hp storage

january 2003

invent

technical blueprint

# hp StorageWorks remote mirroring for Oracle

## table of contents

## executive summary

This paper discusses performance and best practices for using the disaster-tolerant capabilities of HP StorageWorks Data Replication Manager (DRM) MA/EMA with Oracle databases and HP StorageWorks Modular Array 8000 (MA 8000) and Enterprise Modular Array 12000 (EMA 12000) to attain the best reliability, performance, and ease of management.

DRM provides remote mirroring of large, heavily loaded databases at distances up to 100 km with virtually no performance degradation. Prior to this current work, through a series of 21 tests provided by the Oracle Storage Compatibility Program (OSCP), HP has demonstrated the compatibility of DRM remote mirroring technologies when used with high-performance Oracle databases.

This technical blueprint includes the following topics:

- overview of DRM and Oracle
- performance test setup, procedures, and results
- synchronous versus asynchronous replication
- mirroring only the redo logs versus an entire database
- mirroring large databases across multiple storage arrays

## business needs

In a changing business environment, data availability, maintenance, and recovery are crucial to sustaining critical business operations. Uptime requirements in service level agreements now specify “five-nines” (99.999%) of application availability. As a result, data centers are responsible not only for maintaining data integrity, but also for providing the ability to recover from disruptions and disasters quickly and seamlessly. DRM provides the ability to mirror data to remote/standby data centers, ensuring data is online and available.

This technical blueprint is intended to help storage and database administrators get the most out of remotely mirrored Oracle implementation.

## component review

This section provides a brief overview of DRM and the OSCP.

### hp StorageWorks data replication manager

DRM is an ideal solution for mirroring data online and in real-time to remote locations through a local or an extended Storage Area Network (SAN). Using DRM software, data replication is performed at the storage system level and in the background to any host activity.

DRM provides the following benefits:

- Performs real-time data replication locally, regionally, or globally to ensure data protection and business continuance
- Scales your SAN with unlimited distance capability with FC-IP and FC-to-ATM gateways
- Offers direct Fibre Channel distances up to 100 km
- Provides the highest levels of flexibility over long distances through synchronous and asynchronous data transfers
- Supports replication of the HP StorageWorks OpenView Storage Virtual Replicator storage pools
- Has Clone and Snapshot management with HP StorageWorks Business Copy upgrade
- Has multipath support across multiple operating systems with HP StorageWorks Secure Path
- Provides efficient data resynchronization from short outages through write history logging (MiniMerge Reconstruction)
- Supports HP Tru64 UNIX TruCluster, HP OpenVMS cluster, HP MC/Service guard Microsoft Windows 2000/Windows NT Stretched Cluster Services, Microsoft Server Cluster Software, VERITAS Clusters, NetWare Cluster Services, and IBM HACMP, taking high availability to the next level
- Supports Tru64 UNIX, OpenVMS, HP-UX, Windows 2000/Windows NT, Sun Solaris, Novell NetWare and IBM AIX

HP StorageWorks EMA12000 and MA 8000 RAID Arrays running DRM can replicate data up to 100 km through direct Fibre Channel links at full Fibre Channel speeds (100 MB/sec), connect their SAN islands through fiber optic Wave Division Multiplexing in a Metro public or private network, or go unlimited distances with Fibre Channel over IP networks and Fibre Channel-to-ATM gateways. DRM gives customers the widest choice of communication networks, bandwidth, distance, and availability options to best meet their enterprise-level network storage requirements.

### Oracle Storage Compatibility Program

The OSCP provides tests that measure remote mirroring technologies to ensure their compatibility with Oracle databases. The OSCP self-test suite is available to qualified vendors. As a member of the OSCP, HP implemented these tests using DRM.

The 21 tests in the current test suite verify both synchronous and asynchronous mirroring by simulating power and network failures. For complete details regarding the OSCP remote mirroring tests, refer to the Oracle Storage Compatibility Testing – Remote Mirroring white paper at

http://h18006.www1.hp.com/storage/softwhitepapers.html

HP successfully completed all test requirements stated in the Oracle remote mirroring test suite using DRM for both synchronous and asynchronous modes. The results were submitted to Oracle for verification and approved for entry into the program.

## solution-specific configuration

This section provides test suite hardware and software components and configuration.

The goal of this work was to determine the impact of remote mirroring using DRM on a large, heavily loaded database, and to develop recommended best practices for implementing Oracle 9*i* with DRM. To accomplish this we installed and configured a balanced set of servers, storage, and SAN infrastructure, and laid out the Oracle database on the equipment in a configuration that optimized both operational I/O and, for full recovery after an outage, remote mirror normalization.

### solution test setup

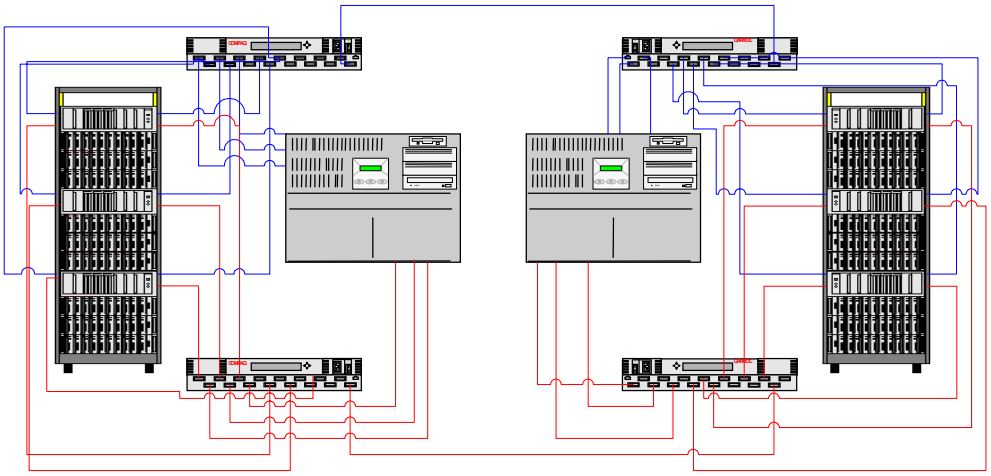

**figure 1. servers and storage**

## intersite links

Port 15 was configured as the intersite link for each of the SAN switches. For testing at 0 km, a shortwave Gigabit Interface Converter (GBIC-SW) was used in Port 15. For the distance of 100 km, a Very Long Distance GBIC was used.

The SAN switches were installed with Extended Fabric Licenses, which allow assignment of one port per switch to be a *long-distance port*. The long-distance port was configured in the switch with 60 Buffer-to-Buffer Credits to optimize performance for long-distance transmission.

For long-distance testing, Wavefront Technologies, Inc. NetSim spool boxes were configured to provide a maximum distance for each intersite link of 100 km.

**table 1. software used in the tested configuration**

| |
|---|
| ES45 |
| Tru64Unix 5.1A |
| Oracle 9*i* (9.0.1) |
| HSG80 Fibre Channel-based storage arrays |
| ACS 8.6P-4 |
| 16 Port SAN switches |
| Brocade firmware v2.6 |

### storage layout

Storage components were configured to optimize the HSG80 for both performance and normalizations. In general, 3 x 2 striped-mirror sets were used for each Logical Unit Number (LUN). The tablespace partitions and archive logs were placed on 3 x 2 x 36-GB stripe sets. The redo logs were placed on 3 x 2 x 18-GB 15k-rpm drives. The ACS initialize capacity switch was used to minimize the amount of normalization needed for the redo log disks and to improve write performance by placing the redo logs on the outer tracks of the physical disks. The storage layout was configured as follows for each of the three HSG80 pairs.

**table 2. storage layout**

| | scsi bus 1 | scsi bus 2 | scsi bus 3 | scsi bus 4 | scsi bus 5 | scsi bus 6 |
|---|---|---|---|---|---|---|
| shelf 0 | rcs10 on primary1, rcs20 on primary2, rcs30 on primary3 | | | | | |
| | d10 on primary1/standby1, d20 on primary2/standby2, d30 on primary3/standby3 | | | | | |
| | s0 | | | | | |
| | m0 | | m1 | | m2 | |
| | disk10000 | disk20000 | disk3000 | disk40000 | disk50000 | disk60000 |
| | 18GB/15k | 18GB/15k | 18GB/15k | 18GB/15k | 18GB/15k | 18GB/15k |
| shelf 1 | rcs11 on primary1, rcs21 on primary2, rcs31 on primary3 | | | | | |
| | d11 on primary1/standby1, d21 on primary2/standby2, d31 on primary3/standby3 | | | | | |
| | s1 | | | | | |
| | m3 | | m4 | | m5 | |
| | disk10100 | disk20100 | disk30100 | disk40100 | disk50100 | disk60100 |
| | 18GB/15k | 18GB/15k | 18GB/15k | 18GB/15k | 18GB/15k | 18GB/15k |
| shelf 2 | rcs12 on primary1, rcs22 on primary2, rcs32 on primary3 | | | | | |
| | d12 on primary1/standby1, d22 on primary2/standby2, d32 on primary3/standby3 | | | | | |
| | s2 | | | | | |
| | m6 | | m7 | | m8 | |
| | disk10200 | disk20200 | disk30200 | disk40200 | disk50200 | disk60200 |
| | 36GB/10k | 36GB/10k | 36GB/10k | 36GB/10k | 36GB/10k | 36GB/10k |
| shelf 3 | rcs13 on primary1, rcs23 on primary2, rcs33 on primary3 | | | | | |
| | d13 on primary1/standby1, d23 on primary2/standby2, d33 on primary3/standby3 | | | | | |
| | s3 | | | | | |
| | m9 | | m10 | | m11 | |
| | disk10300 | disk20300 | disk30300 | disk40300 | disk50300 | disk60300 |
| | 36GB/10k | 36GB/10k | 36GB/10k | 36GB/10k | 36GB/10k | 36GB/10k |
| shelf 4 | rcs14 on primary1, rcs24 on primary2, rcs34 on primary3 | | | | | |
| | d14 on primary1/standby1, d24 on primary2/standby2, d34 on primary3/standby3 | | | | | |
| | s4 | | | | | |
| | m12 | | m13 | | m14 | |
| | disk10400 | disk20400 | disk30400 | disk40400 | disk50400 | disk60400 |
| | 36GB/10k | 36GB/10k | 36GB/10k | 36GB/10k | 36GB/10k | 36GB/10k |
| shelf 5 | rcs15 on primary1, rcs25 on primary2, rcs35 on primary3 | | | | | |
| | d15 on primary1/standby1, d25 on primary2/standby2, d35 on primary3/standby3 | | | | | |
| | s5 | | | | | |
| | m15 | | m16 | | m17 | |
| | disk10500 | disk20500 | disk30500 | disk40500 | disk50500 | disk60500 |
| | 36GB/10k | 36GB/10k | 36GB/10k | 36GB/10k | 36GB/10k | 36GB/10k |

Table 2 shows six LUNs on each controller pair, and each LUN in a corresponding remote copy set (RCS). The LUNs were composed of three mirror sets, each containing two physical drives. The mirror sets were then combined into a single stripe set, and the stripe set was presented to the server as a LUN.

In Table 2:
- disk10100, disk10200, and so on are the physical drives attached to the HSG80.
- m1, m2, and so on are the mirror sets containing the physical drives.
- s1, s2, and so on are the stripe sets containing multiple mirror sets.
- d10, d11, and so on are the LUNs.
- rcs10, rcs11, and so on are the RCS.

For example, rcs15 contains LUN d15, which contains stripe set s5. Stripe set s5 contains mirror sets m5, m6, and m7 (3 x 2 RAID 1+0 device), which contain physical disks disk10500, disk20500, disk30500, disk40500, disk50500, and disk60500.

## solution design and design rules

The initiator site is the original site where the Oracle application is running. The target site is where the second copy of the data is stored. The target site does not have to be dedicated to the Oracle application. The process of failing-over an Oracle application to the target node involves making the application's replicated data accessible and starting instances on the target node to restore application availability. The solution functionality includes the following features:

- All data can be mirrored between storage elements in two different storage arrays that can be in separate geographical locations.
- Each I/O write access is sent to both storage locations, and reads occur only at the local storage location.
- DRM copies data online and in real time to remote locations through a local or extended SAN.

## optimizing for normalization

The disk configuration in Table 2 is balanced from an I/O perspective, and, more importantly, has been optimized for full normalizations. The LUNs were distributed equally on two fabrics. Three HSG80s (not pairs) have the bandwidth to completely fill one 100-MB channel during normalization. With DRM using dual-redundant fabrics, both fabrics can be simultaneously active. Hence, the normalization is optimized over both fabrics. Table 4 shows normalization times when normalizing 1.3 TB of storage under different circumstances.

**table 3. normalization times**

| Distance | 0 km | 100 km, with extended fabric license | 100 km, without extended fabric license | 100 km under load |
|---|---|---|---|---|
| normalization times (hours and minutes) | 2:41 | 3:11 | 3:31 | 5:31 |

Note: All times shown are on an unloaded system, except as specified. The time to normalize under load is highly variable, depending on the activity rate. The load used in testing varied from 100% query to 100% update and, on average, was approximately 50% query, 50% update with 1000 users.

## storage layout best practice

Oracle recommends using a Stripe and Mirror Everything (SAME) approach to storage layout. As a result, all LUNS were configured as RAID 1+0. Each physical drive in the mirror pairs is on a separate SCSI channel to minimize contention and optimize availability.

Writeback caching *should* be used on all Oracle data and redo log LUNs to improve performance. Writeback caching is not necessary on the archive log LUNs. HP does not recommend using writeback caching with large logs.

Use a naming convention that is simple and easy to follow. HP recommends that RCS names correspond to their unit names. RCS names might start with rcs and correspond to the unit. For example, rcs10 contains the unit d10, and rcs15 contains d15.

## test database specification

The experimental design goal was to construct an Oracle database using the TPC-C-based workload tools provided by Oracle as part of the overall workload generator tool kit. The goal was to simulate a large database of approximately 1 TB. To support this target volume for customer data, additional tablespaces were configured for system tables, indexes, undo segments, redo logs, and archive logs as needed. In total, the on-disk size of all database files was approximately 1 TB on the initiator storage site.

All database volumes were protected with RAID 1+0; therefore, the total raw disk space used by the database was approximately 2.6 TB on the initiator site storage, yielding approximately 1.3 TB of usable space.

For all tests that required remote mirroring of the whole database, all of the storage was remotely mirrored. Remote mirroring required that the same approximately 2.6 TB be configured identically at the target site.

The customer table was partitioned into 12 equal parts. Twelve was chosen for two reasons. One, this number kept the tablespaces to a manageable size of slightly less than 8 GB each. Secondly, 12 partitions allows for distributing the tablespaces for each schema across six disks in a well-balanced approach.

Note that the TPC-C load program is based on a concept of districts, warehouses, and customers. When running the TPC-C load process, the number of warehouses is an input parameter from which total customers are calculated using 30,000 customers per warehouse. Table 4 shows a summary of the size of each customer table in the test.

**table 4. customer tables**

| **schema** | cust80 (80-byte rows) | cust800 (800-byte rows) | cust8000 (8000-byte rows) |
|---|---|---|---|
| **number of rows/partition** | 55,920,000 | 10,140,000 | 510,000 |
| **number of warehouses/partition** | 1,864 | 338 | 17 |
| **total warehouses (12 partitions)** | 22,368 | 4,056 | 204 |
| **total rows** | 671,040,000 | 121,680,000 | 6,120,000 |

The 8000-byte rows measure approximately 8,200 bytes, so they no longer fit into a single 8-k Oracle block. The resulting chaining caused each row to use two blocks; therefore, calculations were based on 16 k (two Oracle blocks) per row, even though the rows only use 8,200 bytes each. An appropriate size of16-k blocks was available with the Oracle 9*i* ability to handle different block sizes on an individual tablespace basis. However, for consistency and backward compatibility with Oracle 8*i*, the tablespace was set up with the default 8-k blocks.

The customer table for each schema was range partitioned on the c_abs_id column. This column was chosen because it is the lookup column for the workload generator. Local partitioned indexes for each table were then created on the table. A local partitioned index uses the same column as the table partition, hence the use of c_abs_id as the partition key. Most of the index partitions used approximately 1 GB of space, except for the 80-byte row tables, where each partition was 1.5 GB. The difference is because of the greater number of rows to index in the 80-byte row tables.

To avoid having one huge schema containing 1.3 billion customers, two sets of identical schemas were used in the final database. Because the TPC-C load process does not support multiple schemas, the process was used to populate the first three schemas, and then a set of PL/SQL procedures was used to copy the data from the primary schemas (tpcc_80, tpcc_800, and tpcc_8000) to the second set of schemas (tpccb_80, tpccb_800, and tpccb_8000).

## database layout

As in real-world situations, the test environment could have used more disks to distribute data across. However, given that real-world scenarios do not have an unlimited number of disks, the environment was configured with an adequate amount of storage and spindles.

There were four 40-MB redo log groups, each with members spread across all three HSG80 pairs. Redo log groups 1 and 3 were placed on one LUN and redo log groups 2 and 4 on a different LUN. This arrangement prevented disk contention between the LGWR and ARC processes after log switches. We used 40-MB logs, even though this choice caused more frequent log switches, in order to "smooth out" the overall performance. Log switches are very costly events, and the larger the log, the more impact it has on the system. Choosing smaller log sizes yielded better and more predictable performance.

The customer table is the main table in the database. The different schemas used nearly 600 GB of space. The 72 x 8-GB total partitions were spread across six LUNs. Corresponding indexes were spread across three LUNs that were shared with other tablespaces, such as *system* and *undo*. The overall layout was set up as follows.

**table 5. customer table**

| preferred controller/fabric | primary1/standby1 | primary2/standby2 | primary3/standby3 |
|---|---|---|---|
| upper | redo log 1a<br>redo log 3a<br>d10/dsk10 | redo log 1b<br>redo log 3b<br>d20/dsk20 | redo log 1c<br>redo log 3c<br>d30/dsk30 |
| lower | redo log 2a<br>redo log 4a<br>d11/dsk11 | redo log 2b<br>redo log 4b<br>d21/dsk21 | redo log 2c<br>redo log 4c<br>d31/dsk31 |
| upper | partitions 1 and 7<br>d12/dsk12 | partitions 2 and 8<br>d22/dsk22 | partitions 3 and 9<br>d32/dsk32 |
| lower | partitions 4 and 10<br>d13/dsk13 | partitions 5 and 11<br>d23/dsk23 | partitions 6 and 12<br>d33/dsk33 |
| upper | undo<br>d14/dsk14 | undo and indexes<br>d24/dsk24 | indexes and system<br>d34/dsk34 |
| lower | indexes<br>d15/dsk15 | temp and undo<br>d25/dsk25 | archive logs<br>d35/dsk35 |

In Table 5, d10 though d35 refer to the LUNs from the storage layout sections. Dsk10 through dsk35 refer to the UNIX device special file names and were set up to correspond to the LUN numbers for ease of management.

Note: Undo management in Oracle 9*i* can be automatic or use traditional rollback segments. Both configurations were tested, and we found no difference in performance. Automatic undo management was chosen for the tests because of its simpler management.

## scaling-growth-flexibility

The following sections show how the database was set up, the actual test results (failover and failback best practices), and performance achieved under various loads and conditions.

Our conclusion is that the impact of remotely mirroring an entire database using DRM has no appreciable impact on running applications.

### oracle database layouts for DRM best practices

Database files need to be balanced on the storage between the two different fabrics. For normalization purposes, this means to balance by capacity. For the database, this means to also balance by load. I/O load balancing is key to achieving maximum performance both with HSG80s in general and with DRM in particular. Load balancing can be accomplished by setting the preferred path appropriately between controllers. Use HP SANworks Network View to determine the load on the HSG80s. To set the preferred path to one fabric or the other, use the "set *unit* preferred_path" command.
For example:

> set d10 preferred_path=this

> set d11 preferred_path=other

Redo logs need members on each of the separate HSG80 controller pairs. If only one controller pair is mirrored, then the second member can be on the same controller pair.

Redo log LUNs can be initialized to a smaller capacity both to improve performance and to minimize normalization times. The following is an example of the commands to initialize a disk to 1 GB and then mirror it:

> initialize disk10000 capacity=2048000

> initialize disk20000 capacity=2048000

> add mirror m0 disk10000 disk20000

## configuring DRM

The following commands were used to set up the remote copy sets.

From primary1:
> add remote_copy_set rcs10 d10 standby1\d10
> add remote_copy_set rcs11 d11 standby1\d11
> add remote_copy_set rcs12 d12 standby1\d12
> add remote_copy_set rcs13 d13 standby1\d13
> add remote_copy_set rcs14 d14 standby1\d14
> add remote_copy_set rcs15 d15 standby1\d15

From primary2:
> add remote_copy_set rcs20 d20 standby2\d20
> add remote_copy_set rcs21 d21 standby2\d21
> add remote_copy_set rcs22 d22 standby2\d22
> add remote_copy_set rcs23 d23 standby2\d23
> add remote_copy_set rcs24 d24 standby2\d24
> add remote_copy_set rcs25 d25 standby2\d25

From primary3:
> add remote_copy_set rcs30 d30 standby3\d30
> add remote_copy_set rcs31 d31 standby3\d31
> add remote_copy_set rcs32 d32 standby3\d32
> add remote_copy_set rcs33 d33 standby3\d33
> add remote_copy_set rcs34 d34 standby3\d34
> add remote_copy_set rcs35 d35 standby3\d35

Failsafe error mode is recommended when mirroring databases. Failsafe is especially recommended when mirroring an entire database across multiple controller pairs to guarantee that LUNs on the various controller pairs stay in sync under all conditions. When using multiple controller pairs, failsafe guarantees that the standby site will always be a point-in-time image of the primary, and all transactions written on the initiator site will be mirrored to the target site as well.
For example:> set remote_copy_set rcs10 error_mode=failsafeNote: When using multiple controller pairs, all remote copy sets should be set to failsafe on all controllers to ensure a proper point-in-time image at the remote site.

For detailed configuration information, refer to the DRM Configuration Guide available at http://h18006.www1.hp.com/products/sanworks/drm/documentation.html

**notes on performance tests**

There are numerous ways to examine the performance of a storage array subsystem. What is the maximum performance that can be achieved? In other words, how many transactions can be transferred through the DRM environment (saturation testing)? Or, the question that is perhaps even more relevant to users: What performance can the subsystem achieve for typical real-world applications?

Saturation test results are measured as requests per second for random transfer patterns and as megabytes per second for sequential patterns. Saturation tests ignore the often-lengthy response times that can result when a subsystem is under heavy I/O load. But because application workload tests explore the subsystem's suitability for deployment of real-world applications, response times cannot be ignored—in this test, response times are the key measurement. As such, the test goal is to determine how much load the subsystem can withstand and still maintain a reasonable response time.

In the test suite, a response time limit was placed at less than two seconds for all schemas on the heavy update tests of 25% query and 75% update. The tests simulated 1,000 users loading the system, and inter-transaction wait time varied until the response times met the design goal. An average of five seconds was determined to be a suitable value for the inter-transaction wait time.

The next issue for this testing was reproducibility of the results. The goal for reproducibility was a standard deviation of less than 10 ms between runs of the same test on the 8000-byte schema. The 8000-byte schema was chosen because it has fewer rows than the other schemas, and the indexes could be cached in the system global area (SGA) of the database. The large SGA eliminated the variability of index reads on the performance of DRM. Because of this standard, results within 10 ms (one standard deviation) are considered statistically insignificant for the 8000-byte rows. Higher variability can be expected for the 800- and 80-byte row schemas, and differences of under 20 ms should be considered statistically insignificant.

## performance tests conditions

There were two sets of tests: *normal operating conditions* and *fault conditions*. Tests were run against the three schemas (8000-, 800-, and 80-byte rows) noted in Table 4.

The normal conditions tested were:

1. Baseline database without mirroring
2. Entire database mirrored at 0 km
3. Entire database mirrored at 100 km, synchronously
4. Redo logs only mirrored with DRM and archive logs transported using Oracle 9*i* DataGuard (DG)
5. Entire database mirrored at 100 km, asynchronously

**table 6. response times**

| response times (seconds) | normal conditions | | | | |
|---|---|---|---|---|---|
| **8000-byte schema** | **no DRM** | **DRM 0 k** | **DRM 100 k sync** | **DG redo logs only** | **DRM 100 k async** |
| 25% update | 0.220 | 0.229 | 0.238 | 0.228 | 0.224 |
| 75% update | 1.099 | 1.092 | 1.097 | 1.115 | 1.222 |
| **800-byte schema** | **no DRM** | **DRM 0 k** | **DRM 100 k** | **DG redo logs only** | **DRM 100 k async** |
| 25% update | 0.285 | 0.292 | 0.296 | 0.289 | 0.285 |
| 75% update | 1.547 | 1.588 | 1.609 | 1.563 | 1.671 |
| **80-byte schema** | **no DRM** | **DRM 0 k** | **DRM 100 k** | **DG redo logs only** | **DRM 100 k async** |
| 25% update | 0.361 | 0.373 | 0.376 | 0.369 | 0.360 |
| 75% update | 1.809 | 1.824 | 1.816 | 1.796 | 1.857 |

Note: From the user's perspective, there is almost never a significant degradation of response time. Where there is a slight degradation (as in 1.547 to 1.609), this instance can be attributed to DRM using one of the two ports on each controller exclusively for mirroring traffic.

The fault conditions tested were:

1. Entire database mirrored, all subsystems normalizing
2. Entire database mirrored, one link down
3. Entire database mirrored, both links down, write logging enabled

**table 7. response times**

| response times (seconds) | normal conditions | fault conditions | | |
|---|---|---|---|---|
| **8000-byte schema** | **DRM 100 km** | **normalization 100 km** | **single link 100km** | **write logging 100 km** |
| 25% update | 0.238 | 0.240 | 0.253 | 0.273 |
| 75% update | 1.097 | 1.597 | 1.471 | 1.227 |
| **800-byte schema** | **DRM 100 km** | **normalization** | **single link** | **write logging** |
| 25% update | 0.296 | 0.306 | 0.324 | 0.314 |
| 75% update | 1.609 | 2.125 | 1.922 | 1.688 |
| **80-byte schema** | **DRM 100 km** | **normalization** | **single link** | **write logging** |
| 25% update | 0.376 | 0.398 | 0.465 | 0.409 |
| 75% update | 1.816 | 2.489 | 2.340 | 1.856 |

This data is provided to demonstrate robustness. Even under fault conditions, most tests have minimal response time degradation.

## remote mirroring best practices

**Mirroring the entire database** provides the simplest setup, easiest ongoing maintenance, and best performance for Oracle databases mirrored at distances up to 100 km. Ongoing maintenance is generally simpler than with other means of mirroring.

**Synchronous replication** is recommended in all circumstances. Synchronous replication provides the best data integrity with no performance penalty. In most cases, synchronous replication is faster.

**All LUNs should be mirrored using failsafe error mode**. This setting guarantees that all LUNs at the target site are always a point-in-time-image of the primary site.

All RCS using LUNs preferred to the same controller should be in **a common association set**. This will failsafe lock all affected RCS in the event of an error.

The **Extended Fabric License** should be used for the SANswitches when mirroring over distances beyond 50 km and sometimes even less depending on the database load.

## failover operations

DRM makes both failover and failback operations simple to perform and—using HP StorageWorks Command Scripter—easy to automate.

Failover essentially consists of two steps:
1. Site fail the RCS to the target site.
2. Present the storage at the target site to the standby host.

For example:

> site_failover rcs10

> site_failover rcs11

> set d10 enable=(host2a, host2b, host 2c, host2d)

> set d11 enable=(host2a, host2b, host 2c, host2d)

This site failover example enables access to LUNs d10 and d11 on host2.

## failback operations

Like failover operations, failback operations are relatively simple to set up and initiate. Failback requires that the RCS be normalized before attempting to set the initiator back to its original role. The normalization state can be determined by doing a show remote_copy_set *rcs_name*.

For example:

> show remote_copy_set rcs35

Name Uses Used by

---

RCS35 remote copy D35

…

Target state:

PRIMARY3\D35 is **NORMAL**

Note that the state is normal. The failback is operationally very similar to the failover.

1. Shut down the database(s) at the standby site.
2. Disable access from the storage to the standby servers.
3. Set the RCS to point to the original initiator.
4. Enable access at the initiator site to the primary server.

For example, after shutting down the database:

From the standby (target) site:

> set d35 disable=(host2a,host2b,host2c,host2d)

> set rcs35 initiator=primary3\d35

and at the primary (initiator) site:

> set d35 enable=(host1a,host1b,host1c,host1d)

## failover and failback best practices

Command Scripter should be used on both the initiator and target site to automate all of the functions necessary for failover and failback. Use of Command Scripter simplifies storage management and lowers the likelihood of human error during operations. The necessary failback and failover commands (see previous sections) can be placed in a script that should be run in the event of a failover or failback. For complete details on failover and failback, refer to the *Data Replication Manager HSG80 ACS 8.7P Failover/Failback Procedures Guide* (Part Number: AA-RPJ0D-TE), available at http://h18006.www1.hp.com/products/sanworks/drm/documentation.html.

## bill of materials

| Site Equipment List |
|---|
| Database Server: 1 x ES45 with |
| 4 x 1 GHz EV68 Alpha processors |
| 32 GB RAM |
| 6 x KGPSA HBAs |
| 1 x 10/100 MB NIC |
| 1 x 1 GB NIC |
| 1 x KZPCA SCSI controller |
| 6 x 18 GB 10k Ultra SCSI-3 internal disk drives |
| Storage Arrays: 3 x MA8000 Fibre Channel storage arrays |
| 3 pairs HSG80 controllers |
| 12 x 18 GB 15k RPM Ultra SCSI3 disks (containing redo logs) |
| 24 x 36 GB 10k (containing Oracle data and archive logs) |
| 6 x 72 GB 10k RPM Ultra SCSI-3 disks (for backup purposes) |
| SAN: 2 x 16 Port SANswitches |
| Brocade firmware v2.6 |
| Brocade Extended Fabric Licenses |

## conclusions

The results of Oracle's remote mirroring OSCP test suite show that DRM and the MA8000 and EMA12000 work effectively together to provide business continuity and disaster tolerance that Oracle users can have confidence in.

These current tests prove that even on large, heavily loaded Oracle databases, DRM can mirror the data at distances up to100 km with virtually no performance impact. In addition, Command Scripter allows for "single-click" failover and failback, greatly simplifying storage management operations. Like any good tool, DRM can best help Oracle users achieve their goals when its use is well-planned, utilizing the best practices identified in this paper.

The value of the two demonstrations together is that Oracle users can confidently plan their installations to provide business continuity and disaster tolerance with HP StorageWorks Data Replication Manager MA/EMA, and can achieve that security with no penalty to the performance of their applications.

## glossary

**Association set**—A group of one or more logical units treated as a single entity for failover and logging purposes when running in synchronous normal mode. Association sets also provide failsafe properties across multiple remote copy sets.

**Asynchronous replication**—A method of deploying disaster-tolerant solutions. With asynchronous replication, data updates are locally implemented, locally stored, and later forwarded to all other remote copies

**Fabric (Fibre Channel)**—One of the three Fibre Channel topologies. In the Fabric topology, N_Ports are connected to F_Ports on a switch. The fabric routes or switches the data frames from the source node to the destination node. Depending on vendor support, fabric switches may be interconnected to support up to 16 million+ N_Ports on a single network.

**Failback**—The process of restoring a primary service from the secondary after a failover.

**Failover**—When a primary service (path, fabric, or entire site) fails, the backup service takes over. Only one service is operational at any time.

**Failsafe error mode**—A synchronous error mode that prevents writes on the primary site should the secondary site become unavailable. Oracle refers to this as a manual split mode in the OSCP documentation on remote mirroring.

**Initiator**—The site that is the primary source of information. In the event of a system outage, the database would be recovered from the target system.

**Logical Unit Number (LUN)**—A unit is made up of one or more disk devices. In StorageWorks software terminology, a unit made up of more than one disk device is either a RAID-0, RAID-1, RAID1+0, or RAID-5 storageset.

**Normal error mode**—A synchronous error mode that allows writes on the primary site should the secondary site become unavailable. In normal mode, write history logging may be enabled to speed normalization in the event of a short network outage. Oracle refers to this as an automatic split mode in the OSCP documentation on remote mirroring.

**Normalization**—The process of assuring that mirrored drives have consistent data. A full normalization copies the entire logical unit. A fast-failback or mini-merge happens when the changes are logged to a log unit in an association set. Only the changes since the failure are copied to the mirror members as opposed to the entire logical unit.

**RAID (Redundant Array of Independent Disks)**—A way of coordinating multiple disk drives within a single housing developed to improve performance, availability, or both.

**Remote copy set (RCS)**—A set of two or more LUNs that contain the same data, with one LUN on an initiator and the others on the targets. ACS V8.7P allows 12 RCS per controller pair.

**Storage Area Network (SAN)**—A network that connects storage devices, such as disk, tape, and CD ROM drives to all types of computing devices. Based on the Fibre Channel interconnect, SANs provide high-speed, fault-tolerant access to data for client, server, and host computers. These computing devices can be as simple as a small workstation or a large mainframe system.

**SCSI (Small Computer Systems Interface)**— A PC bus interface standard that defines standard physical and electrical connections for devices. This standard enables many different kinds of devices, such as disk drives, magneto optical disk, CD-ROM drives, and tape drives to interface with the host computer.

**Synchronous data replication**—A method of deploying disaster-tolerant solutions. Synchronous data replication ensures that data copies are always identical to prevent critical data loss in the event of a failure or disaster. In this mode, data is written simultaneously to the cache of the local subsystem and the remote subsystems, in real time, before the application I/O is completed, thereby ensuring the highest possible data consistency.

**Target**—The destination site for a remote copy set. Information is copied from the initiator to the target.

# appendix

## sample init.ora used for testing

background_dump_dest = /var/oracle/901/admin/cust/bdump

compatible = 9.0.0

control_files = ('/oradata/system/control01.ctl', '/oradata/index/control03.ctl', '/oradata/archive/control02.ctl')

core_dump_dest = /var/oracle/901/admin/cust/cdump

db_block_size = 8192

db_cache_size = 10G

DB_RECYCLE_CACHE_SIZE=1G

db_domain = ''

db_name = cust

db_writer_processes = 10

fast_start_mttr_target = 300

instance_name = cust

java_pool_size = 117440512

large_pool_size = 104857600

log_archive_dest_1 = 'LOCATION=/oradata/archive/'

log_archive_format = '%t_%s.arc'

log_archive_start = TRUE

log_buffer = 2097152

log_checkpoint_interval = 0

log_checkpoint_timeout = 0

open_cursors = 300

parallel_max_servers = 8

processes = 2500

remote_login_passwordfile = EXCLUSIVE

# resource_manager_plan = 'SYSTEM_PLAN'

sga_max_size = 15G

shared_pool_size = 117440512

sort_area_size = 1048576

timed_statistics = TRUE

undo_management = AUTO

undo_retention = 120

undo_tablespace = UNDOTBS1

user_dump_dest = /var/oracle/901/admin/cust/udump

cursor_space_for_time = true

### sample sysconfigtab for large databases

```
ipc:
 shm_mni=1024
 shm_seg = 256
 shm_max=17179869184
 ssm_threshold=0
proc:
 max_proc_per_user=2048
 maxusers=4096
 max_per_proc_data_size=17179869184
 max_per_proc_address_space=17179869184
 per_proc_data_size=17179869184
 per_proc_address_space=17179869184
```

## why hp

As a full service provider of servers, storage, software and infrastructure, only HP can:

- ensure continuous uptime by systematically eliminating single points of failure across the board—from the hardware right up to the Oracle application level. HP solutions for Oracle provide instant, automated switchover for hardware, operating system, database, and Oracle components, ensuring service continuity in the event of failure.
- deliver rock-solid security with solutions that perform encryption for secure transactions and instant authorization/authentication checks.
- enable rapid deployment of Oracle solutions because HP consultants, cooperating closely with Oracle, use a structured approach based on Oracle best practices to help customers design and implement the IT infrastructure that their enterprise needs to ensure a smooth, speedy rollout of Oracle solutions.
- provide end-to-end control of the entire Oracle environment with management tools and support services that manage every component, from hardware to application—even in distributed, Internet-based system environments.
- enable faster recovery time to ensure that the customer's Oracle environment is restored with minimal impact to their business.
- provide the highest level of storage performance in the industry, which contributes to higher productivity of the customer's Oracle resources.

## for more information

To learn more about HP storage area networks, contact your local HP sales representative or visit our Web site at http://www.hp.com/go/san.

For more information about DRM visit our Web site at http://h18006.www1.hp.com/products/sanworks/drm/index.html

For more information about DRM and the Oracle Storage Compatibility Program visit our Web site at http://h18006.www1.hp.com/products/sanworks/drm/compatibility.html

Oracle Corporation provides further information regarding the use of remote mirroring with Oracle databases and Oracle Storage Compatibility Program testing in two papers available from Oracle:

Guidelines for Using Remote Mirroring Storage Systems for Oracle Databases
by Bill Lee
Guidelines for Testing Remote Mirroring Storage Systems for Oracle Databases by Bill Lee and Cynthia Yip.

Let us know what you think about the technical information in this document. Your feedback is valuable and will help us structure future communications. Send your comments to: Oracle_Storage_Solutions@hp.com

For the HP sales office nearest you, refer to your local phone directory, or call the following appropriate HP regional office.

**corporate and North America headquarters**
Hewlett-Packard
3000 Hanover Street
Palo Alto, CA 94304-1185
Phone: (650) 857-1501
Fax: (650) 857-5518

**regional headquarters**
*Latin America*
Hewlett-Packard
Waterford Building, 9th Floor
5200 Blue Lagoon Drive
Miami, FL 33126 USA
Phone: (305) 267-4220

*Europe, Africa, Middle East*
Hewlett-Packard
Route du Nant-d'Avril 150
CH-1217 Meyrin 2
Geneva, Switzerland
Phone: (41 22) 780-8111

*Asia Pacific*
Hewlett-Packard Asia Pacific Ltd.
Hewlett-Packard Hong Kong Ltd.
19/F, Cityplaza One
1111 King's Road
Taikoo Shing
Hong Kong
Phone: (852) 2599-7777
HP business solutions, products, and supplies can be purchased at your local HP partner (www.hp.com/go/locator) or bought directly from HP at www.buy.hp.com or by calling toll-free 1-800-613-2222.



01/2003