Building a
Highly Scalable and Available Data Environment for Oracle9i RAC

To provide a highly scalable and available database environment for Oracle9i™ Real Application Clusters (RAC), administrators must establish a highly available storage infrastructure. A storage area network (SAN) can provide redundant paths to storage, and running EMC® PowerPath® can leverage this redundancy by providing a mechanism for path failover in an Oracle9i RAC infrastructure.

In today’s business environment, high availability is required for mission-critical applications. The Oracle9i™ relational database management system (RDBMS) is highly available and scalable. The Oracle9i Real Application Clusters (RAC) option enables the Oracle9i RDBMS to be configured in a cluster database architecture where multiple nodes share the same storage. Oracle9i RAC provides high availability; if one node fails, the others take over and provide uninterrupted access to the database. However, if only one I/O path exists from each node to the shared storage in such an environment, this I/O path potentially becomes a single point of failure.

A typical Oracle9i configuration includes a storage area network (SAN), which can help provide a highly available data infrastructure by using redundant components to ensure that no component becomes a single point of failure. A Fibre Channel–based SAN fabric supports multipath routing between SAN switches. In a typical topology, a node has multiple Fibre Channel host bus adapters (HBAs), each of which are connected to the same SAN, resulting in multiple paths to the same device (see Figure 1). SAN storage devices can also accept multiple Fibre Channel connections.

Redundant paths in a SAN provide failover capability when any component in the data path fails. Multiple paths also enhance efficiency, allowing administrators to load balance SAN traffic by considering I/O across all available paths. In doing so, the SAN takes advantage of the additional bandwidth provided by each physical connection.

Although redundant I/O paths are beneficial for load balancing and link failover, they can create complications. Because each device on a SAN appears as a SCSI ID on each HBA that is connected to the SAN, a system with multiple HBAs connected to the SAN will behave as if each device on each path is a separate SCSI device. Thus, the operating system behaves as though multiple
storage resources exist when in fact there are only multiple paths to the same resource.

Pointing to the same storage device along different paths could potentially cause data corruption and system crashes. To prevent such problems, administrators can install path management software, such as EMC® PowerPath®, on each node in the SAN. PowerPath enables multiple I/O paths to the shared storage by masking the paths, and presents the operating system with the appearance of a single SCSI connection. This masking ensures that the node receives a single view of the storage devices across multiple HBAs. PowerPath also automatically detects available paths, restores failed paths, and load balances I/O across all paths. When integrated in an Oracle9i RAC environment, PowerPath provides highly available, scalable, and fault-tolerant shared storage.

Using PowerPath for path failure detection
EMC PowerPath helps provide high availability by automatically detecting and restoring failed paths while storage arrays, nodes, and applications remain available. Should an HBA, a storage processor, or a cable fail, PowerPath completes an I/O request through another available channel, helping prevent the interruption of data to an application. PowerPath also provides automatic online path recovery after the path is repaired, which can reduce planned outages to restore services.

PowerPath virtual devices map paths to storage
PowerPath resides on a node as a software component between Oracle9i RAC, the Oracle® cluster file system (OCFS), and the HBA device driver layer (see Figure 2). PowerPath operates independently of applications, the RDBMS, management utilities, and file systems, allowing administrators to install and configure PowerPath without modifying existing software.

The PowerPath driver resides on the node, above the HBA driver. The node has multiple HBAs so that it can provide path failover through the PowerPath driver. The PowerPath driver enables virtual devices, which provide failure-resistant and load-balanced paths to the Dell|EMC storage array. An application references a PowerPath virtual device; in turn, the PowerPath driver manages path allocation to the storage array.

In the following example, four logical unit numbers (LUNs) are configured on a Dell|EMC storage array, which uses two storage processors to connect—through two separate Fibre Channel switches—to a node containing two HBAs. The resulting paths and their mappings to the Power Path virtual devices provide a total of 16 paths from a node to storage.

Having 16 paths to storage would ordinarily result in 16 logical devices being visible to the node. However, the PowerPath driver creates four PowerPath virtual devices, and each of these virtual devices maps four paths to a logical device on the storage array, as indicated in Figure 3.

Powermt management utility facilitates path management
For automatic failure detection and recovery, PowerPath provides an administration utility called Powermt, which provides a command
line interface to the PowerPath environment. PowerPath periodically tests the paths for failure detection according to a built-in algorithm. Using the Powermt utility, administrators can set up a host node to perform autorecovery on failed paths by using the following command-line instruction to the PowerPath driver:

```
powermt set periodic_autorestore=on|off
```

For convenient management of a UNIX® or Linux® node configured with PowerPath, the Powermt utility provides several features, including:

- **Checking a PowerPath configuration**: The `powermt check` command checks the specified paths and, if desired, removes from the PowerPath configuration any paths marked dead.
- **Configuring paths to logical devices**: The `powermt config` command configures all detected logical devices as PowerPath devices and adds these devices to the PowerPath configuration, creating devices as required.
- **Removing paths from PowerPath management**: The `powermt remove` command deletes the specified path (or paths) from PowerPath’s list of configured paths. It does not delete the logical device to which the paths refer.

### Integrating PowerPath in an Oracle9i RAC environment

To perform I/O, Oracle9i RAC uses PowerPath virtual devices (such as `/dev/emcpowera` and `/dev/emcpowerb, shown in Figure 3) instead of LUNs, or logical devices (such as `/dev/sdb` and `/dev/sdc`). If an actual device path fails, the PowerPath driver routes I/O to an alternative path without causing any interruption to the RDBMS functionality.

Integration of PowerPath with Oracle9i RAC is a simple process. PowerPath may be deployed in either a new or existing Oracle9i RAC implementation, as explained in the following sections.

#### Integrating PowerPath with a new Oracle9i RAC implementation

Oracle9i RAC uses shared storage on the SAN for its database, redo log, and control files. The RDBMS engine from each node in the cluster must have direct access to this storage to create or update any of these files. Without PowerPath, the RDBMS engine would directly access the LUNs on the shared storage to create the required files. With PowerPath integrated, the procedure is the same, with one important exception. The partitions and OCFS are created on the PowerPath virtual devices rather than on the partitioned LUNs:

1. Partition the PowerPath virtual devices on the shared storage array according to the database sizing requirements:
   ```
fdisk /dev/emcpowera
fdisk /dev/emcpowerb
fdisk /dev/emcpowerc
fdisk /dev/emcpowerd
```

2. Create the OCFS on the new partitions and mount the file system:
   ```
   mkfs.ocfs -F -b 128 -L u01 -m /u01 -u 200 -g 300 -p 0775 /dev/emcpowera1
   mkfs.ocfs -F -b 128 -L u02 -m /u02 -u 200 -g 300 -p 0775 /dev/emcpowerb1
   mkfs.ocfs -F -b 128 -L u03 -m /u03 -u 200 -g 300 -p 0775 /dev/emcpowerc1
   mkfs.ocfs -F -b 128 -L u04 -m /u04 -u 200 -g 300 -p 0775 /dev/emcpowerd1
   ```

3. Set up the clusterware on the OCFS, and create the Oracle9i RAC database on the shared SAN storage.

   The integration of PowerPath is transparent to the Oracle9i RAC database engine, as is the fact that each PowerPath virtual device points to multiple physical I/O paths.

#### Integrating PowerPath with an existing Oracle9i RAC implementation

In an existing Oracle9i RAC database that is set up without PowerPath, the database engine uses the LUNs on the shared SAN storage to create and update the database files. Integrating PowerPath is a straightforward process using the following steps:

1. Shut down all Oracle services in the cluster, including database listeners, Oracle Intelligent Agents, and Oracle Cluster Manager.
2. Shut down the database on all cluster nodes.
3. Use the `umount` command to unmount all OCFS volumes.
4. Configure the Dell|EMC storage array to support PowerPath, and install the PowerPath software on all cluster nodes. On each node, the PowerPath software will automatically create PowerPath devices that point to the existing LUNs.
5. Modify the `/etc/fstab` configuration file and replace the device names of each LUN (for example, `/dev/sdb`, `/dev/sdf`, `/dev/sdj`, `/dev/sdn`) on the storage array with the PowerPath virtual device name (for example, `/dev/emcpowera`, `/dev/emcpowerb`, `/dev/emcpowerc`, `/dev/emcpowerd`).
6. Restart the OCFS and use the `mount` command to mount the file system.
7. Start up Oracle services, clusterware, and Oracle9i RAC database instances on all cluster nodes.

**Using PowerPath to maintain availability**

As reliable access to information becomes a critical mission for today’s data centers, IT architects and administrators must use centralized, scalable storage to help create a highly available data infrastructure across the enterprise. SANs provide centralized data storage, and EMC PowerPath complements SAN architecture by helping manage redundant SAN paths to provide a high-availability environment for Oracle9i RAC.

Factors such as I/O load, effect of downtime, and availability of administrator maintenance time can help determine whether an organization’s networked Dell|EMC storage would benefit from PowerPath. PowerPath offers fault tolerance to help eliminate downtime, load balancing to enable more efficient I/O traffic, and automatic detection and repair of paths that have failed, thereby helping to reduce administrative overhead. By integrating PowerPath with Oracle9i RAC, administrators can improve SAN uptime and provide a stable infrastructure for mission-critical data.

Zafar Mahmood (zafar_mahmood@dell.com) is a software engineer in the Dell Database and Application Engineering Deployment Department of the Dell Product Group. He has been involved in database performance optimization, database systems, and database clustering for more than six years. He is currently working on Oracle9i RAC implementations at Dell. Zafar has an M.S. in Electrical Engineering with a specialization in Computer Communications from the City University of New York.

Paul Rad (paul_rad@dell.com) is a senior software engineer in the Dell Database and Application Engineering Department of the Dell Product Group. Paul has master’s degrees in both Computer Science and Computer Engineering from the University of Texas at San Antonio.

Robert Nadon (robert_nadon@dell.com) is a software engineering consultant in the Factory Installation Development Group of the Dell Product Group. He focuses on factory installation for UNIX-based (Linux and BSDi) operating systems and for the Oracle9i database. Another area of interest is Dell software RAID offerings. Robert has a B.S. in Computer Engineering from Texas A&M University. He is a Red Hat Certified Engineer (RHCE) for Red Hat Linux 6.2 and 7.0.

---

FOR MORE INFORMATION

Dell | EMC storage: [http://www.dell.com/emc](http://www.dell.com/emc)
Dell and Oracle: [http://www.dell.com/oracle](http://www.dell.com/oracle)
Oracle database: [http://www.oracle.com/oracle9i](http://www.oracle.com/oracle9i)