
Integration of EMC SRDF and TimeFinder with Sun Cluster 3

Abstract

This paper describes how SRDF and TimeFinder devices may be used in Sun Cluster 3 environments.

Published 1/13/2005

Copyright © 2004 EMC Corporation. All rights reserved.

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

THE INFORMATION IN THIS PUBLICATION IS PROVIDED “AS IS.” EMC CORPORATION 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 EMC software described in this publication requires an applicable software license.

EMC², EMC, Symmetrix, Celerra, CLARiiON, CLARAlert, Documentum, HighRoad, Legato, Navisphere, PowerPath, ResourcePak, SnapView/IP, SRDF, TimeFinder, VisualSAN, and where information lives are registered trademarks and EMC Automated Networked Storage, EMC ControlCenter, EMC Developers Program, EMC OnCourse, EMC Proven, EMC Snap, Access Logix, AutoAdvice, Automated Resource Manager, AutoSwap, AVALONidm, C-Clip, Celerra Replicator, Centera, CentraStar, CLARevent, Connectrix, CopyCross, CopyPoint, DatabaseXtender, Direct Matrix, Direct Matrix Architecture, EDM, E-Lab, Enginuity, FarPoint, FLARE, GeoSpan, InfoMover, MirrorView, NetWin, OnAlert, OpenScale, Powerlink, PowerVolume, RepliCare, SafeLine, SAN Architect, SAN Copy, SAN Manager, SDMS, SnapSure, SnapView, StorageScope, SupportMate, SymmAPI, SymmEnabler, Symmetrix DMX, Universal Data Tone, and VisualSRM are trademarks of EMC Corporation. All other trademarks used herein are the property of their respective owners.

Part Number H1294

Table of Contents

Overview	4
SCSI-3 Persistent Reservations on SRDF and TimeFinder Devices	4
Important Considerations and Configuration Essentials	4
TimeFinder and SRDF Synchronization and Restore Operations	6
TimeFinder Replication from within a Cluster to a Nonclustered Host.....	6
TimeFinder Replication within a Cluster.....	7
SRDF from Cluster to Cluster	8
SRDF from Cluster to Nonclustered Host.....	9
SRDF within a Cluster	9
Conclusion	11

Overview

The purpose of this document is to describe how data replication using EMC® TimeFinder® and EMC SRDF® may be used with Sun Cluster 3.

It is assumed that you have a good working knowledge of TimeFinder, SRDF, and Sun Cluster 3.

SCSI-3 Persistent Reservations on SRDF and TimeFinder Devices

Sun Cluster 3 uses SCSI persistent reservations (PRs) for device fencing. Persistent reservations are set on a per-device basis. Each device, regardless of whether it is a normal Symmetrix® volume, an SRDF device, or a TimeFinder device, will have its own persistent reservation. Persistent reservations are never replicated from one device to another. In other words, persistent reservations are not transferred across the link from R1 to R2 devices, nor are persistent reservations copied from the standard device to a business continuance volume (BCV) device during TimeFinder operations.

Simply put, persistent reservations are never transported and the Sun Cluster framework will treat an SRDF or BCV device just as it would any other device, without any special considerations.

The intent of this white paper is to advise how to ensure data integrity and proper device fencing with Sun Cluster 3 when using SRDF and TimeFinder.

Important Considerations and Configuration Essentials

Sun Cluster 3 does not offer integrated support for local or remote replication of data using SRDF or TimeFinder. Therefore, management of TimeFinder and SRDF devices cannot be done through Sun Cluster commands, and replication commands are not automated.

Clusters must be designed to avoid the following limitations, and system administrators must implement administrative actions to ensure data availability of replicated devices:

- Replication of VxVM disk groups: VxVM disk groups may not be replicated to the same or other cluster nodes that can see the source devices. The objective is to avoid having more than one logical copy of a VxVM disk group on any node. In other words, a BCV or SRDF copy of a disk group may not be visible to the node that has the original disk group, or to any other host that can see the original disk group. If any node can see both the source (R1 or STD) and target devices (R2 or BCV), that node is not in a supported configuration. The reason for this is that VxVM 3.X architecture is unable to distinguish between the original copy of a disk group and a byte-for-byte copy of that disk group. A few things could happen if this configuration rule is ignored. First, the wrong copy of a disk group may be imported. Second, disks from the original and copied disk groups may become mixed together when imported, and the result will be data loss due to volume corruption.

Note: VxVM 4.0 does allow replication of VxVM disk groups within the same host with an EMC Symmetrix array, but at the time of publication, VxVM 4.0 was not yet supported with Sun Cluster 3.0, and EMC has not qualified that functionality within clusters.

- Offline BCV devices: BCV devices that are being established could be tagged offline by the Solaris operating system if an attempt is made to read or write to the BCV device. As a result, after BCV split, VxVM will not be able to find BCV copies of disk groups unless those VERITAS Dynamic Multi-Pathing (DMP) paths have been reenabled immediately after the split. Furthermore, the DMP paths must be re-enabled on all nodes within the cluster; this must be done by the controlling application, and will not be done automatically by the cluster software. Failure to reenable DMP paths to BCV devices will result in unsuccessful failover of BCV device groups to other nodes. A disadvantage of running `vxdctl enable` on all nodes is that should disk group failover be necessary, the failover will be delayed until the “enable” has completed. Therefore, the failover time window could be significantly enlarged and is dependent on the number of device paths on each node.

- System architects should consider that the node with controlling EMC SYMAPI software may become unavailable, leaving the cluster without a means to control the SRDF and/or BCV devices. Provisions must be made to deal with a situation when such a node is unavailable.

Table 1. Matrix of Qualified Sun Cluster 3 Configurations with SRDF and TimeFinder

Volume Manager/EMC Data Replication Product	Disk Group Replication from Cluster A to Cluster B	Disk Group Replication within the Same Cluster¹	Disk Group Replication from within a Cluster to a Nonclustered Host	Sharing of a Clustered SRDF or BCV Device with a Stand-alone Node or Other Cluster/s
VxVM with SRDF	Supported	Supported with restrictions ¹	Supported	Not supported ²
VxVM with TimeFinder	Supported	Supported with restrictions ¹	Supported	Not supported ²
SVM/Disksuite with SRDF	Not Supported	Not Supported	Not Supported	Not Supported
SVM/Disksuite with TimeFinder	Not supported	Not Supported	Not Supported	Not supported

¹ VxVM disk groups may be replicated such that the source and copy (SRDF or BCV copy) of a disk group will not be visible to any node that can see the source devices at the same time.

Example 1: In a three-node cluster, if VxVM disk group A is visible to all three nodes, then none of the nodes may be mapped to the BCV copy of disk group A at any time. In this example, the BCV copy of disk group A may not be used in the cluster at all.

Example 2: In a three-node cluster, if VxVM disk group A is visible to node #1 and node #2, but not visible to node #3, then only node #3 may have visibility to the BCV copy of disk group A.

Example 3: In a two-node cluster, if VxVM disk group A is visible to node #1 and node #2, then the BCV copy of the disk group must be mapped outside the cluster.

² Any device imported within Sun Cluster 3.X will have persistent reservations placed on that device. When a persistent reservation has been set on a device, only nodes within the cluster will have read/write access for that device. If any node outside the cluster attempts to access such devices in violation of the reservation, a reservation conflict will occur and I/O will be denied from any host outside the cluster that owns the PR. Therefore, devices under the protection of Sun Cluster 3.X cannot be shared with other clusters or nonclustered hosts.

TimeFinder and SRDF Synchronization and Restore Operations

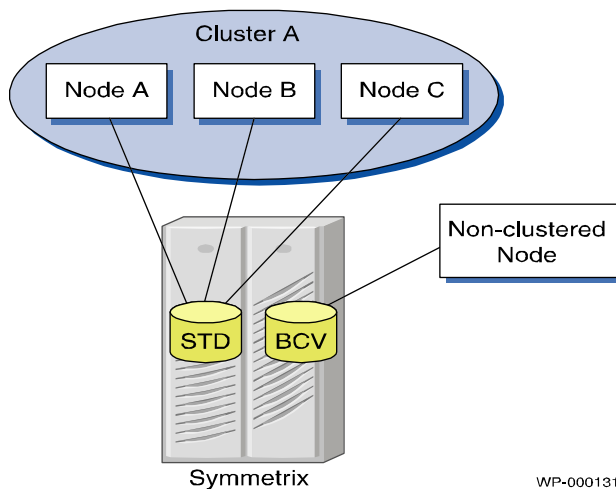
Normal data synchronization operations—where data is copied from the STD device to a BCV device—are transparent to the Sun Cluster 3.X framework. The Sun Cluster software does not detect SRDF and TimeFinder device state changes, such as splits, restores and establish operations. Therefore, whenever SRDF or TimeFinder devices are not in a mountable state, these device groups must be tagged as being offline and put into maintenance state by issuing a `scswitch -m -D <dgname>` command.

See the following non-exhaustive list of examples of where it is necessary to put a device group into maintenance state before initiating an SRDF or BCV data synchronization:

- When BCVs are being established, the BCV disk resource must be put into maintenance state because those BCV devices are not accessible by any nodes and are therefore not available to the cluster.
- R2 devices during normal R1-to-R2 device synchronization. In this case, the R2 devices cannot be imported because they will be in a read-only state. The R2 devices must be unmounted and put into maintenance state prior to an SRDF establish.
- R2 devices during an SRDF restore operation. This is similar to the previous item in that the R2 devices cannot be mounted while the SRDF restore is in progress.
- During a BCV restore, the BCV devices are not ready and must therefore be unmounted and placed into maintenance state prior to starting a BCV restore.

TimeFinder Replication from within a Cluster to a Nonclustered Host

In Figure 1, BCV devices are visible only to some other nonclustered node or nodes. This configuration is typically used to back up clustered data. As persistent reservations are not copied onto the BCV, the BCV device can be mounted on any other host.



WP-000131

Figure 1. Replication of a Clustered Device to a Nonclustered Host Using TimeFinder

There are no special considerations for operation within cluster, except that prior to initiating a BCV restore, the disk resource in the cluster must be put into maintenance mode prior to initiating the restore (e.g., `scswitch -m -D <dgname>`).

TimeFinder Replication within a Cluster

The example in Figure 2 is supported because the BCV copy of the disk is not visible to the nodes that can see the STD copy of the disk group. This ensures that the VxVM disk group will not experience device mapping corruption as mentioned previously.

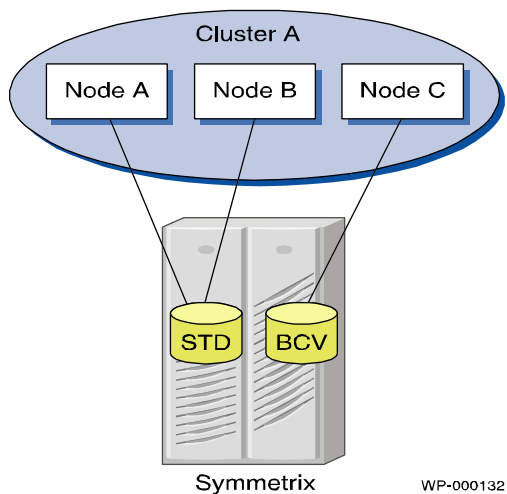


Figure 2. Replication of a Clustered Device within a Cluster Using TimeFinder

When the BCV is synchronized with the standard, you should follow this procedure:

1. Take the BCV disk group offline:


```
scswitch -m -D <dgname>
```
2. Initiate the establish:


```
symmir -g <symdgname> est
```
3. After synchronization is complete, split the BCV from the STD:


```
symmir -g <symdgname> split
```
4. Rescan the devices into VxVM:


```
vxdctl enable
```

Note: It is not always necessary to rescan the devices. However, VxVM will take BCV devices out of the import list if a scan is done while the BCVs are established. Therefore, to ensure the BCVs will import after the split, run `vxdctl enable` on all nodes that can see the BCV devices. This is especially important in a clustered environment when more than one node can experience offline devices.

5. Bring the device group online:


```
scswitch -z -D <dgname> -h <nodename>
```

The following configuration is not supported because the BCV device has been mapped to Node A and Node C, and both those nodes can see both the STD and BCV device.

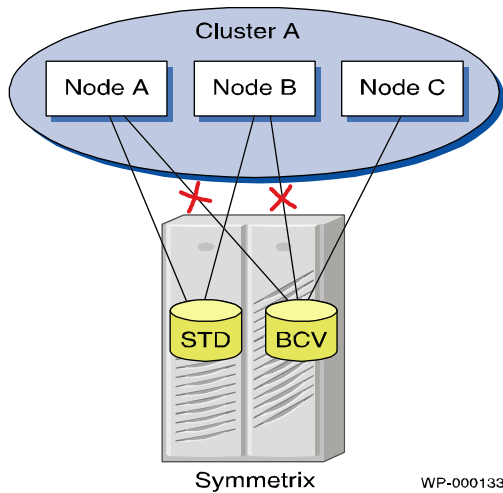


Figure 3. Node A and B Can See Both the STD and BCV Devices, Therefore This Is an Unsupported Configuration

SRDF from Cluster to Cluster

Figure 4 is of SRDF between two clusters. As the R1 devices are mapped inside Cluster A, and the R2 devices are mapped inside Cluster B, both the R1 and R2 devices will be owned by their respective clusters and thus have their own persistent reservations assigned by each cluster. During normal SRDF operations where data is being synced from R1 to R2, the R2 device group will have to be offline and in maintenance state. Sun Cluster does not provide automatic failover of an SRDF disk group from Cluster A to Cluster B; a custom solution must be developed to manage failover and ensure appropriate use of R1 and R2 devices.

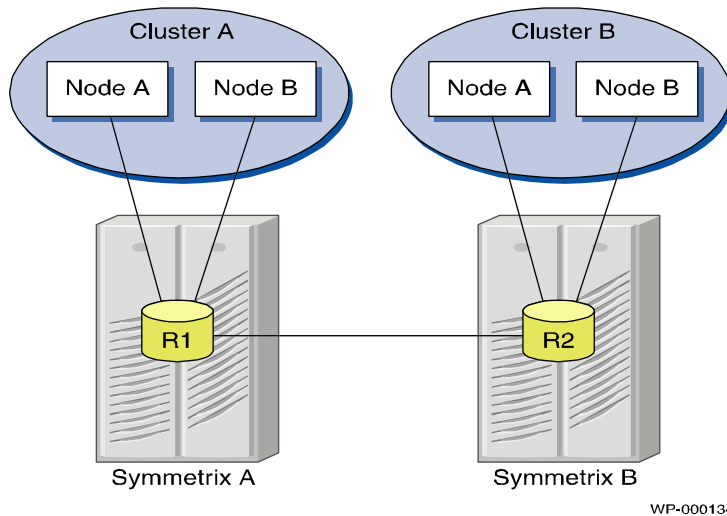


Figure 4. SRDF Replication between Two Different Clusters

SRDF from Cluster to Nonclustered Host

In Figure 5, the R1 device is managed by Sun Cluster and will therefore have persistent reservations. The R2 device, which is not managed by a cluster, can be mapped to any node.

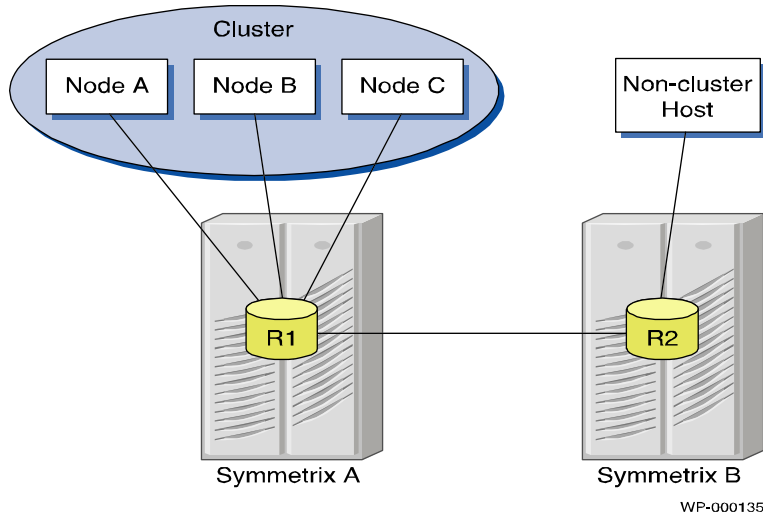


Figure 5. SRDF from a Cluster to Nonclustered Host

As with BCV restore, when SRDF restore is initiated, the R1 devices must be offline to the cluster (e.g., `scswitch -m -D rdfdg`). The persistent reservations on the R1 device will not be affected by the SRDF restore operation.

SRDF within a Cluster

Figure 6 is an example of an unsupported configuration of SRDF replication within a cluster. The reason the configuration is not supported is that Node C can see both the R1 and R2 copies of the same disk group. This could result in data corruption due to cross-mapped devices inside a VxVM disk group.

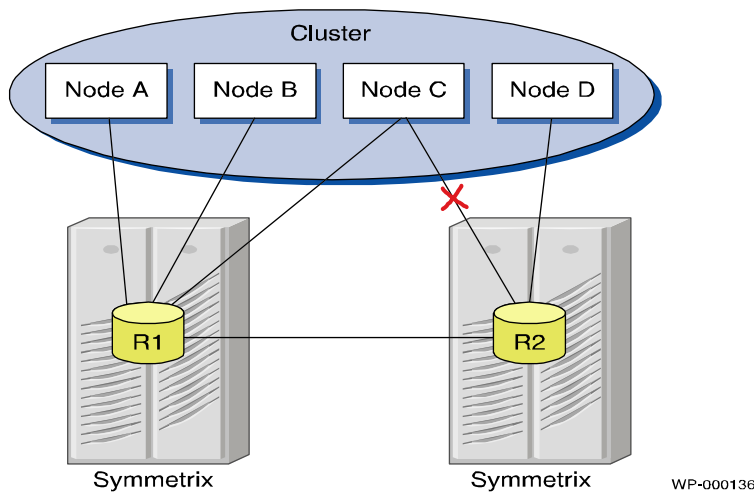


Figure 6. An Unsupported Configuration because Node C Can See Both the R1 and the R2 Copy of a Device

Figure 7 shows another example of a configuration that adheres to the same rule—the R1 and R2 disk groups are not visible from the same nodes.

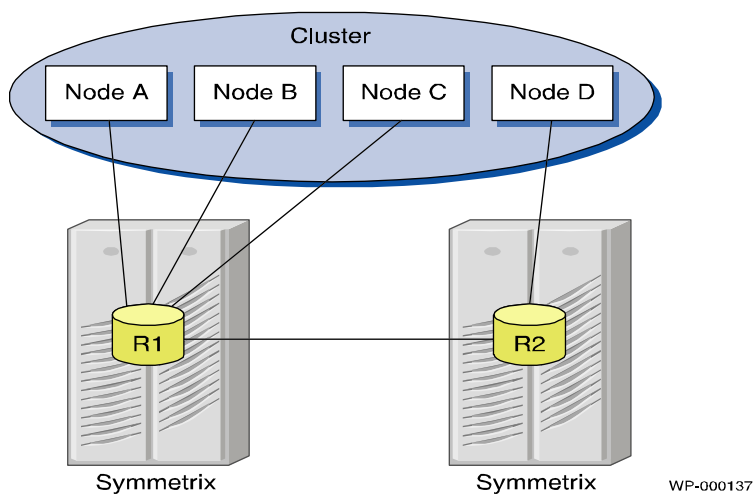


Figure 7. Example of How SRDF May Be Used within the Same Campus Cluster

Note that none of the nodes can see both the R1 and R2 devices.

Conclusion

Replication of VxVM disk groups is supported as described in Table 1.

Replication of Solaris Volume Manager (SVM) metavolumes is not supported.

The basic rules are:

- Do not mix copies of the same disk group on the same host.
- Ensure that a device being updated from either the remote SRDF device, or TimeFinder device, is deported and in cluster maintenance state.
- If BCV devices are accessed by VxVM while they are being established, those BCV devices will be taken offline to the host—DMP will close all paths to those devices. Ensure that the devices are reintroduced to VxVM and the cluster framework by running `vxddctl enable` on all nodes and `scgdevs` on at least one node after the BCV has been split.
- Device group failover will be delayed if a `vxddctl enable` is in progress. The time consequences should be quantified during cluster acceptance testing.
- Ensure that SYMCLI can operate from some other node/s or hosts should the controlling host become unavailable.

Refer to the *EMC Support Matrix* for minimum required microcode levels for support of SRDF and TimeFinder inside Sun Clusters. Refer to the Sun Cluster section.