Adaptive Resync in vSAN

Table of contents

Adaptive Resync in vSAN

Executive Summary

Delivering consistent performance while maintaining data resiliency is a key tenet behind enterprise storage solutions. VMware vSAN is the industry leading distributed storage system built right into VMware vSphere, and is designed to offer the highest level of resiliency and performance, with the maximum amount of agility should hardware faults occur, or demands of the environment change. vSAN 6.7 and newer contains a sophisticated method to balance the use of resources during times of resynchronization activities. Adaptive Resynchronization, or Adaptive Resync, optimizes vSAN's ability to dynamically, and transparently adapt I/O to best accommodate the needs of the virtual machines (VMs) against the tasks of the underlying storage system. This feature requires no manual configuration or intervention. The mechanisms behind Adaptive Resync will be discussed in more detail in this document.

vSAN's Approach to Data Placement and Management

VMware vSAN is a distributed storage system that uses physical storage devices on each ESXi host in a cluster to contribute to the vSAN storage system. vSAN presents storage as a single, distributed datastore visible to all hosts in the vSphere cluster.

vSAN is an object-based storage system, meaning that virtual machines that live on vSAN storage are comprised of several storage objects. VMDKs, VM home namespace, VM swap areas, snapshot delta disks, and snapshot memory maps are all examples of storage objects in vSAN. These objects are made up of smaller chunks of data, known as components. vSAN intelligently places object components across the hosts in the vSAN cluster to ensure the level of redundancy by the assigned storage policy. Figure 1 illustrates the relationship between objects, components, and replicas.

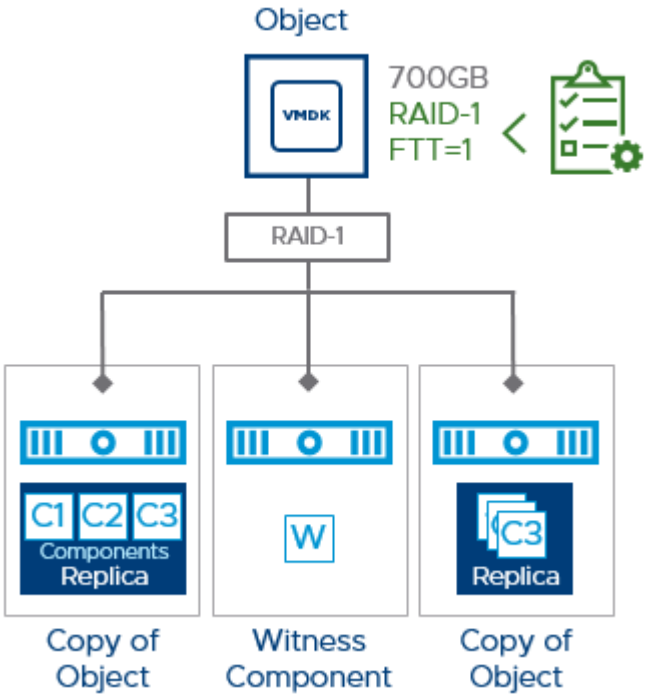

Figure 1. The relationship between an object, components, and replicas

vSAN contains all of the intelligence necessary to automatically manage the distribution of object components across the hosts that constitute a vSAN cluster, and will actively rebuild or resynchronize components when VM objects are not currently adhering to their defined storage policies, severely imbalanced, or in the event of operational changes in the environment. Figure 2 illustrates the components that make up a replica object being moved to another host by way of resynchronization.

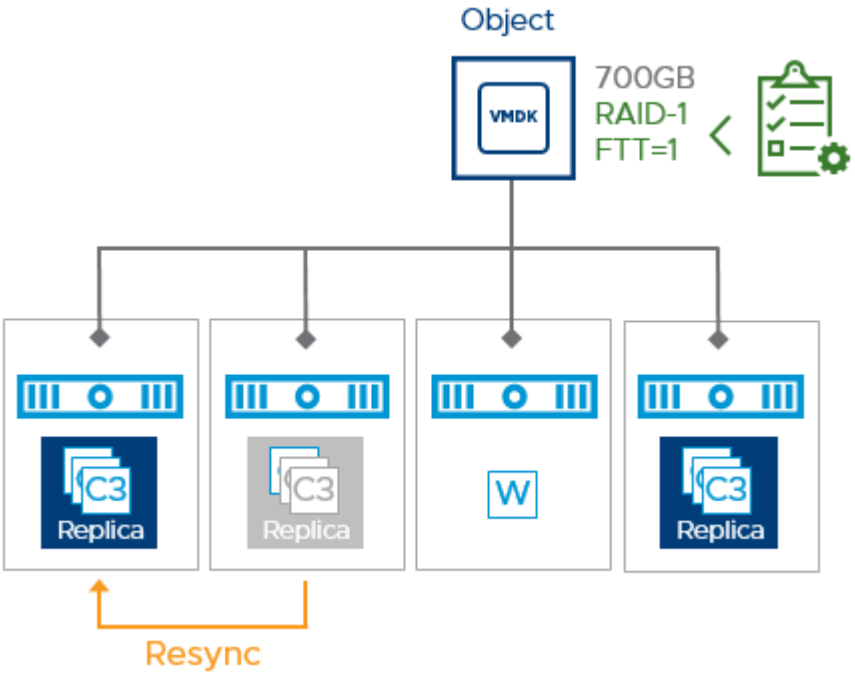

Figure 2. Components that comprise a replica being moved to another host using resynchronization

While the objective of resynchronizations is to restore or ensure the level of resiliency defined for a given VMDK, VM or collection of VMs, the cause for a resynchronization can vary. Some of the reasons for resynchronizations include:

- Object policy changes
- Host or disk group evacuations
- Host upgrades (Hypervisor, on-disk format)
- Object or component rebalancing
- Object or component repairs

Resynchronization traffic shares the same physical fabric and logical data path used by VM I/O, as shown in Figure 3.

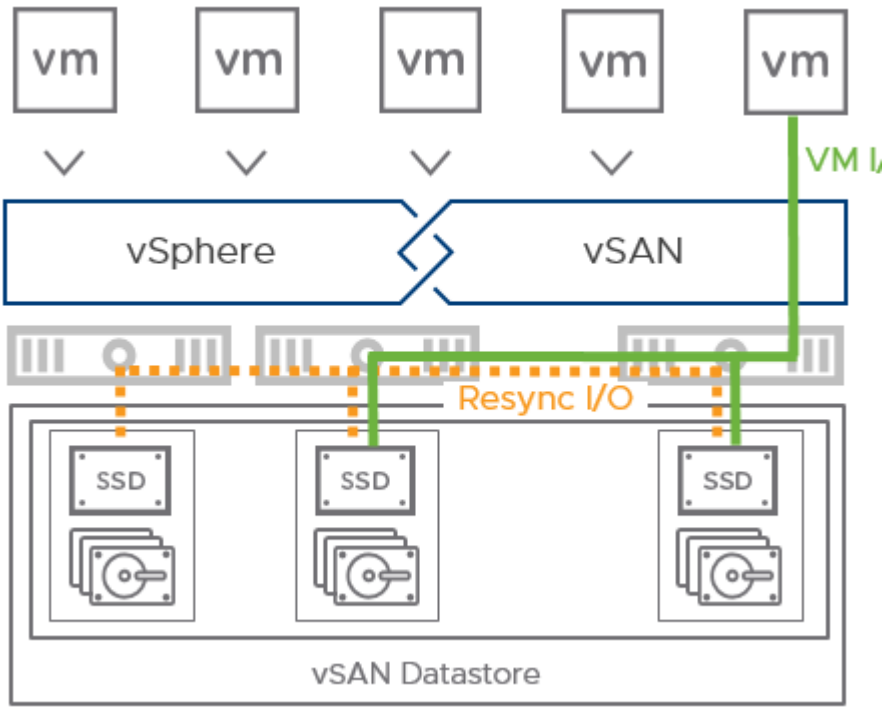

Figure 3. Front end VM I/O, and back end resynchronization traffic

Resynchronization traffic can impact the performance of VMs running in a vSAN cluster, and improvements have been made to minimize this impact. Intelligent rebuilds are a collection of improvements over several versions of vSAN that focused on reducing the amount of data that needed to be resynchronized during decommissioning, rebalancing, and repairs. Improvements made in vSAN 6.7 improved the ability to properly regulate the flow of resynchronization data versus VM I/O during periods of contention - ensuring a fair distribution of resources, and more predictable performance and behavior of VMs powered by vSAN.

Past versions of vSAN made iterative improvements in controlling resync activity. vSAN 5.5 through 6.0 provided some control, but was limited in I/O management capabilities, and only available through the CLI. In vSAN 6.6, a manual throttling mechanism was introduced in the UI allowing a user to define a static limit on resync I/O. Much like previous versions, the mechanism was primarily designed to address temporary conditions. It required manual intervention, knowledge of performance metrics, had limited abilities to control I/O types at specific points in the storage stack. **This manual throttling mechanism has been removed from more recent editions of vSAN in favor of the automated Adaptive Resync method described below.** vSAN 6.6.1 introduced an automated mechanism to reduce starvation of I/Os to VMs during heavy resync activity. While this did reduce conditions of total saturation of resync I/Os, the underlying scheduler could not properly control the types of I/O in an optimal way.

Adaptive Resync

Adaptive Resync found in vSAN 6.7 and newer is a mechanism to properly regulate the flow of resynchronization I/O versus VM I/O during periods of contention. It represents a new level of capabilities and control that allows vSAN to share storage bandwidth resources by I/O class.

In order to provide a mechanism to ensure the proper priority and fairness of I/Os under a variety of conditions, the solution needed to be able to achieve the following:

- Establish and identify different classes of I/O as they occur.
- Have a way to regulate bandwidth for classes of I/O at various layers in the storage stack.
- Have logic that can intelligently decipher conditions of I/O types, to dynamically place back pressure and regulate as needed.

The following details primary design elements that make up Adaptive Resync in vSAN.

Awareness and Control of I/O Types

vSAN distinguishes four types of I/O, and has a pending queue for each I/O class, as shown in Figure 4.

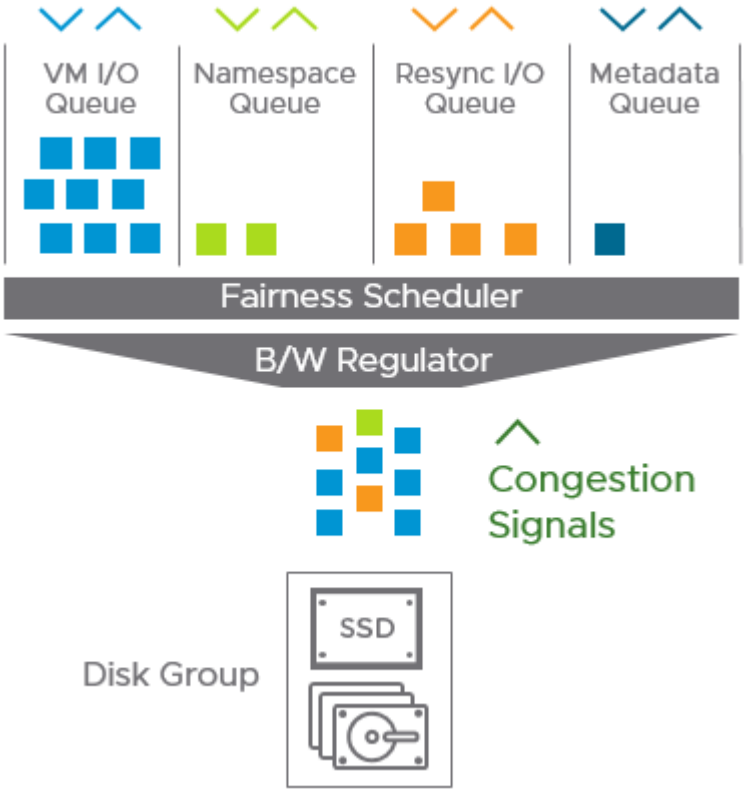

Figure 4. Adaptive Resync found in vSAN

VM I/O represents all read and write requests from the guest VM. This includes VM I/Os that are committed or fetched from multiple hosts as a part of the assigned storage policy. Resync I/O represents all I/O activities that occur as the result of resynchronization, rebuild, or repair operations. Metadata I/O represents all I/O related to the management of the objects that comprise a VM, such as witness traffic, as well as cluster monitoring and membership activities. Namespace I/O relates to all I/O related to the namespace object of the VM.

Measuring Bandwidth

Adaptive Resync is able to determine the sustainable bandwidth of disk groups, and is part of the "Bandwidth Regulator." The bandwidth measurement is derived from the rate at which that data can be destaged from the disk group's write buffer device to the capacity devices. The determined level of bandwidth is not a static entry, but rather, **a continuous feedback loop based on congestion signals associated with each I/O** . Each I/O provides a signal to tell vSAN the level of congestion for the resource indicating the highest level of contention. These signals can come as a result of write buffer crossing certain capacity

thresholds, or a disk controller's queue depths being exhausted. Separate bandwidth regulators exist for reads and writes to provide more granular control of scheduling of I/Os.

Congestion Control

vSAN employs a sophisticated, highly adaptive congestion control scheme to manage I/O from one or more resources. It is used as a feedback mechanism to measure contention, and take appropriate action. vSAN's congestion control accommodates for the contention that may be occurring at various layers within the distributed storage system. It can redistribute the congestion in such a way to limit the impact on performance across the entire system, while avoiding the need to impart artificial latency to relieve the congestion.

This type of a congestion control scheme is used because simple queueing found on most traditional, non-distributed I/O systems would not be able to identify and control the points of contention. Some of the congestion control activity in vSAN can be viewed courtesy of the performance service. The congestions metric can be found in the vCenter Server UI, under the vSAN tab, at a cluster level, host, disk group, disk, and VM levels.

vSAN has two distinct types of congestion to help regulate I/O, improving upon the single congestion type found in older versions of vSAN (vSAN 6.6 and earlier). This allows for greater control in the shifting of latency to one of two locations higher up in the stack.

- **Bandwidth congestion.** This type of congestion can come from the feedback loop in the “bandwidth regulator” described above, and is used to tell the vSAN layer on the host that manages vSAN components the speed at which to process I/O. Bandwidth congestion is visible in the UI by highlighting the host, clicking Monitor > vSAN Performance > Disks, and selecting the disk group.
- **Backpressure congestion.** This type of congestion can come as the result of the pending queues (shown in Figure 4) for the various I/O classes filling to capacity. Backpressure congestion can shift this delay up to the highest layer queue within vSAN. This is the queue that is associated with a specific object, such as a VMDK. Backpressure congestion is visible in the UI by highlighting the cluster, clicking Monitor > vSAN > Performance, and selecting the “VM” category.

The benefit to this optimized congestion control method is the ability to better isolate the impact of congestions and improve resource utilization. Shifting latency to another layer higher in the stack can avoid exhausting memory and disk resources lower in the stack, and avoid imparting additional, artificial latency beyond what the system is already seeing. Adaptive Resync in vSAN makes use of the congestion control scheme as described.

RECOMMENDATION: Use the performance service to help determine where contention may be occurring. Resynchronization performance is subject to the performance capabilities of the underlying hardware, and fabric that connect the hosts comprising the vSAN cluster.

The Dispatch/Fairness Scheduler

The dispatch scheduler is at the heart of vSAN’s ability to manage and regulate I/O based on the conditions of the environment. In vSAN 6.7 and later, the scheduler driving Adaptive Resync was designed to handle the inherent parallelism found across a distributed storage environment. When combined with other elements of vSAN's architecture as described here, the dispatch scheduler now performs the following:

- Maintain a pending queue for each I/O.
- Listen to the bandwidth regulator for permission to issue the next I/O.
- Pull the next I/O from the relevant queue to ensure fairness.
- Initiate back-pressure congestion to the highest layer in the stack when a queue is full.

The ability for the scheduler to manage each individual I/O class is important. The separate queues for the I/O classes allow vSAN to prioritize new incoming I/O that may have an inherently higher level of priority over existing I/O waiting to be processed. An example would be an environment that had a significant amount of resync activity occurring with very few VM I/Os, followed by an influx of new VM I/Os. This new scheduler found in vSAN is able to immediately provide the appropriate level of priority for processing the VM I/Os.

The scheduler works in concert with the bandwidth regulator to determine when new I/Os can be processed. Based on the logic built into the scheduler, **it will freely process any I/O types as long as the aggregate bandwidth of the I/O types do not exceed the available bandwidth advertised by the bandwidth scheduler** . This allows for the maximum use of resources under periods in which there is no contention. When I/Os exceed the advertised available bandwidth, the scheduler will ensure that resync I/Os are assigned no less than approximately 20% of the available bandwidth, leaving the remaining 80% for VM and other I/O classes.

During periods of contention, the scheduler can determine from the bandwidth regulator the rate of ingress, and can continue to reduce the inflow of I/Os until the congestion signals stop growing. The scheduler will balance these different classes of I/O based on the activity in each queue, bandwidth capabilities courtesy of the congestion signaling, and be able to apply back pressure at the appropriate layer in the stack to ensure fairness of the I/O. The location in the stack in which it places backpressure is dependent on the conditions observed by vSAN.

Results

Adaptive Resync provides the mechanics necessary to implement a fully intelligent, adaptable flow control mechanism for managing resync I/O and VM I/O. The functional behavior of Adaptive Resync could be summarized as:

- When no resync activity is occurring, VM I/Os can consume up to 100% of the available bandwidth.
- When resync and VM I/O activity is occurring, and the aggregate bandwidth of the I/O classes is below the advertised available bandwidth, neither I/O class will be throttled.
- When resync and VM I/O activity is occurring, and the aggregate bandwidth of the I/O classes exceeds the advertised bandwidth, resync I/Os are assigned no less than approximately 20% of the bandwidth, allocating approximately 80% for VM I/Os.

This will maximize the ability for resynchronization traffic to use as much bandwidth as possible while not imparting performance implications on the respective VMs powered by vSAN. Figure 5 represents this behavior, shown as four distinct conditions over a period of time.

- **Period 1:** This represents the VM I/O activity being able to freely use all available bandwidth while no resync activity is occurring.
- **Period 2:** This represents resync activity beginning to occur while VM I/O activity is still busy. As the aggregate bandwidth of both I/O types near 100%, VM I/Os are slowed slightly in order to grant 20% of the bandwidth available for resync activity.
- **Period 3:** This represents VM I/O and resync I/O activity that activity that is still very busy, exceeding the advertised bandwidth, and both maintain their designated allocation.
- **Period 4:** This represents VM I/O activity where VMs are generating less I/O activity. Adaptive Resync allows resync activity to consume the remaining available bandwidth.

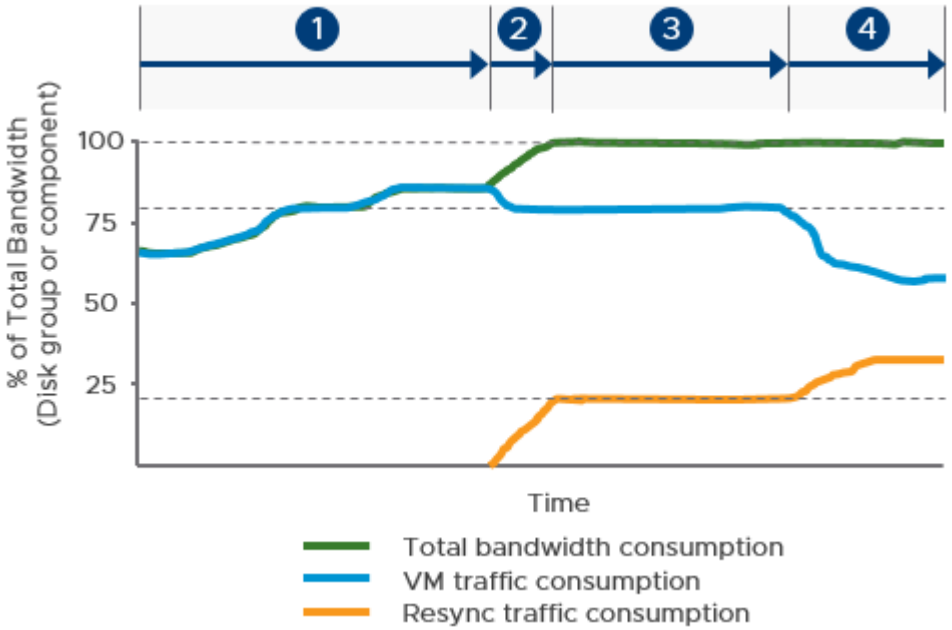

Figure 5. A time-based example of bandwidth consumption with Adaptive Resync.

The key to this logic is that it is highly adaptive based on the current conditions of the environment. Workloads and the VM I/O they generate are inherently bursty. Adaptive Resync accommodates for these conditions. Giving priority to the guest VM I/O while opportunistically allowing vSAN to use as much bandwidth as possible during resync activity.

RECOMMENDATION: Review the vSAN Design and Sizing Guide. The performance of vSAN is heavily dependent on the underlying hardware used for the vSAN cluster. This includes, but is not limited to caching devices, capacity devices, HBAs, Network uplinks, and switch fabrics.

Conclusion

Awareness and control of primary and secondary I/O operations are one of the many benefits of having a distributed storage system like vSAN integrated directly into the hypervisor. Beginning with vSAN 6.7 and continuing through the most recent editions of vSAN, Adaptive Resync takes advantage of VMware's abilities to optimize and tune vSAN in a way that achieves an all new level of consistency of VM performance under a wide variety of conditions.

About the Author

This content in this document was assembled using content from various resources from vSAN Engineering and vSAN Product Management.

Pete Koehler is part of the vSAN Technical Marketing Team in the Cloud Platform Business Unit at VMware, Inc. He specializes in enterprise architectures, data center analytics, software-defined storage, and hyperconverged Infrastructures. Pete provides more insight into the challenges of the data center at vmpete.com , and can also be found on twitter at @vmpete .