosts at the same VMware DPM automation level are favored in order of CPU and memory capacity, with the more critical resource sorted before the other. Larger-capacity hosts are favored for power-on operations and smaller capacity hosts are favored for power-off operations. Hosts at the same automation level and capacity are evaluated for powering off in order of lower virtual machine evacuation cost. Ties are decided by randomizing the order for host power-off operations to spread the selection evenly across hosts. In the future, VMware DPM may evaluate other factors such as host power efficiency in determining host ordering for power-on or power-off evaluation.

The order in which hosts are evaluated by VMware DPM does not determine the actual order in which hosts are selected for power-on or power-off operations. This is simply an ordering of the candidate hosts so VMware DPM and VMware DRS can evaluate them as for powering on and off. A candidate host may be rejected for a number of reasons, based on VMware DRS operating constraints and objectives.

Some example situations limiting selection of a host to be powered off include constraints that make it impossible to evacuate all virtual machines from a candidate host or cases in which the virtual machines to be evacuated can be moved only to hosts that then become too heavily utilized.

Some example situations limiting selection of a host to be powered on include constraints that would prevent virtual machines from moving to a host if it were powered on or situations in which moving the virtual machines to a candidate host is not expected to reduce load on the highly utilized hosts in the cluster.

In addition, VMware DPM does not strictly adhere to its host sort order if doing so would lead to choosing a host with capacity far greater than needed, if a smaller capacity host that can adequately handle the demand is also available.

VMware DPM Advanced Options

The default settings of VMware DPM advanced options are intended to support higher-performing and power-efficient use of cluster resources. It is highly recommend that users do not alter the advanced options on production systems unless they understand the full impact of their actions.

You can change various settings relevant to VMware DPM operation using the VMware DRS advanced options interface. To access the advanced options interface right-click the cluster in vCenter Server and select **Edit Settings**. Select **VMware DRS** in the left pane, and then click Advanced Options at the lower right corner, as shown in Gaining Access to the Advanced Options (See Figure 2) . The **Advanced Options** dialog box appears allowing you to enter option names and option values. In vCenter Server 4, the advanced options will revert to their default values if they are cleared out. Previous to vCenter Server 4, the advanced options had to be explicitly reset to their default values.
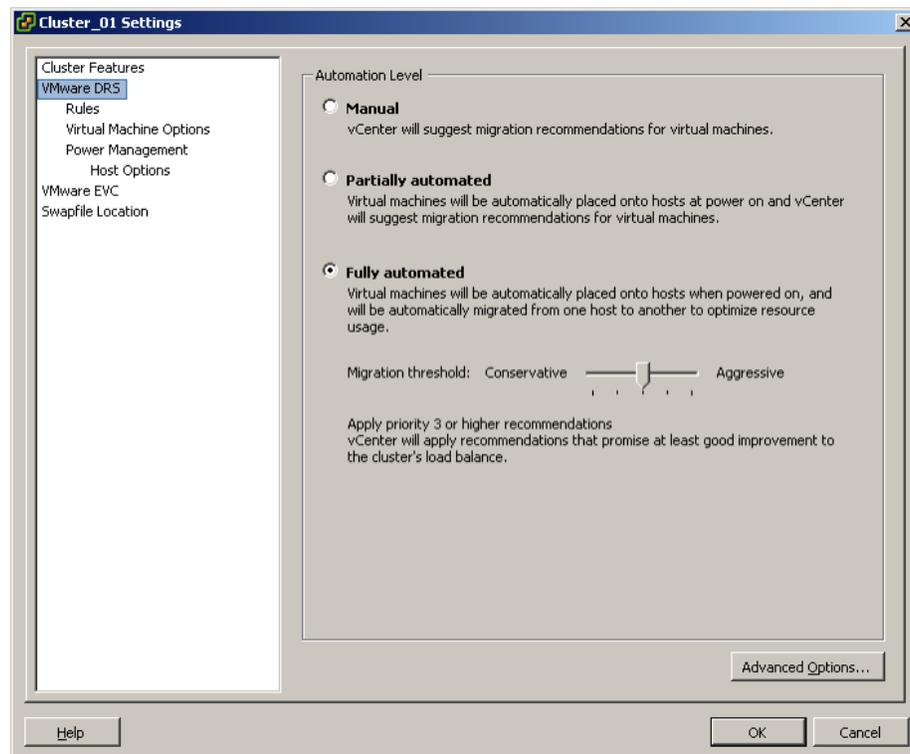

Figure 2. Gaining Access to the Advanced Options

Utilization Options

The utilization options shown in Utilization Options and Settings set the CPU and memory resource utilization targets for each ESX host over a period of time.

OPTION NAME	DEFAULT VALUE	VALUE RANGE
DemandCapacityRatioTarget **Description**: Utilization target for each ESX host.	63%	40 to 90
DemandCapacityRatioToleranceHost **Description**: Tolerance around the utilization target for each ESX host. Used to calculate the utilization range.	18% Note: Default utilization range is 63±18.	10 to 40
VmDemandHistorySecsHostOn **Description**: Period of demand history evaluated for ESX host power-on recommendations. If you increase this value, you risk not powering on hosts in time to address increases in resource demand, thus potentially impacting virtual machine performance.	300 seconds	0 to 3600
VmDemandHistorySecsHostOff **Description**: Period of demand history evaluated for ESX host power-off "recommendations.	2400 seconds	0 to 3600

Table 1. Utilization Options and Settings

Power-off Cost-Benefit Options

The power-off cost-benefit options shown in Power-off Cost-Benefit Options and Settings enable and set the parameters used to analyze when to power off ESX hosts.

OPTION NAME	DEFAULT VALUE	VALUE RANGE
EnablePowerPerformance **Description**: Set to 1 to enable host power-off cost-benefit analysis. Set to 0 to disable it.	1	0 or 1
PowerPerformanceRatio **Description**: Multiplier by which benefit must meet or exceed performance impact. Used by power-off cost-benefit analysis.	40	0 to 500
PowerPerformanceHistorySecs **Description**: Period of demand history evaluated for power-off cost-benefit recommendations.	3600 seconds	0 to 3600
HostsMinUptimeSecs **Description**: Minimum uptime of all hosts before VMware DPM will consider any host as a power-off candidate.	600 seconds	0 to maxint

Table 2. Power-off Cost-Benefit Options and Settings

Powered-on Capacity Options

You do not need to specify a minimum amount of powered-on capacity for correct VMware DRS cluster operation. VMware DRS and VMware DPM recommend that appropriate hosts be powered on when needed and keep hosts powered on to respect VMware HA failover settings, if VMware HA strict admission control is enabled. Your particular usage scenario might make disabling VMware HA strict admission control desirable. You should be aware, however, that doing so (when interoperating with VMware DRS and VMware DPM) can lead to a reduced level of availability and failover protection for your cluster. For more information, see the VMware knowledge base article "Implications of enabling or disabling VMware HA strict admission control when using VMware DRS and VMware DPM" (see Resources for a link).

You can use the advanced options in Powered-on Capacity Options and Settings to specify that a particular minimum amount of CPU or memory capacity, or both, be kept powered on, even when that capacity is not deemed necessary by VMware DRS and VMware DPM. The host capacity kept powered on to satisfy these options is not necessarily compatible with the future needs of some arbitrary virtual machine (for example, it may not match the required CPU characteristics), so these options are most useful in clusters of similar hosts that are compatible with the majority of virtual machines.

OPTION NAME	DEFAULT VALUE	VALUE RANGE
MinPoweredOnCpuCapacity **Description**: Minimum amount of powered-on CPU capacity maintained by VMware DPM.	1MHz	0 to maxint
MinPoweredOnMemCapacity **Description**: Minimum amount of powered-on memory capacity maintained by VMware DPM.	1MB	0 to maxint

Table 3. Powered-on Capacity Options and Settings

VMware DPM and Datacenter Monitoring Tools

Many datacenters use monitoring tools to track the health of datacenter entities such as servers, switches, storage, and cooling units. Some common monitoring tools are:

• BMC Performance Manager (formerly PATROL)

• HP Software (formerly OpenView)

• CA Virtual Performance Management (formerly Unicenter Advanced Systems Management)

• Microsoft System Center Operations Manager

• IBM Tivoli

These monitoring tools are typically set to provide an alarm if a server fails. These alarms are critical in informing datacenter administrators of potential problems that they may need to address.

Server shutdowns controlled by VMware DPM are initiated to reduce power consumption and might falsely appear to monitoring tools as server failures. Although VMware DPM server power-off operations are planned events, whereas server failures are unplanned events, some monitoring tools might not be able to distinguish between them. You probably do not want alarms to be triggered by VMware DPM actions.

The following monitoring logic should be implemented in the datacenter monitoring tool to become aware of VMware DPM:

• When an ESX host is powered off into standby mode (ACPI S5 soft-off), suppress all alarms related to the host being down or unavailable. Alternatively, mark these alarms as not requiring any action or as informational only.

This logic can be implemented using VMware vSphere API 4.0. Specifically, information on VMware DPM actions can be obtained by subscribing to vCenter Server events related to VMware DPM. vCenter Server Events Related to VMware DPM describes the relevant events in greater detail.

• Trigger an alarm when VMware DPM attempts to power on a host and fails. Successfully exiting standby mode by a host is important in order to meet capacity needs of the VMware DRS cluster. Monitor the vCenter Server event called DrsExitStandbyModeFailedEvent for host power-on failures. You should respond by powering the host on using a different method and then disabling VMware DPM on that host until you have resolved the issue.

VMware has been working with the major software management vendors to incorporate this logic into their products. Please contact your monitoring software vendor for information on when this logic will be supported in their products.

DATA OBJECT NAME	DATA OBJECT DESCRIPTION
DrsEnteringStandbyModeEvent	This event records that a host has begun the process of entering standby mode initiated by VMware DPM.
DrsEnteredStandbyModeEvent	This event records that VMware DPM has successfully brought the host into standby mode. A host in this mode has no running virtual machines and no provisioning operations are occurring.
DrsExitingStandbyModeEvent	This event records that a host has begun the process of exiting standby mode initiated by VMware DPM.
DrsExitedStandbyModeEvent	This event records that VMware DPM has successfully brought the host out of standby mode.
DrsExitStandbyModeFailedEvent	This event records that VMware DPM tried to bring a host out from standby mode, but the host failed to exit standby mode.

Table 4. vCenter Server Events Related to VMware DPM

For more information on using these data objects, see the VMware APIs and SDKs Documentation page on the VMware Website. See Resources for link.

VMware DPM Usage Scenario

VMware DPM enables you to power on and shut down hosts automatically. The proper configuration and usage of VMware DPM depend on your power consumption and service level goals.

The basic way to use VMware DPM is to power on and shut down ESX hosts based on utilization patterns during a typical workday or workweek. For example, services such as email, fax, intranet, and database queries are used more during typical business hours from 9 a.m. to 5 p.m. At other times, utilization levels for these services can dip considerably, leaving most of the hosts underutilized. Their main work during these off hours might be performing off-hours backup, archiving, and servicing overseas requests. In this case, consolidating virtual machines onto few hosts and shutting down unneeded hosts during off hours reduces power consumption.

To achieve this usage scenario, consider taking one of the following general approaches:

•Set VMware DPM to automatic mode and let the VMware DPM algorithm dictate when hosts are powered on and shut down. Keep the advanced options set at their default values.

•Tune VMware DPM to be more conservative or more aggressive by shifting the DPM Threshold in the Cluster Settings dialog box or the DemandCapacityRatioTarget advanced option. If VMware DPM performs host power-off and power-on operations too aggressively, shift the priority level of the DPM Threshold to be more conservative—for example, shift it from the default value of priority 3 to a more conservative setting priority 2. VMware DPM would then recommend only the most important (i.e., priority 1 and 2) host power-off and power-on operations.

To save more power by increasing host utilization (that is, consolidating more virtual machines onto few hosts), increase the value of DemandCapacityRatioTarget from the default of 63 percent to, for example, 70 percent. From the guidelines in Evaluating Utilization, the target resource utilization range becomes 52 percent to 88 percent. Using the modified setting, when utilization exceeds 88 percent, more hosts are needed. When utilization dips below 52 percent, VMware DPM can consolidate workloads onto fewer hosts and allow unneeded hosts to be shut down.

•Use VMware DPM to force the powering on of all hosts before business hours and then selectively shut down hosts after the peak workload period. This is a more proactive approach that would avoid any performance impact of waiting for VMware DPM to power on hosts in response to sudden spikes in workload demand.

To take this approach, use VMware vSphere API 4.0 to schedule a task that sets the MinPoweredOnCpuCapacity advanced option to the full cluster CPU capacity in advance of business hours, for example at 8 a.m. This causes VMware DPM to power hosts on before the initial spike at the start of business at 9 a.m. Use the powerOnHosts.pl script found at "Scripts for Proactive DPM" (see Resources for a link) for this task.

Schedule another task to reset MinPoweredOnCpuCapacity shortly after the initial peak period, for example at 10 a.m. This task restores normal VMware DPM behavior back to its previous settings for the remainder of business hours. Use the enableDPM.pl script found at "Scripts for Proactive DPM" (see Resources for a link) for this task. Or, more conservatively, use this script to restore normal VMware DPM behavior at the end of business hours to allow host to be shut down only during off hours.

"Scripts for Proactive DRS" (see Resources for a link) can be used for an expected steep increase in virtual machine demand. These scripts can be used concurrently with scripts for proactive DPM.

The "VMware Distributed Power Mgmt (DPM)" video (see Resources for a link) demonstrates the effects of powering off hosts during low utilization hours and powering them back on during business hours. The demonstration shows virtual machines running at full performance while the data center saves 55 percent in overall power using default VMware DPM advanced option settings. You can automate host power-on and power-off operations with VMware DPM in automatic mode or on a scheduled basis.

Appendix: ACPI Global Power States

ACPI Global Power States shows the global power states defined in the ACPI specification. See Resources for the link to the full specification.

GLOBAL SYSTEM STATE	POWER CONSUMPTION	SOFTWARE RUNS	OS RESTART REQUIRED	COMMENTS
G0 (S0) - Working	Large	Yes	No	Normal running conditions.
G1 (S1-S4) - Sleeping	Smaller	No	No	S1: CPU caches flushed and execution halted. Power is maintained.S2: CPU off.S3: Suspend to RAM.S4: Hibernation.
G2 (S5) - Soft-off	Very near 0	No	Yes	Similar to G3, but some components remain powered on to enable wake-on-LAN.
G3 - Mechanical Off	Real-time clock battery only	No	Yes	Safe to pull plug and disassemble.

Table 5. ACPI Global Power States

Resources

• "Advanced Configuration and Power Interface Specification"
http://www.acpi.info/DOWNLOADS/ACPIspec30b.pdf

• "Implications of enabling or disabling VMware HA strict admission control when using DRS and VMware DPM"
http://kb.vmware.com/kb/1007006

• "Resource Management with VMware DRS"
http://www.vmware.com/resources/techresources/401

• "Scripts for Proactive DPM"
http://communities.vmware.com/docs/DOC-10230

• "Scripts for Proactive DRS"
http://communities.vmware.com/docs/DOC-10231

• VMware APIs and SDKs Documentation
http://www.vmware.com/support/pubs/sdk_pubs.html

• "VMware Distributed Power Mgmt (DPM)" video
http://www.youtube.com/watch?v=7CbRS0GGuNc

- "VMware vSphere API Reference Documentation"
  http://www.vmware.com/support/developer/vc-sdk/visdk400pubs/ReferenceGuide/index.html

- "vSphere Availability Guide"
  http://www.vmware.com/pdf/vsphere4/r40/vsp_40_availability.pdf

- "vSphere Resource Management Guide"
  http://www.vmware.com/pdf/vsphere4/r40/vsp_40_resource_mgmt.pdf

- "Wake-On-LAN Compatibility as a Prerequisite for Distributed Power Management"http://kb.vmware.com/kb/1003373