DELL EMC VxRAIL™ BEST PRACTICES FOR DATA SERVICES

ABSTRACT
This white paper discusses guidance for data services features Deduplication & Compression, Erasure Coding for storage space efficiency and Quality-of-Service limit, Software Checksum for data integrity, and Fault Domains for high availability pertaining to VxRail appliance.

December 2016
The information in this publication is provided “as is.” Dell Inc. makes no representations or warranties of any kind with respect to the information in this publication, and specifically disclaims implied warranties of merchantability or fitness for a particular purpose.

Use, copying, and distribution of any software described in this publication requires an applicable software license.

Copyright © 2016 Dell Inc. or its subsidiaries. All Rights Reserved. Dell, EMC, and other trademarks are trademarks of Dell Inc. or its subsidiaries. Other trademarks may be the property of their respective owners. Published in the USA 12/16 Technical White Paper H15662.

Dell EMC believes the information in this document is accurate as of its publication date. The information is subject to change without notice.
# Table of Contents

Intended Use and Audience ................................................. 5
Introduction to VxRail ....................................................... 5
Introduction to vSAN ......................................................... 6

Deduplication and Compression ........................................ 6
   Enabling Deduplication and Compression .......................... 7
   Disabling Deduplication and Compression ......................... 8
   Deduplication and Compression Monitoring ....................... 9
   VxRail Guidance for Deduplication and Compression ............ 10
   Additional Information on Deduplication and Compression .... 10

Erasure Coding ................................................................. 10
   RAID-5 Erasure Coding ................................................ 11
   RAID-6 Erasure Coding ................................................ 11
   VxRail Considerations for Erasure Coding ....................... 13
   Additional Information on Erasure Coding ....................... 13

Quality of Service Limits .................................................. 13
   Defining IOPS Limits .................................................. 14
   VxRail Considerations for Quality of Service .................... 14
   IOPS limit set to 1000 ............................................... 15
   Additional Information on Quality of Service .................... 15

Software Checksum ......................................................... 15
   Disabling Software Checksum ....................................... 15
   VxRail Considerations for Software Checksum .................... 16
   Additional Information on Software Checksum .................... 16

Fault Domains ................................................................. 16
   vSAN Data Placement ................................................ 16
   Fault Domains Considerations for VxRail ........................ 17

   Scenario 1: Fault Domains with S, P, V, or E Series Appliances 19

   Scenario 2: Fault Domains with VxRail G Series ................. 19
Conclusion

Additional Resources
INTENDED USE AND AUDIENCE

This guide discusses essential details for how and when to use data services on VxRail appliances. It has been prepared for anyone involved in planning, installing, and maintaining VxRail, including EMC field engineers and customer system and network administrators. Please work with your EMC or Partner implementation representative to perform the actual installation.

INTRODUCTION TO VXRAIL

VxRail is a jointly VMware and Dell EMC developed hyper-converged infrastructure family of appliances built on the foundations of VMware Hyper-Converged Software (vSAN, vSphere and vCenter). VxRail comes bundled with vSAN, vCenter, Log Insight and a full suite of industry-leading data services from EMC and VMware, including recovery, data protection, cloud tiering and backup. With a single point of support for the hardware and software in a multiple form factors (2U 4-node, 1U 1-node, and 2U 1-node), VxRail is the simplest building block for the Software Defined Data Center (SDDC). A quick view of the software that is inside VxRail is shown in the below image:

What’s Inside VxRail

VxRail is a turnkey and curated vSAN experience. Each appliance of VxRail 3.5 starts with 4-node (meaning 4 ESXi servers), and the new VxRail 4.0 can start with as little as 3 nodes. Scaling can be as granular as drives or in single node increments up to 64 nodes. Each node in VxRail appliance is configured with at least one disk group (a main unit of storage on a node). Each disk group includes one SSD for caching and one or multiple HDDs/SSDs for capacity depending vSAN hybrid or all-flash configuration. vSAN aggregates these disk groups from to create a single datastore and each host can have more than one disk group (up to 5 in each node).
INTRODUCTION TO VSAN

VxFrail is powered by VMware vSAN for radically simple, enterprise-class native storage in a hyper-converged appliance. Uniquely embedded in the hypervisor, vSAN delivers high performance, flash-optimized hyper-converged storage for any virtualized application. Flash devices and/or hard disks are attached to the server to provide an optimized and highly resilient shared datastore that is suitable for a variety of workloads.

High Level VMware vSAN Architecture

vSAN provides many enterprise-ready storage data services for Availability, I/O Acceleration, Data Security etc. In this document we will discuss some of these vSAN core features pertaining to VxFrail appliances. In this document we will talk about following features:

- **Deduplication and Compression**: You can enable deduplication and compression on a vSAN cluster to eliminate duplicate data and reduce the amount of space needed to store data.

- **Erasure Coding**: You can set the failure tolerance method policy attribute on VMs to use RAID 5 or RAID 6 erasure coding. Erasure coding can protect your data while using less storage space than the default RAID 1 mirroring.

- **Quality of Service Limits**: You can set an IOPS limit for an object, such as VMDK. If a disk’s IOPS exceeds the limit, I/O operations are throttled.

- **Software Checksum**: By default, checksums are enabled to ensure data integrity. Each object calculates checksum information. You can disable checksums.

- **Fault Domains**: A high availability feature to ensure protection against physical device or rack failures.

DEDUPLICATION AND COMPRESSION
vSAN provides VxRail appliances with storage deduplication and compression to provide cost savings for your storage needs in all-flash clusters. It cannot be used in a hybrid disk environment. These two features work together to provide operational efficiency with fewer storage resources.

vSAN can perform block-level deduplication and compression to save storage space. Enabling deduplication and compression on a vSAN all-flash cluster reduces redundant data within the disk group.

Deduplication removes redundant data blocks, whereas compression removes additional redundant data within each data block. These techniques work together to reduce the amount of space required to store the data. vSAN applies deduplication and then compression as it moves data from cache to the capacity tier, as illustrated in the following figure.

Deduplication and compression are applied to a disk group. The Health page of the vSphere Web Client provides detailed reporting of deduplication and compression data. vCenter Server management features allow you to enable/disable this feature for the entire VxRail cluster.

**Enabling Deduplication and Compression**

To enable, go to the vSphere Web Client to edit vSAN settings. Change the “add disks to storage” setting from automatic to manual and enable deduplication and compression.

When the VxRail disk group is enabled for deduplication and compression, all data must be evacuated. The upgrade of the disk format is done in a rolling fashion with no downtime; however, performance may be impacted.
Disabling Deduplication and Compression

To disable, go to the vSphere Web Client and change the deduplication and compression setting to disabled. Deduplication and compression can be disabled on a live cluster but got some implications. Disabling deduplication and compression will evacuate all of data going through each disk group and reformatting the disk groups. It may take a while for this operation depending on cluster size and number of disk groups.
Disabling deduplication and compression

**Deduplication and Compression Monitoring**

Go to the vSphere Web Client to see how much storage you are saving with deduplication and compression. The capacity overview section shows utilization broken down by disk, objects, and overhead for the vSAN cluster. The used capacity breakdown section shows how much data is used compared to the overhead for duplicate data storage and metadata. You can also discern VMDK data from other objects.

You can expect up to 5% overhead of your raw storage capacity used to store metadata. CPU and memory require minor overhead as well.

**Storage capacity, savings, and overhead**
VxRail Guidance for Deduplication and Compression

- Only VxRail all-flash appliances support deduplication and compression and by default, deduplication and compression are disabled on VxRail.

- If you want to use deduplication and compression, we recommend enabling them before VMs are deployed. If turned on after VMs are deployed, all data must be reprocessed to apply the deduplication and compression algorithms. This can be a highly resource and time intensive activity.

- Deduplication and compression are enabled together. There is no way to enable them separately. vSAN will only store compressed data if a unique 4K block can be reduced to 2K or less.

- When using deduplication and compression, it is recommended to have at least 4 nodes in the cluster to maintain high availability under all conditions. A single disk failure when deduplication and compression are enabled on nodes with a single disk group will cause the entire disk group to be unavailable until the disk is replaced. Having 4 or more nodes provides redundancy needed to maintain high availability in this scenario.

- Adding a disk to a disk group is permitted in VxRail. When deduplication and compression are enabled, the process of adding or removing capacity devices to a disk group, will require the disk group to be recreated.

- Enabling deduplication and compression consumes a small amount of capacity for its metadata. The space consumed by this metadata is relative to the size of the vSAN datastore and is typically around 5% of the total capacity.

- We do not recommend using vSAN deduplication if you are using encryption offerings that apply algorithms at the block level.

- Object space reservation must be 0% or 100% of all storage policies, although 100% is not efficient for deduplication. This number represents the logical size of the VM object that must be allocated, or thick provisioned when deploying virtual machines. Setting it to zero means that the virtual disk will be thin provisioned.

Additional Information on Deduplication and Compression

To know more about these features, refer to following documents:

- [VMware vSAN 6.2 Space Efficiency Technologies](#)
- [Deduplication and Compression Blog](#)

ERASURE CODING

vSAN 6.2 provides data protection for VM’s running on VxRail in two ways, “erasure coding (RAID-5/6)” and “mirroring (RAID-1)”. Mirroring creates an exact copy of VM object and keep it in a separate host so that if one fails it can be served from other copy. RAID 5 or RAID 6 erasure coding provides an alternate way to protect against data loss and increase storage efficiency. Erasure coding can provide the same level of data protection as mirroring (RAID-1) while using less storage capacity, though it does have increased technical requirements. Data protection can be configured on a per-VM basis allowing each application and its objects to have differing protection based on its requirements.

You can configure RAID-5 on all-flash clusters with four or more VxRail nodes. You can configure RAID-6 on all-flash clusters with six or more VxRail nodes. RAID-5 protects against one failure, and RAID-6 protects against two
concurrent failures. Erasure coding is compatible with vSAN's deduplication/compression, quality of service, and checksum features.

To enable erasure coding, go to vSphere Web Client to edit the VM Storage Policy. Select erasure coding instead of mirroring.

Erasure Coding Policy Setting

It is recommended to create different policies such as “RAID-1” or “RAID-5/6” based on VM performance or capacity requirement and apply accordingly to VM objects.

RAID-5 Erasure Coding

RAID-5 provides protection against one failure by configuring erasure coding **Number of failures to tolerate** to one (FTT=1). A minimum of four nodes (3+1) are required, but the overhead is 1.33x instead of 2x with mirroring. In other words, a 20GB disk only needs about 27GB to provide data protection, instead of the 40GB needed for mirroring.

FTT=1 and RAID-5 (Erasure Coding)

RAID-6 Erasure Coding
RAID-6 provides protection against two failures by configuring erasure coding **Number of failures to tolerate** to two (FTT=2). A minimum of six nodes (4+2) are required, but the overhead is 1.5x instead of 3x with mirroring. In other words, a 20GB disk only needs about 30GB to provide data protection, instead of the 60GB needed for mirroring.

**FTT=2 and RAID-6 (Erasure Coding)**
VxRail Considerations for Erasure Coding

The following table contrasts the requirements and recommendations for data protection via mirroring versus RAID 5/6 erasure coding. Converting between the two is driven by storage policy and requires additional capacity during conversion.

<table>
<thead>
<tr>
<th>Failures to Tolerate (FTT)</th>
<th>RAID-1 (Mirroring)</th>
<th>RAID-5/6 (Erasure Coding)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Node or Fault Domain Minimum</td>
<td>Required Nodes</td>
</tr>
<tr>
<td></td>
<td>For policy compliance without failures</td>
<td>To allow in-place rebuilds</td>
</tr>
<tr>
<td>1</td>
<td>3*</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>3</td>
<td>7</td>
<td>8</td>
</tr>
</tbody>
</table>

*VxRail 3.5 starts with 4 hosts as minimum.

Additionally, we need to keep in mind the performance impact when erasure coding is utilized. This is irrespective of VxRail and is applicable to any vSAN deployment leveraging this feature. E.g. using erasure coding impacts random write performance due to the required read-modify-write steps. In addition, there is CPU overhead to compute parity. VxRail users can expect the same overhead when leveraging this feature.

One final point to remember is that vSAN Stretched Clusters do not support the use of Erasure Coding in vSAN 6.2, as only 3 fault domains are supported.

Additional Information on Erasure Coding

To know more about these features, refer to following documents:

- [VMware vSAN 6.2 Space Efficiency Technologies](#)
- [Erasure Coding Blog 1](#)
- [Erasure Coding Blog 2](#)

QUALITY OF SERVICE LIMITS

Quality of Service (QoS) is available on a per VM/VMDK basis by setting an IOPS (input/output operations per second) limit. If a disk’s IOPS exceed the specified limit, I/O operations are throttled. If the IOPS limit for an object is set to 0, IOPS limits are not enforced.

These limits allow for different levels of service for different applications. Service Level Agreements (SLAs) can be provided for different workloads and to protect Tier-1 application performance. IOPS limits can prevent noisy neighbor workloads from impacting higher priority workloads. This feature is transparent to the VxRail Manager and fully compatible with other vSAN features.

To determine QoS levels, IOPS are calculated on all read/write operations for a VM/VMDK, including its snapshot. Read and write operations are regarded as equivalent. The vSphere Web Client reports IOPS and compliance with the specified limits on the Health page of the vSphere Web Client. There is also an IOPS assessment tool to set up initial limits for QoS.
Defining IOPS Limits
To enable an IOPS limit for an object, create a new storage policy rule and set the desired limit. The value for the limit is normalized to a block size of 32KB. Thus, a 64KB read or write is treated as two separate operations, leading to half of the configured IOPS limit being the number of operations performed. If there is no limit for I/O, use the default value of zero.

Defining the IOPS limit for a disk

VxRail Considerations for Quality of Service
Best practices for QoS are not to create policies for all workloads, nor even your priority workloads. Instead, use QoS to avoid the "noisy neighbor" issue. Create limits for those lower priority applications which you don't want to consume excess resources.

The following examples show performance with and without setting an IOPS limit.
IOPS limit set to 1000

Additional Information on Quality of Service
To know more about these features, refer to following documents:

- Quality of Service Blog

SOFTWARE CHECKSUM
End-to-end checksums can be enabled to ensure the integrity of data by confirming that each copy of a file on VxRail is the same as the source file. Checksums provide an additional level of data integrity that is transparent to the VxRail Manager. Checksums are used to detect and resolve silent disk errors and corruption in memory and disk components. When checksum verification fails, data is fetched from the redundant copy. If a checksum mismatch is detected, vSAN automatically repairs the data by overwriting the incorrect data with the correct data. Checksum calculation and error-correction are performed as background operations.

Disabling Software Checksum
Checksums are enabled by default in VxRail. It is recommended practice to keep software checksum enabled. But they can be turned off in the vSphere Web Client if application already provides data integrity mechanism. To disable checksums, edit the storage policy rule set to “Disable object checksum.” Note that turning off checksums is an immediate operation, but turning checksums back on requires a full data copy to apply the checksum. This can be a resource and time intensive activity.
Storage policy rule for software checksum

VxRail Considerations for Software Checksum

As a best practice for VxRail, keep software checksum enabled whenever you create a new workload. If you disable it on initial creation, all data will need to be rewritten once the policy is changed to enable checksum. This process creates substantial CPU and resource overhead if there is a significant amount of data.

Checksums are implemented in software and accelerated by Intel crc32c instructions. Checksums are calculated as early as possible and verified as late as possible to provide end-to-end validation. Both network and disk errors can be detected. vSAN checksums protect both hybrid and all-flash VxRail appliances and nodes.

Additional Information on Software Checksum

To know more about these features, refer to following documents:

- [Software Checksum Blog](#)

FAULT DOMAINS

Fault Domains is a feature for advanced resiliency whereby VxRail appliance clusters can be segmented into fault domains to provide added level of protection across physical components. This can help minimize virtual machine downtime when an entire physical rack fails or goes down in the environment. It achieves this by defining a logical separation that spans a physical construct, for example physical racks. In this example, nodes in a rack could be assigned to a fault domain. vSAN would place the redundant VM objects/components in separate fault domains, across racks in different fault domains, to maintain availability of the VM object/component in the event of a rack failure.

vSAN Data Placement
Before we discuss fault domains feature in detail, it is important to understand how vSAN places its VM data across different nodes in the cluster.

vSAN implements Storage Policy-Based Management, and each virtual machine deployed in a vSAN datastore has at least one assigned policy. When the VM is created and assigned a storage policy, the policy requirements are pushed to the vSAN layer. These policies determine how storage objects are provisioned and allocated within the datastore to guarantee the required level of service; i.e., the number of failures that you want the VM to be able to tolerate, also known as “NumberOfFailuresToTolerate (FTT)”

Let us take an example of Number of failures to tolerate (FTT) set as “1”.

vSAN Policy: Number of Failures to Tolerate (FTT) = 1

In this example, we can think of these four hosts as one VxRail appliance. When we create a VM with FTT=1 policy, vSAN creates a copy of the disk and places it in another node. In case one node fails, the available copy can serve any data requests. This is at very high level and many things go in back ground to achieve this resiliency.

Alone, the FTT policy can only prevent “host failures.” When you scale, or when you want to make sure that VM data is available beyond “host failures” vSAN fault domains provides the functionality you need.

**Fault Domains Considerations for VxRail**

VxRail comes with different form factors and node types. Based on different configurations, nodes can be deployed within a single chassis or across multiple chassis. A minimum of three nodes is required for a cluster, this could be 3 appliances consisting of a single node or a 3 node G series appliance. When we define fault domains for VxRail appliance, we need to consider what appliance types and configurations are used in the VxRail cluster to determine how to layout the appliances across racks so that we can create rack level resiliency. The G Series appliances can hold up to 4 nodes in a chassis so all nodes contained would belong to a single rack. The E, S, P and V Series appliances have each node in an individual chassis and they can either be placed in single rack or across various racks within the datacenter to create VxRail cluster.

The following recommendations for fault domains are specific to VxRail clusters:
• To allow for complete chassis failure, all nodes(s) in the same VxRail chassis should be part of the same fault domains.

• Nodes requirement will vary based on failure to tolerate definition and placement of nodes when fault domains configured.

• We recommend that each fault domains have the same number of nodes to ensure that enough resources are available to provide full coverage in the event of a failure.

• Fault domains can be defined for VxRail infrastructure software as well as additional services components, including vCenter Server and Log Insight, according to the protection plan for the data center.

For G Series appliance in a single chassis VxRail cluster, use the following guidelines for fault domains:

<table>
<thead>
<tr>
<th>NumberOfFailuresToTolerate</th>
<th>Method*</th>
<th>Minimum Number of VxRail Nodes</th>
<th>Minimum Number of VxRail G Series Appliances</th>
<th>Number of Fault Domains Required</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>RAID-1</td>
<td>9</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>1</td>
<td>RAID-5</td>
<td>12</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>RAID-1</td>
<td>15</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>RAID-6</td>
<td>15</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>RAID-1</td>
<td>21</td>
<td>7</td>
<td>7</td>
</tr>
</tbody>
</table>

*FTT can be defined with mirroring or with erasure coding, RAID5/6. For more information, refer to VMware vSAN Storage Efficiency Services Best Practices Guide.

For all other appliance Series in single chassis for each node VxRail cluster, use following guidelines:

<table>
<thead>
<tr>
<th>NumberOfFailuresToTolerate</th>
<th>Method*</th>
<th>Minimum Number of VxRail Nodes</th>
<th>Recommended Number of VxRail Nodes for Self Healing</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>RAID-1</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>1</td>
<td>RAID-5</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>RAID-1</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td>RAID-6</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>3</td>
<td>RAID-1</td>
<td>7</td>
<td>8</td>
</tr>
</tbody>
</table>
SCENARIO 1: FAULT DOMAINS WITH S, P, V, OR E SERIES APPLIANCES

In this example, we are demonstrating another scenario where three nodes of the VxRail appliances are placed across multiple racks. The same fault domains feature can now be configured in different way. Instead of multiple nodes in same chassis within single rack, we can configure the nodes in different racks for the VxRail appliance and provide similar level of resiliency.

Fault Domains configured for one VxRail appliance using single node chassis series models

SCENARIO 2: FAULT DOMAINS WITH VXRAIL G SERIES

Let us take an example of a VxRail cluster where there are three 4-node G Series VxRail appliances, each in a separate rack. When a virtual machine that tolerates one failure is deployed, it is possible for both replicas to be deployed to different hosts in the same rack when fault domains are not used. Thus, if the rack, chassis, or network switch goes down, the VM is not protected.

However, if fault domains are enabled, the hosts in the same chassis can be grouped together to form a fault domains. vSAN never places a copy (replica) of a virtual machine’s data in the same fault domain as the original. To calculate the number of fault domains required to tolerate a specific number of failures, use the table above. Each fault domain contains one or more VxRail appliance chassis.

In this example, we configured three fault domains according to the recommended best practices.
Fault domains configured in three G Series VxRail appliances distributed across chassis

We used the vSphere Web Client to configure the following fault domains:

- Fault Domains 1 contains VxRail appliance 1 with three to four nodes (rack 1)
- Fault Domains 2 contains VxRail appliance 2 with three to four nodes (rack 2)
- Fault Domains 3 contains VxRail appliance 2 with three to four nodes (rack 3)
CONCLUSION

VxRail is powered by vSAN which provides many capabilities for data protection, efficiency and quality of service that apply to different use cases. Each of these capabilities have specific best practices that can help you maximize the use of your VxRail appliances when used. When using any of these features make sure that you understand the intended use and implications and follow the guidance laid out in this document and in supporting material.

ADDITIONAL RESOURCES

For further reading, check out following materials:

For additional information about VxRail, vSphere, and vSAN, use the following resources:

- VxRail Tech Book
- vSphere 6.0 Documentation
- vSAN 6.2 Design and Sizing Guide
- VxRail Fault Domains Solution Brief
- vSAN Data Placement Blog

For additional information about vSAN fault domains, see these resources:

- Designing and Sizing vSAN Fault Domains
- Managing Fault Domains in vSAN Clusters
- Fault Domains Blog 1
- Fault Domains Blog 2
- Fault Domains Blog 3