

# EMC GREENPLUM DATA COMPUTING APPLIANCE USING SAN MIRROR AND EMC SYMMETRIX VMAX FOR DISASTER RECOVERY

## A Detailed Review



## EMC SOLUTIONS GROUP

### Abstract

This white paper explains how customers can leverage EMC<sup>®</sup> Symmetrix<sup>®</sup> VMAX<sup>™</sup> SAN mirror and EMC Symmetrix Remote Data Facility (SRDF<sup>®</sup>) for data replication between two sites in synchronous mode to yield a reliable remote disaster recovery solution for EMC Greenplum<sup>®</sup> Data Computing Appliance data analytics deployments.

October 2011

Copyright © 2011 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.

For the most up-to-date listing of EMC product names, see EMC Corporation Trademarks on EMC.com.

All trademarks used herein are the property of their respective owners.

Part Number H8827.1

# Table of contents

<b>Executive summary .....</b>	<b>6</b>
Business case .....	6
Solution overview .....	7
Key recommendations.....	7
<b>Introduction .....</b>	<b>8</b>
Purpose .....	8
Scope .....	8
Audience.....	8
Terminology .....	8
<b>Overview of components .....</b>	<b>9</b>
Introduction to the components .....	9
EMC Greenplum DCA .....	9
Key features of the DCA .....	9
EMC Symmetrix VMAX .....	11
Symmetrix Management Application Server.....	11
Symmetrix Management Console .....	11
Auto provisioning devices .....	11
Storage groups .....	12
Port groups.....	12
Initiator groups.....	12
Masking view .....	12
EMC Symmetrix Remote Data Facility.....	12
Setting up SRDF consistency groups on VMAX using ECA.....	13
RDF consistency group operations .....	13
EMC TimeFinder/Snap.....	13
VDEV .....	14
Save devices .....	14
EMC PowerPath.....	14
EMC Solutions Enabler .....	15
<b>Solution overview .....</b>	<b>16</b>
Overview .....	16
Solution architecture.....	16
Hardware resources .....	17
Software resources .....	17

<b>Connecting the Symmetrix VMAX to a Greenplum DCA.....</b>	<b>18</b>
SAN configuration .....	18
Hardware components .....	18
VMAX to DCA—SAN configuration .....	18
VMAX configuration .....	19
DCA with SAN mirrors.....	20
Segment failure on local Site A.....	22
Allocation and mounting of SAN devices on the DCA.....	23
Moving mirrors.....	25
<b>SAN mirror SRDF consistency group - SRDF/S .....</b>	<b>27</b>
SRDF setup .....	27
<b>SAN mirror rotating snapshots.....</b>	<b>29</b>
TimeFinder/Snap on remote VMAX.....	29
Scripts .....	29
Prerequisites.....	30
Setting up the remote site to support rotating snapshots .....	32
Rotating snapshots script .....	33
Mount-validated snapshot script .....	35
<b>Failover.....</b>	<b>37</b>
Failover to the remote site.....	37
Restoring validated snapshots to the R2 devices .....	40
Using restoring snapshots.....	40
Database consistency check .....	40
Determining the current Master Server .....	41
Starting the database in utility mode.....	42
Checking the synchronization and consistency states .....	42
Reversing the Segment Server roles.....	42
Checking the catalog.....	43
<b>Failback.....</b>	<b>44</b>
R2 to R1 with changes.....	44
<b>SAN Mirror performance results .....</b>	<b>47</b>
Overview.....	47
DCA-only testing .....	47
DCA with VMAX SAN Mirror testing .....	48
DCA with VMAX SAN and SRDF testing.....	48
Comparison .....	48
Performance summary .....	49

**Automating the solution** ..... 50  
    Using customized scripts ..... 50

**Conclusion** ..... 56  
    Summary ..... 56  
    Findings ..... 56

**References**..... 57  
    White papers ..... 57  
    Product documentation..... 57  
    Other documentation ..... 57

## Executive summary

### Business case

Today's data warehouses are required to address an increasingly broader range of capabilities, driven by changing business requirements and enabled by advances in technology:

- Businesses are being driven to respond to realtime events to improve operational efficiency, to meet or exceed service level agreements, to respond to realtime market conditions, or to predict the next change. Analytics performed on the most recent data are of the greatest value to these organizations. New platforms, such as the EMC® Greenplum® Data Computing Appliance (DCA), are capable of loading data at very high speeds and in near real time, avoiding the need for overnight or weekly “snapshots”. As a result, the data in the warehouse is more current and of higher value to the organization.
- Changes to regulatory requirements have had a major impact on data warehouses, requiring broader protection and security of business and customer data.

Data warehouses are experiencing exponential data growth, driven in part by the addition of unstructured or semi-structured data (such as Internet clickstream data), with market, financial, and customer data. The cleansing and transformation of data, as it is loaded into the warehouse, further increases the value of this data to an organization. Businesses cannot afford to lose this data or the value that was added in the process.

Data warehouses have become mission-critical systems for the enterprise, and the need for enterprise-class protection of the data in the data warehouse has become an implicit business requirement. The size of the database, including historical data, and the critical nature of the data introduces new challenges to back-up and recovery of a data warehouse. Customers require solutions that:

- Provide effective protection and security, including the ability to replicate data off-site as necessary
- Can recover data quickly, without the need to recreate current or historical data
- Are simple and efficient to deploy and manage, and nondisruptive to data warehouse operations
- Can perform data replication without impact on the data load or query performance

## Solution overview

EMC Greenplum has created a purpose-built data analytics and business intelligence (BI) platform, the DCA. The DCA is offered in a variety of configurations to meet different customer sizing and performance needs.

This DCA SAN mirror solution uses EMC's market-leading disaster recovery (DR) solutions to ensure robust and reliable remote data protection for the DCA data analytics environment. The solution also allows companies to test the availability and reliability of their DR data periodically, without interrupting their primary site analytics processing and replication to the remote site, using remote site snapshot technology.

This white paper describes a DCA solution that uses EMC Symmetrix® VMAX™ SAN mirror and EMC Symmetrix Remote Data Facility (SRDF®) for data replication between two sites in synchronous mode. The focus of this white paper is the solution's ability to deliver a reliable and robust DR system for the DCA data analytics environment.

## Key recommendations

The purpose of this solution is to provide a recoverable instance of Greenplum Database™ at a remote site in the event of a disaster, and to provide a remote, testable image of Greenplum Database.

The key recommendations of this solution include the use of:

- SRDF in synchronous mode together with EMC Symmetrix Engenuity Consistency Assist (ECA) to enable a consistent and recoverable database image on a remote site.
- Rotating EMC TimeFinder®/Snap snapshots to ensure that continuous database checking and validation of the remote DCA database can take place. This is provided in this solution so that customers can be confident of a consistent restartable database on the remote site.
- An automated process to fail over the DCA to a remote site, allowing customers to run their operations on the remote DCA in case of a disaster.
- The capability on the remote site for other processing work such as test and development.
- An automated process to fail back the DCA to the primary site and allow the resumption of processing on the local site DCA.
- Online offloading of the Segment Server mirror instances to the VMAX SAN devices. This has the benefit of increasing the available capacity on the DCA.

**Note:** In a situation where a Segment Server on the primary DCA fails, remote database consistency cannot be ensured on the remote array.

# Introduction

## Purpose

The purpose of this white paper is to describe the validation of a solution for the EMC Greenplum DCA using an EMC Symmetrix VMAX array and SRDF for data replication between two sites in synchronous mode. This white paper details the functionality of the solution, showing how the SAN mirror configuration is set up and how failover of the Greenplum Massively Parallel Processing (MPP) architecture to a DR site is achieved.

## Scope

The scope of this white paper is to:

- Document the configuration of each of the components used in this solution.
- Identify some best practices that can be used by EMC Professional Services when implementing customer-specific DCA DR configurations.
- Provide flowcharts and details of the scripts used to automate the processes used for the DR functionality outlined in this white paper.

## Audience

The primary audience of this white paper is EMC Professional Services and EMC customers looking to understand how a DR solution for the Greenplum DCA can be achieved.

It is assumed that readers of this document are familiar with the DCA, Symmetrix SRDF, and TimeFinder/Snap.

## Terminology

This paper includes the following terminology.

**Table 1. Terminology**

Term	Definition
Business intelligence (BI)	The effective use of information assets to improve the profitability, productivity, or efficiency of a business. IT professionals use this term to refer to the business applications and tools that enable such information usage. The source of information is frequently the data warehouse.
Data warehousing (DW)	The process of organizing and managing information assets of an enterprise. IT professionals often refer to the physically stored data content in some databases managed by database management software as the data warehouse. They refer to applications that manipulate the data stored in such databases as DW applications.
Recovery point objective (RPO)	The maximum acceptable time period between the last available consistent image and a disaster or failure.
Recovery time objective (RTO)	The maximum acceptable time to bring a system or application back to operational state after a failure or disaster.

## Overview of components

### Introduction to the components

This section identifies and briefly describes the components deployed in the solution environment. The components used are:

- EMC Greenplum DCA
- EMC Symmetrix VMAX
- Auto-provisioning devices
- EMC Symmetrix Remote Data Facility (SRDF)
- EMC TimeFinder/Snap
- EMC PowerPath®
- EMC Solutions Enabler

### EMC Greenplum DCA

The DCA is a purpose-built, highly scalable, parallel data DW appliance that architecturally integrates database, computing, storage, and network resources into an enterprise-class, easy-to-implement system. The DCA offers the power of MPP architecture, delivers the fastest data loading capacity, and the best price to performance ratio without the complexity and constraints of proprietary hardware.

#### Key features of the DCA

- The DCA uses Greenplum Database software, which is based on MPP architecture. MPP harnesses the combined power of all available compute servers to ensure maximum performance.
- The base architecture of the DCA is designed with scalability and growth in mind. This enables organizations to easily extend their DW/BI capability in a modular way; linear gains in capacity and performance are achieved by expanding from a Greenplum Database Module (quarter-rack) up to twelve full racks with minimal downtime.
- Greenplum Database software supports incremental growth (scale out) of the data warehouse through its ability to automatically redistribute existing data across newly added computing resources.
- The DCA employs a high-speed Interconnect Bus that provides database-level communication between all servers in the DCA. It is designed to accommodate access for rapid backup and recovery and data load rates (also known as ingest).
- Excellent performance is provided by effective use of the combined power of servers, software, network, and storage.
- The DCA can be installed and available on-site within 24 hours of the customer receiving delivery.
- The DCA uses cutting-edge industry-standard commodity hardware rather than specialized or proprietary hardware.
- The DCA is offered in multiple-rack-appliance configurations to achieve the maximum flexibility and scalability for organizations faced with terabyte- to petabyte-scale data opportunities.

The DCA includes the following modules:

- Greenplum Database Standard Module
- Greenplum Database High Capacity Module
- Greenplum HD Module
- Greenplum Data Integration Accelerator (DIA) Module

Each module consists of four servers as described in Table 2. A DCA is configured and scaled according to application requirements:

- A DCA configuration starts with one rack containing one Greenplum DB Standard Module or one Greenplum DB High Capacity Module
- Up to a total of four modules can be configured per rack
- Additional racks can be configured, up to a total of twelve racks

The DCA is a self-contained data warehouse solution that integrates all the database software, servers, and switches that are required to perform enterprise-scale data analytics workloads. The DCA is delivered racked and ready for immediate data loading and query execution. It provides everything needed to run a complete Greenplum Database environment within a single rack.

Table 2 briefly describes the main components of the DCA.

**Table 2. Components of the DCA**

Item	Description
Greenplum Database	Greenplum Database is an MPP database server, based on PostgreSQL open-source technology. It is explicitly designed to support business intelligence (BI) applications and large, multi-terabyte data warehouses.
Greenplum Database system	An associated set of Segment Instances and a Master Instance running on an array, which can be composed of one or more hosts.
Master Servers	The servers that run the master database, responsible for the automatic parallelization of queries.
Segment Servers	The servers that run the Segment Instances and perform the real work of processing and analyzing the data.
Interconnect Bus	The Interconnect Bus provides high-speed communication between Master and Segment Servers. It consists of two switches to communicate requests from the Master to the Segments, between Segments, and to provide high-speed access to the Segment Servers for quick parallel loading of data across all Segment Servers.
Admin Switch	The Admin Switch provides the management interface between the servers and additional racks.

Item	Description
Greenplum DB Standard Module	A DCA database module running the Greenplum Database and consisting of four Segment Servers, each with 12 600 GB SAS disk drives.
Greenplum DB High Capacity Module	A DCA database module running the Greenplum database and consisting of four Segment Servers, each with twelve 2 TB SATA disk drives.
Greenplum HD Module	A DCA module consisting of four Segment Servers running the Greenplum Hadoop Community Edition (CE) software.
Greenplum DIA Module	A module designed for fast data integration and parallel, batch, and micro-batch data loading. A DIA consists of four servers utilizing certified partner software.

For more information about the DCA, refer to the following white paper: *EMC Greenplum Data Computing Appliance: Performance and Capacity for Data Warehousing and Business Intelligence – A Detailed Review*.

## EMC Symmetrix VMAX

Built on the strategy of simple, intelligent, modular storage, EMC Symmetrix VMAX incorporates a scalable Virtual Matrix™ interconnect that connects all shared resources across all VMAX Engines, enabling the storage array to grow seamlessly and cost-effectively from an entry-level configuration into the world's largest storage system. Symmetrix VMAX provides improved performance and scalability for demanding enterprise storage environments while maintaining support for EMC's broad portfolio of platform software offerings.

### Symmetrix Management Application Server

Symmetrix Management Application Server (SMAS) is a combined product installer of the Symmetrix Management Console (SMC) and Symmetrix Performance Analyzer (SPA).

### Symmetrix Management Console

Symmetrix Management Console (SMC) is a powerful and intuitive application that configures and manages multiple Symmetrix arrays. It presents the functionality of the Symmetrix Solutions Enabler SYMCLI (command line interface) in a browser-based GUI and simplifies storage administration tasks through the use of built-in wizards. It includes the functionality for configuring Fully Automated Storage Tiering for Virtual Pools (FAST VP). For more information, refer to the *EMC Symmetrix Management Console Product Guide*.

## Auto provisioning devices

Storage provisioning with the **symaccess** command enables you to create a group of devices, a group of director ports, a group of host initiators, and with one command, associate them in a masking view. Once a masking view exists, devices, ports, and initiators can be easily added or removed from their respective groups. This feature reduces the number of commands needed for masking devices, and allows for easy management of the masking view. The **symaccess** command is used to create and manage the groups and views.

### Storage groups

Storage group names can be up to 64 characters and are not case sensitive. Group names must be unique per group type, but different group types can share the same name.

### Port groups

Port groups may contain any number of valid front-end ports. A port can belong to more than one port group. Only Fibre Channel (FC) and GbE ports on front-end directors can be added to a port group. Port groups can have mixed port types.

Front-end ports must have the Access ControlLogix (ACLX) flag enabled to be added to a port group.

### Initiator groups

An initiator group is a container of one or more host initiators (FC or iSCSI). Each initiator group can contain up to 32 entries. An initiator group may also include the name of another initiator group to allow the groups to be cascaded to a depth of one. A host bus adapter (HBA) may only belong to one group, but may have masking views for both an upper and lower group if cascaded.

You can create an initiator group using the HBA's worldwide name (WWN), iSCSI, a file containing the WWN or iSCSI names, or another initiator group name.

### Masking view

A masking view is a container of a storage group, a port group, and an initiator group. When you create a masking view, the devices in the storage group become visible to the host. The devices are masked and mapped automatically. Volume dynamic addressing is enabled by default. The Symmetrix array assigns the next available LUN address on the front-end port when the masking view is created. The LUN assigned on the front-end port does not necessarily match the masking LUN. The groups being used must already exist and contain some entries (the initiator group can be empty) so that a complete view can actually be created.

## EMC Symmetrix Remote Data Facility

EMC SRDF remote replication software is a field-proven, widely deployed, array-based disaster restart solution with tens of thousands of licenses shipped to the most-demanding customer environments. Using the industry-leading high-end Symmetrix system, SRDF offers the choice and flexibility to meet any service-level requirement.

The SRDF family of software provides remote data replication, independent of the host and operating system, application, and database. SRDF helps companies manage planned and unplanned outages, enabling 24x7x365 data availability. It allows businesses to focus on maximizing revenue-generation and customer-support opportunities, improve productivity, and control or reduce costs for increased competitive advantage.

One member of the SRDF family of software is SRDF/Synchronous (SRDF/S), which maintains realtime synchronous remote data replication from one Symmetrix production site to one or more Symmetrix systems, located within campus, metropolitan, or regional distances, and provides a recovery point objective (RPO) of zero data loss. SRDF/S supports maximum distance of 125 miles or 200 kilometers.

## Setting up SRDF consistency groups on VMAX using ECA

An RDF consistency group (SRDF/CG) is a composite group of Symmetrix SRDF devices (RDF1, RDF2, or RDF21) enabled for remote database consistency. The devices in the consistency group are configured to act in unison to maintain the integrity of a database when distributed across multiple Symmetrix arrays or across multiple devices within an array.

RDF consistency protection software preserves the dependent write consistency of devices within the group by monitoring data propagation from source devices to their corresponding target devices. If a source R1 device in the consistency group cannot propagate data to its corresponding R2 device, the RDF consistency software suspends data propagation from all the R1 devices in the group. This enables you to quickly recover from certain types of failures or physical disasters by retaining a consistent, DBMS-restartable copy of your database. RDF consistency group protection is available for both SRDF/S and SRDF/A.

RDF consistency protection for synchronous devices is provided using ECA. ECA provides consistency protection for synchronous mode devices by performing suspend operations across all SRDF/S devices in a consistency group or a named subset of devices in a composite group. ECA is supported by the SRDF daemon (**storrd**) that performs monitoring.

### RDF consistency group operations

If any source (R1) devices in an SRDF/S consistency group cannot propagate data to their corresponding target (R2) devices, the SRDF daemon suspends data propagation from all R1 devices in the consistency group, halting all data flow to the R2 targets. This ensures a consistent R2 data copy of the database exists at the point in time at which an interruption occurs. The SRDF daemon monitors data copy operations and coordinates the suspension of R1 to R2 data propagation in the event that consistency protection is triggered.

The SRDF/CG feature is used in SRDF/S solutions to guarantee that a dependent, write-consistent image of production data on the R1 devices is created across the SRDF links.

The TimeFinder/Consistency Group (TimeFinder/CG) feature guarantees that a consistent, point-in-time image of data written across multiple local devices (TimeFinder source devices) is created on another set of local devices (TimeFinder target devices). SRDF/CG and TimeFinder/CG both use the ECA infrastructure. By using TimeFinder/CG in an SRDF configuration, you can create dependent, write-consistent local and remote images of production data across multiple devices and Symmetrix systems.

## EMC TimeFinder/Snap

EMC TimeFinder/Snap functionality provides a space-efficient mechanism to create pointer-based replicas of production systems. Rather than creating completely independent copies of production volumes, which carry the expense of provisioning a full-sized set of target volumes, TimeFinder/Snap uses a pointer-based mechanism that only requires storage for changed data. This provides a complete point-in-time replica of the source device while reducing the amount of additional storage required. The storage pool for the changed data is called the *save pool*. Multiple save pools can be created and snapshot devices share storage in the save pool with which they are associated.

The TimeFinder/Snap command (**symsnap**) creates virtual device copy sessions between a source device and multiple virtual target devices. These virtual devices (VDEVs) only store pointers to changed data blocks from the source devices, rather than a full copy of the data. TimeFinder/Snap software allows you to make multiple pointer-based, space-saving copies of data simultaneously on multiple target devices from a single source device. The resulting point-in-time copy data is available to the target host for instant access.

Up to 16 virtual point-in-time copies of a single source device can be created, which are known as basic virtual sessions. Multi-virtual sessions allow up to 128 virtual point-in-time copies.

Since a virtual session is associated with a point-in-time copy and a particular virtual device, keep a virtual session active as long as you need its snapshot copy. When a virtual session is terminated, the associated point-in-time copy is removed because Enginuity automatically releases the storage space in the associated save device pool.

### VDEV

A VDEV is a Symmetrix host-addressable cache device used in TimeFinder/Snap operations to keep pointers to point-in-time copies of the source device. VDEVs are space efficient because they only contain address pointers to the actual data tracks stored on the source device, or in a pool of save devices. VDEVs can be metadevices once the source is also a metadvice.

During a TimeFinder/Snap session, a VDEV pointer indicates either the source track or the save device track, depending on whether or not the track has been copied to the save device.

### Save devices

A save device is a Symmetrix device that is not accessible to the host and can only be accessed through virtual devices that point to it. Save devices provide pooled physical storage and are configured with any Symmetrix supported RAID scheme. Save devices cannot be metadevices. They store either source data copied to the save pool during the TimeFinder/Snap session or updates from the host mapped to the VDEV.

Since snapshot operations are designed to create point-in-time copies of the source device when only a fraction of the source device changes over time, the save device pool storage capacity can be much smaller than the capacity of the source device.

## EMC PowerPath

EMC PowerPath is server-resident software that enhances performance and application availability. It works with the storage system to intelligently manage I/O paths, and supports multiple paths to a logical device. PowerPath provides automated failover and recovery in the event of a hardware failure by detecting a path failure, and redirecting the I/O to another path, and subsequently putting the faulty path back into service once it has been repaired. PowerPath also does nondisruptive, transparent load-based testing to ensure that a data path is capable of carrying the workload presented to it.

## EMC Solutions Enabler

The EMC Solutions Enabler kit contains the base management software that, with SYMCLI commands and APIs, provides a host to configure Symmetrix and to control Symmetrix operations for SRDF and TimeFinder.

SYMCLI resides on the host system to monitor and control operations on Symmetrix storage arrays. SYMCLI commands are invoked from the host operating system through the command line or through scripts. SYMCLI commands invoke low-level channel communications to specialized *gatekeeper* devices on the Symmetrix.

SYMCLI is required to control TimeFinder and SRDF operations from the DCA. Because of this, Solutions Enabler should be loaded on the database Master Servers in the DCA.

## Solution overview

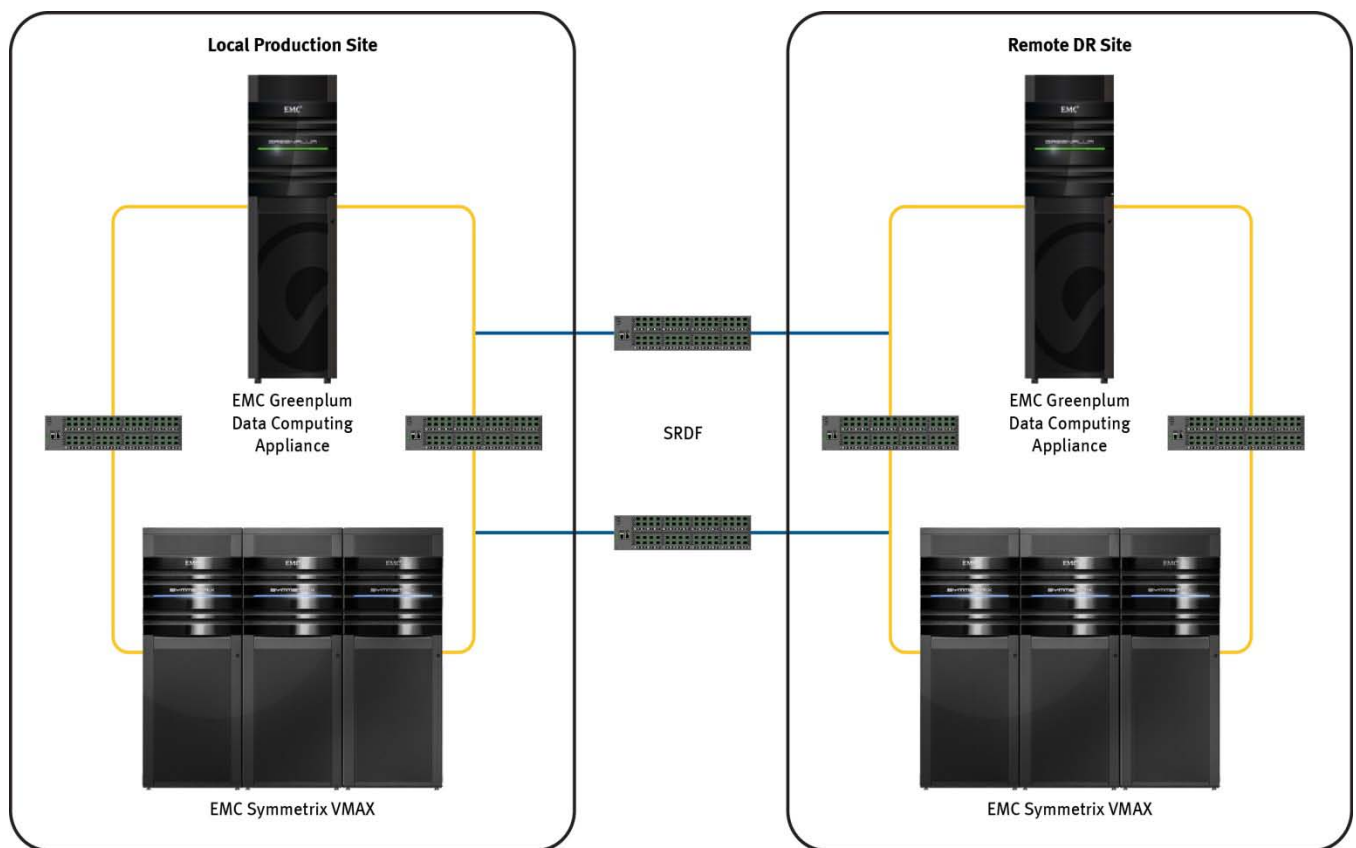
### Overview

This section illustrates the architectural layout of the DCA with Symmetrix VMAX SAN mirror solution. It also provides details of the hardware and software resources that were used in this solution:

- Solution architecture
- Hardware resources
- Software resources

### Solution architecture

Figure 1 illustrates the architectural layout of this solution.



GEN-001766

Figure 1. Solution architecture

## Hardware resources

The hardware used to validate the solution is listed in Table 3.

**Table 3. Hardware components**

Equipment	Quantity	Configuration
EMC Greenplum DCA	2	Full-rack DCA with Greenplum Database modules
EMC Symmetrix VMAX	2	4-Engine

**Software resources** The software used to validate the solution is listed in Table 4.

**Table 4. Software components**

Software	Version
EMC Greenplum Database	4.1.1.1
EMC VMAX Enginuity	5875.198
EMC Solutions Enabler	7.3
EMC PowerPath	5.6
EMC SRDF	7.3
EMC Symmetrix Management Console	7.3

## Connecting the Symmetrix VMAX to a Greenplum DCA

**SAN configuration** Figure 2 illustrates how one VMAX engine is connected to the DCA through the SAN.

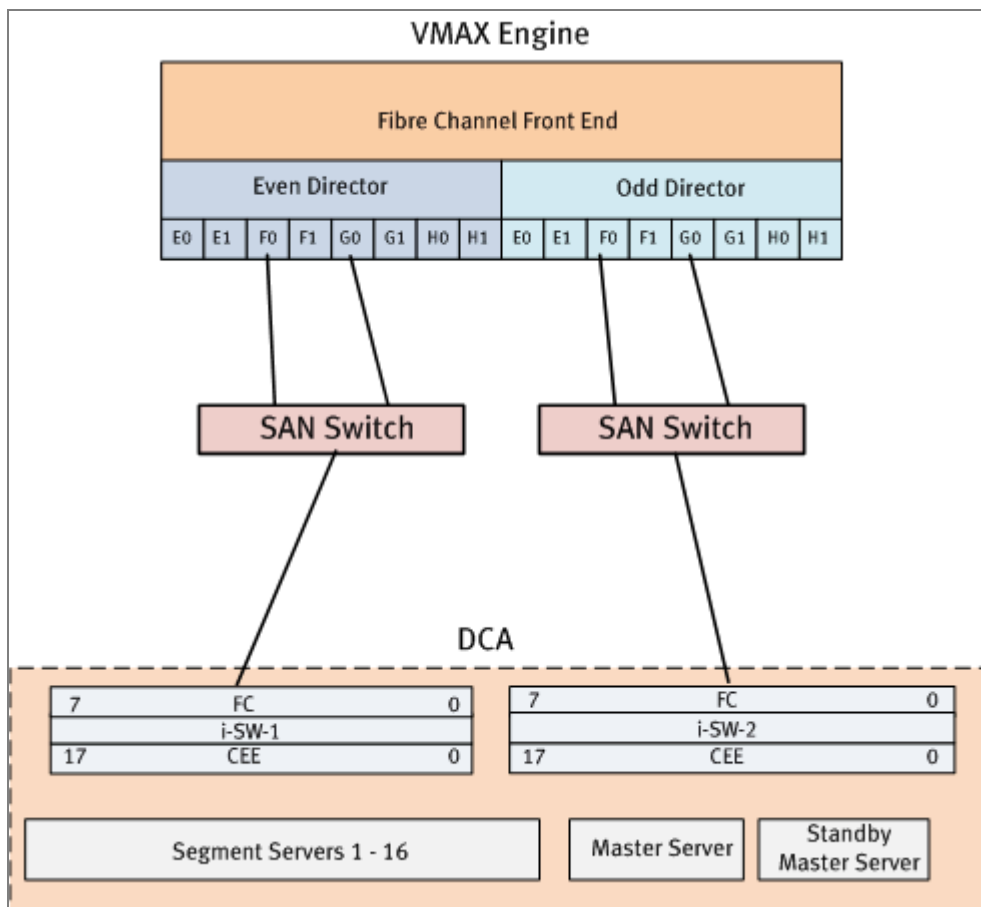


Figure 2. VMAX to DCA connectivity

### Hardware components

The hardware resources for local SAN configuration in this solution are detailed in Table 5.

Table 5. SAN configuration—hardware resources

Equipment	Quantity	Specification
DCA Interconnect Bus*	2	10 GbE and 8 Gb FC
SAN switch	6	8 Gb FC

\*The existing Greenplum DCA Interconnect Bus is designed to accommodate the SAN connectivity requirements.

### VMAX to DCA—SAN configuration

Figure 2 shows the connectivity from one VMAX engine to the DCA through two independent SAN switches. Internal to the DCA are two interconnect switches, **i-sw-1** and **i-sw-2**, through which both the Segment and Master Servers connect to the

converged enhanced Ethernet (CEE) ports 0 to 17 on each. The FC ports 0 to 7 on both interconnect switches connect to the SAN switches. Each FC director had the E, F, and G ports available for connecting to the SAN. The customer workload and required bandwidth determine how many VMAX ports to configure and connect.

Prior to setting up and configuring a VMAX array, the customer's workload and bandwidth requirements must be fully analyzed. Consult local EMC performance specialists about designing and implementing this solution.

This solution used a total of 16 FC ports from each VMAX to the SAN switches on each site. The H ports were reconfigured on both arrays to support the SRDF traffic required between the arrays. The RF ports were connected to two different external FC switches that were solely used to replicate from one site to the other site. It is also possible to connect the DCA directly to the VMAX array.

Within the DCA, 16 FC ports are available for connection to a remote SAN or a direct connection. These 16 FC ports span across two different internal interconnect switches.

The DCA interconnect switch configuration was modified to support the SAN mirror solution. The standard Ethernet configuration was changed to a converged, enhanced Ethernet configuration. This was necessary to support Fibre Channel over Ethernet (FCoE) between the servers and the SAN. To carry out these functions, the following steps were performed:

1. Added a VLAN classifier rule to dynamically classify the Ethernet packets on an untagged interface into the VLANs.
2. Added rules to the VLAN classifier groups.
3. Created a CEE map and configured the bandwidth for each group.
4. Enabled FCoE on the VLAN interface.
5. Configured the interfaces to converged mode.
6. Activated the VLAN classified group.
7. Set FCoE priorities.
8. Applied the CEE provision map.

Once the configuration was updated, the Segment Servers and the Master Servers were then zoned to the FAs on the VMAX.

## VMAX configuration

Prior to designing or configuring a VMAX array to support the SAN mirror solution, it is extremely important that the customer's workload and requirements are fully analyzed. EMC recommends using EMC Symmetrix performance specialists to ensure that expectations are met and the best configuration for the customer's needs is achieved.

To fully utilize the storage capacity of the internal disks of a DCA in a SAN mirror setup, the LUN size was created to be the same size as the data partitions on the DCA Segment Servers. Table 6 shows the amount of storage required by the full-rack DCA in this solution.

**Table 6. Full-rack DCA storage requirements**

	DCA	VMAX
Each Segment Server	2 x 2.7 TB	2 x 2.7 TB
Master Server	1 x 2.1 TB	1 x 2.1 TB
Standby Master Server	1 x 2.1 TB	1 x 2.1 TB

All LUNs presented from the VMAX array for the database operations were created on 450 GB 15k rpm FC disks. A total of 280 physical disks were used and these were all configured as RAID 5 (7+1). One VMAX disk group was used for all these disks to spread the workload evenly across all components of the array.

Metadevices had to be used due to the size of the LUNs required. In this setup, striped metadevices were used instead of concatenated metadevices. Striped metadevices are a preferred configuration for sequential writes. The specific type of metadevice to use is determined by the customer workload.

Each of the created metadevices consisted of 16 members. Each member was created as a 168 GB split on the physical drive.

**Note:** Ensure that the customer workload profile is taken into account prior to designing or implementing a SAN mirror configuration on a VMAX.

To support TimeFinder/Snap operations on the remote site, there was a requirement to create a save pool. Sixty-four 1 TB SATA disks were configured as RAID 6 devices. These were all used to create one save pool that stored any changed tracks during a snapshot operation. The size and configuration of the pool is determined by the customer workload profile, as well as the frequency of the snapshots taken on the array.

When designing a save pool for a customer, it is important to follow best practices and to consult with a local EMC Symmetrix specialist.

#### DCA with SAN mirrors

In normal DCA configurations, the storage for both the primary and mirror database instances resides on the local storage in the appliance. During normal operations, queries are always serviced by the primary instance while writes are replicated to the mirrors for the purpose of resiliency.

Figure 3 shows both the primaries (P0, P1, P2, P3, P4, and P5) and mirrors (M11, M16, M21, M8, M14, and M18) on local storage on a single DCA Segment Server with local mirroring.

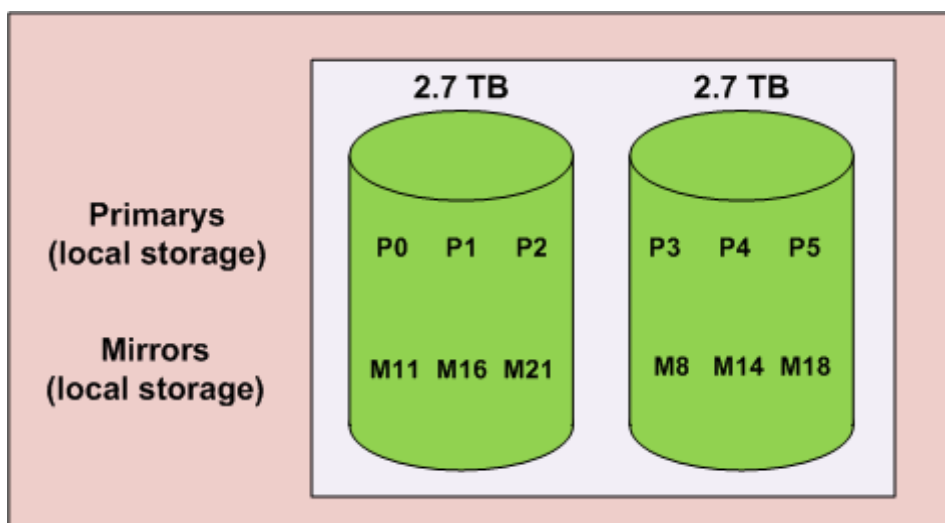


Figure 3. Primaries and mirrors on local storage for a Segment Server

Figure 4 shows the primaries on local storage and the mirrors on SAN storage on a DCA Segment Server configured for the SAN mirror.

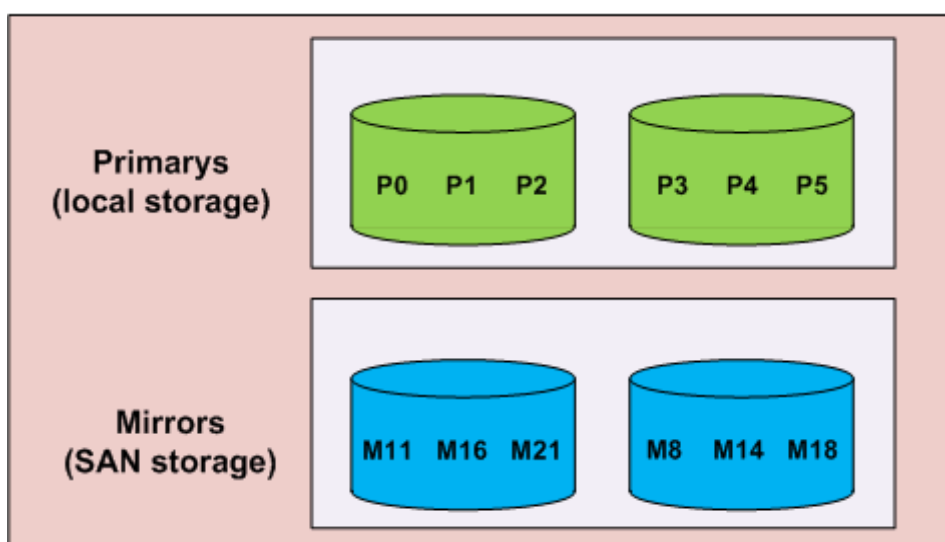


Figure 4. Primaries on local storage and mirrors on SAN storage for a Segment Server

In a DCA configured for the SAN mirror, all data for the primary database instances remains on the DCA. Therefore, all database queries are serviced locally without the need to utilize the SAN. During normal write operations, such as an ingest of data, the DCA services write operations to its primaries on internal storage, and also synchronizes the data to its mirrors on the SAN devices.

Figure 5 provides a high-level image of a DCA with a Master Server, standby Master Server, and four Segment Servers. Segment Server 1 primaries (Primary0 and Primary1) on local storage have associated mirrors (Mirror0 and Mirror1) on the SAN on Segment Servers 2 and 3. For illustration purposes, only two primaries and two mirrors for each Segment Server are shown. Typically, a Segment Server has six primaries and six mirrors.

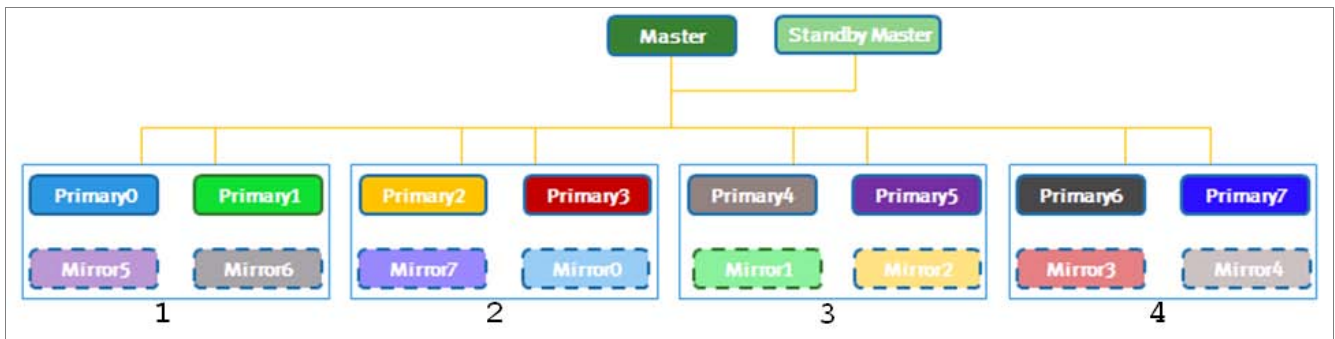


Figure 5. Segment Servers with primaries and associated mirrors

In the event of a Segment Server failure, the primary instances on the failed Segment Server become unavailable. When this occurs, the active Master Server detects this and the associated mirror instances of the failed primary instances are promoted to the role of primaries. These mirror instances that have now been promoted to primaries have their storage serviced by the VMAX. When the failed Segment Server is restored, the instances are recovered and the data is synchronized from the corresponding primary. Figure 6 provides an example of this, where Segment Server 1 fails and its mirrors on Segment Server 2 and 3 are promoted to primaries and continue servicing the workload.

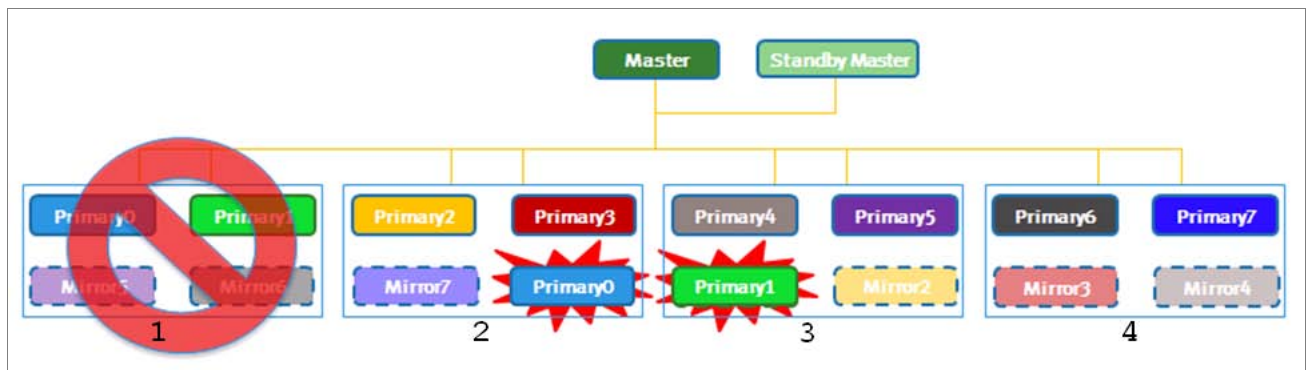


Figure 6. Segment Server failure

#### Segment failure on local Site A

In the case of a Segment Server failure on Site A, the six primaries that resided on the failed Segment Server fail over to their six mirrors that reside on the SAN devices spread across three other Segment Servers in the same HA failure group. The database is still fully operational.

Due to losing a Segment Server on the local site, the primaries for the three other Segment Servers have now lost their mirrors. As only the mirrors are replicated to the remote site, the database is now inconsistent on the remote site only and cannot be restarted.

The rotating snapshot continues to run on Site B and the validated snapshot remains in place. The rotating snapshots keep operating on the other two sets of snapshot devices that are available, but continue to fail the database consistency check. This is expected behavior as some of the mirror devices are unavailable due to the Segment Server that has failed.

Once the issue with the failed Segment Server on the primary site is resolved, some synchronizing is required on Site A. Refer to the *Greenplum Database 4.1 Administrator Guide* for the specific commands to both start and monitor the recovery of a failed Segment Server.

While in the resynchronizing state, the remote database continues to report as inconsistent until the resynchronization process is completed. When the failed Segment Server is fully recovered and back to its normal operating state, the rotating snapshots successfully complete the consistency check and begin rotating around the three sets of snapshot devices, while again always keeping a validated snapshot in place.

### Allocation and mounting of SAN devices on the DCA

Using the Solutions Enabler SYMCLI command line interface, storage provisioning with **symaccess** can be used to create a group of devices, a group of director ports, a group of host initiators, and with one command, associate them in a masking view.

The **symaccess** command is used to create and manage the groups and views. A host-visible (RW) gatekeeper device must be created with the ACLX device attribute. In addition, the ACLX flag must be enabled on the Symmetrix port.

The steps for creating a masking view are:

1. Create a storage group (one or more devices).

In this example, devices 034D and 035D are added to a newly created storage group “sdw1”:

```
symaccess -sid 1836 create -name sdw1 -type storage devs
034D,035D
```

2. Create a port group (one or more director or port combinations).

The following command created a port group “vmaxsw1” and contained eight front-end ports on the VMAX:

```
symaccess -sid 1836 create -name vmaxsw1 -type port -dirport
5F:0,6F:0,7F:0,8F:0,9F:0,10F:0,11F:0,12F:0
```

3. Create an initiator group (one or more host WWNs or iSCSIs).

The following command created an initiator group called “sdw1” and contains its corresponding FCoE WWN found on the server, using the `bcu port -list` command:

```
symaccess -sid 1836 create -name sdw1 -type initiator -wwn
100000051e74804b
```

Once the initiator group has been created, it can be updated to include further WWNs of the FCoE cards contained in the servers as follows:

```
symaccess -sid 1836 -type initiator add -name sdw1 -wwn
100000051e74804c
```

4. Create a masking view containing the storage group, port group, and initiator group created previously. When a masking view is created, the devices are automatically masked and mapped to the appropriate front-end ports.

```
symaccess -sid 1836 create view -name sdw1 -pg vmaxsw1 -ig  
sdw1 -sg sdw1
```

The VMAX devices presented to each Segment Server need some preliminary work prior to being used by the DCA. Once the storage is accessible on each Segment Server, EMC PowerPath was used to manage and perform load balancing on the SAN devices. Although all devices had a number of SAN paths available, EMC PowerPath controls and manages this by using one PowerPath pseudo name. In the following example, **/dev/emcpowera** was used.

Using the Linux OS parted tool, a partition table and partition are created on the VMAX LUNs with an offset of 1,024 KB. Once the partitions are created, an XFS file system is created. The final step in configuring the DCA and SAN devices is to mount them, using the following specific parameters:

1. Create a GUID partition table on the PowerPath pseudo device:

```
parted /dev/emcpowera mklabel gpt
```

2. Create a primary partition for an XFS filesystem with an offset of 1,024 KB on the PowerPath pseudo device:

```
parted /dev/emcpowera mkpart xfs 1024KB 2755GB
```

3. Create an XFS filesystem on the newly created partition on the PowerPath pseudo device:

```
mkfs -t xfs -f /dev/emcpowera1
```

4. Create a directory to which the PowerPath pseudo device can mount to:

```
mkdir /data1/san_mirror
```

5. Mount the PowerPath pseudo device to /data1/san\_mirror:

```
mount -o noatime,inode64,allocsize=16m /dev/emcpowera1  
/data1/san_mirror
```

6. Set the read ahead:

```
blockdev --setra 16384 /dev/emcpowera1
```

On the R1 site, for devices to remain persistent and automatically mount following a reboot, the `/etc/fstab` needs to be updated to include the PowerPath pseudo name of the presented SAN devices on each individual server. This is not a requirement on the R2 site.

## Moving mirrors

Greenplum Database is initialized on a DCA using **dca\_setup** with the option of having the mirrors on either local storage or SAN storage. A Greenplum Database utility—**gpmovemirrors**—can move the mirrors to the SAN for a Greenplum Database on a DCA previously initialized with mirrors on local storage. This is an online process for the Segment Servers. Before using the **gpmovemirrors** utility, run **dca\_setup** to create a config file for **gpmovemirrors**. The config file specifies the host address, port, and system file space location of the current mirror and host address, port, replication port, and system file space location of the new mirror in the following format:

```
[<filesystem1_fsname>[:<filesystem2_fsname>:...]<old_address>:<port>:<system_filespace_location>[<new_address:port>:<replication_port>:<system_filespace_location>[:<fselocation>:...]]
```

For example:

```
sdw1-1:50000:/data1/mirror/gpseg11 sdw1-1:50000:51000:/data1/san_mirror/gpseg11
```

The host address, ports, and system filesystem location information can be found in the **gp\_segment\_configuration** and **pg\_filespace** tables. For more information, refer to the *Greenplum Database 4.1 Administrator Guide*.

**Note:** The allocation and mounting of SAN mirrors on the Segment Servers can be done online.

As the Master Server and standby Master Server also require offloading of their operations to the VMAX, the following procedure must be followed to complete this. This requires a short period of downtime as the database requires a restart during these operations.

1. Remove the standby Master Server (smdw):  

```
gpinitstandby -r
```
2. Mount the SAN device on the standby Master Server using the following syntax (all one line):  

```
mount -o noatime,inode64,allocsize=16m /dev/emcpower1 /data/master
```
3. Initialize the standby Master Server:  

```
gpinitstandby -s smdw
```
4. Activate the standby Master Server:  

```
gpactivatestandby -f -d /data/master/gpseg-1/
```
5. Delete the master data directory on the Master Server (mdw):  

```
rm -r /data/master/*
```
6. Mount the SAN device on the Master Server using the following syntax (all one line):  

```
mount -o noatime,inode64,allocsize=16m /dev/emcpower1 /data/master
```

7. Initialize mdw as a standby:  
`gpinitstandby -s mdw`
8. Activate the Master Server (mdw):  
`gpactivatestandby -f -d /data/master/gpseg-1/`
9. Initialize the standby Master Server (smdw):  
`gpinitstandby -s smdw`

## SAN mirror SRDF consistency group - SRDF/S

### SRDF setup

A total of eight SRDF ports were configured on each VMAX array to support the replication of data from the source to the remote site. All 34 VMAX LUNs were assigned with either an R1 or R2 attribute using Solutions Enabler. Modifying this device attribute requires an online configuration change on the array with no impact to the customer. The same number of devices must also be created on the remote array and must be the same size.

On the source site, using Solutions Enabler, a consistency group named **sanmirrdf** was created using the following command:

```
symcgs create sanmirrdf -rdf_consistency -type rdf1
```

Once the consistency group is created, the next step is to add devices to the group. If there are multiple SRDF groups created on the array, use the following command to determine which one is to be used for consistency:

```
symcfg list -rdfig all
```

To add all the devices from the RDF group identified in the previous command, use the following command:

```
symcgs -cg sanmirrdf -sid 55 addall dev -rdfig 1
```

This command adds all the devices contained in **rdfig 1** to the consistency group **sanmirrdf**.

Using a simple command, all consistency groups created can be viewed to see how many devices and what type of devices are contained in the group as follows:

```
symcgs list
```

```
      C O M P O S I T E      G R O U P S

Name          Number of      Number of
              Type  Valid Symms  RAGs  DGs  Devs  BCVs  VDEVs  TGTs
sanmirrdf    RDF1  Yes   1     1     0   34   0    0     0
```

Once the consistency group has been created, a full synchronization from the source to the remote array must be completed. This sets up the relationship between the local and remote devices. The initial full synchronization is performed using the following command:

```
symrdf -cg sanmirrdf establish
```

The database now running on the local array devices copies the data to the remote array. The number of SAN devices being used determines the amount of time taken to replicate the data. Customers also have to ensure that SRDF links have the required amount of bandwidth to sustain the SRDF traffic on the link.

Using a **symrdf -cg sanmirrdf** query command enables you to view the current status of the SRDF links and how much data has been, and still needs to be, replicated.

When the query returns an RDF Pair state as “Synchronized” the following command can be used to enable ECA consistency on the devices:

```
symcg -cg sanmirrdf enable
```

```
Execute a consistency 'Enable' operation for composite  
group 'sanmirrdf' (y/[n]) ?
```

SRDF replication between both sites is initialized, and data is replicated from the local to the remote array consistently.

For more information on SRDF and its operations, refer to *EMC Solutions Enabler Symmetrix SRDF 188 Family CLI Version 7.3 Product Guide*.

## SAN mirror rotating snapshots

### TimeFinder/Snap on remote VMAX

The scripts in this solution provide a DR solution for a DCA that uses VMAX for the mirror devices, by performing TimeFinder/Snap operations on these devices to validate the Greenplum Database on a remote VMAX and DCA. Without interrupting normal SRDF replication, the script is used to mount, check, and validate the Greenplum Database on the remote DCA every 10 to 15 minutes. The last validated snapshot is always kept in place until another successful run has been completed.

In normal operating conditions, the remote VMAX devices (R2s) are in a read-only state and cannot be mounted directly to the remote DCA for database consistency checking. The SRDF synchronous state would have to be changed to either “split” or “failed over” to check the database consistency on the R2 site. This can halt replication, leaving the business unprotected in case of a disaster on the local R1 site.

Using TimeFinder/Snap technology, customers can take a consistent point-in-time snapshot of the R2 devices and mount the VDEVs to the remote DCA. As the VDEVs are read- and write-enabled, they are available to be mounted directly to the DCA while the SRDF replication remains intact. This solution used three individual VDEV snapshots for a single R2 device, which allowed for three different point-in-time copies to be taken.

### Scripts

**Note:** Scripts were created for all the testing carried out in this solution. The content of the scripts is determined by specific customer needs. Workflow examples are provided in the section **Automating the solution**. For more information about configuring, optimizing, and installing these scripts, consult EMC Professional Services.

The mounting process and the Greenplum Database check were automated through the use of preconfigured scripts.

Solutions Enabler, along with PowerPath, was used to control the automatic taking of snapshots, as well as mounting of the devices on each Segment Server, Master Server, and standby Master Server.

Specific scripts were designed for each of the following scenarios:

- Managing three rotating snapshots
- Finding and mounting the correct devices in the correct order on each host
- Determining the current Master Server and checking the database
- Auto-mounting the valid snapshot
- Auto-mounting the R2 devices
- Restoring the validated snapshot to the R2 devices and mount R2 devices
- Failover production to the remote site
- Failback production to the local site

The environment consisted of a production DCA and VMAX, along with a remote DCA and VMAX. SRDF was used for replication from one site to the other as shown in Figure 7.

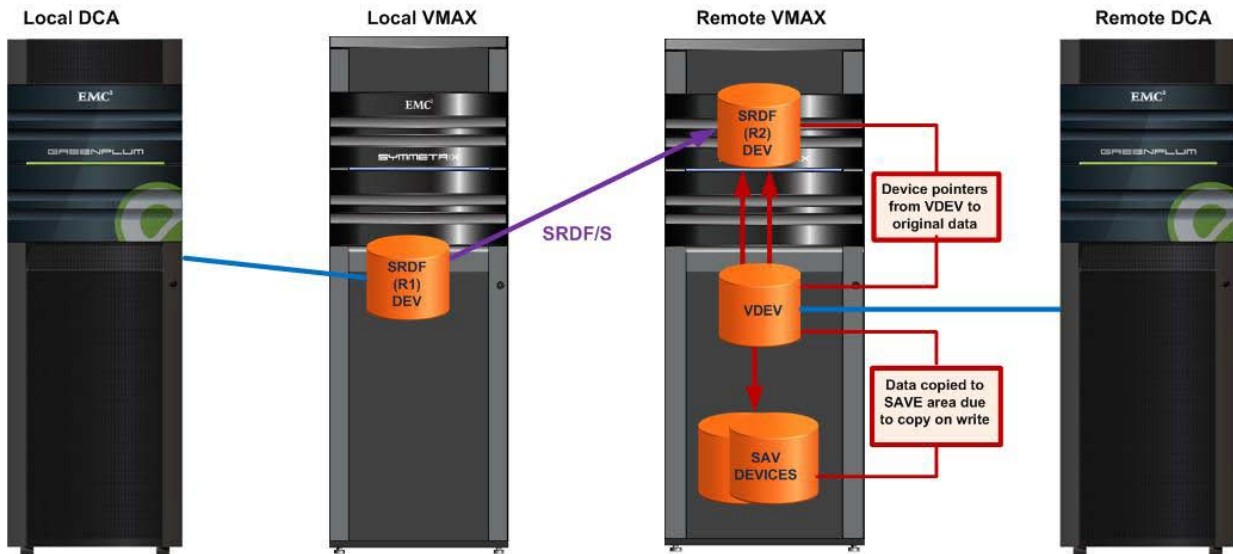


Figure 7. Solution environment

### Prerequisites

PowerPath is a prerequisite for the scripts to run and was installed on all of the Segment Servers, the Master Server, and the standby Master Server. All SAN devices presented to the servers were under the control of and managed by PowerPath.

Solutions Enabler was required on the Master Server and standby Master Server only. Solutions Enabler is used to communicate directly with the VMAX array.

To use the TimeFinder/Snap scripts, some files need to be created and copied to the remote DCA. On the remote DCA's standby Master Server, a directory specific to TimeFinder/Snap snapshots was created. This stored the scripts and the required files for each individual customer. The content of the directory and its files is determined by the DCA used in the solution—quarter, half, or full rack.

In this white paper, **san\_snaps** is the base directory where all required files and directories are created and stored.

A host file must also be created containing a list of the hostnames used in the solution. These must be valid hostnames and be accessible to the Master Server and standby Master Server through **ssh** using public keys. The VMAX also has storage groups using these names, which the scripts can use to automate the masking of devices. The entries in this file are also used to access the snap control files in the **snap\_files** directory.

Another directory—**snap\_files**—must be created within the directory referenced by **san\_snaps**. This is where the files used to set up the snapshot relationship between devices in the VMAX are stored. Each file in this directory should be named as follows:

- **hostname\_snapN**, where the hostname corresponds to the specific hostname of the servers used in the host file, and N corresponds to the snapshot number 1, 2, and 3.

For example, if working with Segment Server `sdw7` and R2 and snapshot devices for `snap2`, the file would be named `sdw7_snap2`.

In the following example, 0175 and 0185 are R2 devices on the remote VMAX array. For all snapshots on `sdw7_snapN`, these devices remain static. 0405 and 0415 are the snapshot devices created on the array and form a direct relationship with the R2s for `snap2`.

```
[root@mdw snap_files]# cat sdw7_snap1
0175    0405
0185    0415
```

Solutions Enabler **`symsnap`** and **`symaccess`** commands are used to carry out specific snapshot functions on these files.

Two other files must also be created within the base directory:

- `validated_snap`
- `snap_queue`

### **validated\_snap**

This file stores the specific snapshot number that was recorded as the last valid run after taking a point-in-time copy, as well as running database consistency checks. For more information, see the section **Database consistency check**.

### **snap\_queue**

This file contains 1, 2, and 3, which determines the running order of the appropriate snapshot based on the last run. This file is updated automatically following a successful or unsuccessful run of the rotating snapshots.

Within the base directory, the directory structure is as follows:

- `/root/san_snaps`
- `rotatingsnap.sh`
- `power-path-script.sh`
- `validated_snap`
  - `snap_queue`
  - `snap_files`
    - `sdw1`
    - `sdw2`
    - `sdwX`

The particular type or size of the DCA determines the number of individual **`sdw`** files that need to be created. In this specific solution, a full-rack DCA was used.

The `/etc/sudoers` file on the Master Server and the standby Master Server required editing by commenting out the line “Defaults requiretty”. This change was needed as the script runs as **`root`** due to the requirement to mount and unmount filesystems, but

## Setting up the remote site to support rotating snapshots

also runs certain commands as the **gpadmin user** for database-specific tasks. When these commands are run through **ssh**, there is no **tty** available so disabling the **requiretty** line circumvents this.

For an example of the workflow for the rotating snapshot, refer to Figure 13 in the section **Automating the solution**.

Once all the prerequisites are completed on both the remote VMAX and the DCA, some initial setup tasks must be carried out to set up the relationship between the source R2 devices and the three different snap VDEV devices. These tasks must be carried out before attempting to run the rotating snapshots.

If this is not a first time setup, an option is provided at the beginning of the script to keep the last validated snapshot in place. This ensures that there is a point-in-time copy available, regardless of the RTO.

The basic tasks involved in setting up are as follows:

- **File creation**  
In accordance with the prerequisite section for each Segment and Master Server, three individual files were created manually. All these files are now grouped together to produce three larger files, which are then used to create the relationships between the R2 devices and the snapshot devices. This is a timesaving feature—instead of issuing **symsnap** commands individually to all of the 18 servers, only one **symsnap** command is required.
- **Validated snapshot**  
If the customer has already run rotating snapshots and there is a validated snapshot in place, the option to keep this snapshot is provided, guaranteeing that a consistent copy remains in place.
- **Unmount the SAN partitions**  
All current SAN mirror partitions that are presented and accessible to the remote DCA are unmounted.
- **TimeFinder/Snap operations**  
Using each of the three snapshot files, the following commands are run to activate the pairing between the R2s and the VDEVs:  

```
symsnap create -sid 036 -f snapX -svp snap
```

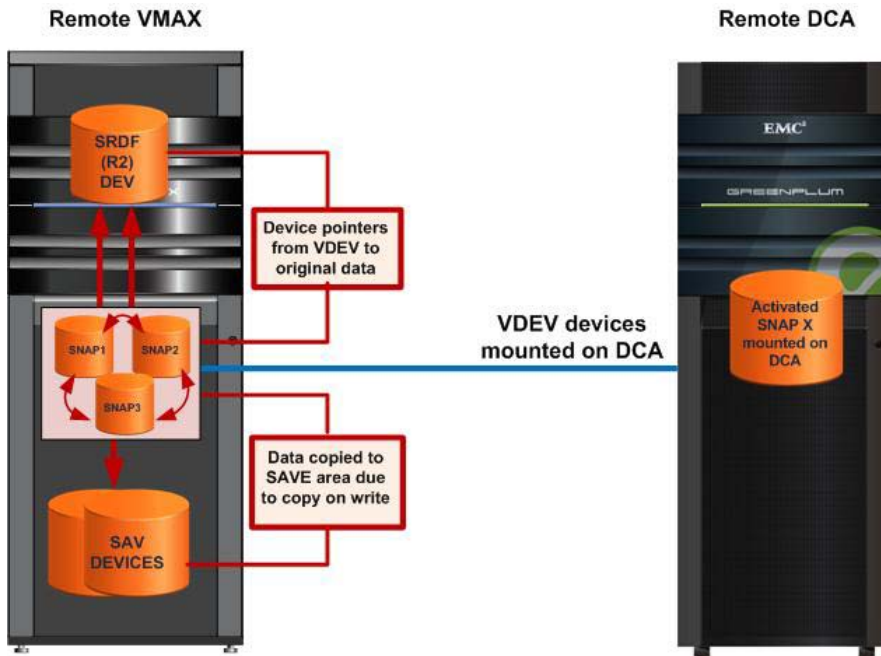
<b>Where:</b>	<b>Is:</b>
The VMAX serial number	036
The file being used	snapX
The name of the snapshot pool to be used to store updated information on the array	Snap

The section **Restoring validated snapshots to the R2 devices** provides details of the tasks carried out in restoring the last validated snap. If, for example, this restore operation is carried out on Snap1, then Snaps 2 and 3 need to be recreated using

**symsnap** command, once the restore completes. This allows you to return to taking rotating snapshots.

**Rotating snapshots script**

As noted previously, R2 devices are not readily presentable to the remote DCA for database consistency checking while synchronizing over the SRDF links. For this reason, TimeFinder/Snap snapshots were used to take consistent point-in-time copies of the R2 devices. This enabled SRDF synchronizing to continue to copy data from the R1 to R2 sites as illustrated in Figure 8.



**Figure 8. Rotating snapshots**

For Segment Server **sdw1**, the following three files were created manually. The R2 devices remained static across all three files and the target VDEVs changed for each file:

sdw1_snap1		sdw1_snap2		sdw1_snap3	
R2	snap1	R2	snap2	R2	snap3
00B5	034D	00B5	054D	00B5	074D
00C5	035D	00C5	055D	00C5	075D

Files of the same format were created for all 16 Segment Servers as well as the Master Server and standby Master Server. For a full DCA with 18 serves, a total of 54 files were required.

Once all the files were created and the prerequisites were completed, the rotating snap script ran as follows:

1. Determining the snapshot devices to be used

Checking both the contents of **validated\_snap** and **snap\_queue** determines the specific files to be used that contain the required VDEVs.

## 2. Masking of snapshot devices to each Segment Server and Master Server

The particular snapshot number in use determines which devices need to be visible from the Symmetrix VMAX to the DCA servers. As Symmetrix masking storage groups have already been created in the prerequisite section, it is only the storage group that needs to be changed each time a different set of devices is accessed by the servers.

Solutions Enabler **symaccess** commands are used to update the storage masking groups with the required VDEVs.

## 3. Scanning the SCSI bus

As snapshot devices are presented and removed quite regularly on all of the DCA servers, there is a requirement to rescan the internal SCSI buses to configure the newly added LUNS in Linux. This operation is carried out by checking the content of **/sys/class/scsi\_host**. All the Segment Servers in the DCA are connected to the SAN using host 0, host1, and host2. The Master and standby Master Servers use host 3 and host 4.

The following commands are used to update the LUNs presented:

```
echo "-- --" > /sys/class/scsi_host/host0/scan
echo "-- --" > /sys/class/scsi_host/host1/scan
echo "-- --" > /sys/class/scsi_host/host2/scan
```

## 4. PowerPath discovery

Using PowerPath-specific **powermt** commands to discover the new devices being presented and their associated PowerPath names.

Host-specific commands are executed in parallel across each host to reduce the amount of time for the snap script to run.

## 5. Creating the TimeFinder/Snap snapshot

Using TimeFinder **symsnap** commands, the association between the R2 and the VDEVs is created.

```
Symsnap -sid 36 -f snapx create -svp snap
```

## 6. Activating the snapshot

Using TimeFinder/Snap **symsnap** command, the particular session for each file is activated using the **-consistent** option. On entering this command the VDEVs are read- and write-enabled as is the point-in-time copy.

```
Symsnap -sid 36 -f snapx activate
```

## 7. Mounting the snapshot

The VDEVs are now mounted with their required options and are available on the DCA Segment Servers.

## 8. Greenplum Database validation and consistency check

This is described in more detail in the section **Database consistency check**.

## 9. Unmounting the VDEVs from the DCA

Whether the database check is successful or not, the VDEVs are unmounted at the end of the script but the point-in-time snapshot session is still present.

The host-specific steps for scanning devices, PowerPath discovery, and mounting devices can be seen in Figure 14 in the **Automating the solution** section.

After completing one successful execution of the rotating snap script, a validated snapshot is available that can be used to restore from, if required. The **validated\_snap** file is also updated to reflect the successful run. This file stores 1, 2, or 3, which corresponds to the successful snapshot.

### Time to complete one Snap operation

A 6 TB (~4.5 TB on disk) database is created on the local R1 site and synchronized to the remote R2 site. The rotating snapshot script is run and each pass through the script takes on average 15 minutes to complete. Each run of the snapshot script carries out a full database check to ensure consistency.

If, for some reason, the snapshot and database check is not consistent, then a **symsnap recreate** command is run against the devices, which allows them to be reused when next requested by the **snap\_queue** file. If a TimeFinder/Snap session is terminated, then the relationship between the R2 and VDEV is lost and any content relating to this session is cleared from the save pool. For this reason, the last validated snapshot remains in a “copy-on-write” state.

Auto-mounting of the valid snapshot is illustrated in Figure 9.

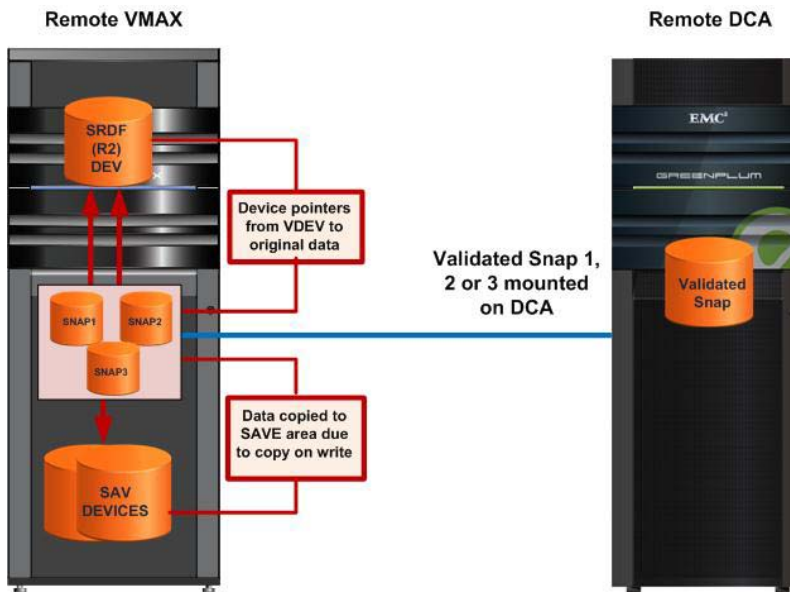


Figure 9. Auto-mounting of the valid snapshot

### Mount-validated snapshot script

Running this script mounts the last validated snapshot and checks that the database is still consistent. The script uses the content of the **validated\_snap** file to determine which snapshot was last run successfully. The script updates the VMAX storage group masking information with the valid VDEVs for each server. It also utilizes some components of the **rotating snap** script to update information on the SCSI bus and on PowerPath.

In a disaster recovery scenario, this script can be used to verify that the Greenplum Database can be restarted, prior to attempting to mount the R2 devices.

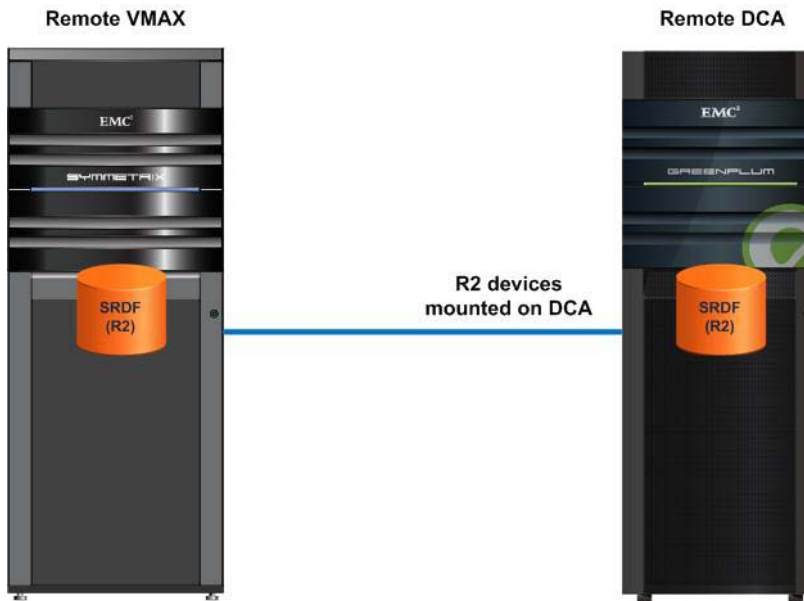
EMC recommends that you do not run production from the snapshot devices for a prolonged period of time because of the design of TimeFinder/Snap snapshots. The save pool can become full in the case of a heavy write workload. Also, database application response times are slower when running from snapshot devices.

Customers can also use this script to mount a valid copy of Greenplum Database to run tests on possible software upgrades, patches, or any other DCA or Greenplum-related tests.

# Failover

## Failover to the remote site

Figure 10 illustrates a failover operation to the remote site.



**Figure 10. Failover to R2**

In the case of a disaster or site outage on the R1 site, customers may be required to run their normal operations from the remote R2 site for a prolonged period of time. Using Solutions Enabler, SRDF customers are able to make the move from the local production site to the remote site. The failover scenario in this solution is scripted to enable customers to recover on the remote site as quickly as possible. There are different scenarios in which a customer might use the script developed for this solution. Some examples are described in Table 7:

**Table 7. Failover scenarios**

Scenario	Example
Planned failover (for testing or maintenance)	A planned failover is a controlled failover operation to test the robustness of the disaster restart solution. Production is temporarily moved to the secondary site during a planned failover.
Unplanned failover	Production processing is moved from the primary to the secondary site due to an unexpected failure of the production DCA at the primary site, the primary Symmetrix system, or both. The secondary site temporarily becomes the primary/production site.

In this solution, the SRDF links are manually set to “failed over” through the use of manual SRDF commands. Once in this state, the R2 devices automatically become read- and write-enabled and can be mounted to the remote DCA for validation and Greenplum Database checking. At the beginning of the failover operation, all SAN devices are unmounted from the remote DCA.

The **failover** script is executed as follows:

1. Terminate the **rotating snap** script

Since it is not recommended to take remote snapshots while in a DR situation, the automated **rotating snap** script is halted. Although the script is halted, a validated snapshot still remains in place.

The **rotating snap** script checks for a lock file on start-up. To stop the script from running, this lock file must be created by issuing the **touch /tmp/.san\_mirror\_lock** command.

2. Unmount the SAN devices

Automatically unmount all the SAN devices currently being used by the remote DCA.

3. Initiate the SRDF failover command

In a true disaster scenario, the R1 site would be down, so the failover command was used to switch operations to the R2 site. Using the **symrdf -cg sanmirrdf failover** command, the R2 device mode of operation automatically changes to read- and write-enabled.

4. Scanning the SCSI bus

As snapshot devices are presented and removed quite regularly on all of the DCA servers, there is a requirement to rescan the internal SCSI buses to configure the newly added LUNS in Linux. This operation is carried out by checking the content of `/sys/class/scsi_host`. All the Segment Servers in the DCA are connected to the SAN using host 0, host1, and host2. The Master and standby Master Servers use host 3 and host 4.

The following commands are used to update the LUNs presented:

```
echo "-- --" > /sys/class/scsi_host/host0/scan
echo "-- --" > /sys/class/scsi_host/host1/scan
echo "-- --" > /sys/class/scsi_host/host2/scan
```

5. PowerPath discovery

PowerPath-specific commands to discover the new devices being presented and their associated PowerPath names.

6. Mount the R2 devices

The specific R2 devices for each Segment Server, Master Server and standby Master Server are masked through the use of Solutions Enabler **symaccess** commands.

7. Greenplum Database checked and verified

For more information, refer to the section **Database consistency check**.

8. Synchronize the mirror and the primary Segment Servers.

Run **gprecoverseg -F**

9. Return the primary and mirror instances to their preferred roles as the roles have been changed as a result of executing the database consistency check

Run **gprecoverseg -r -a**

10. GPDB check fails

If the consistency check fails on the R2 devices, then a restore operation from the last validated snapshot to the R2 devices is required.

**Note:** Manual inputs are necessary to ensure the operations being carried out by the failure script are as required.

The last validated snapshot is a point-in-time copy which was mounted previously and left in the active state. The **validated\_snap** file contains the specific number of the snapshot devices that are held in a copy-on-write state.

Depending on the time between the last validated snapshot and the present time, the restore operation determines the recovery time objective (RTO) and RPO for the customer. It also determines how much data needs to be restored from the save pool to the R2 devices.

When the restore operation is completed, the database on the R2 devices can be accessed and verified for consistency. When verified, normal operations can run for an extended period of time on the R2 devices while the local site is being recovered.

#### **Time taken to complete an SRDF failover operation**

During testing, a 6 TB (~4.5 TB on disk) database is created and an automated failover operation is initiated from the remote R2 site DCA. In the case of a true disaster on the local R1 site, it must be expected that communications to the R1 DCA and VMAX will be unavailable. After the failover operation starts, it takes on average 48 minutes for the remote DCA to come online and service queries with full performance with:

- Primary instances on local disks
- Mirror instances on SAN R2 devices

Refer to the flow chart in Figure 17 for failover details of the exact steps that are used in the failover script.

## Restoring validated snapshots to the R2 devices

Figure 11 illustrates how validated snapshots are restored to the R2 devices.

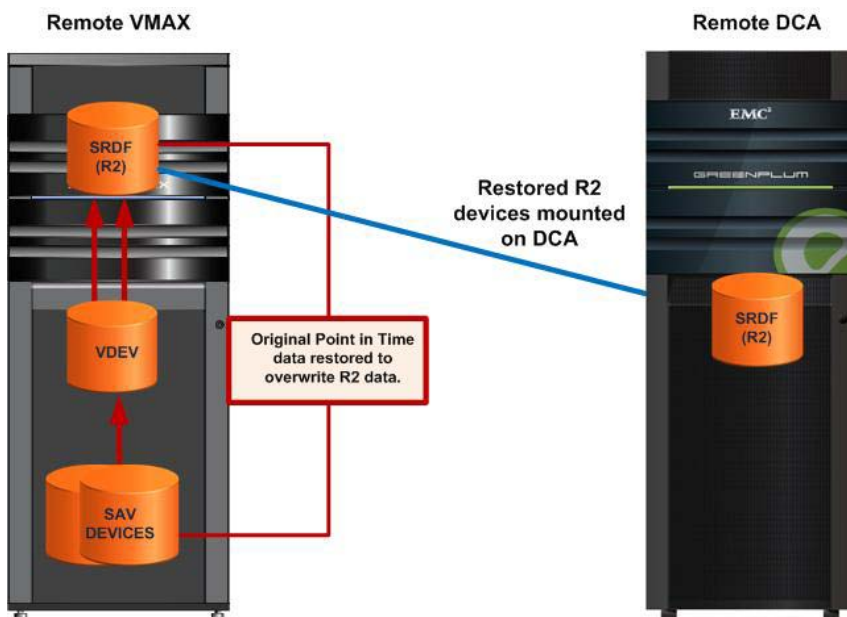


Figure 11. Restoring snapshots to R2 devices

### Using restoring snapshots

If there are issues in starting or recovering the database after mounting the R2 devices on the remote DCA, then the next step is to restore from the last validated snapshot to the R2 devices. The last validated snapshot is the point-in-time copy that was mounted previously and left in the active state. The **validated\_snap** file contains the specific number of the run, which was retained.

Depending on the time between the last validated snapshot and the present time, the restore operation determines the RTO and RPO for the customer. This also determines how much data needs to be copied from the save pool to the R2 devices. Once the restore operation is complete, customers can bring up the database on the R2 devices and operate for an extended period of time.

For an example of an SRDF failover workflow, refer to Figure 15 in the section **Automating the solution**.

### Time taken to complete an SNAP restore operation

A simulated test is run to test and time the restore operation from the last validated SNAP to the R2 devices. This takes three hours to complete. After completion, the DCA is running off its local disks as primaries and SAN Mirrors are on the SRDF R2 devices.

A full database consistency check is carried out on the validated snap before the restore operation and also on the R2 devices before the script is marked as completed.

## Database consistency check

When one of the three rotating snapshots or R2 devices is independently allocated and mounted on the remote DCA, a database consistency and validation check is required. The specific commands and required steps to do this are provided below.

All of these tasks have been included in an automated script for this specific environment. You can obtain the script by contacting EMC Professional Services.

It is not possible to check the validity or consistency of the database directly from the R2 devices while they are in their normal synchronous mode of operation. When involved in a copy operation, TimeFinder/Snap devices are available to be mounted on the DCA. Because of this, the rotating snapshots are used on the remote VMAX and mounted on the remote DCA. The database is checked for consistency. If the database check fails, then this snapshot is marked as failed and a new snapshot is created. The last validated snap remains in place and its corresponding number is stored in the **validated\_snap** file. This ensures that the customer has a valid point-in-time copy on the remote array that has successfully completed the database check.

To check the database, an automated script was created to carry out the following tasks:

1. Determine the current Master Server
2. Start the database in utility mode
3. Check the synchronization state
4. Check the consistency state
5. Reverse the roles of the primaries and mirrors so that the primaries are running off the SAN
6. Check the catalog

#### Determining the current Master Server

In normal operating mode on the local site, one of the Master Servers is the *active* server from the database's perspective. When the database is replicated to the remote site, it is necessary to determine the current Master Server on the remote DCA as it is possible for the standby Master Server to be the active Master Server at the time of running the database check.

The commands required for starting, stopping, and querying the database must be run against the active Master Server.

The current active Master Server is found by comparing the **dbid** and **standby\_dbid** entries in the text file **\$MASTER\_DATA\_DIRECTORY/gp\_dbid** on both the Master Server (**mdw**) and standby Master Servers (**smdw**).

If:	Then the current Master Server is:
The "original master" has dbid==1 and standby_dbid does not exist	The "original master"
The "original master standby" has dbid==1 and standby_dbid does not exist	The "original master standby"
The "original master" and the "original master standby" have standby_dbid, and if the "original master" has dbid==1	The "original master"
The "original master" and the "original master standby" have standby_dbid, and if the "original master standby" has dbid==1	The "original master"

### Starting the database in utility mode

The database is started in utility mode so its consistency and synchronization state can be checked. To start the database in utility mode, use the following command:

```
gpstart -m -a
```

Utility mode is used because it starts only the Master Server and only the data on the Master Server is required for carrying out these checks.

### Checking the synchronization and consistency states

The synchronization state refers to the synchronization status of a Segment Server with its mirror copy. Since snapshots of the R1 mirror devices only are being taken, the database needs to be fully synchronized for it to be usable on the R2 side.

The consistency state refers to the synchronization state of the standby and primary Master Servers. If zero rows are returned from the following query, then the database is deemed to be in a consistent state:

```
Select * from gp_segment_configuration where preferred_role is 'm' and (status='d' or (status='u' and mode='r'));
```

```
[gpadmin@mdw ~]$ psql template1 -c "select * from gp_segment_configuration where preferred_role='m' and (status='d' or (status='u' and mode='r'));"
dbid | content | role | preferred_role | mode | status | port |
-----+-----+-----+-----+-----+-----+-----+
hostname | address | replication_port | san_mounts
-----+-----+-----+-----+-----+-----+-----+
(0 rows)
```

If one row is returned from the following query, then the database is deemed to be in a synchronized state:

```
Select * from gp_master_mirroring where summary_state is 'Synchronized';
```

```
[gpadmin@mdw ~]$ psql template1 -c "select * from gp_master_mirroring where summary_state='Synchronized';"
summary_state | detail_state | log_time | error_message
-----+-----+-----+-----+-----+-----+-----+
Synchronized | | 2011-06-29 02:08:08-07 |
(1 row)
```

### Reversing the Segment Server roles

While still in utility mode, the roles of the Segment Servers need to be changed so that the primaries are running off the SAN devices instead of the mirrors. This change needs to be completed as only the mirror data replicated from the R1 site is available.

The Greenplum Database tracks the roles of the Segment Servers in the **gp\_segment\_configuration** table.

The columns to be altered are:

- Mode where:
  - s = synchronized

- c = change logging
- r = resynching
- Status where:
  - u = up
  - d = down
- Role where:
  - p = primary
  - m = mirror

The mode, status, and roles are altered for each of the Segment Servers with the following commands:

```
update gp_segment_configuration set mode='c',status='u',role='p'
where preferred_role='m' and content>-1;
update gp_segment_configuration set mode='s',status='d',role='m'
where preferred_role='p' and content>-1;
```

This enables the database to be restarted in full mode to carry out the catalog check using the **gpcheckcat** utility. To restart the database, the following commands are used:

```
gpstop -a -M immediate
gpstart -a
```

### Checking the catalog

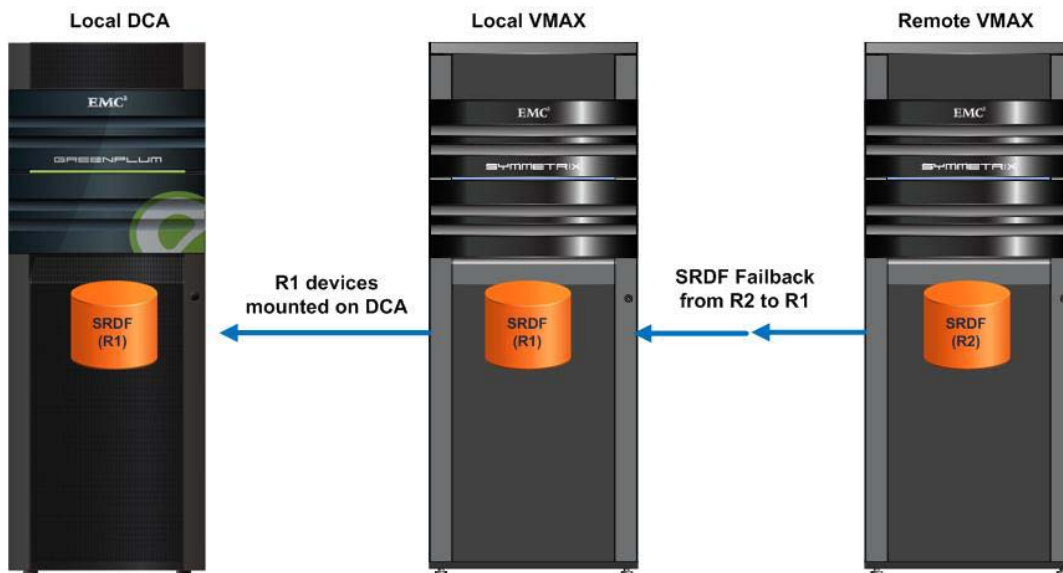
The **gpcheckcat** utility is called with the option **-A**, which causes it to check all the catalogs of all the databases. If this check does not return any issues, along with the consistency and synchronization checks, then the database is deemed to be okay.

For an example of the database check workflow, refer to Figure 17 in the section **Automating the solution**.

# Failback

## R2 to R1 with changes

Figure 12 illustrates failback from the remote site.



**Figure 12. Failback from R2 to R1**

A failback, or source (R1) device takeover, is performed when you are ready to resume normal SRDF operations by initiating read-write operations on the source (R1) devices, and stopping read-write operations on the target (R2) devices. The target (R2) devices become read-only to their local hosts while the source (R1) devices are read- and write-enabled to their local hosts.

If the R2 devices on the remote site have been used for production processing for a prolonged period of time, a large number of invalid tracks (TBs in size) may need to be copied to the R1 devices. In this situation, customers can use EMC host-based SRDF control software to resynchronize the primary and secondary devices while the secondary host continues production processing. When there is a relatively small number of invalid tracks to be copied to the R1 devices, the failback process can be initiated.

When a failback is initiated for each specified SRDF pair in a device group, the following occurs:

1. The target (R2) device is write-disabled to its local hosts.
2. Traffic is suspended on the SRDF links.
3. The invalid tracks on the source (R1) side are refreshed from the target R2 side.
4. The track tables are merged between the R1 and R2 sides.
5. Traffic is resumed on the SRDF links.
6. The source (R1) device is read-write enabled to its local hosts.

If required, customers can keep the original R1 copy of the database as the running database by using specific SRDF commands to keep the data and preserve the R1 copy.

Some operations for the failback must be run from both the local and remote DCA. The scenario described below is for true disaster recovery where the customer needs to keep the updated R2 information and overwrite the local R1 site.

For R2 site operations:

1. Stop database on remote DCA.
2. Unmount the SAN devices on the remote DCA.

For R1 site operations:

**Note:** SRDF consistency must be disabled before running the **symrdf failback** command.

1. Stop database on the R1 site.

In a disaster scenario, the database should already be stopped.

2. Unmount the R1 devices if they are mounted.

As the local DCA Segment Servers still have knowledge of the original R1 devices, forcing an unmount operation ensures that the new updated R1 devices can be mounted.

3. Run the **symrdf failback** command.

When the operation is complete, the SRDF pair query shows as synchronized, with any updates on the R1 site replicating to the remote R2 site.

4. Mount the R1 devices to the DCA servers.
5. Bring the database up in **Admin** mode only.
6. Switch roles.
7. Shutdown the database.
8. Bring up the database in full mode.
9. Check for database consistency.

10. Log in as **gpadmin**.

11. Start the database using **gpstart -a**.

12. Run **gprecoverseg -F** to synchronize the mirror and the primary Segment Servers.

13. Run **gprecoverseg -r -a** to return the primary and mirror instances to their preferred roles.

SRDF traffic now synchronizes from the R1 to the R2 devices again so the ECA consistency is re-enabled.

For an example of an SRDF failback workflow, refer to Figure 16 in the **Automating the solution** section.

### Time taken for an SRDF failback operation

After the failover operation is completed, the remote R2 site is the primary location for the customer operations. Customers can remain operating in this location for a sustained period of time. Once the local site is either repaired or restored, the updated database on the R2 site must be synchronized back to the R1 site. This is performed by using the SRDF failback operation. To carry out the failback operation, a failback-prep script must be run on the R2 site.

Preparation takes on average six minutes to complete.

After the failback-prep completes on the R2 site, run the failback script on the local R1 site. This initially mounts the R1 devices to the DCA with the role of primary. After the database is checked and validated, the data on the SAN devices (role of primaries) is synchronized to the local internal DCA disks (role of mirrors). After this sync completes, the roles of both the primaries and mirrors are reversed. Customer can now operate in their original setup with the SAN devices as mirrors, and replicating to the remote site.

On a 6 TB (~4.5 TB on disk) database, the average times recorded are:

- Failback-prep: 6 minutes
- Failback R1site: 30 minutes
- Total failback time: 36 minutes

## SAN Mirror performance results

### Overview

A benchmark that attempts to cover all the different types of workloads that customers would run on a DCA is used:

- **Long running queries:** Queries run on partitioned AO QuickLZ tables
- **Medium running queries:** TPC-H like queries run on schema where the fact tables (orders, line item) are partitioned AO QuickLZ tables, and the dimension tables are of storage type Heap
- **Short running queries:** Deep-sliced queries (ranging from 10 to 40 slices) on Heap storage tables
- **Data ingestion:** Loading data into AO QuickLZ TPC-H line item table with the source data fed by two gpfdists from one etl host
- **OLTP benchmark:** PGBENCH TPC-B transactions (inserts, selects, updates)
- **Trickle feed:** Two streams of single row inserts: one doing one single row insert per connection, and the other doing a batch of 100 inserts within a transaction block per connection

Database features exercised by the benchmark are:

- **Storage features:** Heap, AO Quicklz, Partitioned, Non-partitioned
- **Workload management features:** Resource Queues, Query Prioritization, Memory Quota

### DCA-only testing

When running in a normal full-rack DCA configuration, there is approximately 37 TB (uncompressed) available locally to service all read and write activity.

A 6 TB database is created on the DCA and a benchmark run against it that consists of concurrent mixed read/write workload. This particular workload completed in a total of 754 minutes, which is 12 hours 34 minutes.

Statistics for this test include:

- **DCA available storage:** 37 TB (uncompressed)
- **DB size:** 6 TB (~4.5 TB on disk)
- **Time to complete benchmark:** 12 hours 34 minutes

### DCA with VMAX SAN Mirror testing

After the DCA is connected to the SAN and configured for SAN Mirror, the DCA mirror instances are available on the VMAX SAN storage. In this configuration, no SRDF replication is in place purely for testing purposes.

This configuration increases the usable uncompressed capacity to 65 TB internally on the DCA. When the same workload runs against this configuration, it completes in 691 minutes, which is 11 hours 31 minutes.

Statistics for this test include:

- **DCA available storage:** 65 TB (uncompressed)
- **DB size:** 6 TB (~4.5 TB on disk)
- **Time to complete benchmark:** 11 hours 31 minutes

### DCA with VMAX SAN and SRDF testing

In this test, both local and remote sites are connected and replicating in SRDF/S mode with zero distance between the sites. The internal DCA uncompressed capacity available during this time is still at 65 TB. SRDF adds a very slight increase in the time taken to complete the workload, compared to the previous test run with the local DCA and VMAX only. The time recorded for this test is 747 minutes, which is 12 hours 27 minutes.

Statistics for this test include:

- **DCA available storage:** 65 TB (uncompressed)
- **DB size:** 6 TB (~4.5 TB on disk)
- **Time to complete benchmark:** 12 hours 27 minutes

### Comparison

Figure 13 illustrates the time taken to run benchmark for the different test configurations.

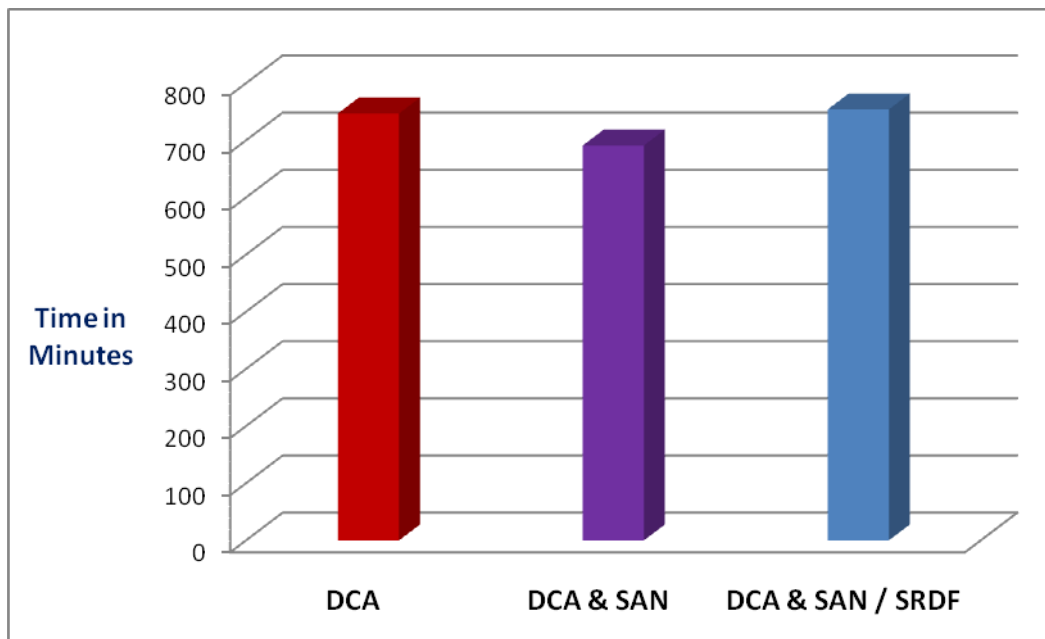


Figure 13. Time taken to run benchmark on 6 TB DB

Figure 14 illustrates the DCA capacity for the different test configurations.

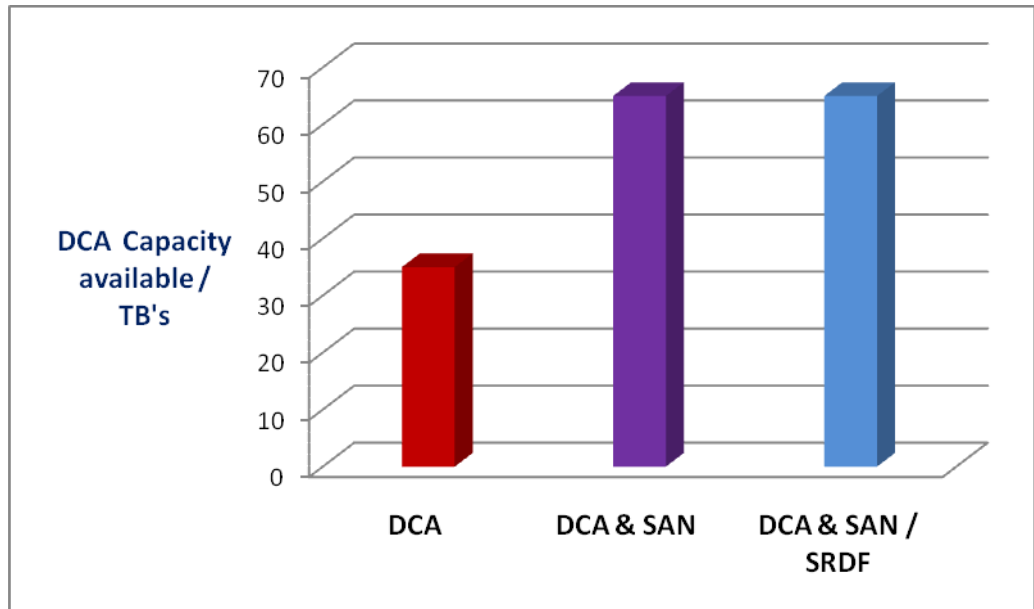


Figure 14. DCA capacity available for different configurations

#### Performance summary

From the tests run on the SAN Mirror setup, the main points of interest are:

- The DCA and SAN setup increased total capacity available on the DCA to service reads and writes to the Greenplum Database.
- The DCA and SAN setup displayed a performance benefit when carrying out mixed workload tests.
- DCA and SAN/SRDF showed no difference in the time taken to carry out the same mixed workload tests compared to the DCA-only setup.

## Automating the solution

### Using customized scripts

Specific customized scripts were created to automate all the scenarios described in this solution. These scripts are available through your local EMC Account teams. An EMC Professional Services engagement may be required to install and customize the scripts, depending on the customer's requirements.

The following flow charts describe the operations for each of the scripts created at a high level:

- Example of a rotating snapshot workflow
- Example of a host device discovery workflow
- Example of an SRDF failover workflow
- Example of an SRDF failback workflow
- Example of a database check workflow

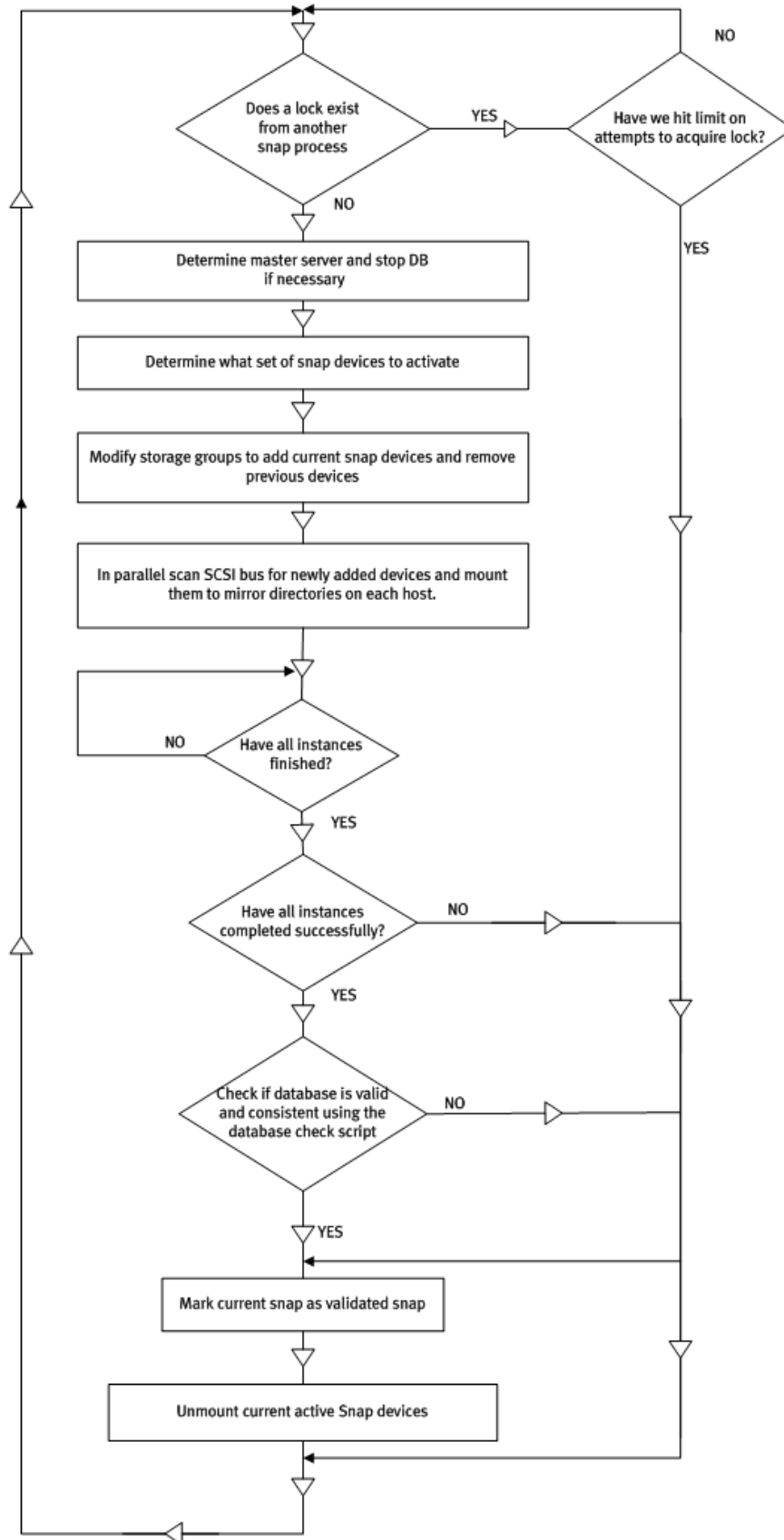


Figure 15. Example of a rotating snapshot workflow

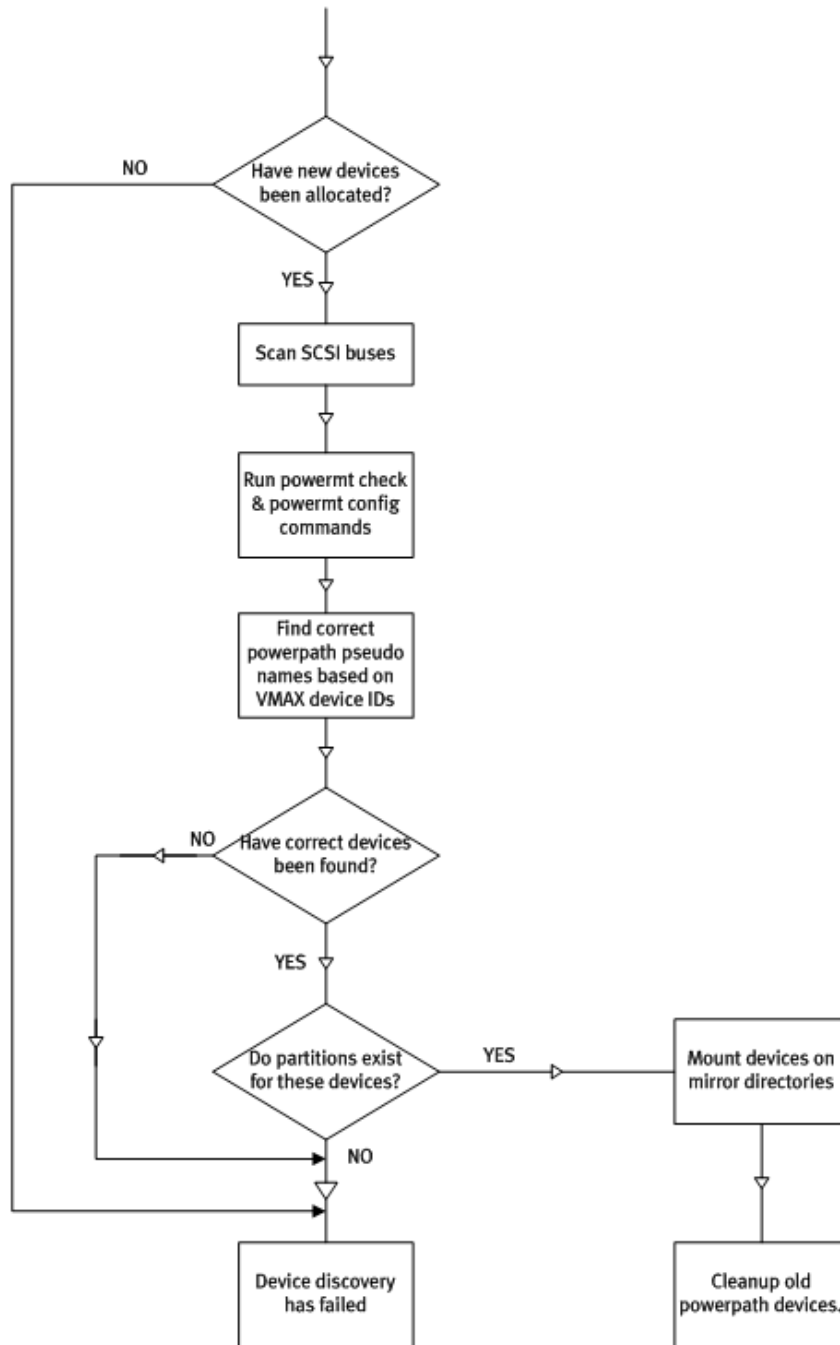


Figure 16. Example of a host device discovery workflow

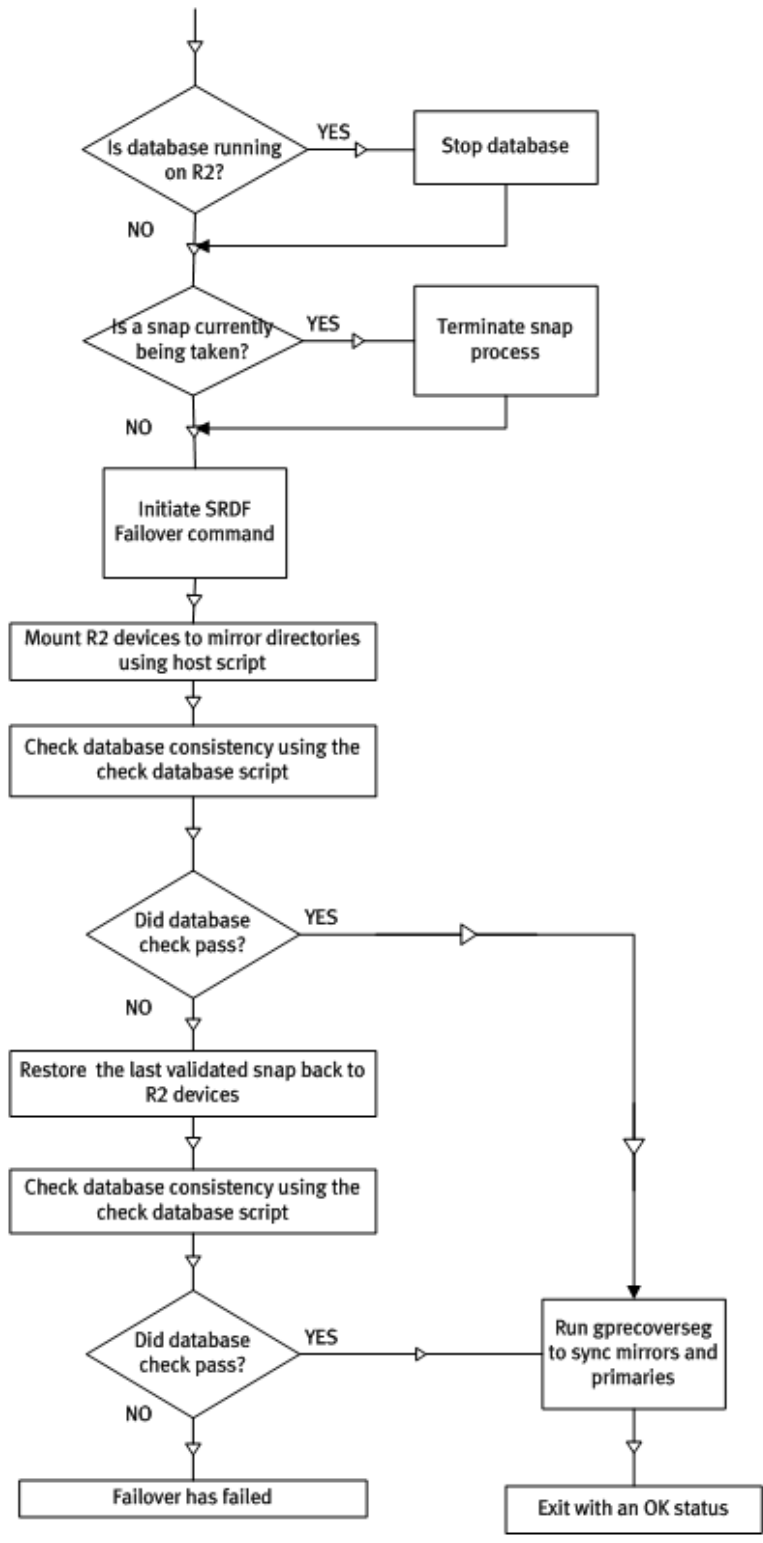


Figure 17. Example of an SRDF failover workflow

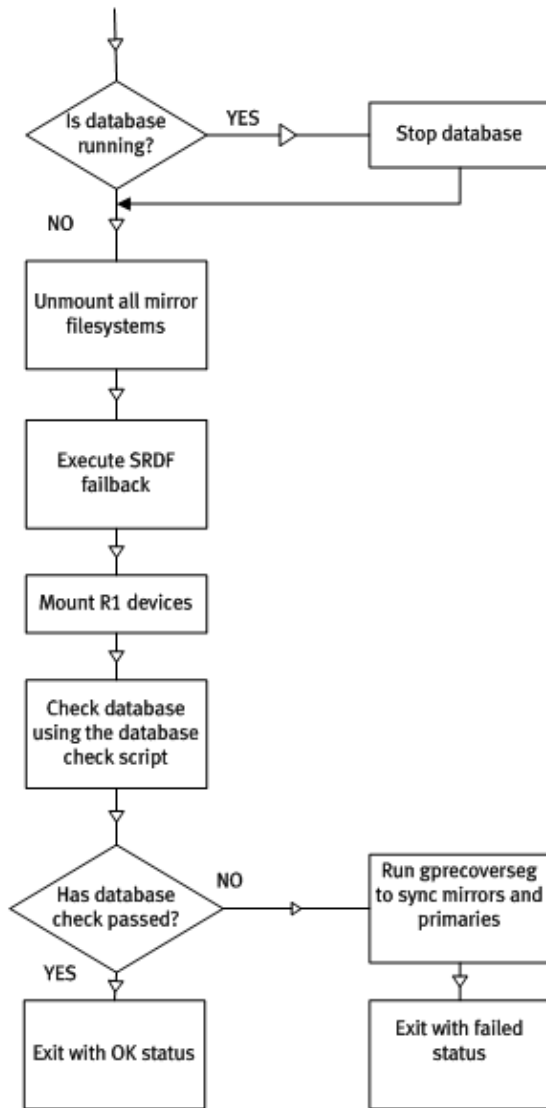


Figure 18. Example of an SRDF failback workflow

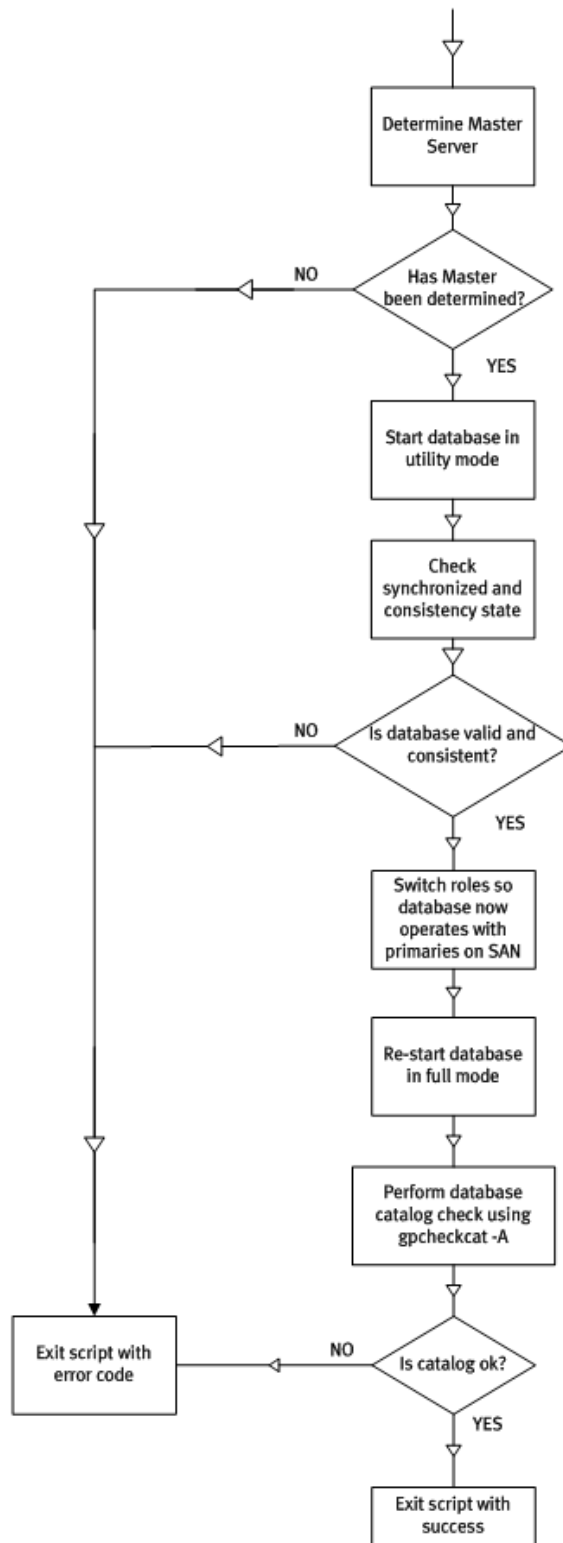


Figure 19. Example of a database check workflow

# Conclusion

## Summary

This solution highlights:

- The use of SRDF in synchronous mode, together with Symmetrix ECA to enable a consistent and recoverable database image on a remote site.
- The use of TimeFinder/Snap snapshots to ensure that continuous database checking and validation of the remote DCA database can take place.
- The use of a clear failover process to fail over DCA processing to a remote site, by placing the DCA mirror partitions on the Symmetrix VMAX.
- The recovery of validated remote snapshots to remote R2 devices.
- The capability on the remote site for other processing work such as test and development.
- The online offloading of the Segment Server mirror instances to the VMAX SAN devices. This has the benefit of increasing the available capacity on the DCA.

Where customers must leverage their existing storage infrastructure practices and methodologies, this solution provides a streamlined approach for the addition of a disaster recovery capability for their mission-critical business intelligence environment.

**Note:** Prior to implementing the SAN mirror solution, it is vital that local EMC performance specialists are consulted and involved.

## Findings

The solution presented in this white paper can help SAN customers to leverage VMAX SAN storage and SRDF/TimeFinder data services to yield a reliable remote disaster recovery capability for Greenplum DCA data analytics deployments.

Key tasks can be automated by utilizing scripts to:

- Create rolling snapshots on the remote site
- Validate the database consistency of the remote site snapshots
- Failover to the remote site from the primary site
- Failback to the primary site

Site A and Site B can operate independently once the SRDF links are split. This solution demonstrates that:

- Site A continues to operate normally with VMAX as mirrors.
- Site B can be used to run validation tests and upgrades without impacting Site A.
- Once testing is complete, the SRDF links can be resumed with updates sent from site A to site B.

## References

### White papers

For additional information, see the white papers listed below.

- *EMC Greenplum Data Computing Appliance: Performance and Capacity for Data Warehousing and Business Intelligence – A Detailed Review*

### Product documentation

For additional information, see the product documents listed below.

- *EMC Solutions Enabler Symmetrix SRDF 188 Family CLI Version 7.3 Product Guide*
- *EMC Symmetrix Management Console Product Guide*

### Other documentation

For additional information, see the documents listed below.

- *Greenplum Database 4.1 Administrator Guide*