# Citrix XenServer ® 7.5 Thin Provisioning for Shared Block Storage Devices

Published July 2018
1.0 Edition

Citrix XenServer ® 7.5 Thin Provisioning for Shared Block Storage Devices


Version: 7.5

Citrix Systems, Inc.
851 West Cypress Creek Road
Fort Lauderdale, FL 33309
United States of America

**Disclaimers**



**Trademarks**

Citrix®
XenServer ®
XenCenter ®

CITRIX

# Contents

CITRIX

# Chapter 1. Introduction

This document contains procedures and best practices to guide you through using the experimental feature included in XenServer 7.5 to set up thin provisioning with a shared block storage device.

In previous releases of XenServer, thin provisioning was available only for VDIs stored on local storage (EXT) or file-based shared storage devices. XenServer 7.5 uses GFS2 to make thin provisioning available to customers with block-based storage devices that are accessed through iSCSI software initiator or Hardware HBA.

Thin provisioning optimizes the utilization of available storage by allocating disk storage space to VDIs as data is written to the virtual disk, rather than allocating in advance the full virtual size of the VDI. Using thin provisioning enables you to significantly reduce the amount of space required on a shared storage array and with that your Total Cost of Ownership (TCO).

These storage and cost reductions are especially great where you are using XenServer to host many VMs that are all based on the same initial image. In this case, a pool set up to use thin provisioning can store small difference disks for each of the VDIs that can grow as more data changes in relation to the initial image. Without thin provisioning, the difference disk for each VDI would be allocated as much storage space as the full image.

CITRIX

# Chapter 2. Getting started

This section steps through the process of setting up a clustered pool and connecting it to GFS2 shared block storage device.

## 2.1. Prerequisites

Before you start, ensure that you have the following items:

- XenCenter 7.5.
- A pool of three or more XenServer 7.5 hosts.
  - Dom0 must have at least 2 GiB of RAM.

    If you are working with disks that are greater than 2 TiB, Dom0 must have more than 8 GiB of RAM.
  - All hosts must be in the same site and connected by a low latency network.
  - The IP addresses of these hosts must not change. If you are using DHCP, ensure that you use static assignment for these hosts.
- Block-based storage device that is visible to all hosts.
- A bonded network to use for clustering.

Ensure that high availability is disabled on your pool before beginning.

If you have a firewall between the hosts in your pool, ensure that hosts can communicate on the cluster network using the following ports:

- TCP: 8892, 21064
- UDP: 5404, 5405

For more information about setting up these prerequisites, see Citrix XenServer Administrator's Guide.

## 2.2. Enabling the experimental feature

Thin provisioning for shared block storage devices is an experimental feature. As such, the feature is not enabled by default. To enable this feature, run the following command on all XenServer hosts in the pool you intend to add the GFS2 SR to:

```
xe-enable-experimental-feature corosync
```

## 2.3. Setting up a GFS2 by using the xe CLI (Recommended)

To use the xe CLI to create and configure your pool, run the following commands on a host in your pool:

1. For every PIF that belongs to this network, set `disallow-unplug=true`:

   ```
   xe pif-param-set disallow-unplug=true uuid=<pif_uuid>
   ```
2. Enable clustering on your pool:

   ```
   xe cluster-pool-create network-uuid=<network_uuid>
   ```
3. Create a GFS2 SR. You can use iSCSI or HBA to communicate with the SR:
   - Using iSCSI:

     ```
     xe sr-create type=gfs2 name-label=gfs2 --shared device-config:provider=iscsi
      device-config:ips=<portal_address> device-config:iqns:<target_ign> device-
     config:ScsiId=<lun_scsi_id>
     ```
   - Using ISCSI with CHAP authentication:

```
xe sr-create type=gfs2 name-label=gfs2 --shared device-config:provider=iscsi
 device-config:ips=<portal_address> device-config:iqns:<target_iqn> device-
config:ScsiId=<lun_scsi_id> device-config:chapuser=<chap_user> device-
config:chappassword=<chap_password>
```

- Using HBA:

```
xe sr-create type=gfs2 name-label=gfs2 --shared deviceconfig:provider=hba
 deviceconfig:ScsiId=<LUN SCSI ID>
```

To discover the *<scsiid>* to use, you can run the `xe sr-create` command without providing the `ScsiId` parameter. The command returns a list of ScsiIds that are visible to the host.

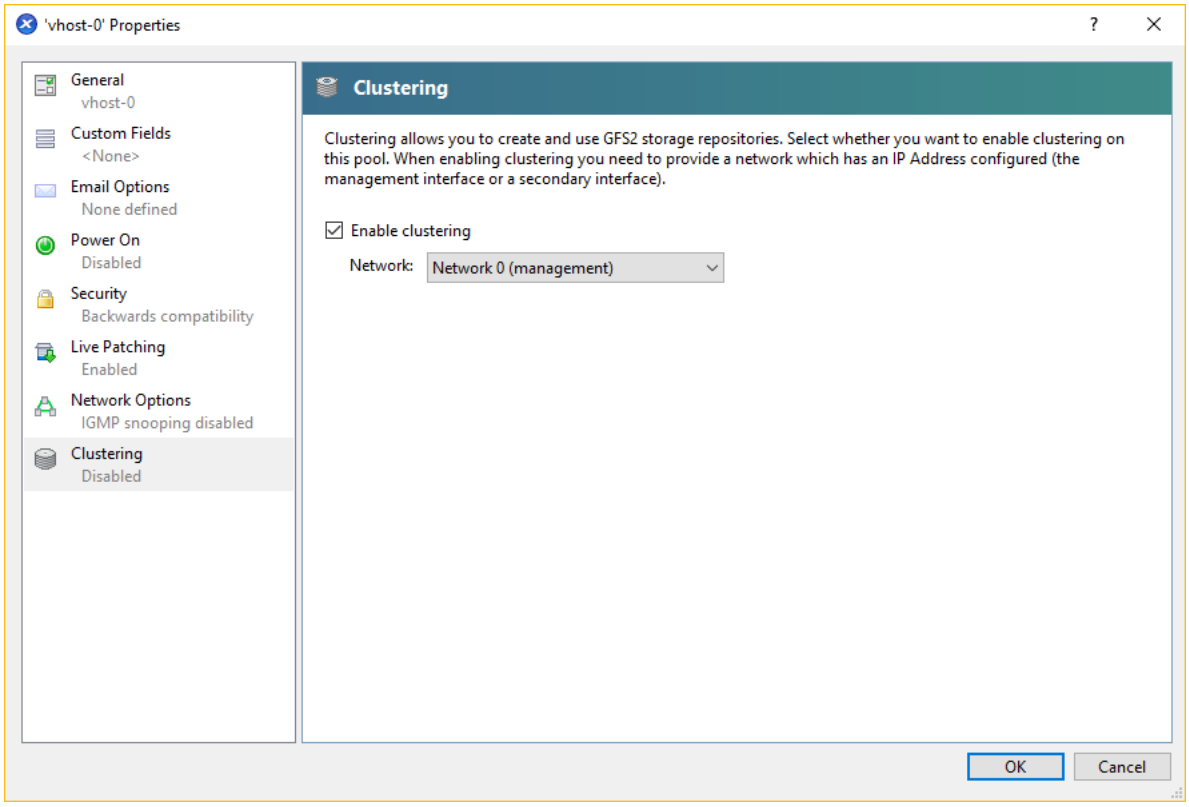
## 2.4. Setting up a GFS2 SR by using XenCenter

To use XenCenter to configure your pool to use a GFS2 SR, complete the following steps:

### Enable clustering

**Important:**We recommend that you use the command line to set up clustering on your pool. If you use XenCenter, the host fencing timeout is incorrect. For more information, see Appendix B, *Known Issues*. After you have used steps 1 and 2 in the preceding section to set up clustering for your pool, you can use XenCenter to add your GFS2 SR.

1. Select your pool in the **Resources** panel and open its **Properties** wizard.
2. Go to the **Clustering** tab.




3. Select **Enable clustering**.
4. Choose the **Network** that clustering uses to coordinate the hosts.

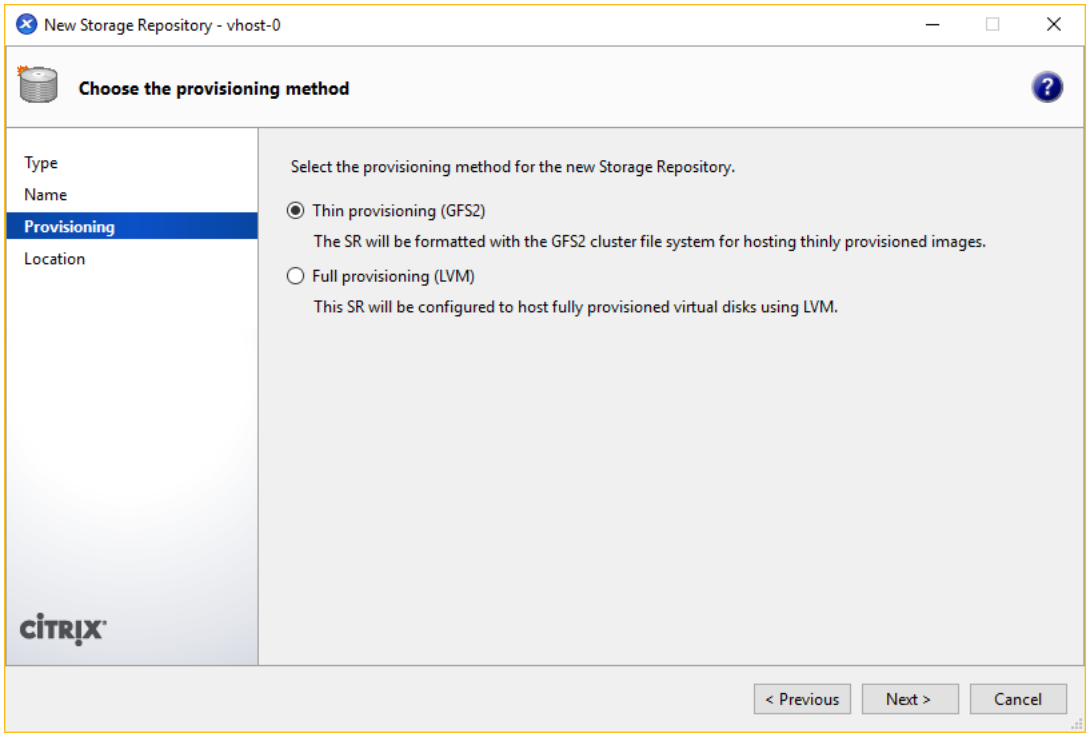
   The network used for clustering must be reliable. If the hosts can not reach the majority of other hosts over the network, they will self-fence. For more information, see Section 3, “Clustering”.

We recommend that you use a physically separate bonded network for clustering. However, if you only have the management network and storage network available to use, choose the management network.

## Add your GFS2 SR

5. From the menu, select **Storage** > **New SR**.
6. On the **Type** page, choose one of the following as the type of your new storage and click **Next**:
   - iSCSI
   - Hardware HBA

   Software FCoE is not supported with this experimental feature.
7. On the **Name** page, provide a **Name** for your SR and click **Next**.
8. On the **Provisioning** page, select **Thin Provisioning (GFS2)** as your provisioning method and click **Next**.




9. On the **Location** page, enter the location of your block-based storage device and provide any other configuration options.
10. Click **Finish** to create the new SR and close the wizard.

CITRIX

# Chapter 3. Key Concepts

XenServer pools that use GFS2 to thin provision their shared block storage behave differently to pools that use shared file-based storage or LVM with shared block storage. It is important to understand the following key concepts and consider them when configuring your XenServer deployment for thin provisioning with shared block storage.

## Clustering

Access to shared block storage from multiple hosts must be coordinated. Block storage devices provide no mechanisms for coordinating access. However, thin-provisioning for shared block storage introduces the ability to cluster the XenServer hosts in your pool. This clustering feature coordinates access to the shared storage. You must use clustering when your pool is connecting to a GFS2 block-based storage device.

A cluster is a pool of XenServer hosts that are more closely connected and coordinated than non-clustered pools. The hosts in the cluster maintain constant communication with each other on a selected network. All hosts in the cluster are aware of the state of every host in the cluster. This host coordination enables the cluster to control access to the contents of the GFS2 SR and protect against the data loss or corruption that might happen without cluster-wide locking.

To ensure the coordination between hosts is maintained, a clustered pool does not allow you to take actions that might cause the cluster state to become out of sync. It is important for each host in a cluster to know about all the other hosts in the cluster. If any of the hosts in your cluster are not contactable, XenServer prevents you from adding or removing hosts in the cluster.

## Quorum

XenServer also takes protective actions to ensure that all hosts in a clustered pool remain coordinated and to preserve your data from corruption. Each host in a cluster must at all times be in communication with at least half of hosts in the cluster (including itself). This is known as a host having *quorum*.

The quorum value for an odd-numbered pool is half of one plus the total number of hosts in the cluster: *(n+1)/2*. The quorum value for an even-numbered pool is half the total number of hosts in the cluster: *n/2*.

For an even-numbered pool, it is possible for the running cluster to split exactly in half. In this case, the running cluster decides which half of the cluster self-fences and which half of the cluster has quorum. However, when a non-running cluster that contains an even number of hosts powers up from a cold start (for example, after a power outage), *(n/2 )+1* hosts must be available before the hosts have quorum and the cluster becomes active.

If a host does not have quorum, that host *self-fences*.

## Self-fencing

If a host detects that it does not have quorum, it self-fences within a few seconds. When a host self-fences, it restarts immediately. All VMs running on the host are killed as the host does a hard shutdown. If HA is enabled, the VMs are restarted on other hosts in the cluster according to their restart configuration. When the host that self-fenced restarts, it attempts to rejoin the cluster.

If the number of live hosts in the cluster falls below the quorum value, all the remaining hosts lose quorum. In this case, all the remaining hosts self-fence and the whole cluster goes offline.

In an ideal scenario, you want to ensure that your clustered pool always has more live hosts than are required for quorum and that XenServer never fences. To ensure this, follow best practices for a resilient deployment and well-managed cluster. For more information, see Chapter 4, *Best Practices*.

CITRIX

# Chapter 4. Best Practices

The benefits of clustering your pool require that all hosts in the pool are able to communicate and coordinate with one another at all times. To increase the probability of this being the case, create a reliable and resilient deployment and ensure that you maintain the pool in a good state.

## 4.1. Creating a Resilient Clustered Pool

Smooth running of your clustered pool depends on having reliable hosts and reliable communication between your hosts. Part of this is the hardware you choose and how you set it up. You can also use existing XenServer features to ensure that your deployment is resilient in the case of hardware or software faults.

### Hardware Redundancy

Set up your hardware to protect against failures.

You can ensure that your servers have an uninterrupted power supply. This can prevent temporary power failures or fluctuations from causing your cluster to lose quorum.

You can also set up independent networks, each with their own network cards and switches. If you create your bonded network from two totally independent networks, you are protected from a failure in a single piece of network hardware.

### Network Bonding

XenServer enables you to bond two, three, or four NICs together in parallel to form one network card. This configuration provides redundancy because if one of the NICs fails another one can continue to send traffic.

A clustered pool needs a reliable network to coordinate between hosts in the cluster. To ensure that reliability, we recommend that the network specified for cluster communications is a bonded network. We also recommend that you use a dedicated network for clustering.

For more information about network bonding, see Citrix XenServer Administrator's Guide.

### Storage Multipathing

If storage multipathing is supported by the storage array, use it to define more than one path between hosts and the GFS2 SR.

Storage multipathing is not required for clustering. However, it is good practice to enable multipathing to provide redundancy in communications between host and SR. If you are also using high availability, we strongly recommend that you use storage multipathing too.

For more information about storage multipathing, see Citrix XenServer Administrator's Guide.

### High Availability

High availability (HA) is a feature that enables your VMs to be restarted on another XenServer host if one fails. Because of the increased possibility of hosts self-fencing to preserve the cluster, we recommend that you configure HA so that the VMs from the fenced host continue to run in your pool.

If you choose to use HA in conjunction with clustering, your heartbeat SR must be GFS2. For more information about high availability, see Citrix XenServer Administrator's Guide.

## 4.2. Managing your Clustered Pool

After you set up your clustered pool and your GFS2 shared block storage, it is important to monitor and maintain the pool to ensure that the risk of the pool losing quorum is minimized.

### Ensure that hosts are shut down cleanly

When a host is cleanly shutdown, it is temporarily removed from the cluster until it is started again. While the host is shut down, it does not count toward the quorum value of the cluster. The host absence does not cause other hosts to lose quorum.

However, if a host is forcibly or unexpectedly shut down, it is not removed from the cluster before it goes offline. In this case, the host does count toward the quorum value of the cluster. Its shutdown can cause other hosts to lose quorum.

### Use maintenance mode

Before taking actions on a host that might cause that host to lose quorum, put the host into maintenance mode. When a host is in maintenance mode, running VMs are migrated off it to another host in the pool. Also, if that host was the pool master, that role is passed to a different host in the pool. If your actions (for example, changing a network cable) cause a host in maintenance mode to self-fence, you do not lose any VMs or lose your XenCenter connection to the pool.

Hosts in maintenance mode are still included in the cluster and count towards the quorum value for the cluster.

You cannot change the IP address of a host that is part of a clustered pool unless you put that host into maintenance mode. Changing the IP address of a host causes the host to leave the cluster. When the IP address has been successfully changed, the host rejoins the cluster. After the host rejoins the cluster, you can take it out of maintenance mode.

To put a host into maintenance mode, in XenCenter, right-click the host from the **Resources** pane and select **Enter Maintenance Mode**.

### Recover hosts that have self-fenced or are offline

It is important to recover hosts that have self-fenced or are offline and restore them to the cluster. While these cluster members are offline they count towards the quorum number for the cluster and decreases the number of cluster members that are contactable. This increases the risk of a subsequent host failure causing the cluster to lose quorum and shut down completely.

Having offline hosts in your cluster also prevents you from performing certain actions. When clustering is enabled on a XenServer pool, every pool membership change must be agreed by every member of the pool before it can be successful. If a cluster member is not contactable, operations that change cluster membership (such as host add or host remove) fail.

For more information about why your XenServer hosts might fail, see Chapter 5, *Troubleshooting*.

For more information about managing and maintaining your XenServer hosts, see Citrix XenServer Administrator's Guide.

### Mark hosts as dead

If one or more offline hosts cannot be recovered, you can mark them as dead to the cluster.

Marking hosts as dead removes them permanently from the cluster. After hosts are marked as dead, they no longer count towards the quorum value. It is important to take this action if you risk losing quorum soon and shutting down the whole cluster. Removing hosts to decrease the total number of cluster members decreases the risk of losing whole cluster.

To mark a single host as dead, run the following command at the console of any host in the cluster:

```
xcli declare-dead dbg '[["IPv4", "203.0.113.4"]]'
```

You can only mark a host as dead if all other hosts in the cluster can be contacted. If some hosts are not contactable, the command fails and you must ensure these hosts are contactable before marking an offline host as dead. If multiple hosts are offline, you must mark all of these hosts as dead at the same time.

To mark multiple hosts as dead, use the following command:

```
xcli declare-dead dbg '[["IPv4", "203.0.113.5"], ["IPv4", "203.0.113.7"]]'
```

> **Warning:**
>
> Using the mark-as-dead capability is a permanent action. If you later recover a host that has been marked as dead, you cannot re-add it to the clustered pool. To add this system back into the pool you must do a fresh install of XenServer. This appears to the cluster as a new host.

After you have marked a host as dead by using the command line, use XenCenter to remove it from the pool.

1. Select the host in the **Resources** panel and do one of the following:
   - Right-click and select **Destroy** in the **Resources** pane shortcut menu.
   - In the **Server** menu, click **Destroy**.
2. Click **Yes, Destroy** to confirm.

> **Warning:**
>
> Destroying a server in a pool is permanent and cannot be undone.

Ensure that any host that has been marked as dead is not rebooted with access to the cluster network. This can cause data corruption.

## 4.3. Use cases for thin provisioning

Consider what use cases you employ your clustered pool and GFS2 SR for. The following use cases benefit from thin provisioning:

- Server Virtualization
- XenApp and XenDesktop (Both persistent and non-persistent)

Thin provisioning with GFS2 is of particular interest in the following cases:

- If you use shared block storage
- If you want increased space efficiency, as images are sparsely and not thickly allocated. (For decreased TCO.)
- If you want to reduce the number of input/output ooperations per second (IOPS) on your storage array, as the GFS2 SR is the first to support Storage Read Caching on shared block storage. (For decreased TCO.)

# Chapter 5. Troubleshooting

This section addresses the common issues that might occur when using a clustered pool and thin provisioning for shared block storage:

### Q: All my hosts can see each other, but I can't create a cluster. Why?

The clustering mechanism uses specific ports. If your hosts cannot communicate on these ports (even if it is able to communicate on other ports), you cannot enable clustering for the pool.

Ensure that the hoss in the pool can communicate on the following ports:

- TCP: 8892, 21064
- UDP: 5404, 5405

If there are any firewalls or similar between the hosts in the pool, ensure that these ports are open.

### Q: What if I already have an LVM partition set up on the block-based storage device?

If you have previously used your block-based storage device for thick provisioning with LVM, this is detected by XenServer. XenCenter gives you the opportunity to use the existing LVM partition or to format the disk and set up a GFS2 partition.

### Q: Why am I getting an error when I try to join a new host to an existing clustered pool?

A: When clustering is enabled on a XenServer pool, every pool membership change must be agreed by every member of the cluster before it can be successful. If a cluster member is not contactable, operations that change cluster membership (such as host add or host remove) fail.

To add your new host to the clustered pool:

- Ensure that all of your hosts are online and contactable. For more information, see Section 5, “Q: Why is my host offline? How can I recover it?”.
- If an offline host cannot be recovered, mark it as dead to remove it from the cluster. For more information, see Section 5, “Q: A host in my clustered pool is offline and I can't recover it. How do I remove the host from my cluster?”.

### Q: How do I know if my host has self-fenced?

If your host self-fenced, it might have rejoined the cluster when it restarted. To see if a host has self-fenced and recovered, you can check the `/var/opt/xapi-clusterd/boot-times` file to see the times the host started. If there are start times in the file that you did not expect to see, the host has self-fenced.

### Q: Why is my host offline? How can I recover it?

There are many possible reasons for a host to go offline. Depending on the reason the host either can be recovered or not.

The following reasons for a host to be offline are more common and can be addressed by using information in the Citrix XenServer Administrator's Guide.

- Clean shutdown
- Forced shutdown
- Temporary power failure

- Reboot

The following reasons for a host to be offline are less common:

- Permanent host hardware failure
- Permanent host power supply failure
- Network partition
- Network switch failure

These issues can be addressed by replacing hardware or by marking failed hosts as dead. For more information, see Section 4.2, “Mark hosts as dead”.

### Q: A host in my clustered pool is offline and I can't recover it. How do I remove the host from my cluster?

A: You can mark a host as dead. This removes the host from the cluster permanently and decreases the number of live hosts required for quorum. For more information, see Section 4.2, “Mark hosts as dead”.

### Q: I have repaired a host that was marked as dead. How do I add it back into my cluster?

A: A XenServer host that has been marked as dead cannot be added back into the cluster. To add this system back into the cluster, you must do a fresh installation of XenServer. This fresh installation appears to the cluster as a new host.

### Q: What do I do if my cluster keeps losing quorum and hosts keep fencing?

If one or more of the XenServer hosts in the cluster gets into a fence loop as a result of continuously losing and gaining quorum, you can boot the host with the `nocluster` kernel command-line argument. To do this, connect to the hosts physical or serial console and edit the boot arguments in grub.

### Q: How do I collect diagnostics for my host or cluster?

A: Collect diagnostic information from all hosts in the cluster. In the case where a single host has self-fenced, the other hosts in the cluster are more likely to have useful information.

If the host is connected to XenCenter, from the menu select **Tools** > **Server Status Report**. Choose the hosts to collect diagnostics from and click **Next**. Choose to collect all available diagnostics. Click **Next**. After the diagnostics have been collected, you can save them to your local system.

If your host has lost connection to XenCenter, you can connect to the host console and use the `xen-bugtool` command to collect diagnostics.

# Appendix A. Constraints

The following constraints are part of the thin provisioning for block storage devices feature:

- Non-shared SRs are not supported. GFS2 SRs must be attached on all members of the pool.
- There is no automated upgrade path to a GFS2 SR from any existing SR type. Upgrade manually by attaching an old SR in parallel with a new SR and moving individual VDIs.
- You cannot resize online VDIs if the VDI is stored on a GFS2 SR.
- LUN resize is not supported.
- If a network has been used for both management and clustering, you cannot separate the management network from clustering without recreating the cluster.
- Changing the IP address of cluster network by using XenCenter causes the clustering and GFS2 to be temporarily disabled.
- VM migration with Storage XenMotion is not supported for VMs whose VDIs are on a GFS2 SR.
- Use of Intellicache is not supported for VMs using a GFS2 SR.
- Using VSS snapshots on GFS2 SRs is not supported.
- The FCoE protocol is not supported with GFS2 SRs.
- You cannot enable clustering on a pool that already has legacy HA enabled. Enable clustering first, then enable HA.
- You cannot use secure CHAP authentication. When using CHAP, the user name and password are transmitted as clear text.
- Disaster Recovery is not supported for pools that use a GFS2 SR.
- You don't receive an alert when the LUN is nearly out of space or actually out of space.
- Trim is not invoked automatically on vdi-delete.

# Appendix B. Known Issues

The following known issues are present in this experimental feature:

- If you enable clustering by using XenCenter, the host fencing timeout is incorrectly set to too large a value (20000s instead of 20000ms). We recommend that you use the CLI to enable clustering.
- If you forcefully remove a host from the cluster by using `Host.destroy`, the cluster configuration is not updated.
- When using HA with clustering, the calculation that HA makes for its failover plan for VM protection is not correct.
- If you shutdown a host, you might encounter a kernel bug that causes dom0 to crash.
- Performance metrics are not available for GFS2 SRs and disks on these SRs (RRDs).
- The multipath status is not properly reflected by XenCenter.
- If the iSCSI target is not reachable while GFS2 filesystems are mounted, some hosts might fence.