5. vSphere DRS

vSphere Distributed Resource Scheduler

5.1 vSphere DRS

Introduction

VMware vSphere Distributed Resource Scheduler (DRS) is a resource management solution for vSphere clusters that allows IT organizations to deliver optimized performance of application workloads.

The primary goal of DRS is to ensure that workloads receive the resources they need to run efficiently. DRS determines the current resource demand of workloads and the current resource availability of the ESXi host that are grouped into a single vSphere cluster. DRS provides recommendations throughout the life-cycle of the workload. From the moment, it is powered-on, to the moment it is powered-down.

DRS operations consist of generating initial placements and load balancing recommendations based on resource demand, business policies and energy-saving settings. It is able to automatically execute the initial placement and load balancing operations without any human interaction, allowing IT-organizations to focus their attention elsewhere.

DRS provides several additional benefits to IT operations:

- Day-to-day IT operations are simplified as staff members are less affected by localized events and dynamic changes in their environment. Loads on individual virtual machines invariably change, but automatic resource optimization and relocation of virtual machines reduce the need for administrators to respond, allowing them to focus on the broader, higher-level tasks of managing their infrastructure.
- DRS simplifies the job of handling new applications and adding new virtual machines. Starting up new virtual machines to run new applications becomes more of a task of high-level resource planning and determining overall resource requirements, than needing to reconfigure and adjust virtual machines settings on individual ESXi hosts.
- DRS simplifies the task of extracting or removing hardware when it is no longer needed or replacing older host machines with newer and larger capacity hardware.
- DRS simplifies grouping of virtual machines to separate workloads for availability requirements or unite virtual machines on the same ESXi host machine for increased performance or to reduce licensing costs while maintaining mobility.

5.2 vSphere Cluster

vSphere Cluster

DRS uses a vSphere cluster as management construct and supports up to 64 ESXi hosts in a single cluster. A vSphere cluster loosely-connects multiple ESXi hosts together and allows for adding and removing resource capacity to a cluster without causing service disruptions to the active workload.

DRS generates recommendations for initial placement of virtual machines on suitable ESXi hosts during power-on operations and generates load balancing recommendations for active workloads between ESXi hosts within the vSphere cluster. DRS leverages VMware vMotion technology for live-migration of virtual machines.

DRS responds to cluster and workload scaling operations and automatically generates resource relocation and optimization decisions as ESXi hosts or virtual machines are added or removed from the cluster. To enable the use of DRS migration recommendations, the ESXi hosts in the vSphere cluster must be part of a vMotion network. If the ESXi hosts are connected to the vMotion network, DRS can still make initial placement recommendations.

Clusters can consist of heterogeneous or homogeneous hardware configured ESXi hosts. ESXi hosts in a cluster can differ in capacity size. DRS allows hosts that have a different number of CPU packages or CPU cores, different memory or network capacity, but also different CPU generations. VMware Enhanced vMotion Compatibility (EVC) allows DRS to live migrate virtual machines between ESXi hosts with different CPU configurations of the same CPU vendor. DRS leverages EVC to manage placement and migration of virtual machines that have Fault Tolerance (FT) enabled.

DRS provides the ability contain virtual machines on selected hosts within the cluster by using VM to Host affinity groups for performance or availability purposes. DRS resource pools allow compartmentalizing cluster CPU and memory resources. A resource pool hierarchy allows resource isolation between resource pools and simultaneous optimal resource sharing within resource pools.

Resource Pools

DRS allows to abstract cluster resources in separate resource pools. This allows IT organizations to isolate resources between resource pools. Resource pools can be used as a unit of access control and delegation, allowing the assigned teams to perform all virtual machine creation and management functions within the boundaries of the resource pool. Resource Pools allow for further separation of resources from hardware. If the cluster is expanded with new resources, by adding new ESXi hosts or scaling-up existing ESXi hosts, the allocated resources remain the same. This separation allows IT organizations to think more about aggregate computing capacity and less about individual hosts.

Distribution of resources amongst resource pools in the cluster is based on the reservation, shares and limit settings of the resource pool and the activity of the child virtual machines within the resource pool and the other sibling resource pools. It is beyond the scope of this paper to expand on this behavior. A separate resource pool whitepaper is published in 2018.

Correct use: Resource pools are an excellent construct to isolate a particular amount of resources for a group of virtual machines without having to micro-manage resource setting for each individual virtual machine. A reservation set at the resource pool level guarantees each virtual machine inside the resource pool access to these resources. Depending on the activity of these virtual machines these virtual machines can operate without any contention.

Incorrect use: Resource pools should not be used as a form of folders within the inventory view of the cluster. Resource pools consume resources from the cluster and distribute these amongst its child objects within the resource pool, this can be additional resource pools and virtual machines. Due to the isolation of resources, using resource pools as folders in a heavily utilized vSphere cluster can lead to an unintended level of performance degradation for some virtual machines inside or outside the resource pool.

Maintenance Mode

Maintenance mode allows IT operation teams to evacuate active workloads off an ESXi host in order to perform maintenance task without interrupting any service. If the ESXi host is a part of a vSphere cluster with DRS enabled, DRS will generate migration recommendations when the ESXi host is placed into maintenance mode. These migration recommendations are based on the currently available resources and the virtual machine demand. DRS aims to distribute the virtual machines across the remaining hosts in the cluster and attempts to provide the resources the workloads require. Depending on the DRS automation levels, the migration recommendations and the live migrations are executed by the IT operations team or autonomously by DRS itself. DRS respects affinity and anti-affinity rules while generating migration recommendations and can impact whether or not a compatible host can be found for the evacuating virtual machines.

5.3 Distributed Resource Scheduling Operations

DRS interoperates with VMware vCenter to provide an overview and management of all resources in the vSphere cluster. A global scheduler runs within vCenter that monitors resource allocation of all

virtual machines and vSphere Integrated Containers running on ESXi hosts that are part of the vSphere cluster.

During the power-on operation of a virtual machine, DRS provides an initial placement recommendation based on the current ESXi host resource consumption. The global scheduling process (DRS invocation) runs every 5 minutes within vCenter and determines the resource load on the ESXi hosts and the virtual machine resource demand. DRS generates recommendations for load balancing operations to improve overall ESXi host resource consumption.

DRS automation levels allow the IT operation team to configure the level of autonomy of DRS.

DRS Automation Levels

Three levels of automation are available, allowing DRS to provide recommendations for initial placement and load balancing operations. DRS can operate in manual mode, partially automated mode and fully automated mode. Allowing the IT operation team to be fully in-control or allow DRS to operate without the requirement of human interaction.

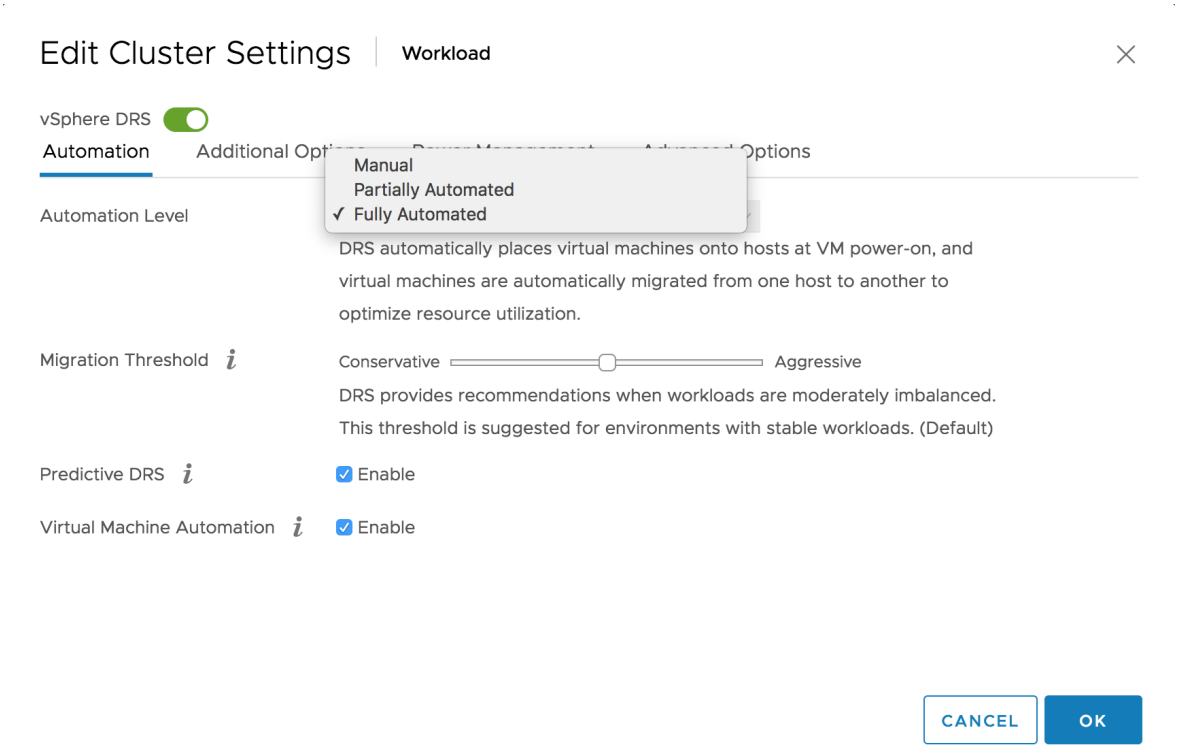


Figure 1: DRS Automation Level

Manual Automation Level

The manual automation level expects the IT operation team to be in complete control. DRS generates initial placement and load balancing recommendations and the IT operation team can choose to ignore the recommendation or to carry out any recommendations.

If a virtual machine is powered-on on a DRS enabled cluster, DRS presents a list of mutually exclusive initial placement recommendations for the virtual machine. If a cluster imbalance is detected during a DRS invocation, DRS presents a list of recommendations of virtual machine migrations to improve the cluster balance. With each subsequent DRS invocation, the state of the cluster is recalculated and a new list of recommendations could be generated.

Partially Automated Level

DRS generates initial placement recommendations and executes them automatically. DRS generates load balancing operations for the IT operation teams to review and execute. Please note that the introduction of a new workload can impact current active workload, which may result in DRS generating load balancing recommendations. It is recommended to review the DRS recommendation list after power-on operations if the DRS cluster is configured to operate in partially automated mode.

Fully Automated Level

DRS operates autonomous in fully automated level mode and requires no human interaction. DRS generates initial placement and load balancing recommendations and executes these automatically. Please note that the migration threshold setting configures the aggressiveness of load balancing migrations.

Per-VM Automation Level

DRS allows Per-VM automation level to customize the automation level for individual virtual machines to override the cluster's default automation level. This allows IT operation teams to still benefit from DRS at the cluster level while isolating particular virtual machines. This can be helpful if some virtual machines are not allowed to move due to licensing or strict performance requirement. DRS still considers the load utilization and requirements to meet the demand of these virtual machines during load balancing and initial placement operations, it just doesn't move them around anymore.

Automation level	Initial Placement	Load Balancing
Manual	Recommended host(s) displayed	Migration recommendation is displayed
Partially Automated	Automatic placement	Migration recommendation is displayed
Fully Automated	Automatic placement	Automatic migration

Table 1: DRS Automation Level Operations

5.4 DRS Migration Threshold

The DRS Migration threshold controls how much imbalance across the ESXi hosts in the cluster is acceptable based on CPU and memory loads. The threshold slider ranges from conservative to aggressive. The more conservative the setting, the more imbalance DRS tolerates between ESXi hosts. The DRS migration threshold impacts the selection of initial placement and load-balance migrations recommendation and pair-wise balance threshold.

Recommendations

During the invocation of DRS, it calculates the imbalance of resource utilization in the cluster and determines which migration of virtual machines can solve the imbalance. To filter ineffective migration recommendations, DRS assigns a priority level to each recommendation. The priority level of the migration recommendation is compared to the migration threshold. If the priority level is less than or equal to the migration threshold, the recommendation is displayed or applied, depending on the automation level of the cluster. If the priority level is above the migration threshold, the recommendations are either not displayed or discarded.

Migration Priority Ratings

DRS uses a scale from priority 1 to 5, in which priority 1 is the highest ranked rating and 5 is the lowest ranked rating.

Priority 1 recommendations are only generated to solve (anti-) affinity rule violations or comply with a maintenance mode request. Priority 1 recommendations are not generated to solve cluster imbalance or virtual machine demand.

Priority 2 to 5 recommendations are used to solve the imbalance of the cluster based. DRS applies a cost, benefit and risk analysis to ensure that migration operations are truly worthwhile.

Cost-Benefit Analysis

DRS is designed to create as little as overhead as possible. Each migration consumes resources. CPU resources are used on both the source and destination host to move memory of the virtual machine. This memory needs to be placed and can possibly displace memory of the currently active virtual machines on the destination host. The transfer of memory consumes network bandwidth. For these reasons, DRS only migrates virtual machines if the benefit outweighs the costs. The more significant the benefit, the higher the priority rating. DRS uses the migration threshold as a filter to determine which priority levels to consider.

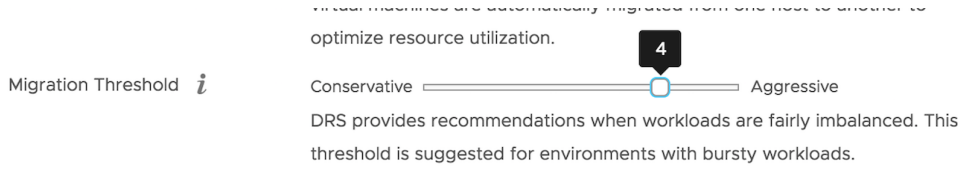


Figure 2: DRS Migration Threshold Setting

Migration Threshold 1 (Conservative)

DRS will only apply recommendations with a priority 1 recommendation. That means that DRS only migrates virtual machines to satisfy cluster constraints like affinity rules and host maintenance. As a result, DRS will not correct host imbalance at this threshold, i.e. no live migrations will be triggered by DRS. Mandatory moves are issued when:

- The ESXi host enters maintenance mode
- The ESXi host enters standby mode
- An (anti-) affinity is violated
- The sum of the reservations of the virtual machines exceeds the capacity of the host

Migration Threshold 2

DRS only provides recommendations when workloads are extremely imbalanced or virtual machine demand is not being satisfied on the current host. DRS considers priority 1 and 2 recommendations. This threshold is suggested for environments that apply a conservative approach to virtual machine mobility. Please note that VM happiness can be impacted by selecting this threshold.

Migration Threshold 3 (Default)

DRS provides recommendations when workloads are moderately imbalanced. DRS considers priority 1,2 and 3 recommendations. This threshold is suggested for environments with a balanced mix of stable and bursty workloads. Many workloads are active at different times and with a different cadence, this setting allows ESXi hosts to cope with the variety of load while avoiding DRS to consume resources for migration. If the load is moderately imbalanced, DRS will take action.

Migration Threshold 4

DRS provides recommendations when workloads are fairly imbalanced. DRS considers priority 1 to 4 recommendations. This threshold is suggested for environments with a high number of bursty workloads.

Migration Threshold 5 (Aggressive)

DRS provides recommendations when workloads are even slightly imbalanced and marginal improvement may be realized. DRS considers priority 1 to 5 recommendations. For dynamic workloads, this may generate frequent load balancing recommendations. This setting can be helpful for clusters that contain primarily bursty CPU-bound workloads.

For more information about this metric and how a recommendation priority level is calculated, see the VMware Knowledge Base article "Calculating the priority level of a VMware DRS migration recommendation." https://kb.vmware.com/s/article/1007485

Pair-Wise Balancing Thresholds

vSphere 6.5 DRS applies an additional load balancing criterion to minimize the load difference between the most utilized and least utilized host pair in the vSphere cluster. The maximum allowed CPU or memory load difference between the most utilized and the least utilized host depends on the cluster migration threshold setting. This is called pair-wise balancing.

For the default migration threshold level of 3, the tolerable pair-wise balance threshold is 20%. If two hosts in the cluster whose CPU or memory usage difference is more than 20%, pair-wise balancing will try to clear the difference by moving virtual machines from a highly utilized host to a lightly utilized host. Please note that DRS might select a virtual machine from any other host to solve the overall load imbalance, it is not restricted to move virtual machines from the most utilized host to the least utilized host. The following table explains how much imbalance will be tolerated for different levels of migration threshold:

Migration Threshold Level	Priority Recommendations	Tolerable Resource Utilization Difference between ESXi Host Pair
1 (Conservative)	1	Not available
2	1,2	30%
3 (Default)	1,2,3	20%
4	1,2,3,4	10%
5	1,2,3,4,5	5%

Table 2: DRS Migration Threshold Level

5.5 DRS Decision Engine

vSphere 6.5 DRS is network-aware and considers CPU, memory and network utilization of ESXi hosts and the CPU, memory and network demand and requirements of virtual machines into account for load balancing and initial placement operations.

During a DRS invocation, DRS generates a list of suitable hosts to run a virtual machine based on CPU and memory resources first, it will determine if any constraints are in play and lastly looks at the ESXi host network utilization before generating a list of preferred ESXi hosts. DRS applies the cost-benefit analysis to understand a move should be made or should not be made.

By adding ESXi host network utilization and virtual machine network requirements to the DRS decision engine, DRS is able to make sure the ESXi host has sufficient network resources to satisfy the requirements of the virtual machine. Please note that network utilization and network requirements are first-class citizens of the load balancing algorithm yet. If DRS determines the vSphere cluster is imbalanced on either CPU or memory, DRS migrates virtual machines around to solve this imbalance, it will not trigger any vMotion if the network utilization within the cluster is imbalanced.

Measured Metrics

In order to get an accurate view of the state of the resource demand and supply state inside the cluster, DRS collect host-level and virtual machine-level metrics every minute. Each host provides an average of three separate 20-second statistics. To provide an extensive list of all the metrics that are collected and monitored is beyond the scope of this paper. However, the key metrics DRS looks into are:

Host-level Resource Reservations

At the host level, DRS looks at the CPU and memory reservations made by the system itself, to ensure the proper execution of critical agents, such as the Fault Domain Manager agent (vSphere High Availability). Please note that these reservations are not the reservations specified on the virtual machines.

Host-level Resource Utilization

DRS collects the utilization metrics of the host. CPU, memory and network utilization. DRS sums the active CPU and memory consumption of the virtual machines per host. The network utilization percentage of a host is the average capacity that is being utilized across all the physical NICs (pNICs) on that host, if the utilization is above 80%, DRS does not consider this host as a valid destination for initial placement and load balancing operations.

Important VM Level Metrics

- CPU active (run, ready and peak)
- Memory overhead (growth rate)
- Active, Consumed and Idle Memory
- Shared memory pages, balloon, swapped

The most important metric to determine a virtual machines' CPU demand is CPU active. CPU active is a collection of multiple stats all morphed into a single stat. One important statistic that is a part of CPU active is CPU ready time. DRS takes ready time into account to understand the demand of the virtual machine. On top of this, DRS considers both peak active as average active in the past five minutes.

The most important metric to determine a virtual machines' memory demand is the active memory and the consumed memory. Another metric that is considered is the page sharing between multiple virtual machines running on the same host, whenever DRS makes a decision it knows about how pages are being shared between their respective virtual machines on that particular host. If a virtual machine is moved away from this host, DRS takes into account the loss of page sharing. This is one of the main reasons, why DRS prefers to move medium-sized workloads over larger sized workloads. Moving a virtual machine to a destination could force the ESXi host to reclaim memory to make room for the in-

transit virtual machine. But there is one metric that supersedes all metrics listed above, and that is virtual machine happiness.

Virtual Machine Happiness

If the CPU and memory demand of the virtual machine is satisfied the entire time, DRS considers that virtual machine to be happy. If a virtual machine is happy why move it? It receives the resources the application demands. Moving a virtual machine to another host would provide exactly the same result if that ESXi host is also capable of satisfying the virtual machine demand. Why move it if there is a slight imbalance in the cluster?

The move of a virtual machine consumes resources that could otherwise be provided to satisfy another virtual machine (application) demand. DRS is designed to avoid incurring cost on the virtual infrastructure.

The virtual machine happiness metric is considered during initial placement and load balancing operations. During initial placement, DRS assess whether placing a particular virtual machine will have any negative impact on the already running virtual machines. Initial placement is not only about trying to power on a virtual machine and finding a good place for its application to perform; it's critical to ensure that the already running virtual machines experience zero or minimal impact. Load balancing follows the same principle when a virtual machine is moved to another host, DRS ensures that the VMs already running on the destination host are not impacted by this incoming virtual machine

5.6 DRS Operation Constraints

In a perfect world, virtual machines could move to any ESXi host in the cluster. However, certain user-configured settings, cluster design or temporary error states can impact initial placement and load balancing operations. There are two types of constraints, explicit and implicit. Explicit constraints are created by user-input, while implicit constraints occur by hardware failure or infrastructure of software limitations

Explicit Constraint

Resource Allocation Settings

vSphere allows IT operation teams to specify the importance of virtual machine or resource pools by setting resource allocation settings on CPU and memory. Reservations, shares and limits, together with the workload activity define the resource entitlement of the virtual machine. The resource entitlement is the primary metric for DRS to establish the happiness level of the virtual machine.

CPU, memory and network reservations define the minimum requirement of the virtual machine to operate. In order to ensure the minimum requirement is met, a process called admission control is active. In a vSphere cluster that has VMware High Availability (HA) and DRS enabled, three admission control processes are operational.

1. The VMware HA admission control ensures enough resources are available to satisfy the minimum requirements of the virtual machines after the configured host failure or percentage resource loss occurs.
2. The DRS admission control determines whether enough unreserved resources are available in the cluster and/or resource pool.
3. The ESXi host admission control determines whether it has enough available unreserved resources are available to run the virtual machine.

The ESXi host admission control informs DRS if it is unable to satisfy the virtual machine requirement and DRS selects another ESXi host for initial placement and load balancing operations. For more information about reservations, limits, and shares, please consult the vSphere 6.5 Resource Management Guide .

During an HA failover, the process that occurs when an ESXi host has failed, HA consults DRS to place the virtual machines on the most suitable ESXi host. The virtual machines are placed on ESXi hosts that are able to satisfy the requirements defined by resource reservation, yet DRS still attempts to take VM happiness into account as much as possible. If a virtual machine needs to be restarted with a very large reservation, it could happen that not a single host in the cluster can satisfy this large reservation. This state is referred to as resource fragmentation. DRS attempts to migrate one or more virtual machines across the cluster to make room for the virtual machine with the large resource reservation.

Please note that virtual machine overhead reservation, the memory required by the ESXi host to run the virtual machine itself, is added on top of the virtual machine memory reservation.

Affinity Rules

DRS allows IT operation teams to control the placement of virtual machines on hosts within a cluster by using affinity rules. Affinity rules constrain the placement decisions of DRS. In essence, all rule sets restrict the number of placement possibilities of virtual machines on ESXi hosts within the cluster. It is highly recommended to reduce the number of affinity rules to a minimum. Two types of rules are available:

- Virtual machine group to ESXi host group (anti-) affinity
- Virtual machine to virtual machine (anti-) affinity

Virtual Machine Group to Host Group Affinity Ruleset

Used to specify affinity or anti-affinity between a group of virtual machines and a group of hosts. An affinity rule specifies that the members of a selected virtual machine DRS group can or must run on the members of a specific ESXi host DRS group. An anti-affinity rule specifies that the members of a selected virtual machine DRS group cannot run on the members of a specific host DRS group. Two types of VM-Host group are available, mandatory (Must run on/Must not run on) and preferential (Should run on/Should not run on).

- A mandatory rule specifies which hosts are compatible to run the listed virtual machines. It limits HA, DRS and the user in such a way that a virtual machine may not be powered on or moved to an ESXi host that does not belong to the associated DRS host group.
- A preferential rule defines a preference to DRS to run a virtual machine on the host specified in the associated DRS host group.

HA and DRS Integration of Preferential Rules

VMware High Availability respects and obeys mandatory rules when placing virtual machines after a host failover. It can only place virtual machines on the available ESXi hosts that are specified in the DRS host group. If no ESXi host is available, the virtual machine will not be restarted until one of the compatible hosts returns to operational state, or until the ruleset is removed or changed to a preferential rule.

Preferential rules are only known to DRS and do not create a restriction when a virtual machine is restarted on one of the remaining hosts in the cluster. Because HA is not aware of these rules, it is unable to select a preferred ESXi host, thereby possibly violating the affinity rule. If a virtual machine is placed on an ESXi host that is outside the ESXi host group, DRS will correct this violation during the next invocation of DRS.

DRS Load Balancing with Preferential Rules

During a DRS invocation, DRS runs the algorithm with preferential rules as mandatory rules and will evaluate the result. If the result contains violations of cluster constraints; such as over-reserving a host

or over-utilizing a host leading to 100% CPU or Memory utilization, the preferential rules will be dropped and the algorithm is run again.

DRS Operations Impact

VM-Host affinity rule restricts the number of hosts on which the virtual machines may be powered-on or to which virtual machines may migrate. Setting VM-Host affinity rules can limit the number of load balancing possibilities, HA failover defragmentation operations and evacuation of virtual machines when an ESXi host is placed into maintenance mode.

Virtual Machine to Virtual Machine Affinity Ruleset
This ruleset is used to specify affinity or anti-affinity between individual virtual machines. A rule specifying affinity causes DRS to try to keep the specified virtual machines together on the same host, for example, for performance reasons. With an anti-affinity rule, DRS tries to keep the specified virtual machines apart, for example, so that when a problem occurs with one host, you do not lose both virtual machines.

When an affinity rule is added or edited, and the cluster's current state is in violation of the rule, the system continues to operate and DRS attempts to correct the violation. For manual and partially automated DRS clusters, migration recommendations based on rule fulfillment and load balancing are presented for approval. You are not required to fulfill the rules, but the corresponding recommendations remain until the rules are fulfilled.

5.7 Implicit Constraints

Number of vMotion Operations

DRS attempts to minimize the number of vMotion because a vMotion process incurs costs to multiple systems. DRS ensures that the number of vMotions on a per-host basis on a per vNIC basis and the total number of vMotion per cluster are under the limit.

When reviewing the load balancing operations, DRS determines the network costs and the processor costs. To calculate the network costs, DRS takes into the network bandwidth into account as well as the memory activity of the virtual machine. If the minimum network connection between the source and destination ESXi host is 1 GB, the vMotion process reserves 25% of a single core on both hosts. If the available bandwidth between the two ESXi hosts is a minimum of 10 GB, 100% of a CPU core is reserved on both hosts. The reservation of CPU resources ensures that enough CPU resources are available for the vMotion process to migrate the virtual machine as quickly as possible. This reservation does impact the overall resource availability of active virtual machines on both ESXi hosts. For this particular reason, DRS needs a good reason to migrate virtual machines around.

These metrics influence the number of load balancing operations per each DRS invocation.

Virtual Machine Memory Activity

DRS performs a what-if analysis on the memory consumption of the virtual machine, especially how fast the virtual machine is writing to memory pages. Knowing the active pages that are being 'dirtied' and also knowing the state of the destination ESXi host provides DRS the insight of the duration of that particular vMotion. This is taking into account during initial placement if a prerequisite move is required or during the load balancing operation.

Datastore Connectivity

VMware recommends connecting all the ESXi hosts inside the cluster to the same set of datastores. This state is considered to be fully connectivity. If an ESXi host is not connected to a particular datastore, either by design or by a failure, there will be partial connectivity. Initial placement and load

balancing operations will take partial connectivity into account and will mark these host as least-favorable.

Network Resource Availability

During initial placement and load balancing operations, DRS applies a Distributed vSwitch port constraint check in which it determines whether the destination ESXi host have enough network ports. DRS also takes physical uplink failures into account and will mark these host as least-favorable.

Agent Virtual Machines

Agents virtual machines play an important role. Agent virtual machines such as HA (Fault Domain Manager) agents are critical to certain workloads. If a virtual machine depends on the availability of an agent virtual machine, DRS will not move this virtual machine to an ESXi host that does not run the required agent virtual machine.

Special Virtual Machines

Virtual machines that have SMP-Fault Tolerance (FT) or latency sensitive enabled to act as an implicit constraint for DRS. Whenever these settings are enabled on a virtual machine, DRS avoids migrating this virtual machine. Whenever an FT primary or secondary fail, DRS provides special treatment for them.

5.8 DRS Behavior

DRS runs every 5 minutes, depending on the migration threshold level it generates a number of migration recommendations to solve the imbalance of the cluster. Typically, migrations that offer the best cost-benefit ratio will occur. In practice, most DRS clusters will predominantly be CPU balanced, this is due to the incorporating idle memory as virtual machine demand metric. *DRS Cluster Additional Options* allow IT operation teams to change the main focus of DRS.

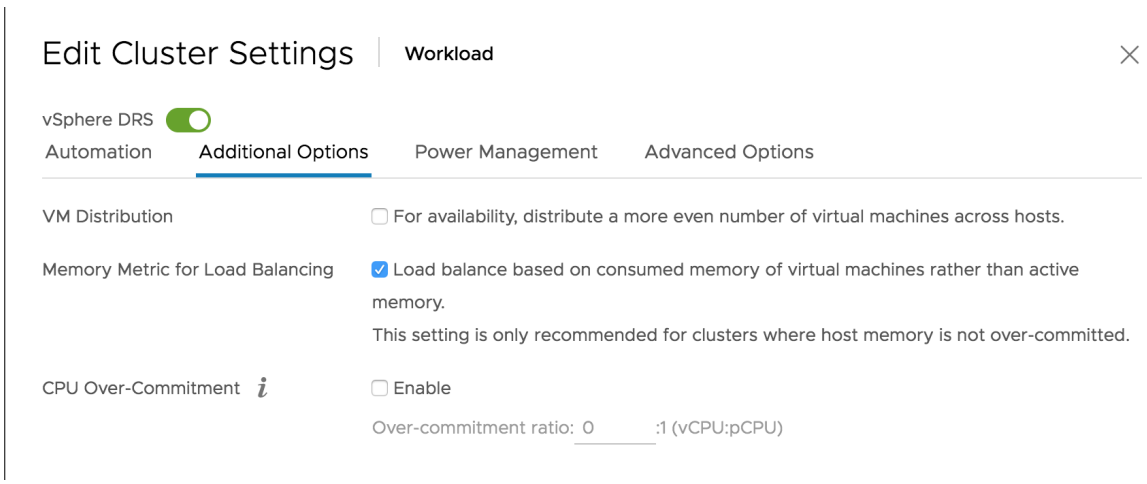


Figure 3: DRS Additional Options

DRS Alignment

DRS is aligned with the premise of virtualization, resource sharing and over-commitment of resources. DRS goal is to efficiently provide compute resources to the active workload to improve workload consolidation on a minimal compute footprint. However, virtualization surpassed the original principle of workload consolidation to provide unprecedented workload mobility and availability.

With this change of focus, many customers do not overcommit on memory. A lot of customers design their clusters to contain enough memory capacity to ensure all running virtual machines have their memory backed by physical memory. In this scenario, DRS behavior should be adjusted as it focusses on active memory use by default.

DRS Default Memory Load Balancing Behavior

During load balancing operation, DRS calculates the active memory demand of the virtual machines in the cluster. The active memory represents the working set of the virtual machine, which signifies the number of active pages in memory. By using the working-set estimation, the memory scheduler determines which of the allocated memory pages are actively used by the virtual machine and which allocated pages are idle. To accommodate a sudden rapid increase of the working set, 25% of idle consumed memory is allowed. Memory demand also includes the virtual machine's memory overhead.

In this example, a 16 GB virtual machine is used to demonstrate how DRS calculates the memory demand. The guest OS running in this virtual machine has touched 75% of its memory size since it was booted, but only 35% of its memory size is active. According to the ESXi host memory management, the virtual machine has consumed 12288 MB and 5734 MB of this is used as active memory.

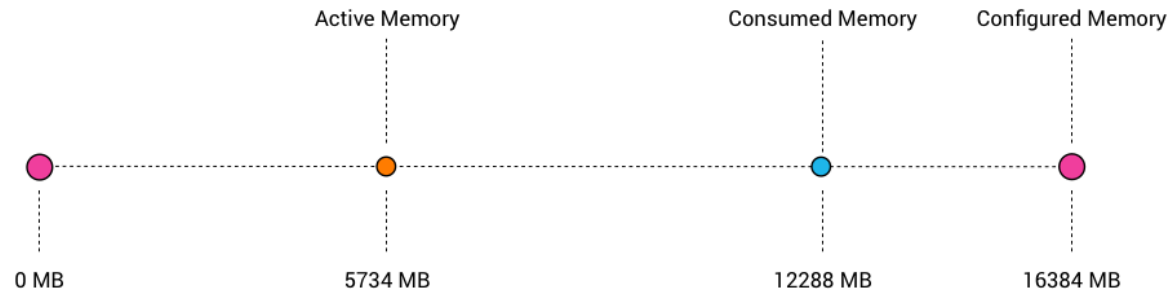


Figure 4: Virtual Machine Memory Demand

DRS accommodate a percentage of the idle consumed memory to be ready for a sudden increase in memory use. To calculate the idle consumed memory, the active memory (5734 MB) is subtracted from the consumed memory (12288 MB), resulting in a total 6554 MB idle consumed memory. By default, DRS includes 25% of the idle consumed memory, i.e. 6554 * 25% = +/- 1639 MB.

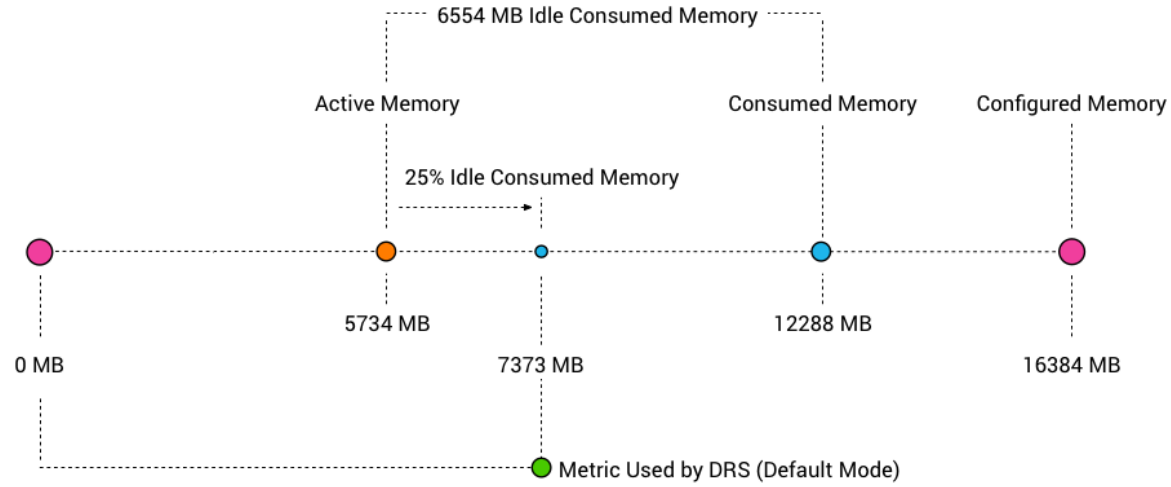


Figure 5: Virtual Machine Idle Consumed Memory

The virtual machine has a memory overhead of 90 MB. The memory demand DRS uses in its load balancing calculation is as follows: 5734 MB + 1639 MB + 90 MB = 7463 MB. As a result, DRS selects a host that has 7463 MB available for this machine if it needs to move this virtual machine to improve the load balance of the cluster.

5.9 DRS Additional Option: Memory Metric for Load Balancing Enabled

When enabling the option “Memory Metric for Load Balancing” DRS takes into account the consumed memory + the memory overhead for load balancing operations. In essence, DRS uses the metric Active + 100% IdleConsumedMemory.

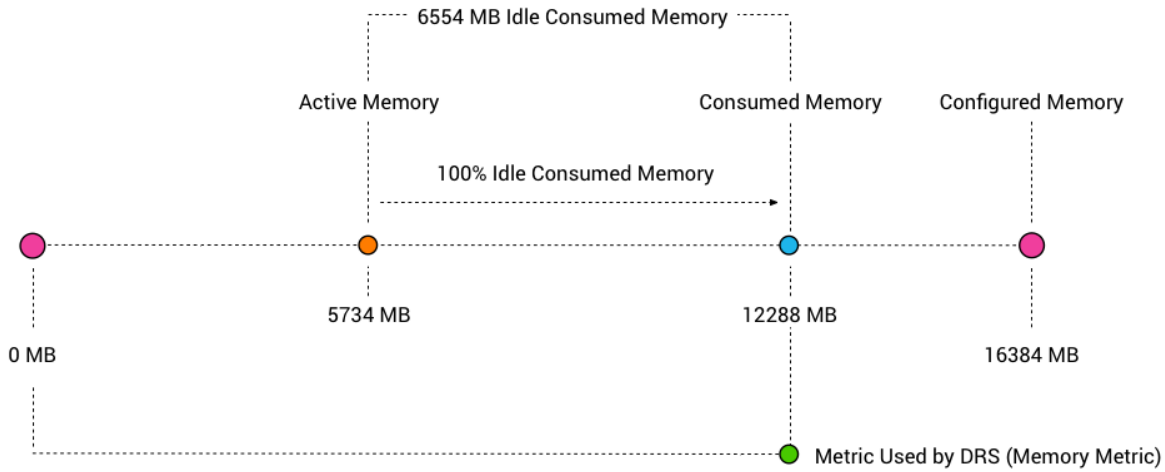


Figure 6: 100% Idle Consumed Memory

vSphere 6.5 update 1d UI client allows you to get better visibility in the memory usage of the virtual machines in the cluster. The memory utilization view can be toggled between active memory and consumed memory.

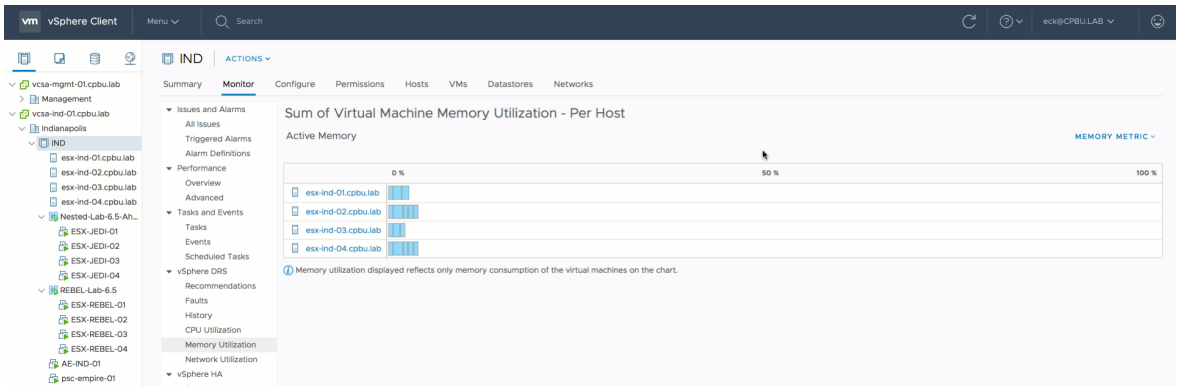


Figure 7: Monitor vSphere DRS Memory Utilization Options

Active versus Consumed Memory Bias

If you design your cluster with no memory over-commitment as guiding principle, it is recommended to test out the vSphere 6.5 DRS option “Memory Metric for Load Balancing”. Conservative IT operation teams should switch DRS to manual mode, to verify the recommendations first.

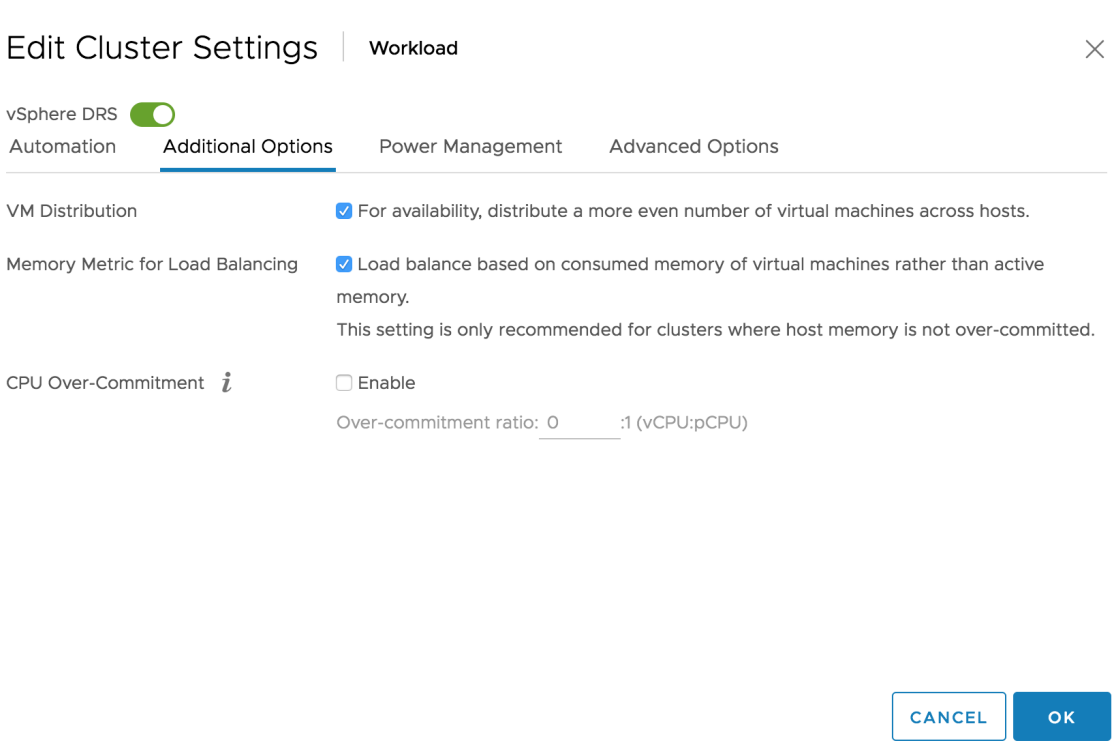


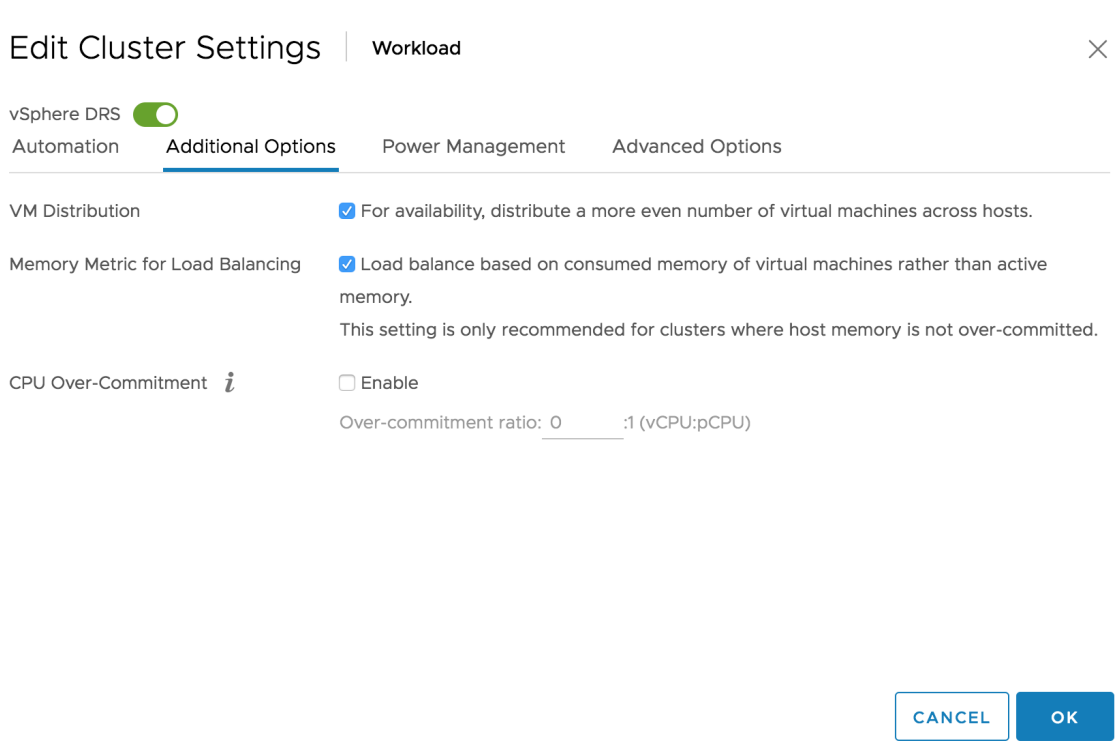
Figure 8: vSphere Cluster DRS Additional Options - Memory Metric for Load Balancing Enabled

5.10 DRS Additional Option: VM Distribution

This setting allows DRS to distribute the virtual machines evenly across the cluster. Whereas in some situations the normal DRS cost-benefit analysis would not be positive, this setting overrules this logic and incur the cost of migration to achieve a more evenly distribution of virtual machines. Please note that this setting will still keep virtual machine happiness in mind, so even distribution of virtual machines is done on a best-effort basis.

This setting aims to have a similar number of virtual machines on each ESXi host, however, if the ESXi hosts differ in physical resource configuration such as CPU cores or total amount of memory, DRS calculates a ratio of virtual machines based on ESXi host capacity.

This setting can be combined with the DRS additional option *memory metric for load balancing* enabled . This setting can be helpful for environments that attempt to minimize the impact of host failures or attempt to balance the load on network IP connections across the ESXi hosts in the cluster. Please note that this setting can increase the number of virtual machine migrations without specifically benefitting application performance.



5.11 DRS Additional Option: CPU Over-Commitment

As previously stated the early premise of virtualization was to share resources efficiently. By default, DRS uses a default CPU over-commit (vCPU to pCPU) ratio that is approximately 80 to 1. Latency sensitive workload can benefit from a lower CPU over-commit ratio by reducing the number of vCPUs waiting to be scheduled. This setting limits the number of vCPUs that can be powered-on in the vSphere cluster. For clusters that run workloads that benefit from lower CPU scheduling times, the CPU over-commitment additional setting is useful. Please note that this setting is geared towards satisfying performance more than providing the best economics.

User Interface Variation

VMware focusses on the development of the H5 client and thus new features are introduced in the H5 client first. The CPU over-commitment setting translates into a different advanced setting when using the H5 UI in vSphere 6.5 Update 1 then using the web client. vSphere versions and updates beyond vSphere 6.5 Update 1 will provide a uniform experience.

Client type	HTML 5 Client	Web Client
Focus	Host-Based vCPU to pCPU ratio	Cluster-wide CPU over-commitment ratio
Advanced Setting	MaxVcpusPerCore	MaxVCPUsPerClusterPct
Minimum value	4	0
Maximum value	32	500

Table 3: User Interface Variation

Maximum vCPU Per Cluster Percentage

This advanced options control is available via the web client and sets the overall cluster-wide vCPU to pCPU overcommitment ratio (i.e. total number of vCPUs in the cluster / total number of pCPUs in the cluster divided by 100 to make it a percentage). The minimum value is 0, which equals to a cluster-wide denial of service, no vCPUs are allowed to consume pCPUs. This could be used by IT operation teams who want to make the cluster unable to accept workloads for a period of time due to upgrade operations. The maximum value is 500, which equals to a 5:1 vCPU to pCPU ratio.

Maximum vCPUs per CPU Core

This advanced option control is available via the vSphere 6.5 U1 H5 client and is enforced at the ESXi host level. No ESXi host in the cluster is allowed to violate this setting. Although the UI allows any setting between 0 and 500, the valid minimum value is 4, while the maximum value is 32. The reason why the maximum value is 32 is that the default vCPU to pCPU limit supported by the ESXi host is 32:1.

In general, it is recommended to use the H5 client as much as possible. However, with this quirky setting, a particular over-commit ratio is only available by using one or the other. If the goal is to attempt to have a vCPU to pCPU ratio of 4:1 or less, use the web client to set the *MaxVCPUsPerClusterPct* option. If the cluster is allowed to have vCPU to pCPU ratio of 4:1 and higher, use the vSphere 6.5 Update, 1 H5 client, to set the *MaxVcpusPerCore* . Please note that vSphere versions and updates beyond vSphere 6.5 Update 1 will provide a uniform experience.

Additional DRS Options Behavior

Using the additional options automatically creates the advanced options seen in the cluster settings overview. In this example, both the additional options Memory Metric for Load Balancing Enabled and VM Distribution are enabled. This results in two advanced settings: *PercentIdleMBMemDemand=100* and *TryBalanceVmsPerHost=1* . Please use the UI settings instead of configuring advanced settings directly.

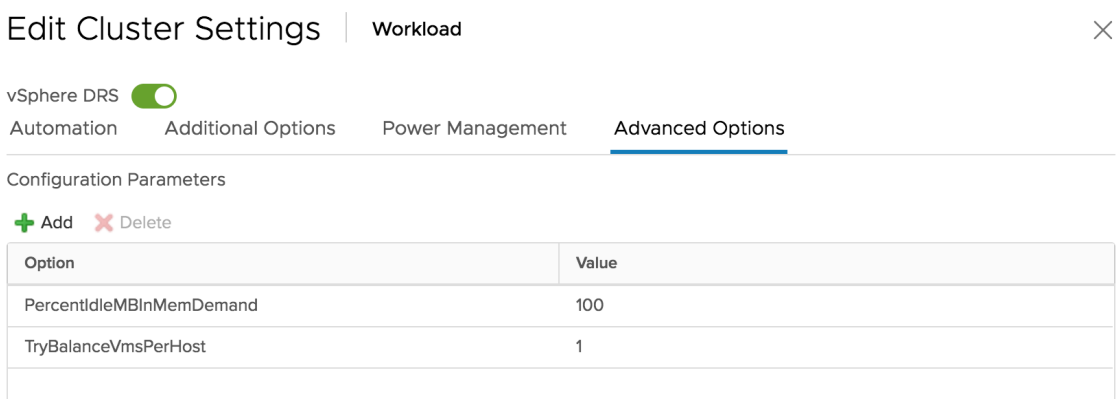


Figure 9: DRS Cluster Advanced Options

Please note that these additional options will override any equivalent cluster advanced options. For example, if you set cluster advanced option PercentIdleMBInMemDemand to some value, and then enable the *memory metric option for load balancing* , the advanced option will be cleared to give precedence to the new memory metric option.

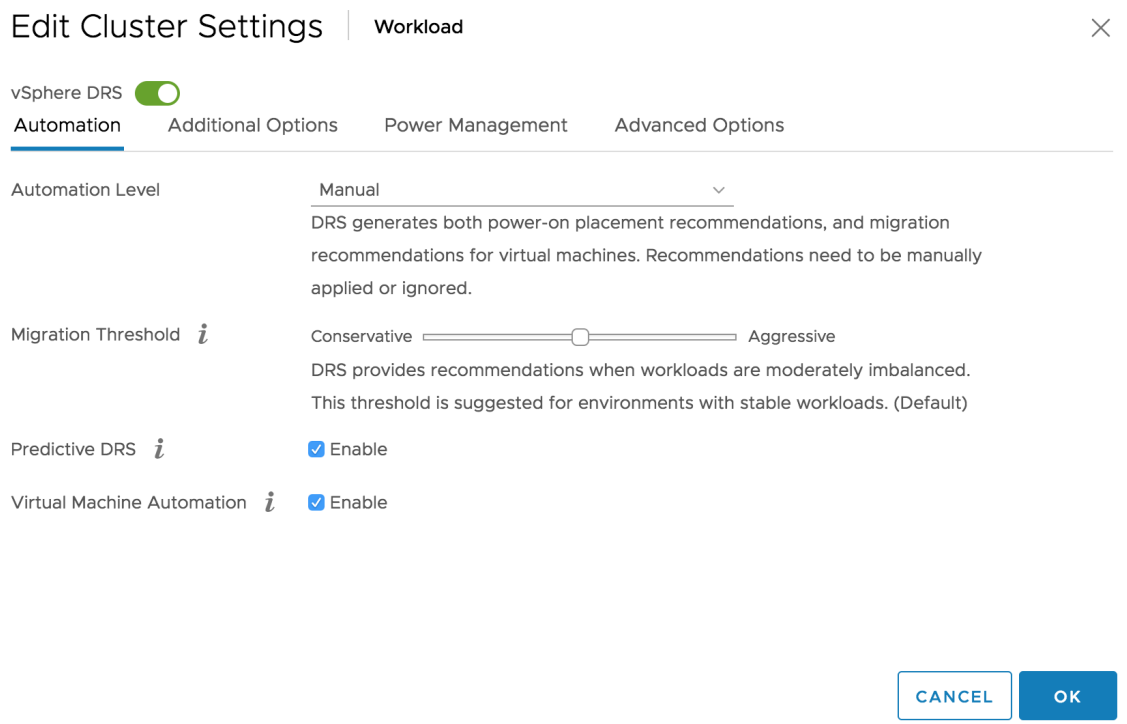
5.12 Predictive DRS

Predictive DRS is a feature that combines the analytics of vRealize Operations Manager 6.4 (and higher) with the logic of vSphere 6.5 DRS. This collaboration between products allows DRS to execute predictive moves based on the predictive data sent by vRealize Operation Manager.

By default, DRS resolves unexpected resource demand by rebalancing the workload across the ESXi host within the vSphere Cluster. This can be considered as a reactive operation. By leveraging trend-analysis offered by vRealize Operations Manager, DRS can rebalance the cluster in order to provide resources for future demand. This can be considered to be predictive.

Combining DRS and vRealize Operation Manager, DRS can avoid the situation of degradation of VM happiness due to (predictable) workload spikes. By proactively redistributing the virtual machines in the cluster to accommodate these workload patterns.

Predictive DRS is configured in two easy steps. One tickbox at vSphere Cluster level and one drop-down menu option in vRealize Operations Manager. Enabling the Predictive DRS option in the vSphere Cluster is by ticking the option at the Automation options view. In vRealize Operations Manager, select the advanced settings of the vSphere object and set "Provide data to vSphere Predictive DRS" to true.


Edit Cluster Settings
Workload
vSphere DRS
Automation
Additional Options
Power Management
Advanced Options
Automation Level
Manual
DRS generates both power-on placement recommendations, and migration recommendations for virtual machines. Recommendations need to be manually applied or ignored.
Migration Threshold
Conservative
Aggressive
DRS provides recommendations when workloads are moderately imbalanced. This threshold is suggested for environments with stable workloads. (Default)
Predictive DRS
Enable
Virtual Machine Automation
Enable
CANCEL
OK

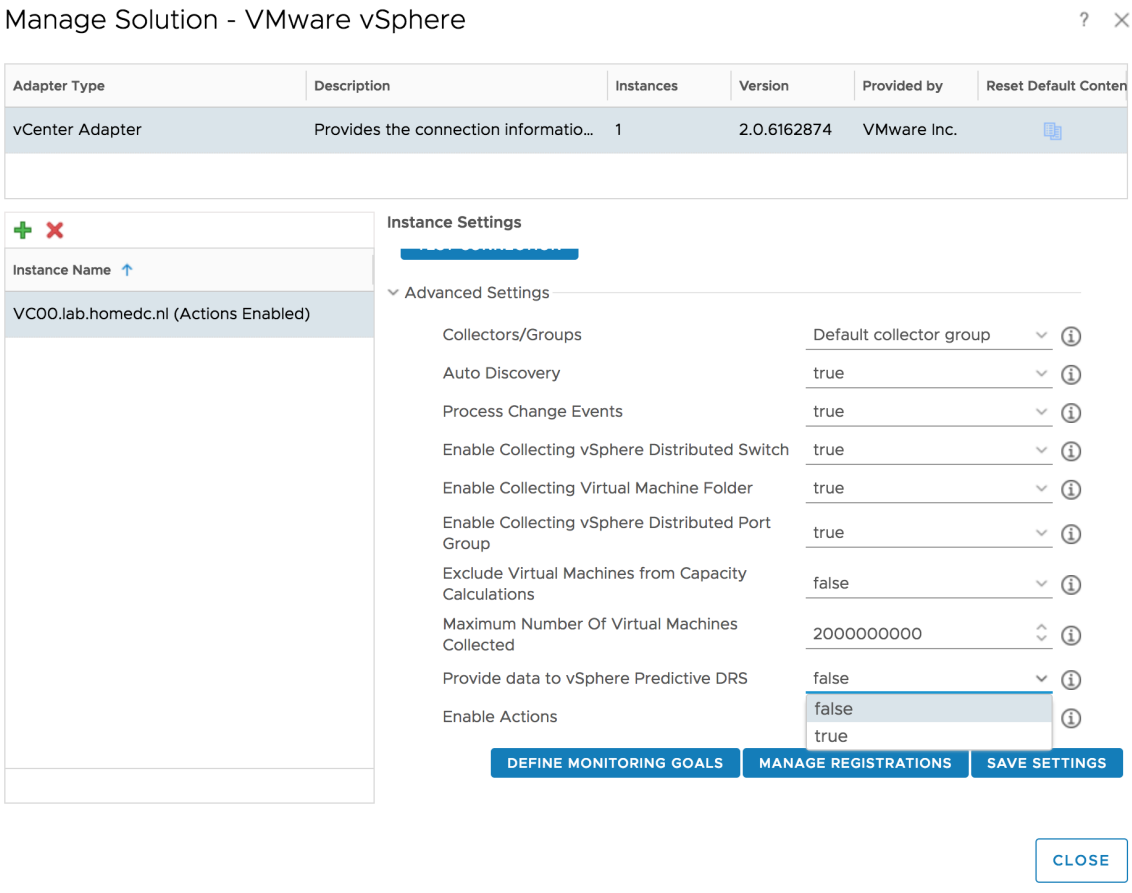

Manage Solution - VMware vSphere
Adapter Type
Description
Instances
Version
Provided by
Reset Default Conten
vCenter Adapter
Provides the connection informatio...
1
2.0.6162874
VMware Inc.
Instance Name
VC00.lab.homedc.nl (Actions Enabled)
Instance Settings
Advanced Settings
Collectors/Groups
Default collector group
Auto Discovery
true
Process Change Events
true
Enable Collecting vSphere Distributed Switch
true
Enable Collecting Virtual Machine Folder
true
Enable Collecting vSphere Distributed Port Group
true
Exclude Virtual Machines from Capacity Calculations
false
Maximum Number Of Virtual Machines Collected
2000000000
Provide data to vSphere Predictive DRS
false
false
true
Enable Actions
DEFINE MONITORING GOALS
MANAGE REGISTRATIONS
SAVE SETTINGS
CLOSE

Figure 10: Enabling Predictive DRS in vCenter and vRealize Operations Manager

Predictive DRS Predictive Threshold

Predictive DRS monitors the behavior of the workloads in the cluster, it collects high-resolution data continuously. It does so for more than hundred metrics across numerous types of objects such as hosts, virtual machines, and datastores. vRealize Operations Manager does not roll up data that can hide important performance behavior, instead, it uses 5-minute granularity of data, allowing it to very intelligently manage peaks.

All this data is the input of the analytics engine that uses multiple algorithms to learn the normal behavior of the workload and then it starts to detect the patterns. This can be daily or monthly. Once the pattern is determined it identifies the upper and lower bounds that shape the Dynamic Threshold for the workload. The dynamic threshold is a function of vRealize Operations Manager, for more information please review the technical paper " Sophisticated Dynamic Thresholds with VMware vCenter Operations ".

A filtering logic transforms the dynamic threshold into a prediction and is ingested by DRS ahead of time. By default, DRS has access to these forecasts of workload behavior for the next 60 minutes. This data is used to distribute the virtual machines in such a way that the cluster will be ready to satisfy the workload demand.

Please note that Predictive DRS is conservative. It will always ensure that the current virtual machine demand will not be by the forecast of the future demand. It will not trade in current VM happiness for future VM happiness. With Predictive DRS VM happiness equals to the maximum of current demand and the future demand.

To be able to create these forecasts, Predictive DRS monitors the behavior of the workloads for 14 days. After 14 days, Predictive DRS will feed DRS with the forecast of the behavior of that particular virtual machine. It is common for dynamic clusters to have virtual machines that are not operational for more than 14 days. For virtual machines that are less 14 days old, predictions are missing and for these virtual machines, the VM happiness equals to the default dataset of DRS of current demand only.

Predictive DRS requires at least 14 days' worth of data to provide a forecast, the longer the period, the more accurate Predictive DRS becomes. Please note that forecasting in Predictive DRS works best for workloads with a periodic usage pattern.

5.13 vRealize Operations Manager Workload Placement

DRS main management construct is the vSphere cluster. DRS is designed to ensure VM happiness for virtual machines within the cluster. The VMware vRealize Operations Manager workload balance can assist IT operation teams that manage multiple vSphere clusters.

The workload placement function monitors the vSphere clusters and is able to determine if the clusters are out of balance. Multiple vSphere clusters can be grouped together into a virtual datacenter that acts as a load-balancing domain within vRealize Operations Manager. This allows IT operation teams to create separate groups of capacity based on performance levels, business requirements or licensing constraints.

vRealize Operations Manager workload balance allow the IT operation team to configure the level of workload consolidation, the level of workload balance, a resource buffer space and automation level of a balance-plan.

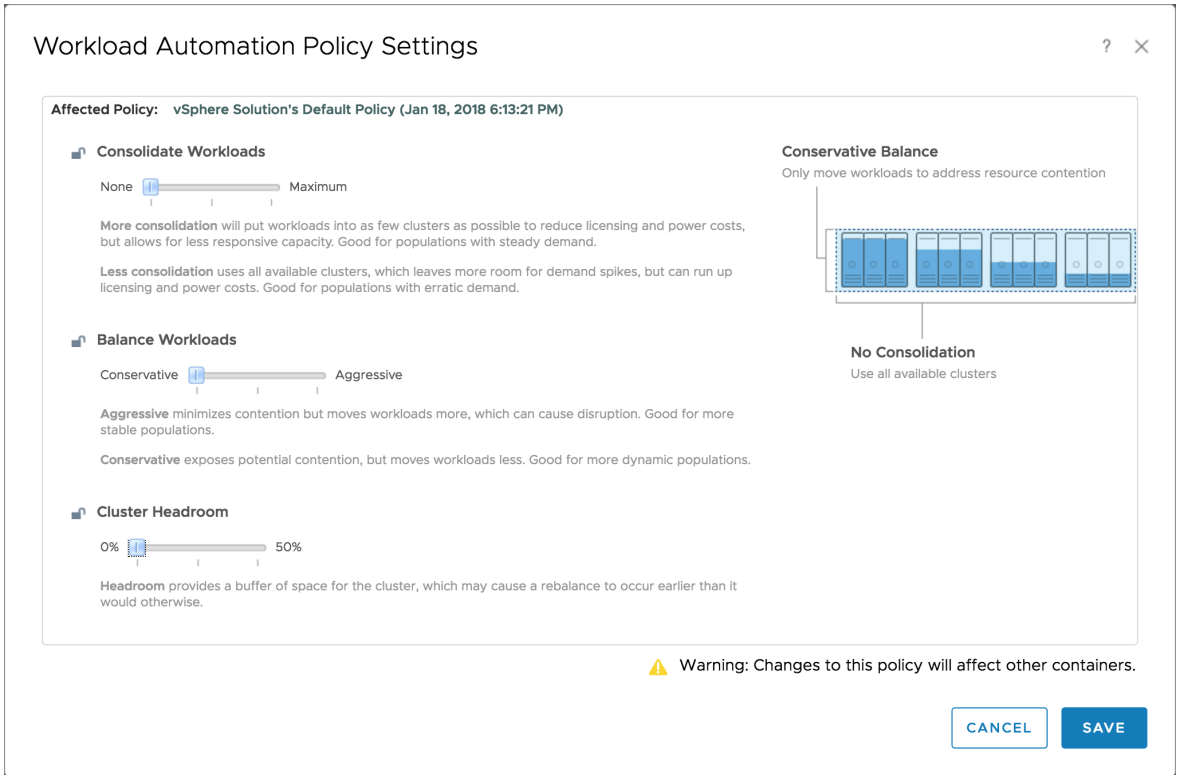


Figure 11: Workload Automation Policy Setting

Consolidate Workloads

The setting consolidate workloads allows specifying the distribution of workloads across the vSphere clusters that are a part of the datacenter or grouped into the custom data center. Less consolidation equals distribution across more vSphere clusters. Aggressive consolidation can be useful for custom datacenters containing vSphere clusters designed to host license constrained workloads.

Balance Workloads

The balance workload setting determines the level of aggressiveness to avoid performance issues.

- The conservative setting restricts the number of migrations and a migration recommendation will only be generated to address resource contention in a vSphere cluster. This setting is useful for vSphere clusters that have a highly dynamic change in demand.
- The moderate setting will generate migration recommendation to avoid performance issues while recommending as few migrations as possible.
- The Aggressive setting minimizes imbalance across clusters, allowing vSphere clusters to have as much headroom as possible to deal with resource spikes. It will typically lead to more migrations between vSphere clusters. This setting is recommended for workloads that generate a stable demand.

The consolidation and balance workload settings influence each other. The workload automation logic has to choose between containing workloads within the desired footprint (consolidate) or to reduce stress within clusters (balance workload). The following table shows the desired state of mixing the two settings.

Reduce stress = distributing workloads with a minimum number of migrations

Balance = distributing workloads across clusters as much as possible

Consolidate = containing workloads on the smallest footprint as possible

Balance Workload	Consolidate Workloads		
	None	Moderate	Maximum
Conservative	Reduce stress	Reduce stress	Consolidate
Moderate	Reduce stress	Reduce stress	Consolidate
Aggressive	Balance	Balance	Consolidate

Table 4: Combined Settings Behavior Result

The user interface of the workload Automation Policy illustrates the outcome of the combined settings. In this example, the consolidate workloads policy is set to maximum and the balance workload policy is set to aggressive. The image indicates that the workload balance domain is minimized as much as possible and that the deviation of utilization between clusters is kept to a minimum.

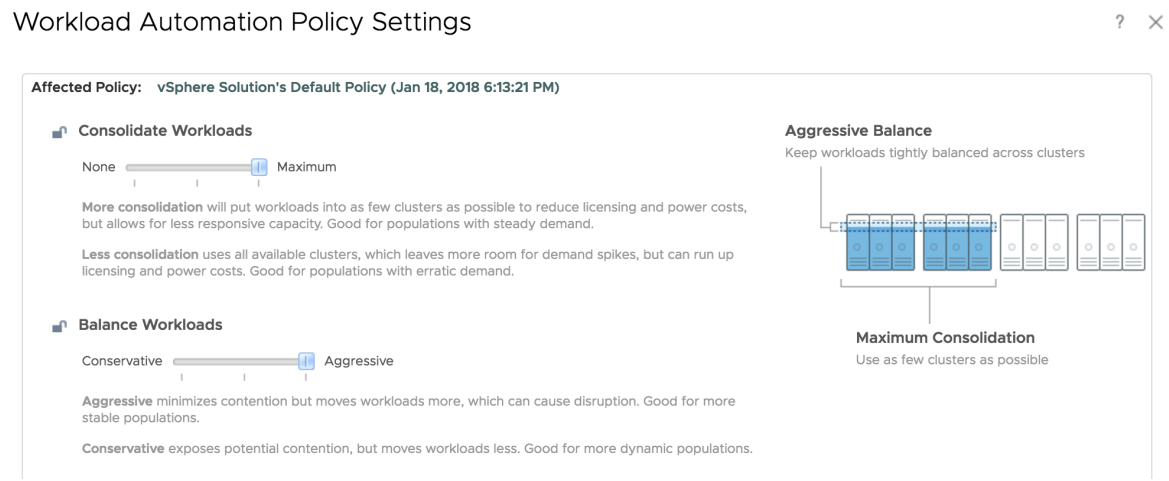


Figure 12: Illustration of combined setting result

Cluster Headroom

The cluster headroom indicates the percentage of resources that should be free that act as a buffer to deal with resource spikes. Migration recommendations for rebalancing are generated when the cluster exceeds the cluster headroom threshold.

Workload Placement Automation Levels

The workload balance process can be triggered manually by the IT operation team or it can be managed by vRealize Operations Manager autonomously. If set to manual, vRealize Operation Manager expects the IT operation team to be in full control and accept or decline the rebalance migration recommendations. When set to automated, all rebalance migration recommendations are accepted and executed automatically.

The scheduled option allows IT operation teams to identify a timeslot in which rebalancing across clusters can occur, for example, the standard maintenance window. Depending on the activity within the datacenter, the recurrence of the rebalance schedule can be set to daily, weekly and monthly.