



DRS Enhancements in vSphere 6.7

Technical White Paper - August 30, 2018



VMware, Inc. 3401 Hillview Avenue Palo Alto CA 94304 USA Tel 877-486-9273 Fax 650-427-5001 www.vmware.com
Copyright © 2018 VMware, Inc. All rights reserved. This product is protected by U.S. and international copyright and intellectual property laws. VMware products are covered by one or more patents listed at <http://www.vmware.com/go/patents>. VMware is a registered trademark or trademark of VMware, Inc. in the United States and/or other jurisdictions. All other marks and names mentioned herein may be trademarks of their respective companies. To comment on this document, go to <https://communities.vmware.com/docs/DOC-38626>.

Table of Contents

Introduction	3
New Initial Placement	3
Host Maintenance Mode Enhancements	3
Enhancements to the Evacuation Workflow	4
Handling Failures.....	4
DRS Support for NVM.....	6
Enhanced Resource Pool Reservation.....	7
Conclusion	9
References.....	9

Introduction

This paper describes the following features updated in 6.7 for VMware vSphere® Distributed Resource Scheduler™ (DRS):

- New initial placement
- Host maintenance mode enhancements
- DRS support for non-volatile memory (NVM)
- Enhanced resource pool reservations

New Initial Placement

As part of our new and enhanced DRS, a new VM initial placement feature was enabled in vSphere 6.5. In earlier versions, DRS took a snapshot of the cluster state to come up with a host recommendation for VM initial placement. In vSphere 6.5, the new algorithm completely avoids snapshotting, so generating recommendations for placement is much faster and the recommendations are also more accurate. VM placement in DRS has the following highlights.

- More even placement of VMs across the cluster upon power-on
- Much faster VM power-on, even with highly concurrent workloads

In vSphere 6.5, the new VM placement was available in limited configurations. For example, it was not available for:

- Clusters where DPM/Proactive HA/HA with admission control is enabled
- Clusters with DRS configured in manual mode
- VMs with manual DRS override setting
- VMs that are FT-enabled
- VMs that are part of a vApp

In vSphere 6.7, the new placement has been made available for all configurations.

More information about the new VM placement can be found in the [DRS 6.5 white paper](#) [1].

Host Maintenance Mode Enhancements

When upgrading the vSphere hosts in your DRS cluster using VMware Update Manager (VUM), it checks with DRS first, to find out which host(s) can be upgraded. DRS in turn runs its algorithm and recommends one or more hosts that can be put into maintenance mode.

Starting with vSphere 6.7, DRS uses the new initial placement algorithm to come up with the recommended list of hosts to be placed in maintenance mode. Further, when evacuating the hosts, DRS uses the new initial placement algorithm to find new destination hosts for outgoing VMs.

Enhancements to the Evacuation Workflow

In vSphere 6.7, DRS has been enhanced to be more efficient in evacuating VMs from a host when it is being put into maintenance mode. In earlier versions, during host evacuation, DRS used to issue vMotion at once for all the powered-on VMs on the host. DRS will now evacuate (vMotion out) powered-on VMs in batches of 8 at a time. The next batch of vMotions will only be issued after the first batch completes. This new model also uses the new initial placement algorithm.

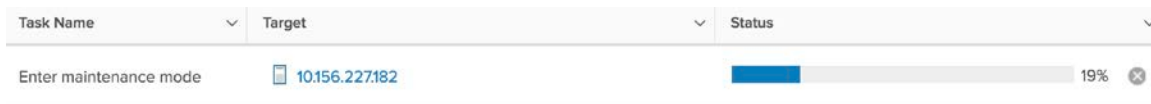
The batching of vMotions makes the entire workflow more controlled and predictable, which makes it easier to estimate the total time to complete the end-to-end workflow. Together with the use of the new initial placement algorithm, the new model results in a very similar distribution of VMs post-evacuation, compared to the old model.

Handling Failures

In vSphere 6.7, when there is a problem while putting a host into maintenance mode, an appropriate fault is generated, which gives detailed information about the problem. The following scenarios showcase the new error handling features.

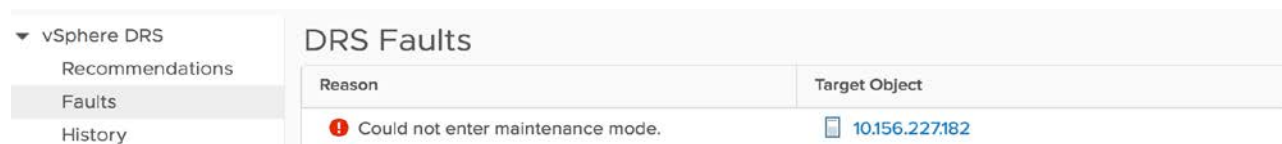
Scenario 1: A VM cannot be evacuated from a host due to a rule violation

We have a VM (VirtualMachine1) that is placed on host 10.156.227.182. The VM has a virtual machine–host affinity rule with the same host. When we try to put the host (10.156.227.182) into maintenance mode, the task halts (Figure 1) and a DRS fault is generated saying the host “Could not enter maintenance mode” (Figure 2).



Task Name	Target	Status
Enter maintenance mode	10.156.227.182	19%

Figure 1 Task - enter host into maintenance mode



Reason	Target Object
Could not enter maintenance mode.	10.156.227.182

Figure 2: DRS fault on host 10.156.227.182

To understand more about how DRS faults work, please refer the VMware documentation about [Faults](#) [2].

When we look at the fault details, it is clear that placing the VM on a different host (than its current one) would violate a virtual machine–host affinity rule, as shown in Figure 3.



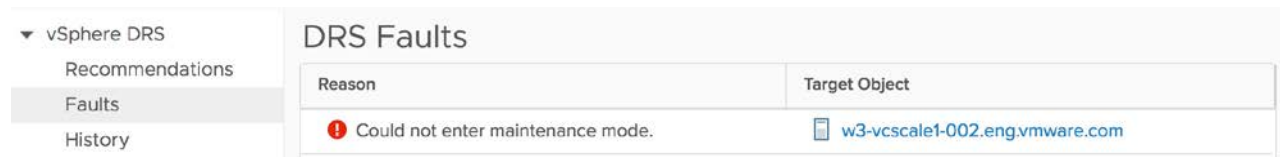
Fault
Virtual machine 'VirtualMachine1' on host '10.156.227.228' would violate a virtual machine - host affinity rule.

Figure 3: DRS Fault describing the virtual machine–host rule violation

We can then cancel the maintenance mode task and fix the problems causing the DRS faults.

Scenario 2: A VM cannot be evacuated from a host because one of its virtual devices is mounted on a local data store of the host

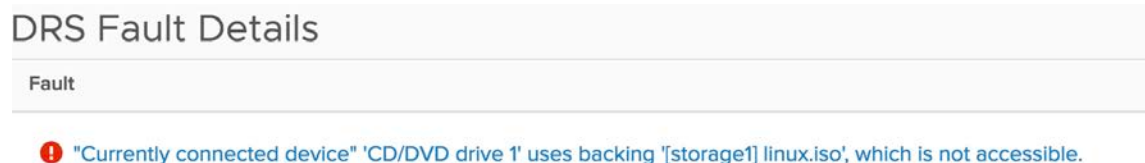
We have a VM (VirtualMachine1) that is placed on host w3-vcscale1-002. The VM has a virtual CD/DVD drive, which is connected and has an ISO image mounted. This ISO image file is placed on a local data store on the host. When we try to put the host (w3-vcscale1-002) into maintenance mode, the task halts and a DRS fault is generated (Figure 4).



vSphere DRS	
Recommendations	
Faults	
History	
DRS Faults	
Reason	Target Object
❗ Could not enter maintenance mode.	📄 w3-vcscale1-002.eng.vmware.com

Figure 4: DRS fault on host w3-vcscale1-002

When we look at the fault details, it is clear that the local file backing of VirtualMachine1's CD/DVD drive would not be accessible if it were evacuated to a different host (Figure 5 and Figure 6).



DRS Fault Details	
Fault	
❗ "Currently connected device" 'CD/DVD drive 1' uses backing '[storage1] linux.iso', which is not accessible.	

Figure 5: DRS fault details



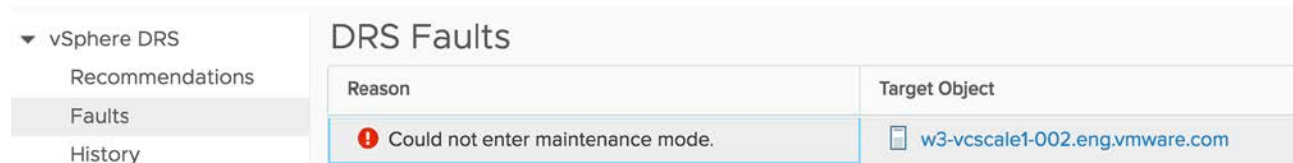
DRS Fault Details	
Prevented Recommendation	
📄 Migrate VirtualMachine1 from w3-vcscale1-002.eng.vmware.com to any host	

Figure 6: DRS fault details showing the recommendation that was prevented

We can then cancel the maintenance mode task and fix the problems causing the DRS faults.

Scenario 3: A VM cannot be evacuated from a host because one of its virtual disks is located on a local data store of the host

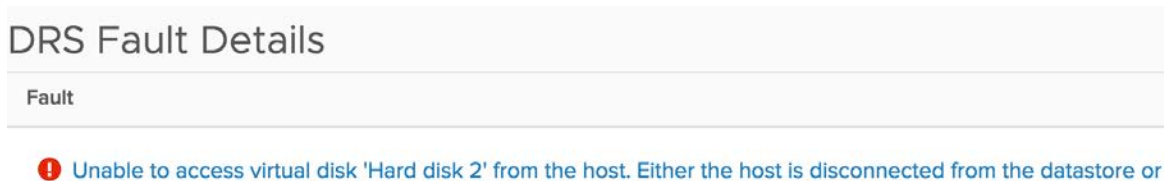
This scenario is very similar to scenario 2, except the VM VirtualMachine1 now has a virtual hard disk drive on the local data store of host w3-vcscale1-002. When we try to put the host (w3-vcscale1-002) into maintenance mode, the task halts and a DRS fault is generated (Figure 7).



Reason	Target Object
❗ Could not enter maintenance mode.	📄 w3-vcscale1-002.eng.vmware.com

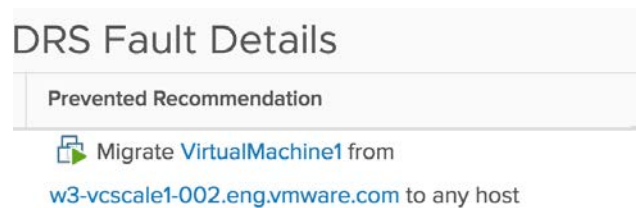
Figure 7: DRS Fault on host w3-vcscale1-002

When we look at the fault details, it is clear that the VM's virtual disk would not be accessible if it were evacuated to a different host (Figure 8 and Figure 9).



Fault
❗ Unable to access virtual disk 'Hard disk 2' from the host. Either the host is disconnected from the datastore or

Figure 8: DRS fault details



Prevented Recommendation
📄 Migrate VirtualMachine1 from w3-vcscale1-002.eng.vmware.com to any host

Figure 9: DRS fault details showing the recommendation that was prevented

In summary, in vSphere 6.7 we have enhanced the host maintenance and VM evacuation workflows to be faster and more efficient. We have also improved the user experience for troubleshooting when we run into issues during these workflows.

DRS Support for NVM

Starting in vSphere 6.7, vSphere DRS can handle VMs running on next generation persistent memory devices, also known as Non-Volatile Memory (NVM) devices.

NVM is exposed as a host-local data store. The VMs could use this datastore in two forms:

- As an NVM device exposed to the guest for its use (Virtual Persistent Memory or vPMem)
- As a location for a virtual machine disk (Virtual Persistent Memory Disk or vPMemDisk)

DRS considers a VM's NVM during the following operations.

- VM power-on placement
- Load balancing
- Mandatory moves

During all of these operations, DRS is aware of NVM devices attached to VMs and guarantees the destination ESXi host of the VM has enough free persistent memory (PMem) to accommodate the VM.

For more information about NVM devices and their support in vSphere, please refer to the [vSphere 6.7 Storage guide](#) [3].

Enhanced Resource Pool Reservation

In vSphere 6.7, we introduced a new two-pass algorithm to allocate a resource pool's resource reservation to its children (also known as divvying).

The old divvying model will not reserve more resources than the current demand in the resource pool, even if the resource pool is configured with a higher reservation. If there is a spike in VM demand after resource divvying is done, DRS will only react when the next resource divvying happens to allocate the remaining reservation. As a result, VMs might suffer from temporary performance issues until the next time resource divvying happens.

In the new model, resource pool reservation is divvied up to its children as much as possible, irrespective of the current demand, thus giving more buffer to account for sudden spikes in VM demand. In the first pass, the resource pool reservation is divvied based on VM demand, limited by each VM's fair share. Then in the second pass, excess reservation is divvied proportionally, limited by the VM's configured size.

Let us consider an example scenario to illustrate the new resource pool reservation divvying model. In this example, we have a resource pool with a memory reservation of 10 GB. It has four VMs of 4 GB RAM each. The VMs are all using only 25% of their configured RAM size (memory demand is 1 GB for each VM). The Reservation, Limit, Shares settings for the resource pool are as shown in Figure 10.

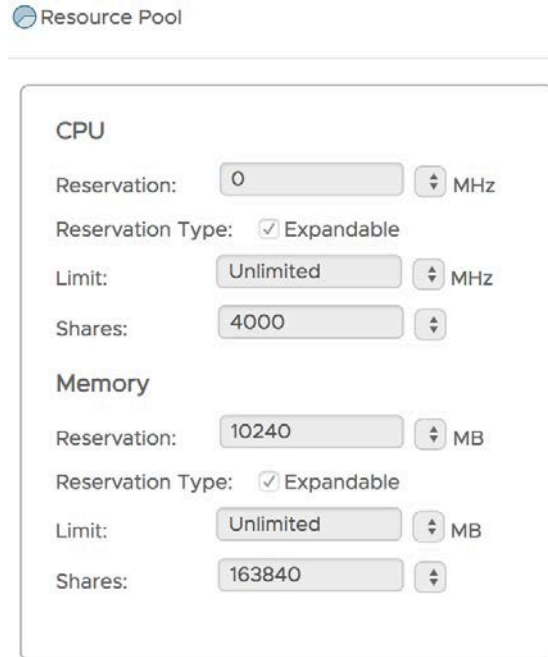


Figure 10: RLS settings for the Resource pool

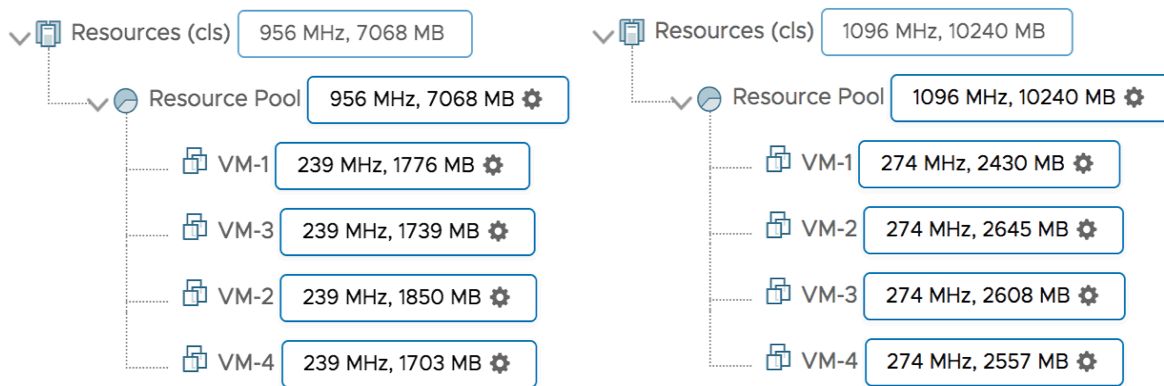


Figure 11: Resource pool and VM entitlements—old and new models

In the old model, DRS would divvy the resource pool reservation to its children based on their demand. As you can see in the left side of Figure 11, in our example, each VM would be entitled to around 1.7 GB (VM demand + some percentage of idle memory + overhead). Whereas in the new model (the right side of Figure 11), when DRS divvies resource pool reservation, it is not limited by the VM demand and each VM gets around 2.5 GB of memory reservation, which means the resource pool reservation has been completely divvied out to its children.

Note: The tree structures in Figure 11 were captured using [DRS Entitlement Viewer](#) [4].

In summary, with the new reservation distribution model, VMs will now have extra room to handle any sudden spikes in workload, avoiding any performance impact. This behavior is particularly helpful for volatile workloads.

Conclusion

vSphere 6.7 comes with a host of enhancements to its DRS algorithm. These enhancements make resource management workflows more robust, predictable, and overall more effective. With the additional support for NVM devices, DRS is also able to handle a more varied set of hardware.

References

- [1] Sai Manohar Inabattini, Vikas Madhusudana , and Adarsh Jagadeeshwaran. (2016, October) DRS Performance in vSphere 6.5
<https://www.vmware.com/content/dam/digitalmarketing/vmware/en/pdf/techpaper/drs-vsphere65-perf.pdf>

- [2] VMware, Inc. (2018, April) Fault Definitions.
<https://docs.vmware.com/en/VMware-vSphere/6.7/com.vmware.vsphere.resmgmt.doc/GUID-AE361248-AE84-4393-9F28-63D573B61A4D.html>

- [3] VMware, Inc. (2018, August) vSphere 6.7 Storage.
<https://docs.vmware.com/en/VMware-vSphere/6.7/vsphere-esxi-vcenter-server-67-storage-guide.pdf>

- [4] Sai Manohar Inabattini, Priyanka Gayam , Vikas Madhusudana, and Avudaiappan Kannan. (2018, August) DRS Entitlement Viewer.
<https://labs.vmware.com/flings/drs-entitlement-viewer>

About the Authors

Sai Manohar Inabattini is a senior performance engineer with the vCenter Server performance group. He works on vCenter Server performance and scalability, with special focus on the Distributed Resource Scheduling (DRS) algorithm. Sai holds a bachelor's degree in computer science from the National Institute of Technology at Warangal, India.

Priyanka Gayam is a performance engineer with the vCenter Server performance group. She works on vCenter Server performance and scalability, with special focus on the Distributed Resource Scheduling (DRS) algorithm. Priyanka holds a bachelor's degree in computer science from the National Institute of Technology at Warangal, India.

Adarsh Jagadeeshwaran works in the vCenter Server performance group in VMware, leading a team that is focused on vCenter Server resource management and scalability. Adarsh holds a master's degree in computer science from Syracuse University, Syracuse, USA.

Acknowledgements

The authors would like to thank Sabareesh Subramaniam, Zhelong Pan, and Fei Guo for their support and guidance for this paper. The authors would also like to thank the DRS team for reviewing this paper, and Julie Brodeur for her help in compiling this paper.