Frequently Asked Questions: Veritas™ Cluster Server I/O Fencing
Introduction

This document is a compilation of common questions about the configuration and use of Veritas Cluster Server (VCS) I/O fencing in data centers. The questions are grouped into categories and include a brief overview of VCS I/O fencing and its components, its role in various Symantec products and other clusterware, recommended best practices, and other functional aspects.

Intended audience

This document is intended to help Symantec Systems Engineers (SE) and Technical Support Engineers (TSE) understand, evaluate, and deploy I/O fencing in production environments.

Note: All references to iSCSI in this document are discussed in the context of the VCS 5.0 MP3 release. Please be aware that iSCSI as a coordination point for VCS I/O fencing will be available starting with VCS 5.0 MP3. It will be mandatory to have iSCSI targets managed by VxVM.
# List of Acronyms

This section provides the complete list of acronyms used in this document.

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>SF</td>
<td>Storage Foundation</td>
</tr>
<tr>
<td>HA</td>
<td>High Availability</td>
</tr>
<tr>
<td>VxVM</td>
<td>Veritas Volume Manager</td>
</tr>
<tr>
<td>CVM</td>
<td>Cluster Volume Manager</td>
</tr>
<tr>
<td>CFS</td>
<td>Cluster File System</td>
</tr>
<tr>
<td>VCS</td>
<td>Veritas Cluster Server</td>
</tr>
<tr>
<td>LLT</td>
<td>Low Latency Transport</td>
</tr>
<tr>
<td>GAB</td>
<td>Group Membership Services/Atomic Broadcast</td>
</tr>
<tr>
<td>VXFEN</td>
<td>Veritas I/O Fencing</td>
</tr>
<tr>
<td>CPS</td>
<td>Coordination Point server</td>
</tr>
<tr>
<td>DMP</td>
<td>Dynamic Multi-Pathing</td>
</tr>
<tr>
<td>DLPI</td>
<td>Data Link Protocol Interface</td>
</tr>
<tr>
<td>PGR/PR</td>
<td>Persistent Group Reservation</td>
</tr>
<tr>
<td>FC</td>
<td>Fiber Channel</td>
</tr>
<tr>
<td>SAN</td>
<td>Storage Area Networks</td>
</tr>
<tr>
<td>CDS</td>
<td>Cross-platform Data Sharing</td>
</tr>
<tr>
<td>EBN</td>
<td>Enclosure Based Naming</td>
</tr>
<tr>
<td>HCL</td>
<td>Hardware Compatibility List</td>
</tr>
<tr>
<td>ASL</td>
<td>Array Support Library</td>
</tr>
<tr>
<td>APM</td>
<td>Array Policy Module</td>
</tr>
<tr>
<td>ALUA</td>
<td>Asymmetric Logical Unit Access</td>
</tr>
<tr>
<td>LVM</td>
<td>Logical Volume Manager</td>
</tr>
<tr>
<td>RAC</td>
<td>Real Application Cluster</td>
</tr>
<tr>
<td>CRS</td>
<td>Cluster Ready Services</td>
</tr>
<tr>
<td>GCO</td>
<td>Global Cluster Options</td>
</tr>
<tr>
<td>DWDM</td>
<td>Dense Wavelength Division Multiplexing</td>
</tr>
</tbody>
</table>
About VCS I/O fencing

1. What is VCS I/O fencing?

VCS I/O fencing prevents corruption of application data in a cluster where nodes share common storage and communicate with one another using private interconnects (a redundant network used for inter-node communication). To maintain consistency of the cluster, it is important for member nodes to be aware of other members in the cluster. This awareness is achieved through the mechanism of cluster membership. When interconnects break down, the method of determining peer failure is compromised, introducing the possibility of a “split-brain” condition. In the event of a split-brain condition, the fencing operation determines the nodes that must retain access to the shared storage and the nodes that must be ejected from the cluster to prevent data corruption.

2. What are the essential components of VCS I/O fencing?

The essential components of VCS I/O fencing are as follows:

- LLT, GAB, and VXFEN drivers
- Coordinator disks with SCSI-3 Persistent Reservations (PR)
- Veritas Volume Manager (VxVM) and Dynamic Multipathing (DMP)
- VCS I/O fencing data disks using CVM/DMP and VCS Disk Group agent

**LLT, GAB, and VXFEN drivers**

VCS uses the following kernel modules (drivers) in UNIX/Linux to support I/O fencing:

- Group Membership Services/Atomic Broadcast (GAB)
  - The GAB driver implements a protocol that ensures delivery of messages atomically to all surviving members of a sub-cluster in the event of cluster reconfigurations.

- Low Latency Transport (LLT)
  - LLT load balances inter-node communication across all available private network links. Cluster communication is evenly distributed across all private (up to 8) network links for performance and fault resilience. If a link fails, the traffic is redirected to the remaining links. LLT operates as a replacement for the IP stack and runs directly on top of the Data Link Protocol Interface (DLPI) layer. LLT can also be configured to run over IP, offering flexibility in deployment. GAB uses the underlying LLT driver to communicate reliably between servers.
• Veritas fencing (VXFEN)

The VXFEN driver works with GAB to provide membership arbitration, thus maintaining tight control over cluster membership. VXFEN uses an arbitration mechanism to determine which sub-cluster survives after a network partition. This arbitration mechanism makes use of special-purpose disks or LUNs that support SCSI-3 Persistent Group Reservation (SCSI-3 PGR or just PR).

Coordinator disks
Coordinator disks are direct or FC SAN-attached disks or iSCSI targets that are shared by all the nodes of a cluster. They are set aside for I/O fencing to use during a cluster membership change. The coordinator disks act as a global lock device during a cluster reconfiguration, akin to a quorum disk. This locking mechanism is used to determine the nodes that can remain in a cluster and the nodes that must be blocked from accessing data drives. The sub-cluster that succeeds in acquiring control of a majority of the coordinator disks is allowed to continue in the cluster whereas the rest of the nodes commit suicide and are ejected from the cluster. Coordinator disks cannot be used for any other purpose in VCS configuration. In VCS global clusters using GCO, each cluster needs an independent set of coordinator disks and they cannot be shared by the global cluster. These disks cannot be used to store user data either. A minimum of three coordinator disks are required per cluster for I/O fencing to work as desired. Any odd numbers of disks greater than three are also supported.

Coordinator disks must adhere to SCSI-3 PR. SCSI-3 PR is an enhanced SCSI specification which is designed to resolve the issues that arise from using SCSI reservations in a modern clustered SAN environment. SCSI-3 PR supports multiple nodes accessing a device while at the same time blocking access to other nodes. PR is persistent across SCSI bus resets and also supports multiple paths from a host to a disk.
**Veritas Volume Manager (VxVM) and Dynamic MultiPathing (DMP)**

The Dynamic Multipathing (DMP) feature of Veritas Volume Manager (VxVM) provides greater reliability and performance by using path failover and load balancing. DMP discovers all the I/O paths available to a LUN in a multi-port array. In the event of a path failure that may result from faults in HBAs, cables, switches or array ports, it dynamically fails over the I/Os to the available paths. DMP also handles failover of SCSI-3 PR commands in a manner understood by a given disk array. It also takes care of installing and removing SCSI-3 PR keys when paths are added or removed. The role of DMP in handling SCSI-3 PR keys is especially significant in the case of Active/Passive arrays. In these arrays, the SCSI-3 PR keys are accepted only on the current active port.

VxVM simplifies placement of SCSI-3 PR keys on LUNs. Disks added to a private or shared disk group are automatically prepared for fencing by automatic installation of SCSI-3 PR keys on those disks. Similarly, removal of disks from a disk group results in clearing the SCSI-3 PR keys from the removed LUNs.

High availability of coordinator disks is crucial for efficient I/O fencing. VxVM (DMP) disks provide this capability for a wide range of disk arrays that are connected over FC SAN or iSCSI. For more information on the supported list of disk arrays, refer to the Hardware Compatibility List (HCL) at the following location:

http://www.symantec.com/business/products/otherresources.jsp?pcid=2245&pvid=203_1

**VCS I/O fencing data disks using CVM/DMP and VCS Disk Group agent**

Data disks are FC SAN-attached LUNs or iSCSI targets that are used for data storage. VXFEN provides full SCSI-3 PR based data protection at the data disk level. Hence, these disks must support SCSI-3 PR for VXFEN to “fence off” data storage from ejected nodes, following split brain and arbitration. Data disks are VxVM disks incorporated in disk groups. VxVM disks automatically benefit from DMP for I/O load balancing and availability.

In a cluster running parallel applications that access shared storage using Clustered Volume Manager (CVM), the shared disk groups are managed by CVM/DMP.

In a cluster running failover applications, the VCS Disk Group agent is responsible for fencing failover disk groups on a disk group basis.

In both the above cases, VXFEN uses the underlying Veritas Volume Manager to enable I/O fencing. Disks added to a private or shared disk group are automatically prepared for fencing, as are new paths discovered to a device. DMP also takes care of handling I/O fencing using alternate paths in the event of a path failure.
Licensing requirements

1. What are the licensing requirements for I/O fencing?
   The VCS I/O fencing license is enabled automatically with the standard VCS license. However, data disk fencing using DMP needs Storage Foundation license keys to be installed.
   To confirm that all the required license keys for VCS I/O fencing are present on the system, run the vxlicrep command as follows:
   ```
   # vxlicrep -e | grep PGR
   PGR#VERITAS Volume Manager=Enabled
   PGR_TRAINING#VERITAS Volume Manager=Enabled
   PGR=Enabled
   PGR_TRAINING=Enabled
   ```

2. What is PGR_TRAINING license?
   The PGR_TRAINING license was provided for Storage Foundation release version 4.1 and earlier. The license is no longer relevant for version 5.0 and later.
   In earlier releases, the discovery of LUNs from certain disk arrays required the Test Unit Ready (TUR) SCSI command to be issued.
   The PGR_TRAINING key enabled VxVM to register the temporary key “zzzzzzzz” during device discovery in the system boot process, issue a TUR command, and then unregister the key.

About split brain

1. What is split brain?
   Split brain is a scenario when the cluster communication mechanism (used to detect the liveness of other nodes) fails, even though those nodes are still running. The unavailability of peer nodes cannot be discerned from broken communication links and the cluster is split into multiple sub-clusters. Nodes in each of the sub-clusters assume that they are the members of the only surviving cluster. Such sub-clusters attempt to take control of the shared data storage in an uncoordinated manner leading to corruption of application data.
2. **What are the common split-brain scenarios?**

The most common split-brain scenarios are as follows:

- **Scenario 1**
  In a cluster where each node is capable of making failover decisions on its own, the failure of private interconnects linking the nodes in the cluster can result in one or more network partitions. The resulting ‘sub-clusters’ assume that they are the only surviving nodes. When they proceed to fail over the applications to their own sub-cluster, data is corrupted.

- **Scenario 2**
  If a cluster node is too busy and does not respond to cluster protocol messages, it is declared dead by the other nodes and removed from their cluster membership. If the node recovers from the non-responsive state (such as a driver finally timing out), it may return to the cluster after it has been excluded by the cluster membership system.

- **Scenario 3**
  Split brain can happen on some systems where the hardware supports a “break” and “resume” function. If the system administrator drops the system to PROM level with a break, the system can be declared as dead. On subsequent resume, the cluster is reformed and the system returns to complete pending I/O, leading to data corruption.

In Scenario 2 and Scenario 3, during the non-responsive as well as the break/resume state, a serious problem occurs when the cluster reforms and the storage resources are already online elsewhere. The returning system may write as soon as it is alive again, before the membership subsystem determines whether it should be excluded from the cluster and cause the node to panic.

3. **How can I prevent split brain?**

Some common approaches toward preventing split brain are as follows:

- **Using redundant private interconnects**
  Using best practices to prevent network failure, such as configuring dedicated, redundant private links between cluster nodes can reduce the risk of split brain to some extent. It is preferred to use switches rather than network hubs. To protect against network switch failures, it is recommended to use independent switches for each link. However, this approach is not fool proof, as it does not address scenario 2 and scenario 3.
• Disabling fault recovery of applications upon communication failure
  One of the ways to prevent uncoordinated access is by disabling recovery of an
  application whenever a node loses communication with another node. However, this
  approach does not serve the required purpose, as the cluster must be capable of taking
  corrective action on the failure of a node in order to provide high availability.

• Using I/O fencing
  I/O fencing is a fool proof solution for preventing split brain.
  The underlying factor behind data corruption is uncoordinated write operations to the
  storage by cluster members. The key to data integrity is in ensuring complete control
  over the write operations on shared storage at the same time.
  To guarantee data integrity, the following important aspects need to be addressed:
  – Ability to determine who remains in the cluster in the event of communication failures
    among cluster members and ensuring that only one sub-cluster survives. This process is
    also referred to as arbitration.
  – Total blocking of access to storage from any system that is not an acknowledged
    member of the surviving sub-cluster. This process is often referred to as “fencing off
    data storage”.
  VCS uses a robust I/O fencing mechanism that provides protection against split-brain
  scenarios by meeting the above objectives.

VCS I/O fencing behavior in the event of split brain
1. Why does VXFEN panic one or more nodes instead of recovering the application in a
   split-brain situation?
   In the event of split brain, neither the state of the application instances nor the state of the
   operating system is known on the nodes that lose the race at the time of I/O fencing. This is
   especially true when applications and operating systems do not respond, leading to split-
   brain detection. There is no guaranteed recovery of the application or the OS from these
   possible states. VCS I/O fencing, therefore, chooses the safest option in favor of data
   protection, that is, to panic the affected nodes instantly. The subsequent reboot allows a
   clean cluster wide recovery of the application.

2. Can I stop VXFEN from ejecting a node out of the cluster and give a warning instead?
   No, ejecting a node out of the cluster prevents data corruption in multiple split-brain
   scenarios. Dependence on the administrator to observe the warning and take action may be
   too late to prevent data corruption.
3. **Does the administrator have control as to which sub-cluster will survive in the event of a split-brain situation?**

   No. Though the VCS 5.0 releases do not support this, application-aware fencing will be available with future releases where the user can specify which applications need to be given preference when I/O fencing is done.

4. **What happens to the evicted nodes when they panic and try to re-join the cluster during reboot?**

   If the evicted node is physically connected to the remaining nodes in the cluster through interconnects when it starts, it joins the existing cluster. Otherwise, VXFEN on the node checks the keys registered on coordinator disks to see if there are nodes that are not accessible to it. If this is the case, the evicted node does not join the cluster and reports a pre-existing split brain. Otherwise, it joins the cluster.

5. **How do I know which sub-cluster will survive in a split-brain situation with Veritas I/O fencing enabled?**

   VXFEN gives preference to the larger sub-cluster in the event of a split-brain situation. Hence, it is highly likely that the larger sub-cluster formed will survive after fencing. The nodes fenced out by VXFEN are halted and rebooted when they find that they are no longer in the cluster.

6. **How is the VXFEN disk race order determined? If there are eight nodes, and it splits into a three-node cluster and a five-node cluster, why does the five-node cluster have a better chance of being available?**

   In a network partition (split-brain) scenario, the VXFEN disk race order is the same for all sub-clusters, that is, they race for the same coordinator disk first before trying to race for the second coordinator disk and so on. In the case where one sub-cluster is larger than the other, the larger sub-cluster has a better chance of being available.

7. **Can a busy node prompt VXFEN to evict the node?**

   A busy node can ultimately result in the inability of that node to heartbeat with its peer nodes. Missing heartbeats can lead to VXFEN evicting the node, as it is seen as a manifestation of split brain. VXFEN evicts the node to prevent data corruption by preventing any pending I/Os by the busy node once it becomes available.
8. How does an administrator know if a node is evicted by VXFEN?

When a node is evicted by VXFEN, it is forcibly halted. The console logs on all of the evicted nodes display the following message from VXFEN before killing the node:

```
v-11-1-20  Local cluster node ejected from cluster to prevent potential data corruption
```

9. Can faulty interconnects prompt VXFEN to evict nodes?

Yes, faulty interconnects are the most common reason for split-brain scenarios and can prompt VXFEN to evict nodes. Faults in interconnects can occur due to failure in NICs (Network Interface Card) or network switches.

10. Does VXFEN need multiple LLT links?

VXFEN does not need extra LLT links, as it uses the available LLT links (at least two LLT links are recommended in a VCS cluster) for its operations.

11. Can VCS survive up to the last node in a split-brain situation?

Yes, VXFEN ensures that at least a minimum number of nodes survive the split brain, thus avoiding application downtime.

12. What is "pre-existing" split brain? How do I resolve it?

When fencing starts on a node, it checks whether there are any keys on the coordinator disks. If the fencing driver finds registrations on coordinator disks from nodes that are not part of this node's GAB membership, the node does not start fencing. This is because the node assumes that there are other nodes in the cluster (corresponding to stale keys) that are not accessible to it. This condition is termed a pre-existing split brain.

Pre-existing split brain occurs because of the following reasons:

1. The private network connection between the current nodes and the rest of the nodes in the running cluster is not functional.

   Resolution: The following steps may be taken to resolve this situation:
   a. Stop GAB and its clients
   b. Resolve the network connectivity issues
   c. Restart the node in order to start GAB and all the other modules the stack.

2. There are no other nodes running fencing currently. This situation may occur if the other nodes leave the cluster without removing their keys from the coordinator disks.

   Resolution: The following procedure may be used to resolve pre-existing split-brain:
   a. Shut down all the other nodes in the cluster that has access to the shared storage.
b. Run the `vxfenclearpre` script.
   ```bash
   # /opt/VRTSvcs/vxfen/bin/vxfenclearpre
   ```
c. Restart all the nodes in the cluster.

13. **What is "cascaded" split brain?**
    On an n-node cluster, the private interconnects may break at multiple points in the cluster leading to multiple sub-clusters. This is called a cascaded split brain. As in a normal split brain situation, VXFEN deals with the scenario by ensuring that no data corruption occurs while also ensuring that at least one sub-cluster survives.

14. **If I disconnect LLT cables to simulate split brain, why is it important to reconnect them before the panicked node becomes available?**
    Unless the LLT cables are reconnected, the node will not be able to join the cluster since it is not detected as a member of the existing cluster. A pre-existing split brain is reported on the joining node and it is not allowed to join the cluster.

15. **What happens to a cluster when one of the coordination points is inaccessible and a split brain occurs?**
    In the absence of I/O fencing, if one of the coordination points is unreachable and a split brain occurs, the cluster splits into sub-clusters. In this scenario, the sub-clusters race for the remaining coordination points individually. I/O fencing ensures that the first sub-cluster that succeeds in attaining a vote on one of the coordination points is given preference to win the other coordination points as well. The sub-cluster that loses the vote for the first coordination point delays for some time before racing for the next coordination point, thus ensuring that the sub-cluster that has won the first coordination point proceeds to win the others as well. Thus, fencing prevents a situation where each sub-cluster attains a single vote and may cause the sub-clusters to panic.

16. **How is I/O fencing in GAB different from that of VXFEN?**
    GAB can initiate an `IOFENCE` signal to eject a node from cluster membership. The `IOFENCE` is invoked if GAB deems it necessary that one or more nodes leave the current cluster membership. After receiving an `IOFENCE` signal, a node may either rejoin the membership or be halted (intentional panic) depending on the GAB client (port) and the nature of the `IOFENCE`. In the event that the node is halted, the panic string will include the GAB `IOFENCE` reason.
Even in these scenarios, the fencing of shared storage is initiated by VXFEN on the surviving nodes when it receives the new membership from GAB, after the IOFENCE.

**Different modes of fencing**

1. **What are the different modes in which fencing can run?**
   Fencing can run in SCSI-3 mode, customized mode, or disabled mode. SCSI-3 mode is the traditional mode for I/O fencing. SCSI-3 PGR compliant disks are used by Veritas I/O fencing driver for arbitration to prevent data corruption in the event of network partitions. The “vxfen_mode” attribute in I/O fencing configuration file (/etc/vxfenmode) has to be set to "scsi3", for fencing to come up in SCSI-3 mode.
   In customized mode, Veritas I/O fencing module provides a framework for plugging in customized mechanisms for configuring coordination points, which help the I/O fencing driver to arrive at a decision during partition arbitration. In this mode, the attribute “vxfen_mode” has to be set to “customized” for fencing to come up in customized mode.
   When fencing is configured in disabled mode, the fencing driver will run but it will not do any actual fencing.

2. **How do I set up fencing in SCSI-3 mode?**
   For information on setting up fencing in SCSI-3 mode, refer to the Veritas™ Cluster Server Installation Guide.

3. **How do I set up fencing in customized mode?**
   For information on setting up fencing in customized mode, refer to the Veritas™ Cluster Server Installation Guide.

4. **How do I set up fencing in disabled mode?**
   For information on setting up fencing in customized mode, refer to the Veritas™ Cluster Server Installation Guide.

**VCS I/O fencing ports**

1. **What is a VXFEN port?**
   VXFEN uses a dedicated port provided by GAB for communication across nodes in the cluster. This can be seen as port ‘b’ when ‘gabconfig –a’ is run on any cluster node. The entry corresponding to port ‘b’ in this membership indicates the existing members in the cluster as viewed by VXFEN.
Example:

GAB Port Memberships
==================================

Port a gen 574802 membership 01
Port b gen 574805 membership 01

2. What are the other GAB ports that need to be active for VCS I/O fencing to work?
   Port “a”, used by GAB for maintaining the cluster membership, must be active for VXFEN to start.

Role of VCS I/O fencing in Symantec product suites and other clusterware

1. What role does VXFEN play in a VCS HA cluster?
   A VCS HA cluster can have parallel or failover service groups configured. If a VCS cluster with parallel service groups encounters a split-brain scenario, VXFEN forces the nodes in one of the sub-clusters to commit suicide in order to prevent data corruption. The sub-cluster that is evicted is not allowed to gain access to the shared storage without re-joining the cluster.
   In the case of failover groups, access to the disk is prevented when VCS fails over the service group to another node.

2. What role does VXFEN play in an SF HA cluster?
   Apart from handling the split-brain scenarios as in a VCS HA cluster, VXFEN capitalizes on the availability and load balancing capabilities provided by DMP for data as well as coordinator disks. The VCS Disk Group agent takes care of I/O fencing the failover disk groups on a disk group basis.

3. What role does VXFEN play in an SF CFS cluster?
   Apart from handling the split-brain scenarios as in an SF HA cluster, VXFEN coordinates the flow of control during cluster reconfiguration between CVM, CFS, and the application in the stack to ensure that recovery at each level (VXFEN, CVM, CFS, application, in that order) is done in the proper sequence.

4. What role does VXFEN play in an SF Oracle RAC cluster?
   In an SF Oracle RAC cluster, VXFEN continues to provide full I/O fencing capabilities as in an SF CFS cluster, independent of Oracle CRS. In the process, it is always ensured that VXFEN acts first on deciding the surviving sub-cluster, thereby avoiding a conflict with the decision taken by CRS.
5. **What role does VXFEN play in Serviceguard CFS/CVM cluster?**

In a Serviceguard CFS/CVM cluster, VXFEN ensures that CVM, CFS, and the application react to cluster membership changes in that order, so that CVM completes mirrored volume consistency recovery before CFS recovery starts. With this sequencing, VXFEN prevents data corruption in the following scenarios:

- CFS starts recovery (reads data x from mirror 1)
- CFS reads same block again (now reads data y from mirror 2)
  
  This will cause data corruption without VXFEN in the stack.
- Even if CFS reads from the same mirror1, CVM recovery could use mirror 2 to recover and overwrite mirror 1 data

VXFEN prevents data corruption by performing sequencing.

In a Serviceguard cluster, VXFEN does not take I/O fencing action and leaves the same to Serviceguard. VXFEN does, however, act as a channel to pass on the cluster membership received by GAB and LLT to Serviceguard.

6. **Can I operate a VCS HA cluster with fencing disabled?**

Yes, the VCS HA cluster can be operated with fencing disabled by setting the vxfen_mode in the /etc/vxfenmode file before starting VXFEN on all the nodes in the cluster. However, this exposes the application to possible data corruption in split-brain scenarios.

7. **Can I operate SF HA cluster with fencing disabled?**

Yes. SF HA cluster can be operated with fencing disabled by setting the vxfen_mode in the /etc/vxfenmode file before starting VXFEN on all the nodes in the cluster. However, this would expose the application to possible data corruption in split brain scenarios.

8. **Can I operate SF Oracle RAC cluster with fencing disabled?**

SF Oracle RAC cluster installation and configuration does not require I/O fencing to be enabled. However, only those clusters configured with fencing enabled are supported in SF Oracle RAC. This is because, Oracle RAC being a parallel application, the chances of data corruption are very high in a split-brain situation.
9. **Can I operate an SF CFS cluster with fencing disabled?**

SF CFS cluster installation and configuration does not require I/O fencing to be enabled. However, only those clusters configured with fencing enabled are supported in SF CFS. This is because, in a clustered file system, the chances of data corruption are very high in a split-brain situation. The behavior of CFS changes depending on whether fencing is configured in enabled or disabled mode. If an SF CFS cluster is configured with fencing enabled, then CFS does not take any action in a split-brain situation; it relies solely on VXFEN to take decisions. However, if an SF CFS cluster is configured with fencing disabled, then CFS disables the CFS mount points on the surviving nodes if it receives a jeopardy membership followed by a node loss.

10. **Can I operate an SF CFS RAC cluster with fencing disabled?**

Yes. In fact, SF CFS RAC is not licensed to support the VCS I/O fencing feature. When you run the product installer or installation scripts, you are not given the option to enable VCS I/O fencing; the I/O fencing driver is disabled. If you install the product manually, use the file /etc/vxen.d/vxenmode_disabled as a template for the /etc/vxenmode configuration file.

If you enable I/O fencing, importing a shared disk group fails with the following error message:

```
VxVM vxdg ERROR V-5-1-10978 Disk group shared_dg: import failed: License has expired or is not available for operation
```

### Enabling or disabling VCS I/O fencing in a cluster

1. **How do I check if fencing is enabled in a cluster?**

The `-d` option used with vxenadm (1M) displays the fencing mode on each cluster node.

Port “b” membership should be present in the output of `gabconfig -a` and the output should list all the nodes in the cluster.

A sample output of “vxenadm -d” command is:

```
vxenadm -d

I/O Fencing Cluster Information:

Fencing Protocol Version: 201
Fencing Mode: customized
Fencing Mechanism: cps
Cluster Members:
  * 0 (system1)
```

RFSM State Information:
2. **What action should I take if the GAB ports needed for I/O fencing are not available?**

LLT and GAB must be started on the node, if port “a” is not visible in the output of `gabconfig --a` command.

The following commands can be used to start LLT and GAB respectively:

To start LLT, on each node, type:

```bash
# /sbin/init.d/llt start
```

If LLT is configured correctly on each node, the console output displays:

```
```

To start GAB, on each node, type:

```bash
# /etc/init.d/gab start
```

If GAB is configured correctly on each node, the console output displays:

```
Apr 5 14:46:29 north gab: GAB:20021: GAB available
Apr 5 14:51:50 north gab: GAB:20026: Port a registration waiting for seed port membership
```

3. **Can I enable and disable VCS I/O fencing without manually rebooting the node?**

Yes. VXFEN can be enabled and disabled without manually rebooting the node. However, it requires stopping the cluster, as the modules in the stack above and including VXFEN must be stopped and restarted for VXFEN to pick up the new mode from the `/etc/vxfenmode` file.

5. **Can I enable and disable VCS I/O fencing without restarting VCS?**

No. VCS must be stopped for enabling or disabling VXFEN. This is because the configuration change in VXFEN requires VCS to be restarted. As all the other modules, except LLT and GAB, run on top of VXFEN in the stack, they need to be brought down in order to enable or disable fencing.

6. **What is the role of “UseFence=SCSI3” entry in the file main.cf?**

The entry “UseFence=SCSI3” in the file main.cf instructs the VCS DiskGroup Agent to perform the following tasks:

- Remove the SCSI-3 PR keys installed on the disks that are part of the failover disk group
- Install fresh keys from the new node at the time of re-importing the failover disk group

This entry does not influence SCSI-3 PR key management on CVM disk groups and is therefore not required in configurations that do not have failover disk groups.

“UseFence=SCSI3” also instructs VCS to call the appropriate ioctl module to register with the fencing driver.
7. **How do I start and stop the I/O fencing driver?**

To configure the vxfen driver:

On Solaris 9 and Linux,
```
# /etc/init.d/vxfen start
```

On Solaris 10:
```
# svcadm enable vxfen
```

On HP-UX:
```
# /sbin/init.d/vxfen start
```

On AIX:
```
# /etc/init.d/vxfen.rc start
```

To stop the I/O fencing driver:

On Solaris 9 and Linux:
```
# /etc/init.d/vxfen stop
```

On Solaris 10:
```
# svcadm disable vxfen
```

On HP-UX:
```
# /sbin/init.d/vxfen stop
```

On AIX:
```
# /etc/init.d/vxfen.rc stop
```

**VCS I/O fencing configuration files**

1. **What are the configuration files in I/O fencing?**

   The **configuration files in I/O fencing** are `/etc/vxfenmode` and `/etc/vxfendg`.

2. **What does the `/etc/vxfenmode` file contain?**

   This file contains the following information:
   
   a. **I/O fencing mode**: The acceptable values are “scsi3”, “customized” and “disabled”.
   
   b. **I/O fencing mechanism**: This information is relevant only for customized fencing. The value must be set to “cps”, if you are using the customized mode.

   c. **Disk policy**: The disk policy determines the way in which I/O fencing communicates with the coordinator disks. The acceptable values are “raw” and
d. Security setting: The security attribute must be set if fencing is enabled in customized mode with CP servers as the coordination points.

e. Coordination point details: This information is present if fencing is enabled in customized mode and CP servers are used as coordination points.

The following is a sample of the /etc/vxfenmode file when fencing is enabled in SCSI-3 mode:

```plaintext
vxfen_mode=scsi3
scsi3_disk_policy=dmp
```

The `vxfen_mode` attribute indicates that fencing is enabled in scsi3 mode and the coordination points are SCSI-3 PR compliant disks. All the coordinator disks are mentioned using a coordinator diskgroup specified in `/etc/vxfendg` file.

The following is a sample of the /etc/vxfenmode file when fencing is enabled in customized mode:

```plaintext
vxfen_mechanism=cps
vxfen_mode=customized
security=1
cps1=[10.209.80.124]:14250
vxfendg=vxfencoorddg
```

In the example, the coordination points used are a combination of CP server and coordinator disks. The CP server based coordination point is specified using the virtual IP of the CP server and the port value. The coordinator disks are specified using the coordinator disk group “vxfencoorddg”.

3. What does the /etc/vxfendg file contain?

The /etc/vxfendg file contains the name of the coordinator disk group that is created with SCSI-3 compliant coordinator disks. This file is valid only if fencing is enabled in SCSI-3 mode.

4. What are the possible values that can be assigned to the UseFence attribute in the main.cf file?

The UseFence attribute can take one of the following values: "SCSI3", "NONE".

If the attribute is set to NONE, VCS does not register with the fencing driver irrespective of whether fencing is enabled or disabled. VCS does not make use of I/O fencing abilities while failing over service groups.
About Coordination Points

1. What are coordination points?
   In the clustering paradigm, a coordination point refers to a point of arbitration. For VXFEN to operate in enabled mode, it needs points of coordination, which are accessible to all the nodes in the cluster, even in the event of a split-brain condition. At present, VXFEN uses SCSI-3 PR compliant FC SAN connected LUNs or iSCSI targets as coordination points. VXFEN also supports remote IP connected servers called CP servers to behave as coordination points.

2. What is the least number of coordination points required for fencing?
   A minimum of three coordination points is required for VXFEN.

3. What is the criteria influencing the number of coordination points?
   The number of coordination points must be odd and greater than or equal to three. Veritas I/O fencing recommends use of three coordination points. More number of coordination points only helps in scenarios where there is a possibility of multiple coordination points going down.

About coordinator disks

1. What are coordinator disks?
   When disks are used as coordination points, they are called as coordinator disks. They are special purpose disks that are SCSI-3 PR compliant. Coordinator disks cannot be used for storing user data. They are VxVM disks from an FC SAN attached storage or iSCSI target. VCS recommends three disks to be used as coordinator disks.

2. Can I use SCSI-2 compliant LUNs as coordinator disks?
   No. The SCSI-2 mechanism has inherent drawbacks such as reservation support for only two hosts, non-compliance with VxVM/DMP, and possible bus resets leading to data corruption. VXFEN therefore mandates SCSI-3 compliant LUNs to serve as coordinator disks for supporting VCS clusters connected to shared data storage.

3. Why do I need SCSI-3 compliant LUNs as coordinator disks?
   VXFEN uses the persistent reservation (PR) feature of SCSI-3 specifications for its operations. SCSI-3 PR supports multiple nodes accessing a device while at the same time blocking access to other nodes. Persistent reservations are persistent across SCSI bus resets and also supports multiple paths from a host to a disk. The VXFEN module on each node
registers a key on the coordinator disks during node startup. In the event of a split-brain situation, VXFEN uses the "pre-empt and abort" feature of SCSI-3, to eject other node’s keys in an atomic and consistent fashion and in doing so, wins the race for that coordination point. This mechanism is used to decide the winning sub-cluster that survives the split brain.

4. **What is the minimum size of LUN required for coordinator disks?**

VXFEN does not impose a hard limit on coordinator disk size. However, since the coordinator disks are controlled by VxVM, the private region size of VxVM disks is one of the factors that decide the minimum size of the coordinator disk. The vxfentsthdw (1M) utility skips the VxVM private region while validating a coordinator disk, thus requiring the disk to be a minimum of 128 MB in size.

5. **Can the coordinator LUNs be allocated from different disk arrays?**

Yes. It is strongly recommended that the coordinator LUNs be distributed across different disk arrays. This ensures that in the event of a disk array failure, the rest of the coordinator disks are accessible. However, this may not completely insulate the cluster from cluster shutdown, if a disk array fails. This is because, the coordinator disks from the remaining disk array may be even in number and this may lead to a complete cluster shutdown.

6. **Is there an alternative to SCSI-3 compliant FC SAN attached LUNs for VXFEN coordination points?**

With VCS 5.0 MP3, software iSCSI initiator-targets are also supported as coordination points for VXFEN. From version 5.1 onwards, Symantec supports the use of Coordination Point Server (CPS) as coordination points. CPS is an IP connected, software-based coordination point, serving multiple VCS clusters.

7. **Is SCSI-3 compliance mandatory for iSCSI targets to be used as coordinator disks?**

Yes. iSCSI targets need to be SCSI-3 compliant to be used as I/O fencing coordinator disks.

8. **How do I decide if an iSCSI target is suitable for use as coordination point?**

You can use the vxfentsthdw (1M) utility, shipped with VCS, to check if the iSCSI target is SCSI-3 compliant. This utility runs on two nodes and determines if the disk can serve as a coordinator disk for use by VXFEN. If there are more than two nodes in the cluster, the utility must be run multiple times on different pairs of nodes such that it is successful at least once on all nodes in the cluster.
9. **How do I check if an FC SAN connected disk is SCSI-3 compliant?**
   You can use the `vxfentsthdw (1M)` utility, shipped with VCS, to check if the disk is SCSI-3 compliant. This utility runs on two nodes and determines if the disk can serve as a coordinator disk for use by VXFEN. If there are more than two nodes in the cluster, the utility must be run multiple times on different pairs of nodes such that it is successful at least once on all nodes in the cluster.

10. **What are the criteria influencing the number of coordination points?**
    The number of coordination points must be odd and greater than or equal to three. Increased number of coordination points help in scenarios where there is a possibility of multiple coordination points going down. Veritas I/O fencing recommends the use of three coordination points.

11. **Can I use a combination of iSCSI targets and FC SAN controlled LUNs as coordinator LUNs?**
    Yes. For example, in the case of a stretch cluster, a local FC SAN connected LUN at each site and an iSCSI target at the third site can be used as three coordination points for the cluster.

12. **Are hardware based iSCSI initiators (iSCSI HBAs) supported as coordinator LUNs?**
    No. Only software-based iSCSI initiators are supported in the use of iSCSI targets as coordinator LUNs, starting with VCS 5.0 MP3.

13. **Does VXFEN read/write to coordinator disks frequently?**
    No. VXFEN does not issue read/write operations to coordinator disks for the purpose of I/O fencing. The coordination points are accessed by VXFEN only during system startup and when there is a cluster reconfiguration. It does not affect the I/O performance or the node failover performance.

**Servers as coordination points**

1. **What is Coordination Point Server?**
   CP server is an alternative arbitration mechanism that integrates with the existing VCS I/O fencing module. The Coordination Point server (CPS) is a software solution running on a remote system or cluster that provides arbitration functionality by allowing client cluster nodes to perform the following tasks:
   a. Self-register to become a member of an active client cluster with access to the data drives
   b. Check for other nodes that are registered as members of this active client cluster
c. Self-unregister from this active client cluster
d. Forcefully unregister other nodes (preempt) as members of this active client cluster

2. **How is CP server different from a coordinator disk?**
   A CP server provides the same functionality as a coordinator disk in an I/O fencing scenario. However, CP server based coordination points provide the following enhanced capabilities:
   a. CP server configurations are scalable, and a configuration with three CP servers can provide I/O fencing for multiple client clusters
   b. Appropriately situated CP servers can eliminate any coordinator disk location bias in the I/O fencing process

3. **What are the possible coordination point configurations to be used in fencing?**
   Symantec recommends the following three coordination point configurations:
   a. Three coordinator disks configuration (traditional I/O fencing)
   b. Two coordinator disks with one CP server configuration
   c. Three CP server configuration

4. **How do I configure a CP server based coordination point to be used by VxFEN?**
   For detailed information on how to configure server-based coordination points, refer to the Veritas™ Cluster Server Installation Guide.

5. **Where does CP server reside?**
   The CP server must reside on a node distinct from the other client cluster nodes. It can also be hosted on an SFHA cluster to provide high availability to the CP server.

6. **In a purely server-based fencing setup, what is the recommended number of CP servers to be used?**
   Symantec recommends the use of three CP servers as coordination points. However, any odd number above three is also acceptable.
Creation of coordinator disks

1. I have a two-node VCS cluster with shared storage but the two nodes see the same LUNs with different names. How do I use them as coordinator LUNs?

   If VxVM is installed on the node, creating a temporary coordinator disk group as a VxVM private disk group on any of the cluster nodes with three disks (visible to all the cluster nodes) helps the rest of the nodes see these disks as VxVM disks. The coordinator disk group can now be deported and the /etc/vxfendg file on each node must be populated with the coordinator disk group name. VXFEN automatically generates the /etc/vxfentab file and populates this file with the list of disks that can be used as coordinator disks on each node.

2. Why shouldn't the coordinator disk group be manually or automatically imported? What will happen if it is shared and imported by CVM on all nodes at the same time?

   The coordinator disk group is created as a private disk group and kept in a deported state so that it is not picked up accidentally for data storage. If the coordinator disk group is imported as a shared disk group, it conflicts with CVM's management of data disks for fencing; CVM treats it as another shared disk group and attempts to install SCSI-3 registration or reservations.

3. What type of disk format should I use for coordinator disks?

   Any of the following disk formats can be used for coordinator disks: Cross-Platform Data Sharing (CDS), Enclosure Based Naming (EBN), native naming, simple disks, or sliced disks. For information on whether a disk array from a specific vendor is supported, please refer to the HCL at the following location:

   http://www.symantec.com/business/products/otherresources.jsp?pcid=2245&pvid=203_1

Replacement of coordinator disks

1. What steps do I follow to replace coordinator disks with new disks?

   Replacing coordinator disks with new disks requires stopping the VCS cluster and I/O fencing components. Please refer to the “Troubleshooting fencing” section of the Veritas Cluster Server Installation Guide for more information. Starting with VCS 5.0 MP3, a utility called vxfenswap (1M) can be used to replace coordinator disks online without bringing down the cluster.
2. **From an I/O fencing point of view, what steps should I take when I migrate storage from one HBA to another?**

   If storage migration includes movement of coordinator disks, then VCS as well as VXFEN needs to be stopped before the migration. They can be made available once the migration is complete at the operating system and VxVM level. Starting with VCS 5.0 MP3, a utility called vxfenswap (1M) can be used to replace coordinator disks online without bringing down the cluster.

3. **How do I move coordinator disks to another array?**

   If coordinator disks need to be moved to another array, then VCS as well as VXFEN needs to be stopped before the migration. They can be made available once the migration is complete at the operating system and VxVM level. Starting with VCS 5.0 MP3, a utility called vxfenswap (1M) can be used to replace coordinator disks online without bringing down the cluster.

4. **How do I replace a single faulty coordinator disk in the coordinator disk group without stopping the cluster using vxfenswap?**

   Perform the following steps to replace a disk in a coordinator disk group when the cluster is online.

   d. Import the coordinator disk group using the vxdg command.
      
      ```
      # cat /etc/vxfendg
      Vxfencoorddg
      # vxdg import vxfencoorddg
      ```

   e. If the disk group was created with the coordinator flag set, reset the flag temporarily.
      
      ```
      # vxdg list vxfencoorddg |grep flags:|grep coordinator
      flags:     cds coordinator
      # vxdg -g vxfencoorddg set coordinator=off
      ```

   f. Remove the faulty disk (ex: c4t0d4) from the coordinator disk group and add the new disk (ex: c4t0d3)
      
      ```
      # vxfenconfig -l
      I/O Fencing Configuration Information:
      ================================
      Single Disk Flag     : 0
      Count                : 3
      Disk List
      ```
### Frequently Asked Questions: VCS I/O Fencing

#### 5. **How do I replace the current coordinator disk group with a new disk group without stopping the cluster using vxfenswap?**

Perform the following steps to replace the coordinator disk group when the cluster is online.

- **a.** View the name of the disk group in the file `/etc/vxfendg`. Typically, the name of the disk group that contains the coordinator disks is `vxfencoorddg`.
  
  ```bash
  # cat /etc/vxfendg
  vxfencoorddg
  ```
- **b.** View the alternative disk groups available for replacement.
  
  ```bash
  # vxdisk -o alldgs list
  ```
- **c.** From one of the nodes, start the vxfenswap utility and specify the new diskgroup with the `–g` option.
  
  ```bash
  # /opt/VRTSvcs/vxfen/bin/vxfenswap -g vxfendg
  ```
- **d.** Review the `/etc/vxfendg` file to confirm that the new coordinator disk group is in use.
  
  ```bash
  # cat /etc/vxfendg
  vxfendg
  ```

#### 6. **How do I change the disk interaction policy online using vxfenswap?**

Perform the following steps to change the disk interaction policy when the cluster is online.

- **a.** View the current disk policy in the file `/etc/vxfenmode`.
  
  ```bash
  # cat /etc/vxfenmode
  vxfen_mode=scsi3
  ```

---

### Disk Information

<table>
<thead>
<tr>
<th>Disk Name</th>
<th>Major</th>
<th>Minor</th>
<th>Serial Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>/dev/vx/rdmp/c4t0d5</td>
<td>32</td>
<td>32</td>
<td>R450 00013154 0313 dmp</td>
</tr>
<tr>
<td>/dev/vx/rdmp/c4t0d6</td>
<td>32</td>
<td>16</td>
<td>R450 00013154 0314 dmp</td>
</tr>
<tr>
<td>/dev/vx/rdmp/c4t0d4</td>
<td>32</td>
<td>48</td>
<td>R450 00013154 0312 dmp</td>
</tr>
</tbody>
</table>

#### g. From one of the cluster nodes, run the vxfenswap utility:

```bash
# /opt/VRTSvcs/vxfen/bin/vxfenswap -g vxfencoorddg
```

#### h. If the coordinator flag was reset in the second step, set it using the following command:

```bash
# vxdg -g vxfencoorddg set coordinator=on
```

The output of `vxfenconfig -l` confirms that the disk is replaced.
b. Confirm the current disk policy in the running cluster.
   # vxfenadm -d
   I/O Fencing Cluster Information:
   =============
   Fencing Protocol Version: 201
   Fencing Mode: SCSI-3
   Fencing SCSI-3 Disk Policy: dmp

c. On each node in the cluster, edit /etc/vxfenmode to include the new disk policy.
   # cat /etc/vxfenmode
   vxfen_mode=scsi3
   scsi3_disk_policy=raw

d. From one of the nodes, start the vxfenswap utility:
   # /opt/VRTSvcs/vxfen/bin/vxfenswap -g diskgroup

e. Run vxfenadm -d to confirm that the new disk policy is in effect.
   # /sbin/vxfenadm -d
   I/O Fencing Cluster Information:
   =============
   Fencing Protocol Version: 201
   Fencing Mode: SCSI-3
   Fencing SCSI-3 Disk Policy: raw

7. How do I refresh lost keys on current coordinator disks without stopping the cluster using vxfenswap?

   Perform the following steps to restore lost keys on coordinator disks when the cluster is online.

   a. Verify that both the nodes are part of the cluster.
      # vxfenadm -d

   b. Verify that the coordinator disks have no keys.
      # vxfenadm -g all -f /etc/vxfentab

   c. From one of the nodes, start the vxfenswap utility.
      # /opt/VRTSvcs/vxfen/bin/vxfenswap -g diskgroup

   d. Confirm that the keys are atomically placed on the coordinator disks.
      # vxfenadm -g all -f /etc/vxfentab
High availability of coordinator disks

1. **Can I use multi-pathed LUNs as coordinator disks?**
   Yes, LUNs with multiple access paths are preferred for use as coordinator LUNs as they provide high availability against single points of failures such as HBA failure or switch failure.

2. **Can I use LUNs controlled by any multipathing solution as coordinator LUNs?**
   No, only VxVM/DMP is supported as the multipathing solution for coordinator LUNs. However, if DMP co-exists with third-party multipathing drivers such as EMC Powerpath, then such LUNs are also supported. For more information on supported third-party driver configurations, refer to the HCL at the following location:

3. **How does DMP shield against path failures to a coordinator LUN?**
   DMP keeps track of all the paths available to coordinator LUNs and automatically fails over to another path if an I/O failure is noticed on the current active path.

4. **What types of arrays are supported for coordinator LUNs?**
   The following types of arrays are supported for coordinator LUNs when VxVM DMP controls them: Active/Active (A/A), Active/Passive (A/P), Asymmetric Logical Unit Access (ALUA). Appropriate Array Support Library (ASL) and Array Policy Module (APM) must be installed on the system for VxVM/DMP to discover the LUNs correctly and prepare them for use as coordinator LUNs.
   For more information on ASLs and APMs compatible with the VxVM version, refer to the HCL at the following location:

5. **Can I use native volume manager (LVM) controlled LUNs as coordinator LUNs?**
   No, VXFEN can only be used with VxVM controlled LUNs.

6. **Do you recommend mirroring of the coordinator disks?**
   Yes, if the array supports hardware mirroring, it is recommended that the coordinator disks be mirrored at the hardware level. This prevents a physical disk failure from affecting the cluster at the time of split brain.

7. **Can software-based mirroring using VxVM be employed to mirror coordinator disks?**
   No. Software-based mirroring using VxVM results in an even number of coordinator disks. VXFEN will not be able to resolve split brain based on majority votes.
8. Can raw disks be used as coordinator LUNs?
Though VXFEN requires having coordinator LUNs placed under VxVM control for high availability, it also allows raw OS paths to be specified as coordinator disks. This is done by setting vxfen_disk_policy to raw in the file /etc/vxfenmode when starting VXFEN. However, in the case of HP-UX 11iv3, only VxVM (DMP) devices can be specified as coordinator disks. Make sure that the attribute vxfen_disk_policy is set to dmp in the /etc/vxfenmode file. iSCSI devices are supported as coordinator disks only when DMP disk policy is used.

9. Can coordinator disks be used as data disks?
No, coordinator disks must be dedicated to VXFEN. They must not be used as data disks as this could interfere with the functionality of VXFEN.

Monitoring registrations on coordination points

1. How do I know if my local node's registrations on the coordination points are intact?
VCS I/O fencing provides an agent called Coordination Point agent, which is a monitor only agent that runs on each node in the cluster. CP agent on a node monitors whether the local node's keys are registered on coordination points specified in /etc/vxfenmode file. When the number of coordination points with missing keys exceeds a predefined threshold value, the agent goes to faulted state otherwise it reports ONLINE status.

2. What is threshold value? Who sets this value?
The threshold value is the fault tolerance attribute of the CP agent and determines when the agent should report ONLINE/FAULTED status. The value is configurable by the user. The default value is 0.

3. What is the recommended value of fault tolerance?
The fault tolerance must be configured depending upon user requirements. For example, set the fault tolerance value to 0 if you want the agent to go to a faulted state when even a single coordination point is either not reachable or is missing the local node's keys. Set the fault tolerance value to 1 if you want the agent to report a faulted state only if two or more failures occur.
4. **Where can I find the Coordination Point agent log messages?**
   
   You can find the Coordination Point agent log messages in the VCS engine log file located at /var/VRTSvcs/log. The VCS engine log file provides information on the coordination point resources, registrations on coordination points, and other information. For detailed information, see the Veritas Cluster Server Administrator’s Guide.

### About SCSI-3 PR keys on coordinator disks and data disks

1. **Who places SCSI-3 PR keys on coordinator disks?**
   
   VXFEN places SCSI-3 PR keys corresponding to each cluster node on the coordinator disks. When the coordinator disks are controlled by VxVM, the SCSI-3 PR keys are placed by DMP on behalf of VXFEN.

2. **Who places SCSI-3 PR keys on data disks?**
   
   In an SF CFS or SF Oracle RAC cluster with I/O fencing enabled, during the disk group import operation, VXFEN directs VxVM/CVM to place SCSI-3 PR keys on the data disks when a node starts or joins a cluster. In an SF HA cluster, when disk groups are imported with ‘–o groupreserve’ option, VxVM places the keys corresponding to the node on the data disks belonging to that disk group. The PR keys are removed from the data disks by VxVM/CVM, when the corresponding disk groups are deported.

3. **Are the SCSI-3 PR keys cleared when one or more node is rebooted?**
   
   When a node is shut down gracefully, VXFEN on that node removes the SCSI-3 PR keys from the coordinator disks. If a node goes down abruptly, the key corresponding to the node is removed by VXFEN from the other nodes. If the last cluster node goes down abruptly, stale keys corresponding to that node remain on the coordinator disks, which need to be cleared by the administrator. The first node that becomes available in the cluster then reports a “pre-existing split brain”.

4. **Are the SCSI-3 PR keys cleared when the array is rebooted?**
   
   Yes, the SCSI-3 PR keys on the disk array are cleared when the array is rebooted.

5. **Are the SCSI-3 PR keys cleared when there is a firmware upgrade on the array?**
   
   Yes, if the firmware upgrade results in the rebooting of the array, the SCSI-3 PR keys are cleared.

6. **Is there a way to detect accidental SCSI-3 PR key removal from coordinator LUNs?**
   
   The keys currently installed on the coordinator disks can be read using the following command:

   ```bash
   # vxfenadm -g all -f /etc/vxfentab
   ```
There should be a key for each node in the operating cluster on each of the coordinator disks for normal cluster operation.

7. **Will dynamic LUN expansion affect SCSI-3 PR keys installed on such LUNs?**
   No. SCSI-3 PR keys are not impacted by dynamic LUN expansion.

8. **Will GAB's IOFENCE affect SCSI-3 PR keys on coordinator disks and data disks in any way?**
   When GAB sends the IOFENCE message to one or more nodes, it will force a system panic on those nodes. VXFEN on the surviving cluster takes the responsibility of removing the PR keys of the ejected nodes from the coordinator disks and coordinates with CVM to eject the nodes from the data disks.

9. **Is there an impact on I/O performance or failover performance or VxVM when SCSI-3 PR keys are installed on data or coordinator disks?**
   No, the SCSI-3 PR keys are installed only during cluster or node startup and accessed during a race following a split brain condition. There is no impact on I/O performance or failover performance.

**About stale SCSI-3 PR keys**

1. **Under what circumstances do I expect coordinator LUNs to have stale SCSI-3 PR keys?**
   If the last node in the cluster goes down ungracefully, its keys may remain on the coordinator LUN when any node in the cluster comes up.

2. **Can stale SCSI-3 PR keys on coordinator disks prevent a node from joining the cluster?**
   Yes, if a node joining the cluster finds stale keys on the coordinator disks, it reports a “pre-existing” split-brain and does not join the cluster. The administrator must manually clear the stale keys before the node can join the cluster.

3. **Under what circumstances do I expect stale SCSI-3 PR keys on data disks?**
   When the last node in the cluster is not brought down cleanly, keys on the data disks and the coordinator disks may persist. The administrator needs to clear the keys on the coordinator disks and data disks. In systems where VxVM controls the data disks, certain failure scenarios, such as the system panic during booting can lead to temporary “zzzzzzzz” keys being left behind on the data disks.
4. Can stale SCSI-3 PR keys on data disks prevent a node from joining the cluster?
Yes. In an SF CFS or SF Oracle RAC environment, if there are stale SCSI-3 PR keys left on the data disks, CVM cannot bring the shared data disks online. Also, shared disk groups cannot be imported. As a result, the node is unable to join the cluster.

Evaluating SCSI-3 PR keys

1. How do I evaluate the number of SCSI-3 PR keys on an FC SAN connected coordinator LUN, if there are multiple paths to the LUN from the hosts?
The utility vxfenadm (1M) can be used to display the keys on the coordinator LUN. The key value identifies the node that corresponds to each key. Each node installs a registration key on all the available paths to the LUN. Thus, in most cases the number of keys registered on each node will be equal to the number of active paths from that node to the LUN. Some specialized SAN topologies may lead to fewer numbers of keys registered on the disks.

2. How do I evaluate the number of SCSI-3 PR keys on an iSCSI target as coordinator LUN?
The utility vxfenadm (1M) can be used to display the keys on the coordinator LUN. The key value identifies the node corresponding to each key. Each node installs a registration key on all the available paths to the LUN. Thus, the total number of registration keys is the sum of keys installed by each node in the above manner.

3. In which environments can I see SCSI-3 PR keys on data disks?
SCSI-3 PR keys are registered on data disks in all environments where I/O fencing is enabled (VCS, SF HA, SF CFS, SF Oracle RAC). However the format of the keys registered on data disks in SF HA differs from those in SF CFS and SF Oracle RAC.

4. What are the options that can be used with the vxfenadm (1M) command to evaluate the SCSI-3 PR keys?
In version 5.0, “vxfenadm -g | -G” can be used to read keys on coordinator disks (using READ_KEYS command).
In version 5.1, “vxfenadm -s” can be used to read the keys on coordinator disks (using READ_KEYS command).
In version 5.0 and 5.1, “vxfenadm –r” returns the reservations on coordinator disks (using READ_RESERVATIONS command).

5. How are the SCSI-3 PR keys on coordinator LUN different from those on data disks?
In 5.0, the SCSI-3 PR keys on coordinator LUNs are placed by VXFEN. The format of the key follows the naming convention wherein ASCII “A” is prefixed to the LLT ID of the system followed by 7 dash characters.
Example:
node 0 uses A--------
node 1 uses B--------

In the VCS environment, the keys on data disks are registered by VCS, the format of which is ASCII “A” prefixed to the LLT ID of the system followed by the characters “VCS”.

Example:
node 0 uses AVCS
node 1 uses BVCS

In an SF Oracle RAC/SF CFS/SF HA environment, VxVM/CVM registers the keys on data disks, the format of which is ASCII “A” prefixed to the LLT ID of the system followed by the characters “PGRxxxx” where ‘xxxx’ = i such that the disk group is the ith shared group to be imported.

Example:
node 0 uses APGR0001 (for the first imported shared group)

In addition to the registration keys, VCS/CVM also installs a reservation key on the data LUN. There is one reservation key per cluster, as only one node can reserve the LUN.

In version 5.1, the format of keys on coordinator disks are modified to make them unique across clusters. Each key is calculated from the cluster ID and node ID of the node.

The key format on coordinator disks is “VF” followed by the cluster ID occupying 4 bytes and the node ID occupying 2 bytes in hexadecimal notation.

An example of a fencing key on a coordinator disk is VF03FF01. In this example, the cluster ID is 03FF and the node ID is 01. The cluster ID is the LLT cluster ID encoded in the hexadecimal format. The node ID is the LLT node ID also encoded in the hexadecimal format.

6. **Why do we need fencing keys on coordinator disks to be unique across clusters? Does VXFEN support multiple clusters using the same coordinator disks in the array?**

No. Currently VXFEN does not support sharing of coordinator disks across multiple clusters. Unique fencing keys have been added so that it is easy to identify the cluster to which a disk belongs and if the disk is zoned to multiple clusters.
Administering SCSI-3 PR keys

1. **What administrative action should I take if I find that the SCSI-3 PR keys are missing from coordinator LUNs?**
   Starting with VCS 5.0 MP3, the utility `vxfenswap(1M)` can be used to replace coordinator disks online with the very same disks without bringing down the cluster. During disk replacement, the missing keys register again without any risk of data corruption.

2. **How do I clear stale keys from a coordinator LUN?**
   The utility `vxfenclearpre(1M)` can be used to clear the stale keys from the coordinator disks. This command can be run on the node that is not part of a live cluster.

3. **How do I identify a faulty coordinator LUN?**
   The utility `vxfentsthdw(1M)` provided with VXFEN can be used to identify faulty coordinator LUNs. This utility must be run from any two nodes in the cluster. The coordinator LUN, which needs to be checked, should be supplied to the utility. Do not run this utility on data disks or coordinator disks on production systems as it over-writes existing keys and data.

4. **Is there a way to swap a data LUN with a coordinator LUN?**
   No, the keys registered on the data LUN and coordinator LUN is different and cannot be swapped. However, starting with VCS 5.0 MP3, in the event of a disk migration or a faulty coordinator LUN, you can perform a swap by using the `vxfenswap(1M)` utility shipped with VCS.

5. **How do I replace a faulty data LUN that has stale SCSI-3 PR keys?**
   The VxVM/CVM disk replacement procedure handles the removal of SCSI-3 PR keys on old data disks and installation of new keys on the replaced LUN. Please refer to the *Veritas Volume Manager Administrator’s Guide* for instructions on replacing the data disks.

6. **Can I remove the SCSI-3 PR keys from data disks without affecting the data?**
   Yes, removing the SCSI-3 PR keys from the data disks does not affect the actual data on the disks.

7. **How do I clear stale SCSI-3 PR keys from data disks without affecting the data?**
   The utility `vxfenclearpre(1M)` can be used to clear stale keys from the data disks. This command can be run on the node that is not part of a live cluster.
8. **Does running vxfenclearpre have a negative impact on other clusters that use fencing?**

*Does it imply that FC SAN zoning should strictly isolate the disks used by a cluster?*

The coordinator disks of a cluster should not be used by any other cluster. It is a good practice to isolate coordinator disks using FC SAN zoning to prevent accidental removal of keys due to human error while using vxfenclearpre. In addition, when a coordinator disk group is created, the `-'o coordination=on' flag should be passed to the ‘vxdg’ command to safeguard coordinator disks from accidental access.

9. **Is a cluster reboot necessary after running vxfenclearpre on coordinator disks?**

The utility vxfenclearpre (1M) is used to clear the stale keys of an evicted node. It cannot be run on the coordinator disks of a live cluster.

### Coordination points in a campus cluster

1. **Why do I need to place an arbitration point in a third site in case of a two-node stretch cluster?**

A coordination point in a third site allows at least a single node cluster to continue operations in the event of a site failure in a two-node stretch cluster. The other site or node can still access the coordination point in the third site, apart from its local coordination point. This allows the surviving node to gain majority vote at the time of racing for coordination points. On the other hand, in the absence of a third coordination point, each site will, in all probability, have won the race for its local coordination point, and with one vote each, each node will commit suicide.

2. **What are the criteria for placement of coordination points in the case of campus clusters?**

In campus clusters, it is recommended that one coordination point be placed in each of the sites and the third coordination point be placed in a third site such that all the nodes in both the primary and secondary sites are connected to it. This ensures that fencing works fine even if one of the sites becomes unavailable.

3. **What are the criteria for the placement of coordination points in the case of two site VCS/SFHA/SF Oracle RAC campus clusters?**

It is advisable to have two coordinator disks placed at two sites and one highly available CP server to be placed at the third site.
4. What are the criteria for the placement of coordination points in the case of large numbers of two-node or four-node VCS/SFHA/SF Oracle RAC clusters?
For a data center with a large number of small clusters, it may not be feasible to allocate two coordinator disks for each cluster. In such situations, we recommend three independent highly available CP servers shared by multiple clusters. The CP servers must be hosted on an SFHA cluster to provide high availability.

VCS I/O fencing logs and configuration files
1. Is it true that /etc/vxfenmode is not required unless fencing needs to be disabled?
No, /etc/vxfenmode is required by VXFEN irrespective of whether it is enabled or disabled. The state (enabled/disabled) should be set appropriately in the file before VXFEN is started on the node.

2. When is /etc/vxfenmode created? What is the significance of the entries in this file?
The file /etc/vxfenmode must be configured by the administrator before VXFEN is started on the node. Sample vfenmode files are available in the directory /etc/vxfen.d. This file is used by VXFEN in deciding whether or not to start in the enabled mode and also in determining whether to use DMP or raw coordinator disks, if it is enabled in SCSI-3 mode.

3. When is /dev/vxfen created?
The file /dev/vxfen gets created when the VXFEN driver is started on the cluster node. Users are not supposed to touch this file.

4. What is the significance of the order of entries in the /etc/vxfentab file given that fencing uses the serial number to order the disks for races?
When VxVM is configured on the system, the /etc/vxfentab file is automatically generated by VXFEN based on the coordinator disk group specified in the /etc/vxfendg file. It contains the list of disks in the VXFEN disk group. At the time of split brain, VXFEN on each node races for the disks in the order listed in the file.

5. Can VCS I/O fencing be assumed to be running normally in the following cases:
   'gabconfig -a' shows the fencing port 'b' enabled on both nodes 'vxfenadm -g all -f /etc/vxfentab' shows the registered keys on the coordinator disks?
Yes, this information is sufficient to infer that fencing is enabled and functioning normally. Alternatively, the 'vxfenadm –d' command can be used to verify that fencing is enabled on all cluster nodes.
6. Which are the log files I need to consult for investigating I/O fencing issues?
   Apart from the console log, the vxfen.log in the /var/VRTSvcs/log directory provides information on I/O fencing related issues.

Best practices for VCS I/O fencing

1. What are the best practices to prevent data corruption in SF CFS RAC cluster during potential split-brain conditions?
   Since SF CFS for Oracle RAC does not include support for the I/O fencing feature, it is necessary to configure two heartbeat links. When a node is reduced to a single heartbeat connection, SF CFS can no longer discriminate between the loss of a system and the loss of the final network connection. This “jeopardy” state affects all applications that use the cluster file system mount points. (Jeopardy cannot be avoided as the I/O fencing feature is not available.) If a node fails after the jeopardy state has been notified, all the cluster nodes cease to be members of the shared disk group, and the mounted cluster file systems are disabled on all nodes in the cluster. This action is taken to prevent potential data corruption as SF CFS is neither aware of nor can it coordinate with the native fencing support feature of Oracle. However, starting with VCS 5.0 MP3, this behavior differs, in that CFS does not take any action if VXFEN is not configured in enabled mode and there is a jeopardy membership followed by node loss.

2. Is VCS I/O fencing an alternative for disk heartbeating?
   No. VCS no longer supports disk heartbeating. With disk heartbeating, there is an additional path available for VCS to do heartbeating if one of the two private interconnects fail. Configuring disk heartbeating does not provide complete protection against all split-brain scenarios. VXFEN provides more capabilities than disk heartbeating to prevent data corruption in the cluster.

3. Can I/O fencing work if vxconfigd is not running?
   Since coordinator disks and data disks are controlled by VxVM, I/O fencing requires vxconfigd to be running in order to function correctly.
**Frequently Asked Questions: VCS I/O Fencing**

**Best practices to prevent split brain**

1. **What are the potential "hot-spots" to look for in avoiding a split-brain situation?**
   
   When the GAB membership (displayed by the gabconfig command) shows a link in jeopardy state, it is an indication of a potential hot-spot, as the failure of another link can lead to a split-brain situation. The administrator should check the cluster interconnects between the nodes concerned when the GAB membership shows a link in jeopardy state.

2. **Does having redundant interconnects alone prevent split brain?**
   
   No. Having redundant interconnects alone cannot prevent split brain from occurring. However, as all the links between any two nodes in the cluster need to go down for a split brain to occur, using redundant interconnects can reduce the chances of split brain. However, this will not shield against split-brain scenarios arising out of non-responding applications and operating systems.

3. **What are the best practices to reduce the risk of private interconnect failures?**
   
   Refer to the following Symantec technote for various options available to lower the risk of split brain:

   [http://support.veritas.com/docs/252635](http://support.veritas.com/docs/252635)

**Best practices for campus clusters**

1. **What are the best practices to prevent host level split brain in campus clusters?**

   To minimize the chances of host-level split-brain, the cross-site LLT network traffic and storage I/O traffic should be routed over the same physical medium using technologies such as Dense wavelength division multiplexing (DWDM). The storage and networks must have redundant-loop access between each node and each storage array. A network infrastructure failure in this case limits the impact to a site-wide split-brain instead of a host-level split-brain.

2. **What are the best practices for placement of coordinator disks in campus clusters?**

   On campus clusters with two sites, if fencing is configured with two disks at one site and one disk at the other site, the site which has more coordinator disks has a higher probability of winning the race (all other factors such as the size of the sub-cluster remaining the same). Critical applications may be hosted on the site having two coordinator disks to minimize the impact. The limitation of this approach is that, if the site with two disks fail, the other site fails to gain majority votes to win the race. As a result, there may be a complete cluster shutdown.
In order to ensure the availability of at least one cluster node after split brain, one coordinator disk can be located on each site whereas the third coordinator disk needs to be placed at a third site. VXFEN offers the flexibility of using an iSCSI target at the third site as the coordinator disk, as against a fiber-connected SAN device.
References
For more information, refer to the following documents:

- VCS I/O fencing concepts and troubleshooting whitepaper
  

- Veritas Cluster Server 5.1 Installation Guide
  
  Available on the SF Oracle RAC 5.1 documentation CD

- Veritas Cluster Server 5.0 Installation Guide
  
  Available on the SF Oracle RAC 5.0 documentation CD

- Veritas Volume Manager 5.1 Administrator’s Guide
  
  Available on the SF Oracle RAC 5.1 documentation CD

- Veritas Volume Manager 5.0 Administrator’s Guide
  
  Available on the SF Oracle RAC 5.0 documentation CD

- Symantec technical note
  
  [http://support.veritas.com/docs/252635](http://support.veritas.com/docs/252635)
About Symantec

Symantec is a global leader in infrastructure software, enabling businesses and consumers to have confidence in a connected world. The company helps customers protect their infrastructure, information, and interactions by delivering software and services that address risks to security, availability, compliance, and performance. Headquartered in Cupertino, Calif., Symantec has operations in 40 countries. More information is available at www.symantec.com.

For specific country offices and contact numbers, please visit our Web site. For product information in the U.S., call toll-free 1 (800) 745 6054.

Symantec Corporation
350 Ellis Street
Mountain View, CA 94043
www.symantec.com

Copyright © 2009 Symantec Corporation. All rights reserved. Symantec and the Symantec logo are trademarks or registered trademarks of Symantec Corporation or its affiliates in the U.S. and other countries. Other names may be trademarks of their respective owners.