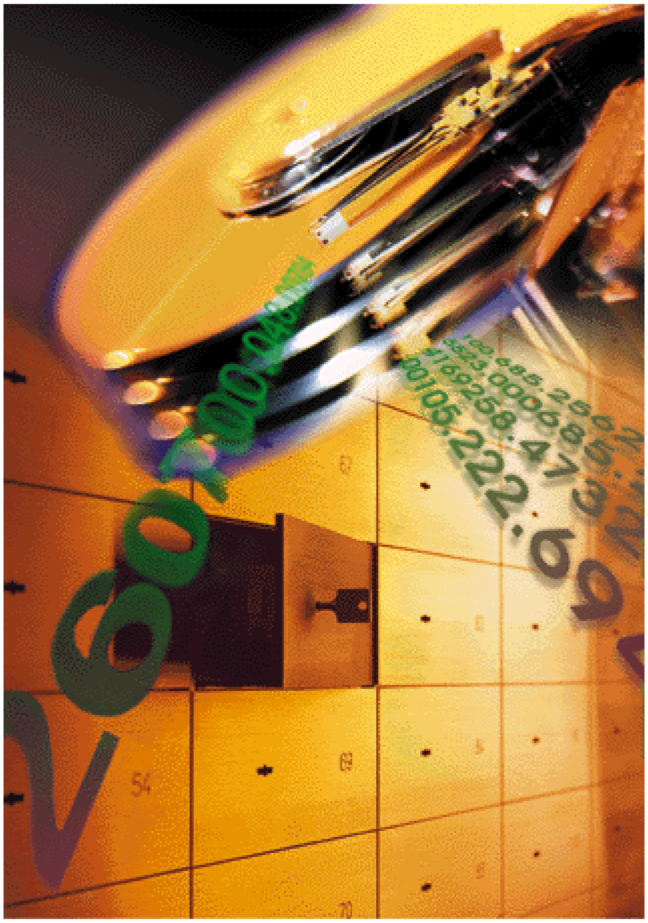
Storage Management for SAP and Oracle: Split Mirror Backup / Recovery With IBM's Enterprise Storage Server (ESS)



Sanjoy Das, Siegfried Schmidt, Peter Pitterling and BalaSanni Godavari

The following terms are trademarks of International Business Machines Corporation in the United States, other countries, or both:

AIX ®
DB2®
DB2 Universal Database®
Enterprise Storage Server (ESS) ®
ESCON®
FlashCopy®
OS/390®
StorWatch®
Tivoli ®
TME 10®

The following terms are trademarks of SAP AG in Germany, in the United States, other countries, or both:

SAP®
SAP Logo®
mySAP.com®
R/3®
ABAP®
SSQJ®
Advanced Technology Group®
OSS®
SAP R/3 Note®

Java® and all Java-based trademarks and logos are trademarks or registered trademarks of Sun Microsystems, Inc. in the United States, other countries, or both.

Legato NetWorker® is a trademark of Legato Systems, Inc. in the United States, Other countries, or both.

Oracle® is a trademark of Oracle Corporation in the United States, Other countries, or both.

Windows® is a trademark of Microsoft Corporation in the United States, Other countries, or both.

Veritas NetBackup® is a trademark of Veritas Software in the United States, Other countries, or both.

OpenView® is a trademark of Hewlett Packard in the United States, Other countries, or both.

BMC® is a trademark of BMC Software in the United States, Other countries, or both.

Other company, product, and service names may be trademarks or service marks of others.

The information provided in this document is distributed "AS IS" basis without any warranty either express or implied. IBM AND SAP DISCLAIM ALL EXPRESS AND IMPLIED WARRANTIES WITH RESPECT TO SUCH INFORMATION, INCLUDING ANY WARRANTIES OF MERCHANTIBILITY OR FITNESS FOR A PARTICULAR PURPOSE. The use of this information or the implementation of any of these techniques is a customer responsibility and depends on the customer's ability to evaluate and integrate them into their operating environment.

While the information contained in this paper has been reviewed by IBM and SAP for accuracy, there is no guarantee that the same or similar results will be obtained elsewhere. Customers attempting to adapt these techniques to their own environments do so at their own risk. The performance data contained herein was obtained in a controlled environment based on the use of specific data. Actual results that may be obtained in other operating environments may vary significantly. These values do not constitute a guarantee of performance.

References in this document to IBM and SAP products, programs, or services do not imply that IBM or SAP intend to make such products available in all countries in which each company operates. Neither IBM nor SAP warrants the others products. Each company's products are warranted in accordance with the agreements under which they are provided.

Contents

Abstract

Recent new products such as SAP's B2B Procurement, CRM and mySAP.com Portals and Trading Exchange markets require an ever-increasing need for continuous system availability.

This paper provides information on an essential component of advanced infrastructure solutions – the High Availability Split Mirror Backup / Recovery **(SMBR)** for SAP R/3 on the Oracle RDBMS and the AIX operating system environments.

The solution described in this paper is intended to deliver a backup with no impact on live R/3 system ("serverless") using the advanced functions of IBM's Enterprise Storage Server (ESS). This "zero" downtime for the live R/3 system means that SAP users typically do not miss a beat while the backup of the live database takes place. No transactions typically are cancelled during the copy process / backup process.

"Instant" availability of a point-in-time copy of the production database using Oracle's HOT BACKUP feature provides the ability to deliver a consistent copy of the database using the log information written during the online backup. The ability to provide consistent copies of the database provides flexibility to place an emergency system at the user's disposal while recovering the live database from a disaster. Beyond Backup / Recovery, a consistent copy of the database may be used for various purposes, such as creation of Reporting, Production-Fix or a Repository instance for a Business Warehouse (BW) system.

Preamble

This white paper, written from the DBA's perspective, addresses the infrastructure design, implementation tasks and techniques required for complex Enterprise Application Integration landscapes for high availability SAP R/3 applications consisting of a database (Oracle), an operating system (AIX) and a Enterprise Storage Subsystem (ESS) which all interoperate to deliver an easy-to-manage backup & recovery solution for SAP customers. Backup & Recovery solutions are mission critical activities in today's world of 7 x 24 computing and are a major focus for IT personnel, for application management and for DBA's. With the exploding growth in storage requirements for SAP application environments, this work touches on major elements of each area of technology spanning critical operating requirements and how this storage-centric solution delivers compelling value for SAP customers.

1 Introduction

Service level agreements increasingly have to reflect that in case of planned downtime such as database Data Manipulation Language (DML) error, backup, hardware / software maintenance, R/3 upgrade and unplanned downtime such as error analysis and restores, the system has to be available within minutes.

SAP's Advanced Technology Group has developed scenarios using live databases that constantly copy or mirror using storage subsystems, allowing business continuation during the split (and resynchronization) of the mirror. Once the logical database mirrors are established, additional copies are created for backup and for use by a standby SAP R/3 System. This solution minimizes the I/O load impact of the live environment and offloads the backup activity away from the live database / storage server to a standby / backup server host. The Split Mirror solution, based on this concept was successfully implemented using Oracle database on the AIX platform using the ESS.

This solution can be implemented for with a single or a dual ESS configuration using the ESS's advanced functions – the local copy function FlashCopy (FC) and the remote synchronous copy function Peer-to-Pee-Remote-Copy (PPRC). The core R/3 system was loaded using SAP - developed SSQJ tool [12] for OLTP volumes.

This solution in a dual ESS configuration is intended to enable customers to implement remote data vaulting and/or to scale to a larger database size.

A solution similar to the one described in the following pages was implemented on IBM's ESS with SAP R/3 on DB2OS390 and first demonstrated in November 1999 (see references [4, 5] for details). Later, a similar solution was validated on the ESS using DB2 UDB on the AIX platform [11]. Both these Split Mirror Backup / Recovery solutions on DB2OS390 and on DB2 UDB / AIX used a special "write Suspend / Resume" function created by IBM initially for DB2OS390 and later for DB2 UDB especially for executing the SAP Split Mirror Backup / Recovery Solution (SMBR). This open systems solution for SAP with Oracle on AIX takes advantage of Oracle's Online or Hot Backup functionality.

2 Customer Requirements

Increasingly, with the rapid trend towards very large databases, accompanied by the need for high availability in a global computing environment, customers now demand that production systems be available on a 7 x 24 basis. This also means that in case of disaster, the system has to be available within minutes or hardly longer than the time needed for the physical reload of the database from secondary or remote storage media. This high availability requirement also implies that backup and the creation of an Emergency system may not cause any downtime of the live production system and all procedures to achieve this must be seamlessly automated.

Customers are also very aware of the fact that software or application logical errors and not hardware failures are the most likely causes for the need for disaster recovery capabilities. Hence for mission critical applications, customers will do everything possible to optimally protect them from a hardware disaster. This means that cost effective, intelligent fault tolerant storage subsystems are an important ingredient of high availability advanced infrastructure solutions, helping to insure against cost-prohibitive downtime possibilities.

Along with the demanding requirements of non-stop, web-centric computing, customers now have to contend with business processes spread over multi-vendor platforms, databases and file systems, where automated interaction between systems provide the essence of competitive productivity. In these emerging environments, the customers, need and demand database availability with current data, fast backup / restore / recovery with no impact on the main production (OLTP) system all enabled with seamless automation through well defined, user friendly management interfaces. With the emerging world of SAP technologies, customer environments have to meet stringent requirements for availability, scalability and flexibility that can handle changing customer requirements based upon SAP instance strategies, data migration requirements and disparate growth rates of databases.

3 SAP Requirements

In order to deliver a solution that matches the complex requirements for high availability backup, recovery and performance, the SAP workload needs to be taken off the live production system and all administrative tasks performed on a copy of the system. The recommended environment is depicted in Figure 1.

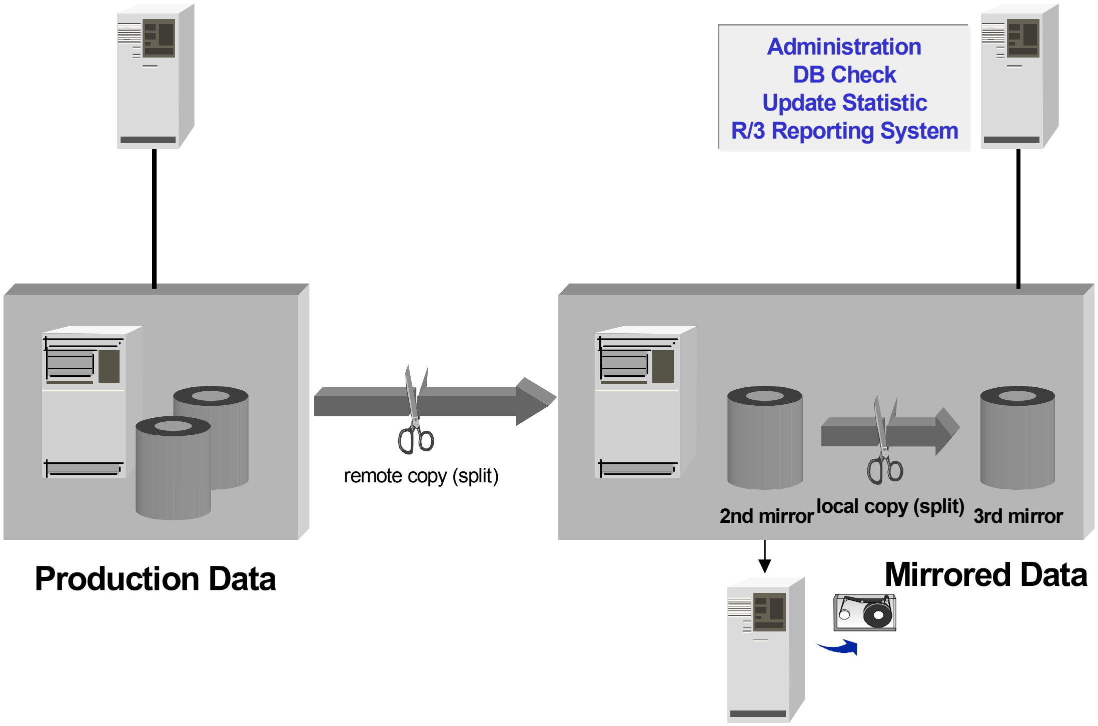


Figure 1: SAP High Availability Advanced Infrastructure

Split Mirror Backup & Recovery Environment

While a production server is connected to a primary fault tolerant storage controller, a mirrored remote copy of the production database is created without the computing support of the production server, in a separate / secondary fault tolerant storage controller located at a spatially separated site, providing high availability and disaster recovery capability for business continuance. This copy can be accessed by a standby / backup server should IT management procedures require its intervention in the event the primary site experiences business interruption. A local copy function within the secondary storage controller produces a consistent point-in-time copy of the production database. This consistent copy, at the option of the user, can be either held inside secondary storage controller or be transferred to tape for remote vaulting. In the event that the primary database or its mirror is not available, the production server or the backup server can be connected to this disk resident copy, helping to dramatically reducing downtime.

The database and SAP Basis administration can be augmented by providing a homogeneous split mirror-based system copy for the purposes of test upgrades, hot packs and database recovery routines. Intelligent storage controllers need to deliver this capability to create near-instant, consistent copies of the database.

- Backup & restore time is database size independent
- Minimum impact on production environment
- Phyical disaster prevention
- Enhanced database administration
- Remote data vaulting
- Offloading backup from production database server
- Reduction of backup-restore-time to minutes

In order to realize this concept, the database management system needs to cooperate with storage subsystems to deliver the results. It should support the creation of a consistent database copy during the application READ/WRITE processing in a manner that exerts no impact on the production system, database or the live production storage server.

4 Oracle and ESS Features to Support the Split Mirror Backup / Recovery Solution

This section describes the key Oracle features required to create a consistent backup database image of the production database using Oracle's Online Hot Backup commands. The backup image includes SAP R/3 objects and File definitions for Oracle as detailed in Section D of the Appendix. This backup process utilizes two advanced functions of the ESS's sub system such as FlashCopy and Peer-To-Peer-Remote-Copy.

4A Oracle Features

For very large Oracle databases in SAP R/3 production environments, the Split Mirror backup capability is essential for the creation of consistent database backup copies without stopping the production system. In order to make this possible, Oracle's features such as "**ALTER TABLESPACE BEGIN BACKUP**" and "**ALTER TABLESPACE END BACKUP**" for Online Hot Backup (with Oracle in ARCHIVELOG mode) capability are used to create backup copies of the production database without any impact on the production OLTP system, database or user activity. During an Online Backup, the Oracle database and the SAP R/3 system remain available. All transactions that are logged in the REDO log files during this backup period are required to be applied to the backup copy of the database to produce a consistent point-in-time copy.

The ALTER TABLESPACE BEGIN BACKUP command will begin logging entire block images on the first change that Oracle encounters on each block owing to the DML activity (such as Insert, Update, Delete). This is accomplished by INIT.ORA parameter setting _LOGBLOCKS_DURING_BACKUP to TRUE .

The ALTER TABLESPACE END BACKUP command creates a redo log record containing the Oracle marker, BEGIN BACKUP checkpoint, also known as System Change Number (SCN) as explained in Section D of the Appendix. This SCN is also recorded in the header of the HOT BACKUP data files and ensures that all the redo generated during the backup has been applied to the data files. During recovery, as mentioned before, the DBA needs to apply at least the redo logs that were generated during the execution of BEGIN BACKUP and END BACKUP commands to make the backup data files consistent. It is also necessary to end HOT BACKUP mode of the tablespace by issuing the ALTER TABLESPACE END BACKUP command.

It is strongly recommended that HOT BACKUPs be taken during periods of low DML activity. Hence the Split Mirror Backup process, as demonstrated in this paper using the ESS's hardware assisted, near-instant local copy (FlashCopy) and remote copy (Peer-To-Peer-Remote-Copy) functions, ensures that the backup process can be executed within a very short period of time, thus minimizing the impact to the production system.

For an overview of the Oracle architecture and other features that take advantage of ESS storage management capabilities, please refer to Section A (Oracle Architectural Overview), Section B (Oracle Database Growth and Impact on Storage) and Section C (Oracle Memory Management) of the Appendix.

4B FlashCopy - ESS's Advanced Local Copy Functions

The ESS Specialist identifies the Logical Unit Numbers (LUNs) by their ESS internal serial numbers. FlashCopy, ESS's "near-instant" local copy function, can be used for all systems that have volumes or LUNs within the same Logical Subsystem (LSS) of an ESS. FlashCopy is set up using the Web interface of the StorWatch ESS Specialist. Then, task selections can be made on the volume pairs – "FlashCopy" with Full or No Copy, and "WITHDRAW" options. See reference [ESS Copy Services] for set up details on FlashCopy.

The NO COPY option in FlashCopy is useful if the copy operation has to complete within a short time so that the source database/application are returned to their normal usage from the end of HOT BACKUP mode. The WITHDRAW command in FlashCopy enables the removal of source and target volume relationships from a previously specified NOCOPY command. The relationship between source and target volumes will automatically end when the physical copy is completed.

Using the ESS Specialist, FlashCopy tasks are created. Then using the Command Line Interface (CLI) rsExecuteTask command for the previously defined tasks, the FlashCopy command with either No Copy or Physical Copy option is executed.

FlashCopy, when set up by the StorWatch ESS Specialist, creates an identical copy of the source volume on to target volumes when appropriate task is initiated using CLI. Volume identification or Disk signatures need to be validated with respect to the host that is connected to the ESS in order for that host to start using the target ESS FlashCopy volumes. In order for a single host to mount both source and target volumes of FlashCopy pairs, AIX provides the RECREATEVG command, which is packaged as a PTF for AIX 4.3.3 in APAR IY10456. It is officially available in:

1. AIX 4.3.3 Recommended Maintenance Level 05 (RML05)
2. AIX 4.3.3 RML06

The RECREATEVG command overcomes the problem of duplicated LVM data structures and identifiers caused by a disk copying process such as FlashCopy. It is used to recreate an AIX Volume Group (VG) on a set of disks that are copied. Then, the normal commands to bring the volume online, i.e. "mounting" can be attempted by the host operating system.

4C Peer-to-Peer-Remote Copy (PPRC) - ESS's Advanced Remote Copy Function

PPRC is a hardware-assisted synchronous remote copy or synchronous mirroring function that can help preserve data integrity. Synchronous mirroring means that an I/O is only completed until after it is acknowledged from the remote site.

PPRC is set up on a volume or LUN basis in two or more ESS's, which are connected by ESCON (Enterprise System Connection) as shown in Figure 2. Updates to a PPRC volume on the local or primary site (primary volume) go first into cache and non-volatile storage (NVS) in the primary storage. The updates are then sent over the ESCON links via larger ESCON frame transmission to the remote or secondary storage control. When the data is in cache and NVS on the secondary site, the receipt of the data is acknowledged and the primary storage control signals to the application the completion of the I/O by a Device End status.

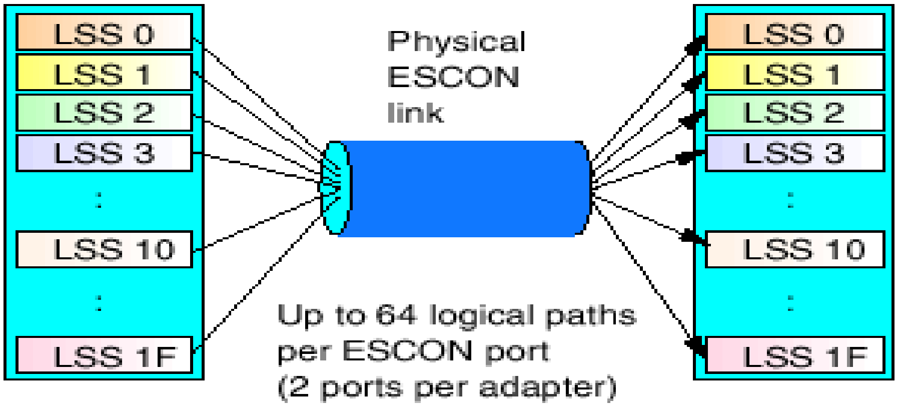


Figure 2: PPRC Connectivity using ESCON Links

The enhancements to ESCON protocol, implemented in ESS micro code as an advanced copy function, allow an extended distance between two ESS's of up to 103 km, when using multi-mode to mono-mode ESCON converters, amplifiers and switches. PPRC can be implemented over longer distances using channel extenders from third party suppliers certified on the ESS.

Up to eight ESCON links are supported between two ESS storage subsystems. The local primary storage control with PPRC source volumes and the remote secondary storage control are connected via ESCON links. One ESS storage control can act as primary and secondary at the same time if it has PPRC source and target volumes. PPRC links are uni-directional, as shown in Figure 3, so that a physical ESCON link can be used to transmit data from the primary storage control to the secondary.

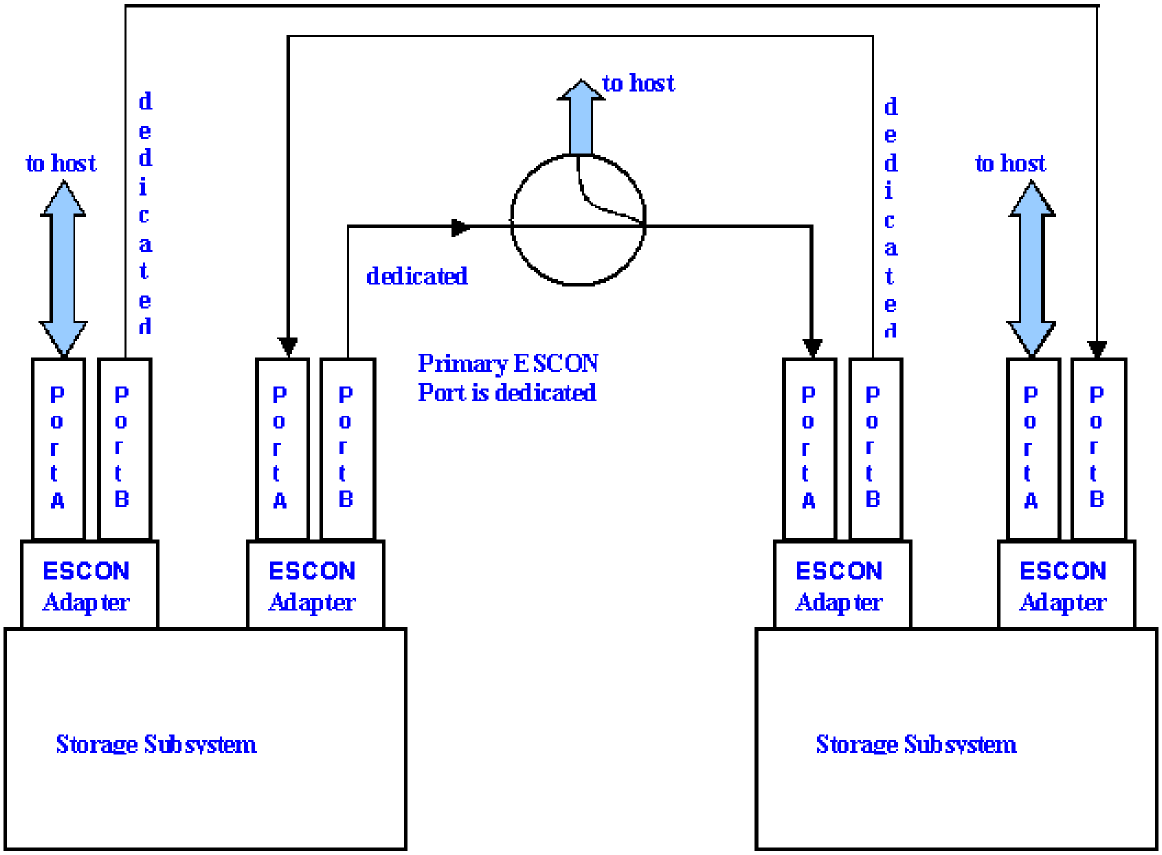


Figure 3: PPRC Links between two ESS Storage Sub Systems

Before PPRC pairs can be established, logical paths must be defined between the logical control unit images. The ESS supports up to 16 logical FB control unit images and up to 32 SCSI/Fiber controller images. Logical paths are established between control unit images of the same type over physical ESCON links that connect Suspending and Resuming Pairs.

During the suspension of a pair, the primary control unit maintains a bitmap in NVS (ESS's Non Volatile Storage located in each of its two symmetric multiprocessing complex that constitute its fault tolerant architecture) with a flag bit for each track that was changed on the primary volume. This allows for a later resynchronization (RESYNCH) of the volume pair while allowing only cylinders flagged in the bitmap table to be copied to the remote site.

As discussed in Section H of the Appendix, the ESS Specialist is a centralized Web-based storage management tool providing flexible user access for storage layout, customization, and task manipulation required for this SMBR solution. In addition to the Specialist, the ESS functions can be administered by the use of CLI at the host level.

As with FlashCopy tasks, PPRC tasks are created via the ESS Specialist and then invoked from the Unix command line using CLI interface. Based on the success of the query - rsExecuteQuery, the SMBR automation process can capture the error code for each of the PPRC tasks.

The features of Oracle and ESS in addition to SAP's requirements to support an automated, "lights-out" backup & recovery necessitate planning, set up and understanding of the underlying application usage and its performance requirements.

5 Split Mirror Backup & Recovery (SMBR) Setup

The SMBR set up involves the physical database design from the SAP installation guide. This includes file systems definitions according to sizing (based upon planned usage statistics, I/O forecasts, number of users etc.). The file systems requirements follow standard SAP installation procedures.

In a full-featured High Availability SMBR solution we recommend to use two physically separated Database hosts and two ESSs, each containing two copies of the production database. In our test environment we used two ESS clusters (refer to figures 6 and 7) like separated storage systems but without reservations this can simply be extended to Two ESS implementation. The installation binaries for SAP kernel and Oracle along with the SSQJ (a load testing tool developed by SAP) file systems are also setup for each of the four copies. SSQJ was developed with ABAP4 in R/3 for benchmarking based on SQL / ABAP statements and table manipulation for performance / throughput analysis.

The ESS LUNs design is based on logical addressing of striped physical volumes in a RAID5 array. The LUN definitions are based on file system requirements and are translated into volume groups via IBM's Subsystems Device Driver (SDD) mapping of virtual paths installed on the AIX host (see Section F of the Appendix).

5A SSQJ: R/3 Load Simulation and Testing Tool [12]

This Split Mirror Backup / Recovery solution utilized the SAP designed SSQJ tool for load simulation. SSQJ is a generic test measurement tool that was developed with ABAP/4 in core R/3. It enables testing and measurement of the functions in the R/3 Basis system such as runtimes of specific functions and their alternatives, resource consumption, query plans in the case of DB statements and other functions. SSQJ runs on all SAP-supported database systems with built in infrastructure for Measurement & Analysis including History of environment measurements.

A number of test suites are included in the current version of the package:

1) SQL statements
2) ABAP statements
3) Large Tables for performance and throughput analysis
4) Archiving Systems
5) Data Conversion
6) TPC/D - Benchmark and
7) Business Warehouse

SSQJ is currently used for SAP Basis Benchmarking, by SAP Database Porting teams, SAP Database Partners and other SAP Hardware Partners.

The SSQJ tool was used as the kernel database for the SAP/ORACLE SMBR testing. SSQJ requires the creation of two new tablespaces, PSAPSSQJD for data and PSAPSSQJI for indexes. The size of the SSQJ tablespaces depends on the required database size. For the initial installation, we require at least 8GB for PSAPSSQJD and 4GB for PSAPSSQJI. The data files for PSAPSSQJD are in /ORACLE/SSD/sapdata3 (total of 8 files of 2GB each), and the data files for PSAPSSQJI are in /ORACLE/SSD/sapdata4 (total of 4 files of 2GB each).

5B SAP / Oracle File System Definition

The file systems required for installing SAP system are created on the ESS volume groups. The volume groups LSS10data, LSS12data, LSS14data, LSS16data, Sapvg and Logvg as shown in Figure 4. First the logical volumes are created using ESS LUNs and then the appropriate options are chosen for the logical volumes under AIX using MIN/MAX policies. Finally the file systems are created on top of the existing logical volumes.

Volume Group	Filesystem Name	Function	Size (in 8MB Physical Partitions)
Sapvg	/usr/sap/SSD	Link to /sapmnt/SSD	0
	/sapmnt/SSD	SAP Executables	60
	/usr/sap/trans	Transport directory	60
	/oracle/SSD	Oracle Executables	130
	/oracle/stage/stage_806	Oracle staging area	80
	/oracle/805_32	Oracle Client	4
	/oracle/SSD/sapreorg	Temp data staging	236
Logvg	/oracle/SSD/origlogA	Online Redo Log files	50
	/oracle/SSD/origlogB	Online Redo Log files	50
	/oracle/SSD/mirrlogA	Mirror of origlogA files	50
	/oracle/SSD/mirrlogB	Mirror of origlogB files	50
	/oracle/SSD/saparch	Archive log files	470
LSS10data	/oracle/SSD/sapdata1	SAP R/3 data files	3810
	/oracle/SSD/sapdata2	SAP R/3 data files	3810
LSS12data	/oracle/SSD/sapdata3	SAP R/3 data files	3810
	/oracle/SSD/sapdata4	SAP R/3 data files	3810
LSS14data	/oracle/SSD/sapdata5	SAP R/3 data files	3810
	/oracle/SSD/sapdata6	SAP R/3 data files	3810
LSS16data	Not Assigned	For future data files	-
Backupvg	/basebackup	SAP backups	2800
	/usr/sap/trans/data	SSQJ data load	700

Figure 4: Oracle File Systems for SAP

As shown in Figure 4, the volume groups LSS10data, LSS12data, LSS14data and LSS16data are 64GB each. The other volume groups are sapvg and logvg comprising of 1GB LUNs. sapvg is used for the SAP and Oracle executables, and logvg holds the online and archive logs.

SAP delivers binary executables in order to help achieve online and offline backups as a part of the kernel installation. The external data management tools like Tivoli Storage Manager (TSM) have products that integrate into SAP R/3 as referred to in Section D of the Appendix. As shown in Figure 13, the SAP R/3 objects – Oracle control files, datafiles, archive log files and online redo log files are required for a consistent database backup and restore / recovery.

Directory	**Meaning**
/oracle/SSD/dbs	SAP and Oracle Profiles
/oracle/SSD/sapdata<n>	Datafiles
/basebackup	BRBACKUP, BRRESTORE logs
/oracle/SSD/saparch	BRARCHIVE logs, Oracle archive
/oracle/SSD/sapcheck	SAPDBA logs(-next, -check, -analyze)
/oracle/SSD/sapreorg	SAPDBA logs (default)
/oracle/SSD/origlogA	Online redo log files
/oracle/SSD/origlogB	Online redo log files
/oracle/SSD/mirrlogB	Online redo log files
/oracle/SSD/mirrlogA	Online redo log files

Figure 5: Oracle Directory structure in SAP R/3

Directory and file names are standardized in the R/3 environment as:

- Tablespace files reside in the sapdata<n> directories
- The online redo log files reside in the origlog and mirrlog directories (mirrored directories)
- The offline redo log files are written to the saparch directory

There should be at least three copies of the Oracle control file.

The profile init<SID>.ora configures the Oracle instance, and resides in directory /Oracle/SSD/dbs:

- The profile init<SID>.sap configures the backup tools BRBACKUP and BRARCHIVE, and resides in directory dbs
- The profile init<SID>.dba configures the SAPDBA tool, and resides in directory dbs
- The Oracle alert file is written to directory saptrace / background
- Trace files of the Oracle shadow processes are written to the directory saptrace / usertrace
- During reorganization, export datasets are written to the directory sapreorg
- The SAP database tools use the directories saparch, sapcheck, sapreorg, and sapbackup.

As shown in Section D (Topic-SAP R/3 Data layout for Oracle) of the Appendix, consideration must be given to sizing recommendations, SAP landscape, and data layout spread evenly across storage systems. ESS RAID5 Disk Layout Considerations, as detailed in Section E of the Appendix (ESS Raid5 Disk Layout Considerations for SAP R/3 Environments), ensure that the "hotspot“ phenomenon common in non-RAID5 environments are eliminated by striping all ORACLE tablespaces across all arrays, thus utilizing the cumulative throughput of all device adapters.

At this point we complete the installation of Oracle 8.0.6.3 and SAP Version 4.6B for AIX host joyjoy. The SAP / Oracle System / Instance name is “SSD” and the system number NR=00.

5C Split Mirror Infrastructure Setup

The SMBR live database, as shown in Figure 6, consists of DB_A for the SAP Instance “SSD “ on the primary ESS attached to host joyjoy. It contains all the Oracle binaries, SAP kernel, transport directory, Oracle Online / Archive log directories, Oracle instance directory, Oracle file systems and SSQJ file systems.

DB_C is the primary PPRC mirror of the live database DB_A. DB_B is the Safety FlashCopy copy of the live database DB_A. This safety copy is created in order to maintain a Point-in-time-copy of the live Production Database DB_A, should there be any problems encountered with DB_A during the SMBR operation.

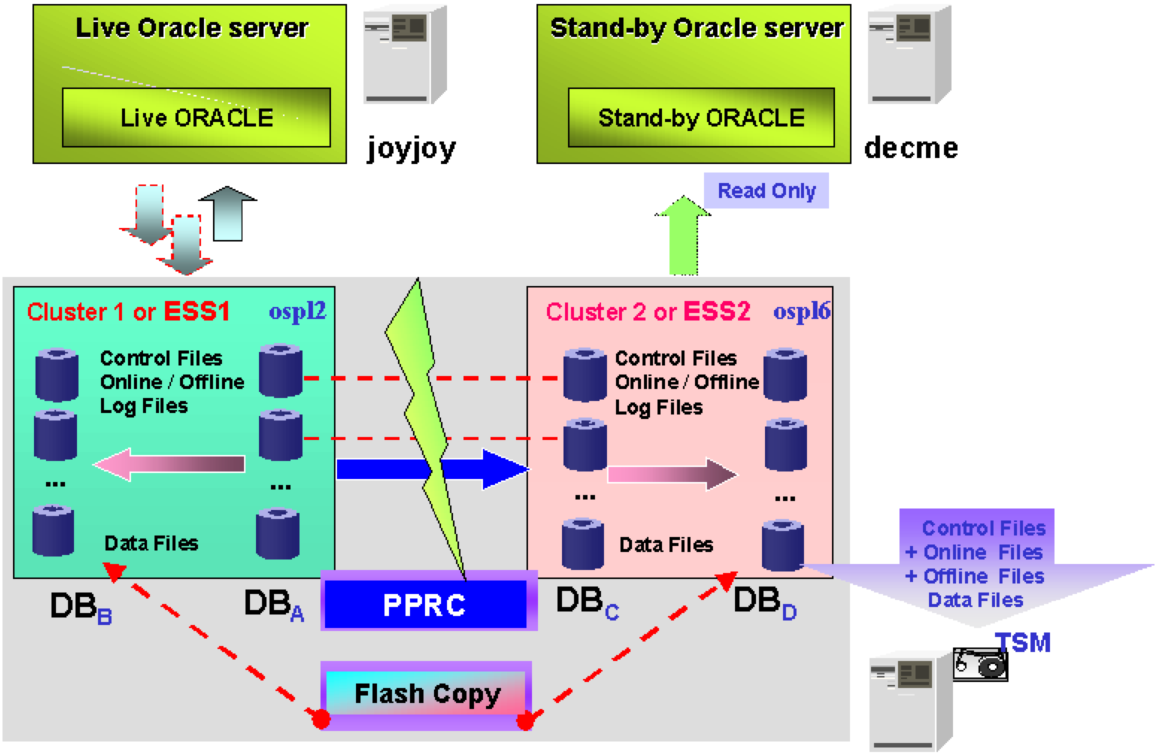


Figure 6: R/3 Split Mirror Backup & Recovery and Backup Host

In situations where the customers' BASIS or DBA group would like to mount the file systems for verification purposes by using the FlashCopy or PPRC target volumes on the same host, the Physical Volume Identifier (PVID)'s need to be renamed using the AIX RECREATEVG command as mentioned earlier in Section 4B.

5D ESS LUN Definition

For all references to LUN and volume group sizes, 1GB translates to 1*1000*1000*1000 bytes. All the LUNS for host "joyjoy" are defined only on cluster 1 for our SMBR solution scenarios using FlashCopy and PPRC. Each of the 8 ranks on Cluster 1 consists of twelve 8GB LUNS. In addition to these, two 1GB LUNs are also defined on each of the ranks. The 8GB LUNS are used for the Oracle data files, while one set of eight 1GB LUN is used for online redolog and archive saparch files. The other set of eight 1GB LUN is used for the SAP R/3 and Oracle binaries and other temporary storage.

OSPL2 and OSPL6 are the production and disaster recovery ESSs in our setup.

For the initial installation, four 8GB LUNs from each rank are used to layout the SAP R/3 Oracle data files. This leaves the other eight - 8GB LUNs for FlashCopy (safety copy) and PPRC.

As shown in Figure 7, there are a total of four LSSs, 4 on each cluster of an ESS. Each LSS on a cluster has two ranks assigned to it. There are two sets of four 8GB LUNs of an LSS that are assigned to one volume group. In our case, that translates to four volume groups of 64GB each, one for each LSS. The volume groups are LSS10data, LSS12data, LSS14data and LSS16data on ESS1 (OSPL2). Similarly, LSS11data, LSS13data, LSS15data and LSS17data volume groups are created on ESS2 (OSPL6).

The other volume groups are sapvg and logvg comprising of 1GB LUNs. sapvg is used for the SAP and Oracle binaries, and logvg holds the online and archive logs.

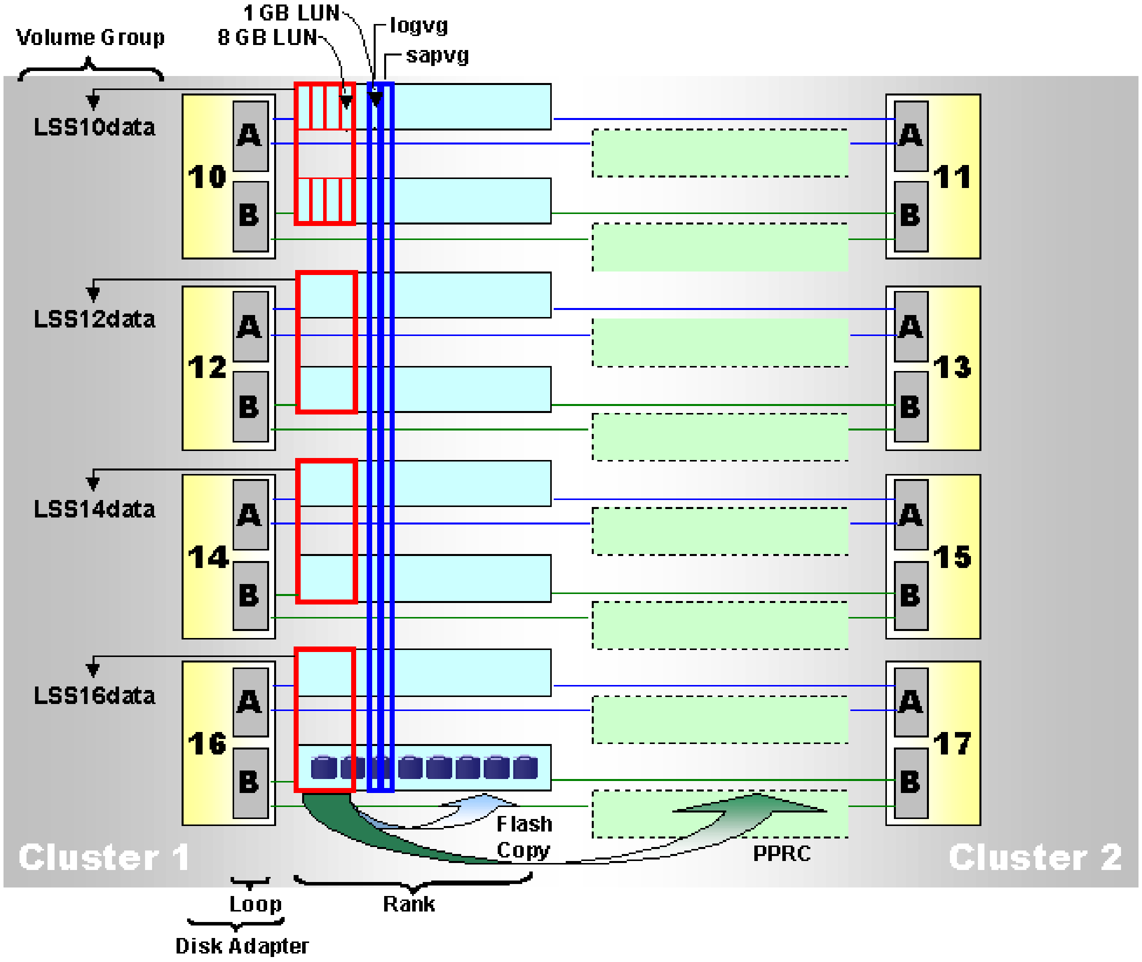


Figure 7: ESS Logical Sub Systems (LSS)

In an optimal database layout where the distribution of tablespaces follows the basic principle – spread the data over as many physical disk / arrays as possible [4, 6, 7] – Tablespaces should be extended by at least one full stripe set across all arrays.

This ensures that tablespaces remain distributed across a stripe set. Allocation of a specific number of arrays for a tablespace facilitates distribution and placement of the data files. The more arrays that are allocated, the better the distribution is. Data file distribution should be based on a round-robin placement across ESS clusters and arrays and the granularity can be achieved using AIX Logical Volume allocation.

In the SMBR Solution validation scenarios, the AIX Logical Volume Manager maps the assigned ESS 8GB LUNs as hdisks. Grouping the LUNs of a single RAID array can create volume groups. Under this traditional route for database layouts, AIX file systems are created over AIX logical volumes that reside on a single array. The data files of a tablespace are then distributed over file systems on different arrays. The criterion for a data file to be placed in these file systems is still the same - the overall I/O activity should be distributed across all available arrays. Creation of volume groups and logical volumes should be within the constraints imposed by the AIX Logical Storage Management.

5E Volume Group Assignment

Once the LUNs are defined on each ESS and assigned to hosts - joyjoy and decme, the LUNs are made available to joyjoy by running the 'cfgmgr' command on that host. The Subsystems Device Driver (SDD) – is a high availability automatic I/O Path Failover Function that provides management of active paths to the LUNs as outlined in Section F of the Appendix. The SDD software needs to be installed on the hosts before running cfgmgr.

The LUNs appear as hdisks on joyjoy. If a LUN is assigned to one path, an hdisk, let's say hdisk6 is defined on joyjoy. As joyjoy has four SCSI adapters, the LUNs are configured such that they can be accessed through all four paths. For each additional path, another hdisk is assigned. So for four paths, we have four hdisks all pointing to the same LUN on the ESS.

When SDD is used, additional data path devices called vpath devices are created. Each hdisk set (a set is based on the number of paths from the host to the LUN) is assigned a vpath device. In our example, vpath1 will consist of hdisk6, hdisk46, hdisk90 and hdisk126.

When a volume group is to be created on joyjoy, two options are available in AIX administration utility smitty:

1) Add a Volume Group, or
2) Add a Volume Group with Data Path Devices.

As we are using SDD, the volume groups are created with the second option.

Volume Layout for Oracle

The volume layout for SAP R/3 system in Oracle for ESS is depicted in Figure 17 (Logical Sub System vpath mapping to hdisks) of Section H of the Appendix.

6 Split Mirror Backup & Recovery Process

6A Start Situation of Split Mirror Backup Process

The live system is in normal READ / WRITE operation state. ONLY the Log Volumes (Online Redo Log Files) are in a constant synchronous PPRC connection between source DB_A and target DB_C across ESS1 and ESS2 as shown in Figure 8.

The constant synchronization of log volumes between ESS1 and ESS2 will ensure that the disaster recovery site is up-to-date with the transactional changes to the database, making any recovery from an apparent failure on the primary database simpler. Furthermore, possible data inconsistencies due to user or program errors are not immediately copied to the remote data files, but with the constantly synchronized Logs we are able to recover to any point in time.

The periodic PPRC re-synchronization of the data volumes during the SMBR activity will ensure that any structural changes (e.g. adding datafiles to tablespaces etc.) including all de-staged data are mirrored on the remote site.

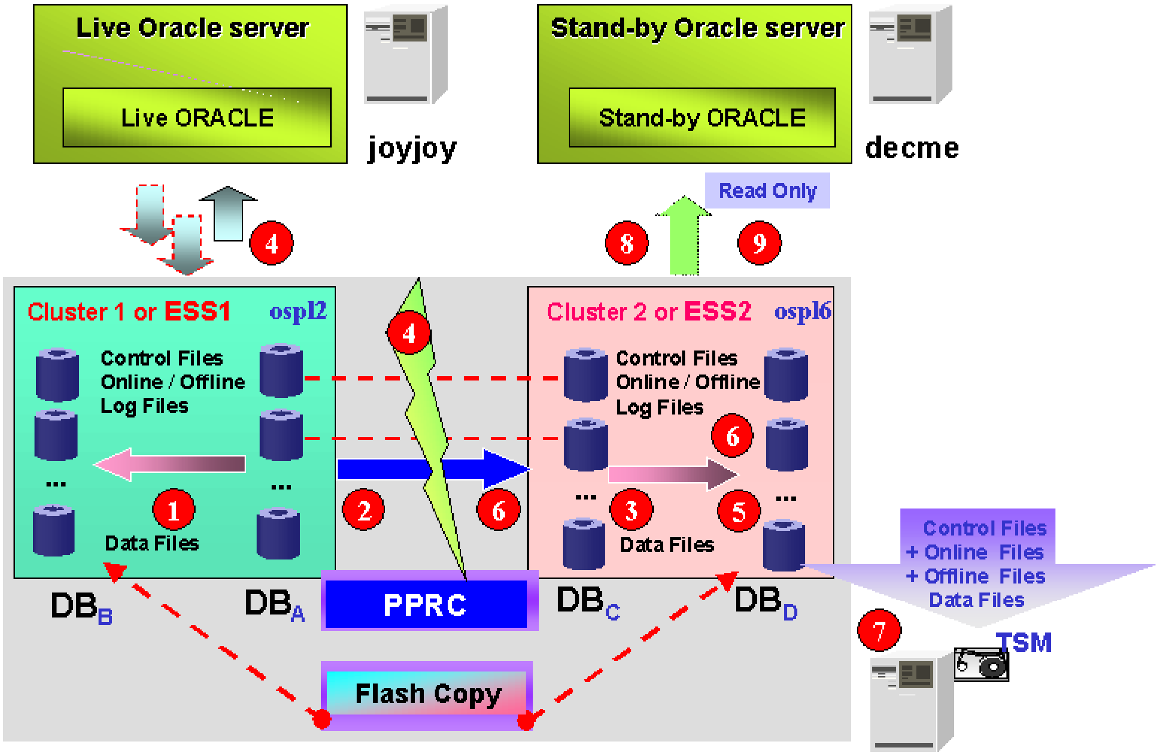


Figure 8: Split Mirror Backup & Recovery and Standby R/3 System

The SMBR process steps (nine) depicted in Figure 8 are listed below:

1: FlashCopy (FC) DB$_A$ to DB$_B$ (Safety Copy)
2: Resync DB$_A$ to DB$_C$
3: Withdraw Prior FC DB$_D$
4: Alter Tablespace Begin Backup on DB$_A$; Freeze PPRC pairs
5: Create Data Volume FC DB$_D$; Alter Tablespace End Backup; Switch Log File
6. Resync PPRC Log Volumes and Create Final FC
7: Backup DB$_D$ to EBT (Enterprise Backup Tool) / TSM
8: Mount FC Volumes to Backup Host "decme"
9: Start R/3 Oracle on Backup Host

After the prior SMBR backup activity, the DB$_D$ instance is in FlashCopy No Copy relationship with DB$_C$ while DB$_B$ is in FlashCopy No Copy relationship with DB$_A$.

DB_D volumes have been used to bring up a SAP instance after successful backup to tape. Using the SAP R/3 Homogeneous System Copy procedures for post copy BASIS administration tasks such as changing the RFC settings, locking user id's, TMS setup and batch job cancellations etc., will ensure that the actual production related activities are deactivated in the test SAP instance on DB_D volumes.

The first step in SMBR process would be to logoff all the users from the Production-Fix / reporting / standby instance i.e. DB_D (on the backup host decme) before stopping respective SAP and Oracle processes. The SAP / Oracle instance on the Safety FlashCopy target volumes DB_B in the primary ESS is also stopped.

The next step is to collect all the current layout details from within the ORACLE database through the dynamic SQL statement (PL/SQL) that can query the latest information on the database structure including the log group, tablespaces, and data file information. Dynamic SQL statements are embedded in the source SQL statement and are stored as character strings input to or built by the source program at run time.

During this step, the file structure information and volume group layout along with device list from the host operating system is obtained. This step will confirm that the current users in the SAP instance SSD on the decme host (probably the Production-Fix system users) are logged off and the Journaled File Systems (JFS) on decme will be unmounted. All the application related file systems, except those that pertain to the AIX host operating system on the decme host, are also unmounted.

General comments on the Backup Process:

1 Begin Phase:

- Check for status of FlashCopy relationships between DB_A & DB_B, DB_C & DB_D
- Verify if database on DB_C was opened and if data volumes between DB_A and DB_C are out of sync (to ensure minimization of error propagation)
- Check that the log volumes between DB_A and DB_C are in synchronous PPRC mode

2 FlashCopy Phase in ESS:

- Withdraw prior FlashCopy relationships between DB_A & DB_B
- For all the tablespaces issue HOT BACKUP commands
- FlashCopy DB_A to DB_B 'Data volumes and Oracle & SAP binaries' in no-copy mode (copy operation of data files, redo log / saparch takes place at different instants in time)
- End HOT BACKUP mode for all tablespaces. Issue log-file switch and control file copy commands
- FlashCopy DB_A to DB_B 'Log Volume group' in no-copy mode

3 PPRC and FlashCopy Phase between ESS1 / ESS2:

- If the suspension of PPRC between data volumes and log / saparch / sapreorg volumes occurs at different instants in time during this step, there is a requirement to execute switch log file / backup control file commands after the end of HOT BACKUP mode
- However, with the PPRC subsystem level 'freeze' command / operation, which is an atomic suspension of all PPRC disk-pairs, the requirement to do a recovery using backup control file / redo log files is eliminated
- FlashCopy DB_c to DB_D 'Data volumes and Oracle & SAP binaries' in no-copy mode (copy operation of data files, redo log / saparch takes place at different instants in time)
- End HOT BACKUP mode for all tablespaces. Issue log-file switch and control file copy commands
- At the end of hot backup process issue log switch / backup control file commands
- Resynchronize PPRC
- FlashCopy DB_C to DB_D 'Log Volume group' (Archive logs) in no-copy mode

4 Copy DB_D to Tape Phase using a backup utility:

- Mount DB_D volumes to host decme. Irrespective of FlashCopy option used (physical complete or no-copy mode), all the volume sets on DB_D are backed up to tape using a backup utility

5 Second Instance Phase:

- Start SAP instance on DB_D using 'SAPDBA -startup'. This verifies that SCN (System Change Number) is consistent with Datafiles and Control files. It also resets the data files to end backup mode and opens the database after rolling back any un-committed transactions between Begin Backup and End Backup modes
- Starting Oracle through 'SAPDBA -startup' command will verify that DB_D is a consistent copy
- Based on customer requirements, the second instance can be provided with a change in SID name using the output from Oracle's 'BACKUP CONTROL FILE TO TRACE' command output

6 New Backup cycle Initialization Phase:

In order to be prepared for a new backup cycle, the systems (DB_B, DB_C and DB_D) must be set back to the start situation

In this SMBR set up, the following tasks were created using ESS Specialist / CLI command sets:

FCAllNoBgCpWESS1: Withdraw previously held FlashCopy NoCopy relationship between FlashCopy source/target pairs DB_A and DB_B in ESS1.

FCAllNoBgCpWESS2: Withdraw previously held FlashCopy NoCopy relationship between FlashCopy source/target pairs DB_C and DB_D in ESS2.

FC1016NoBgCpESS1: FlashCopy source production Data volumes to target Safety copy Data volumes in LSS10, LSS12, LSS14, and LSS16 in ESS1

FC1117NoBgCpESS2: FlashCopy source production Data volumes to target Safety copy Data volumes in LSS11, LSS13, LSS15, and LSS17 in ESS2

FCLogVNoBgCpESS1: FlashCopy source production Log volumes in DB_A to target Safety copy Log volumes (saparch, sapbackup, sapreorg, log directories) in DB_B in ESS1

FCLogVNoBgCpESS2: FlashCopy source production Log volumes in DB_C to target Safety copy Log volumes (saparch, sapbackup, sapreorg, log directories) in DB_D in ESS2

EstALL4PATH: Establish PPRC paths and/or ensure that the paths exist between ESS1 (LSS10, LSS12, LSS14, LSS16) and ESS2 (LSS11, LSS13, LSS15, LSS17)

ResyncPPRCAll: Resynchronize PPRC source and target volumes in between ESS1 & ESS2

ResyncPPRCLogVs: Resynchronize PPRC source and target Log volumes in between ESS1 & ESS2

Frzall: Subsystem level withdrawal of paths and dataflow between PPRC source and target volumes

OSPL2C0 is the Cluster 0 in ESS1

OSPL6C0 is the Cluster 0 in ESS2

rsExecuteTask.sh is a CLI command to be initiated from AIX host

CLI command option -v provides verbose output

6B SMBR Implementation Steps

Each SMBR step of the backup process is described as follows:

Step 1: Create a Safety FlashCopy of the Production Instance

Create Safety copy of the production database DB_A

i. Withdraw previous Safety Copy source / target FlashCopy relationships on the Primary ESS1:

To Withdraw previous Safety Copy source / target FlashCopy relationships on the Primary ESS1, as captured in a Task "FCAllNoBgCpWESS1", the following CLI is initiated against OSPL2C0 from AIX:

```
rsExecuteTask.sh -v -s ospl2c0 FCAllNoBgCpWESS1
```

OSPL2C0 is the cluster 1 of ESS1 that hosts the Production Volumes.

FCAllNoBgCpWESS1 is the predefined task set up between source Production volumes and Target Safety FlashCopy volumes.

ii. Reads and writes to the Production SAP / Oracle database DB_A continue as usual on all the tablespaces while Oracle is put in HOT BACKUP mode using the following Oracle command:

```
ALTER TABLESPACE <PSAPBTABD> BEGIN BACKUP;
```

iii. Flash Copy all DB_A data volumes to DB_B volumes with No Copy option:

```
rsExecuteTask.sh –v –s ospl2c0 FC1016NoBgCpESS1
```

iv. Upon successful completion of FlashCopy, bring the database back to normal mode using Oracle's END BACKUP command. This ends the Online HOT BACKUP mode for Oracle after which checkpoints are completed and a "redo log switch" is initiated for each of the redo log files. The following Oracle commands are executed:

```
ALTER TABLESPACE <PSAPBTABD> END BACKUP;
ALTER SYSTEM SWITCH LOGFILE;
ALTER DATABASE BACKUP CONTROL FILE
                TO  /oracle/SSD/sapreorg/bakCntrlSSD.dbf;
ALTER DATABASE BACKUP CONTROL FILE TO TRACE;
```

The output of the ALTER DATABASE BACKUP CONTROL FILE TO TRACE command creates a trace file that can be used to change SID name of the DB_B volumes.

v. After FlashCopy of the Data volumes and the end of Oracle's HOT BACKUP mode, execute FlashCopy of the log volume group that consists of archive log files, sapbackup & sapreorg directories and redolog files:

```
rsExecuteTask.sh –v –s ospl2c0 FCLogVNoBgCpESS1
```

DB_B now contains the latest Archived log files from DB_A.

This step completes the tasks required to create a safety copy on the source ESS1 storage subsystem. This safety copy will enable us to recover the database with ORACLE recovery command executed on the DB_B volumes, should DB_A volumes become unavailable during this process.

vi. Extract all the system information pertaining to disk storage from within AIX on joyjoy and remote copy those files to host decme. This will help bring up an SAP instance on the DB_B volumes. In case of corruption to DB_A or DB_C volumes during the SMBR activity, the DB_B volumes can be used as a safety measure.

Step 2: Resynchronize DB_A to DB_C

i. Ensure that the ESCON Paths exist (and are active) between PPRC source (DB_A) and target (DB_C) volumes; 'establish' if they do not already exist:

```
rsExecuteTask.sh –v –s ospl2c0 EstALL4PATH
```

ii. Resynchronize all the PPRC source (DB_A) and target (DB_C) volumes (this includes log and data volumes):

```
rsExecuteTask.sh –v –s ospl2c0 ResyncPPRCAll
```

iii. Monitor percentage of Resync completion

Step 3: Withdraw previously held FlashCopy (source / target) Volumes on ESS2 after stopping the R/3 Application and Oracle Database on the Backup host

In ESS2, withdraw previously held FlashCopy No Copy relationship between DB_C and DB_D volumes from the previous backup run:

```
rsExecuteTask.sh -v -s ospl6c0 FCAllNoBgCpWESS2
```

(In the previous backup run, we made a safety copy DB_B in ESS1 and a final flash copy DB_D in ESS2.)

Step 4: Confirm that the Production Database tablespaces are in normal mode; Initiate HOT BACKUP mode; Suspend all the PPRC source (DB_A) and target (DB_C) volume pairs.

Put Oracle in HOT BACKUP mode for all the tablespaces (DB_A) using the following Oracle command:

```
ALTER TABLESPACE <PSAPBTABD> BEGIN BACKUP;
```

This will permit all write activity on DB_A to be written directly to datafiles so that DB_C (PPRC target volumes) can function as a consistent FlashCopy source for the final Split Mirror copy to be created on $DB_{D.}$

In order to create a consistent database copy (FlashCopy) DB_D, the PPRC relationship between DB_A and its mirror DB_C has to be frozen.

The entire subsystem level freeze activity of PPRC pairs between production - DB_A and mirror - DB_C will ensure that the data and log volumes can be suspended in one atomic operation, thus ensuring a consistent database copy. This atomic split capability of PPRC provides critical time saving functionality for high availability operations so that the Disaster Recovery site has the latest synchronous PPRC copy, enabling rapid restoration of this synchronous mirror for production purposes. The execution of the "Freeze" command at the subsystem level instead of the "Suspend" (another PPRC capability) command at LUN level speeds up the operations because it avoids issuance of multiple commands to individual LUN relationships between PPRC source and target volumes.

Thus the Freeze capability of the ESS subsystem enables the rapid creation of a consistent data base copy during the SMBR process.

(a) Freeze all the ESS subsystem level PPRC relationships between ESS1 and ESS2. This activity stops all data flow between paired data and log volumes:

```
rsExecuteTask.sh –v –s ospl2c0 FrzAll
```

(b) Verify the status of the task completion using rsExecuteQuery in a query loop. Obtain the error code and send an automated message upon successful task completion.

NOTE: Maintaining log volumes in continuous synchronous state between ESS1 and ESS2

If constant remote mirroring is an absolute customer requirement for mission critical environments, we recommend that the log volumes be kept in a constant PPRC state. ESS is able to perform a FlashCopy from a PPRC target to another volume without PPRC suspension. In this case, if we still want to use the Freeze command for data volumes, both data and log volume groups must be stored in separate LSSs (as shown in Figure 4).

In this SAP/Oracle implementation, because of the time lapse between the moment of PPRC freeze of all data and log volume pairs at the sub system level, and the completion of the logical FlashCopy (NoCopy) of the data volumes from DB_C onto DB_D, the remote vaulting of log volumes is momentarily suspended during this process. However, as mentioned above, if remote log mirroring via PPRC during a selected SMBR window is an absolute customer requirement, then use of the PPRC SUSPEND command (at the LUN level) only for the data volume pairs, as described in **Step 4**, is recommended.

Step 5: Create Data Volume FlashCopy DB_D; Alter Tablespace End Backup; Switch Log File

Proceed with the FlashCopy of data volumes. Initiate FlashCopy of the source DB_C to target DB_D in NoCopy mode:

```
rsExecuteTask.sh –v –s ospl6c0 FC1117NoBgCpESS2
```

Verify error codes for the FlashCopy. Upon error code of zero, bring the database back to normal mode using Oracle's END BACKUP command. This ends Online HOT BACKUP mode of DB_A after checkpoints are completed and a "redo log switch" is initiated for each of the redo log files. The following Oracle commands are executed:

```
ALTER TABLESPACE <PASAPBTABD> END BACKUP;
ALTER SYSTEM SWITCH LOGFILE;
ALTER DATABASE BACKUP CONTROL FILE
                    TO /oracle/SSD/sapreorg/FcCntrlSSD.dbf;
```

```
ALTER DATABASE BACKUP CONTROL FILE TO TRACE;
```

The output of the ALTER DATABASE BACKUP CONTROL FILE TO TRACE command creates a trace file that can be used to change SID name of the DB_D volumes, thus facilitating the instantiation of a secondary instance.

Step 6: Resynchronize PPRC source and target log volumes and Create Final FlashCopy

After successful completion of FlashCopy (logical copy) and Oracle's END BACKUP command on the Production instance, it is time to reinitiate the PPRC links between ESS1 and ESS2 (which were in Freeze mode in Step 4).

i. Initiate PPRC links and establish paths between OSPL2/ESS1 and OSPL6/ESS2:

```
rsExecuteTask.sh -v -s ospl2c0 EstAll4Path
```

ii. Resynchronize all the incremental redolog activity in ESS1 (as in the execution of Step 5 above) during the period of PPRC freeze:

```
rsExecuteTask.sh -v -s ospl2c0 ResyncPPRCLogVs
```

Keeping the Log & Data volumes in sync at different times and only synchronizing data volumes during the periodic SMBR activity will ensure that the Disaster recovery site (ESS2) has log volumes that can be either rolled forward or rolled backward as per the business requirement, in the event of a disaster.

After FlashCopy of the Data volumes (as in Step 5 above) and at the end of Oracle's HOT BACKUP, execute FlashCopy of the log volume group that consist of latest archive log files, sapbackup & sapreorg directories and redo log files:

```
rsExecuteTask.sh –v –s ospl6c0 FCLogVNoBgCpESS2
```

The FlashCopy of Data volumes completed in Step 5, followed by the FlashCopy of Log Volumes completed during this step, creates the Final Point-in-Time Split Mirror of Production Database.

As mentioned in Section D of the Appendix on the SPLIT BLOCK phenomenon, data file changes, because of update / delete activity in Oracle, will prompt for redolog changes that in turn are not captured in the control file for the SCN changes. As the PPRC subsystem level Freeze is an atomic operation, the control file on ESS1 is merely backed up and is not necessary for any type of recovery. The datafiles captured at the moment of PPRC freeze and log volume group captured after the logswitch/control file copy will provide a consistent database copy. However, during this finite period of SMBR activity, if there are any structural changes (such as adding datafiles, adding tablespaces etc.) made to the database, it is necessary to resynchronize the data volumes as in Step 2. For a successful SMBR completion, ensure that there are no structural changes made to the database during this period.

This step completes the tasks required to create a Split Mirror copy of the Production environment with Oracle's Online backup functionality.

Upon the completion of this step, all SAP production operations landscape is in normal state as at the start situation described in section 6A.

Optional Physical Flash Copy of DB$_D$ Volumes:

If there is a requirement for a complete physical copy of the final FlashCopy (logical copy) DB$_D$ volumes, initiate a FlashCopy with Copy option from DB$_D$ to DB$_E$ (not shown in the Figure 8) in ESS2 as an extension to Step 6.

i. Initiate Physical FlashCopy of DB$_D$ to DB$_E$ in Copy mode.

Now continue with the Steps 7, 8 and 9.

Remote copy all the file systems, volume group, and device lists from the source production host (joyjoy).

Step 7: Backup the Final FlashCopy volumes to a backup utility

At this point using SAPDBA / BRBACKUP / TDP for R/3, Legato or any compatible backup utility program, backup the DB$_{D}$ instance’s SAP / ORACLE objects to tape. We will always save all the DB$_{D}$ volumes to tape media by invoking complete file system backup or a backup utility.

Step 8: Mount FlashCopy volumes (on ESS2 / OSPL6)

Rediscover all the devices that are part of the defined volume groups at the host operating system level. Varyon all the volume groups and mount the file systems required for ORACLE & SAP on host decme.

Varying-on and mounting of all the relevant volume groups is done after rediscovering the physical volumes and remapping those to the virtual paths as seen by the Host operating system via the SDD.

Step 9: Start SAP / Oracle on the Backup Host for verification purposes

In order for us to verify that it is a consistent copy of the live database, this step is optional.

Startup Oracle in nomount mode and bring all the datafiles from BEGIN BACKUP mode to normal mode.

```
ALTER DATABASE DATAFILE ‘/oracle/SSD/sapdata1/btabd_1/btabd.data1’ END
    BACKUP;
```

Startup Oracle using "sapdba -startup" or by server manager.

Startup SAP after checking for Oracle error logs.

Changes to the SMBR Solution for Oracle 8.1.6 and later releases:

Subsequent to our SMBR solutions validation on Oracle 8.0.6, the latest features from Oracle for the 8.1.6 and later releases provide the following two important commands:

a) ALTER SYSTEM SUSPEND: Suspend or stop all I/O to data and control files and restrict access to the database. All ongoing I/O operations are allowed to complete but all new / incoming database accesses are placed in a queued state. This command ***should be preceded*** by the ALTER TABLESPACE <PASAPBTABD> BEGIN BACKPUP command.

b) ALTER SYSTEM RESUME: This command makes the database available for resumption of queries and I/O activity. This command ***should be followed*** by ALTER TABLESAPCE <PASAPBTABD> END BACKUP.

With the availability of these two key functions, the suspension / freeze of the PPRC pairs between ESS1 and ESS 2 will no longer be necessary in our SMBR process.

Accordingly the following changes will be required to the steps in **Section 6** (**Split Mirror Backup & Recovery Process**) of our SMBR process:

In **Step 1** item ii, execute command ALTER SYSTEM SUSPEND ***after*** the ALTER TABLESPACE <PASAPBTABD> BEGIN BACKUP command.

In **Step 1** item iv, execute ALTER SYSTEM RESUME ***before*** ALTER TABLESPACE <PASAPBTABD> END BACKUP command.

In **Step 4**, because of this new SUSPEND feature of Oracle 8.1.6, the PPRC Suspension or Freeze steps are NOT necessary.

Execute ALTER SYSTEM SUSPEND ***after*** ALTER TABLESPACE <PASAPBTABD> BEGIN BACKUP command.

In **Step 5**, execute ALTER SYSTEM RESUME ***before*** ALTER TABLESPACE <PASAPBTABD> END BACKUP.

Execution of **Step 6** is not necessary.

The updated version of SMBR for Oracle on AIX with the ESS will incorporate Oracle's new SUSPEND / RESUME features in detail.

7 Recovery

The backup process described above will provide a consistent copy of the live database. This backup file set will be on tape media and on the DB_D volumes of ESS2 (OSPL6).

7A Recovery Considerations for Split Mirror Backup Process

The database recovery is performed in several steps:

i. For a recovery process, the tape backup needs to be first restored on to the DB_D volumes on ESS2 / OSPL6 and then the DB_D volumes can be FlashCopied on to the PPRC volumes (DB_C) on ESS2 / OSPL6. The recovery steps further involve executing PPRC from source ESS2 / OSPL6 to the target volumes on ESS1 / OSPL2 if DB_A rebuild is required.

ii. Depending on the Service Level Agreements (SLA) consistent with Customer Landscapes and Requirements, we can repeat the SMBR steps in reverse order of execution (Steps 1 through 9 above). This will ensure recovery using a consistent database copy from tape media or from DB_D volumes on ESS2 / OSPL6. It will also provide a completely recovered SAP SSD instance on ESS1 / OSPL2.

iii. During the roll forward process, all committed transactions are reapplied to the database. At the end of the roll forward, a list of open transactions exists. These transactions are still awaiting a commit. For data consistency reasons, open transactions will now be rolled back, to make sure that any un-committed changes are not applied to the database.

While building the recovery scenarios, based upon customer requirements, ESS Specialist-based tasks should be created to execute the recovery processes.

7B General Considerations for SAP / Oracle Recovery Scenarios

The backup process described in Section 6 above will provide a fuzzy copy of the live database, which will be required to be made consistent by the application of the archive logs generated during the period of Oracle's HOT BACKUP state. This backup file set will be on tape media and on the DB_D volumes of ESS2 (OSPL6).

An AUTORECOVERY process for SAP R/3 using ORACLE brings the database to the latest consistent state based on the disk image and the online redo log files. This recovery uses entries in the control file and the online redo log files. If a disk error occurs (in a non-RAID5 sub system) and the contents of the disk are not recoverable, AUTORECOVERY will not work. This applies to the data files and online redo log files of the database. Therefore, as discussed in Section e of the Appendix, "ESS Raid5 Disk Layout Considerations for SAP R/3 Environments", using ESS's RAID5 striping, provides a highly available disk subsystem to protect log files and data files.

The contents of DB_D and DB_C volumes of the SMBR process differ from each other in that the restart of DB_D SAP instance (SSD) will roll back all the transactions that were open at the time of HOT BACKUP mode. Also, because of the RESYNC state in Step 6 of the SMBR process (in Section 6), DB_C and DB_A volumes are a pair of synchronous log mirrors and differ from the DB_D (Point-In-Time copy) volumes.

The DB_D volumes can be used for multiple purposes such as:

- Database Validity Checks
- Database Reorganization dry-run tasks
- Testing SAP support packages
- Training
- Applying of OSS Notes

- Production-Fix system
- Reporting Instance

The DB_D FlashCopy volumes can be also set up to serve as a Standby Database by:

a) mounting the volumes to host decme

b) placing the DB_D in roll-forward pending mode and then roll-forwarding the mirror.

In order to create a second instance to be accessed by SAP R/3 Test Instance the following is suggested:

- Generate a control file script by issuing ALTER DATABASE BACKUP CONTROLFILE TO TRACE right after ending hot backup mode on production database on OSPL2.
- Let the second host decme gain access to the FlashCopy image DB_D in ESS2 by mounting it.
- Make the FlashCopy image area write-able by the second host decme.
- If the second host decme accesses the datafiles in the FlashCopy area with a path different from the primary host, edit the control trace file so that it points to the right data files.
- Create control file with resetlogs option, and recover the database by applying the archived redo logs so that the database has cleared the HOT BACKUP fuzziness.
- Open database with resetlogs option.

8 Process Automation - Solution Integration in Customer Environments

In establishing this SMBR solution for R/3 with SSD and ESS, IBM has created end-to-end automated procedures that will enable this solution to seamlessly integrate into customer environments. While recognizing that each customer will likely implement this solution uniquely, the procedures incorporate use of platform-independent java technology, which will execute the customized set of Split Mirror routines for a particular backup environment. Inherent benefits of Java technology implementation will accrue to the customer in terms of flexibility, security, and location transparency. With this standards-driven approach, customers can integrate this solution with customer-preferred enterprise solutions management consoles (TME10, OpenView, BMC etc).

Acknowledgements

We wish to thank Siegfried Schmidt, and Peter Pitterling, of SAP's Advanced Technology Group (ATG) for their valuable guidance and for sharing SAP's vision and requirements with us to accomplish the work that is reported in this paper. Siegfried's earlier work with us on the Split Mirror Backup & Recovery solution implementation on the DB2OS390 platform with IBM's RVA & ESS and DB2 UDB / AIX on ESS storage subsystems have paved the way for this current implementation with Oracle on the AIX platform.

We would also like to thank a number of IBM colleagues with whose thoughtful guidance, this work is brought forward with the benefit of their applied knowledge in R/3, Oracle and ESS, in delivering high availability, zero downtime, and easy to use Split Mirror Backup and Recovery Solution for mission critical SAP customer environments. Bill Micka, David Short, Murali Godavari, and Sathees B. Kodi have been most helpful with their knowledge of the ESS advanced functions and applied knowledge of Oracle / AIX and have endured this journey with us.

Sanjoy Das,
sanjoy@us.ibm.com

Siegfried Schmidt,
siegfried.schmidt@sap.com

Peter Pitterling
Peter.pitterling@sap.com

BalaSanni Godavari
godavari@us.ibm.com

Appendix

A) Oracle Architectural Overview

An Oracle database system has three processes that write information from the Shared Global Area (SGA) to the appropriate files:

- To accelerate the writing of checkpoints, the Checkpoint Process (CKPT) is started.
- During a checkpoint, the database writer (DBWR) asynchronously writes the changed data blocks from the SGA to the database data files.
- The logwriter (LGWR) synchronously writes the changed log records from the SGA redo log buffer to the currently active online redo log file.

In a production database environment, the database system must always run in ARCHIVELOG mode. An archiver (ARCH) process archives a completed online redo log file into the offline redo log file in the archive directory - /oracle/SSD/saparch

In addition, when an Oracle database instance is started, several other processes are created: SMON, PMON, RECO, and LCK.

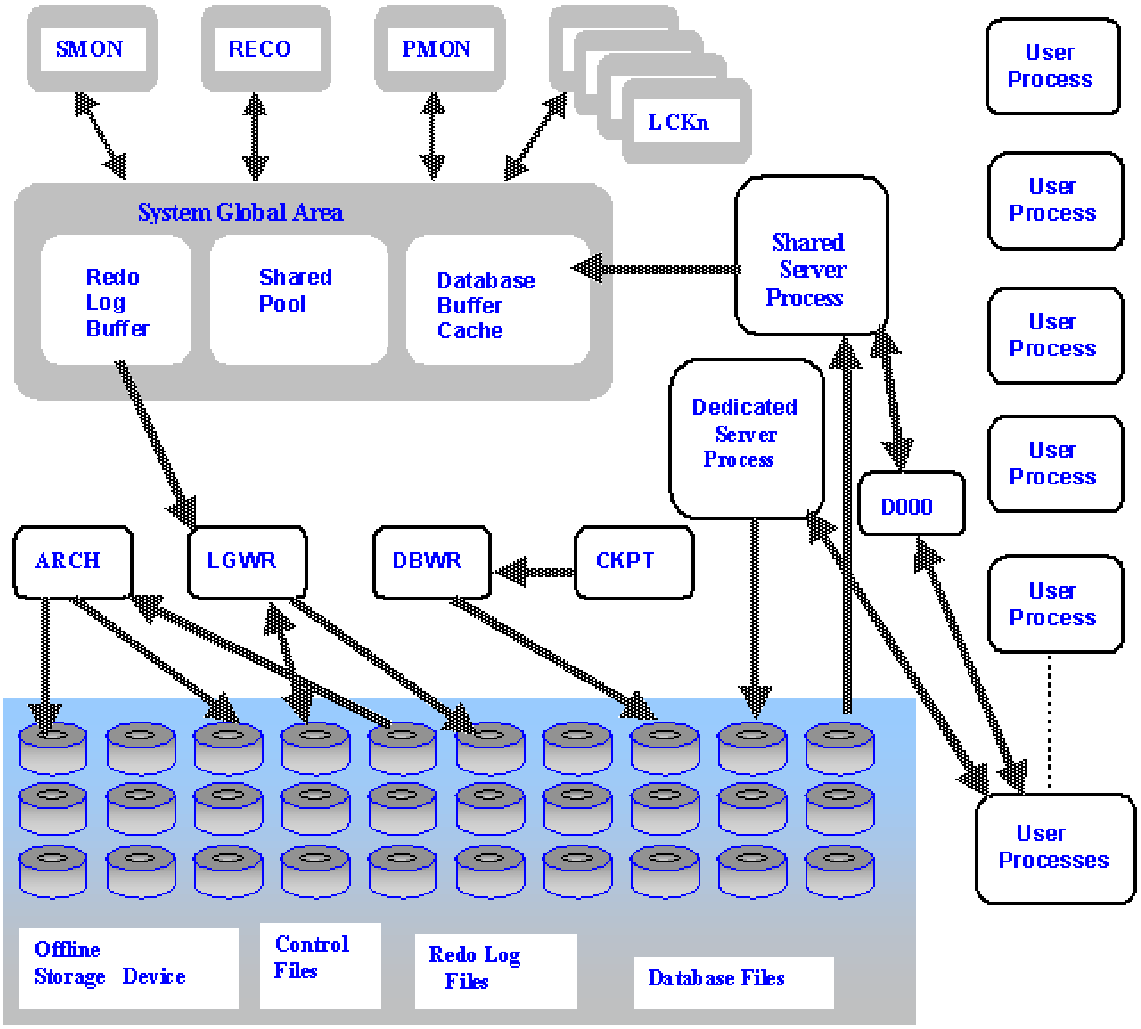


Figure 9: Oracle Overview for SAP R/3

Legend:

SMON : System Monitor	**RECO : Recoverer**
PMON : Process Monitor	**LCK : Lock**
CKPT : Checkpoint	**ARCH : Archiver**
LGWR : Log Writer	**DBWR : Database Writer**

The Oracle database is stored in 4 / 8 / 16KB blocks in data files on disk. In order to accelerate read / write access to the data, these data blocks are cached in the database buffer pool in production CPU main memory. The Oracle database management system holds the executable SQL Statements in the Shared SQL Area, which is part of the shared pool.

SAP/Oracle Work process management:

Each SAP R/3 Dialog, Batch, Update, and Spool work processes in an SAP instance use TCP/IP and TNS listener to:

- Connect to the database as one SAP R/3 user
- Handle database requests for the different R/3 system users
- Communicate with a corresponding shadow process on the database

Oracle's Dedicated Shadow Processes are created when a new user establishes a session with the R/3 system. The Shared Processes are required by the DBMS systems to function and perform various database management tasks.

The various Oracle processes that impact disks are depicted in Figure 10:

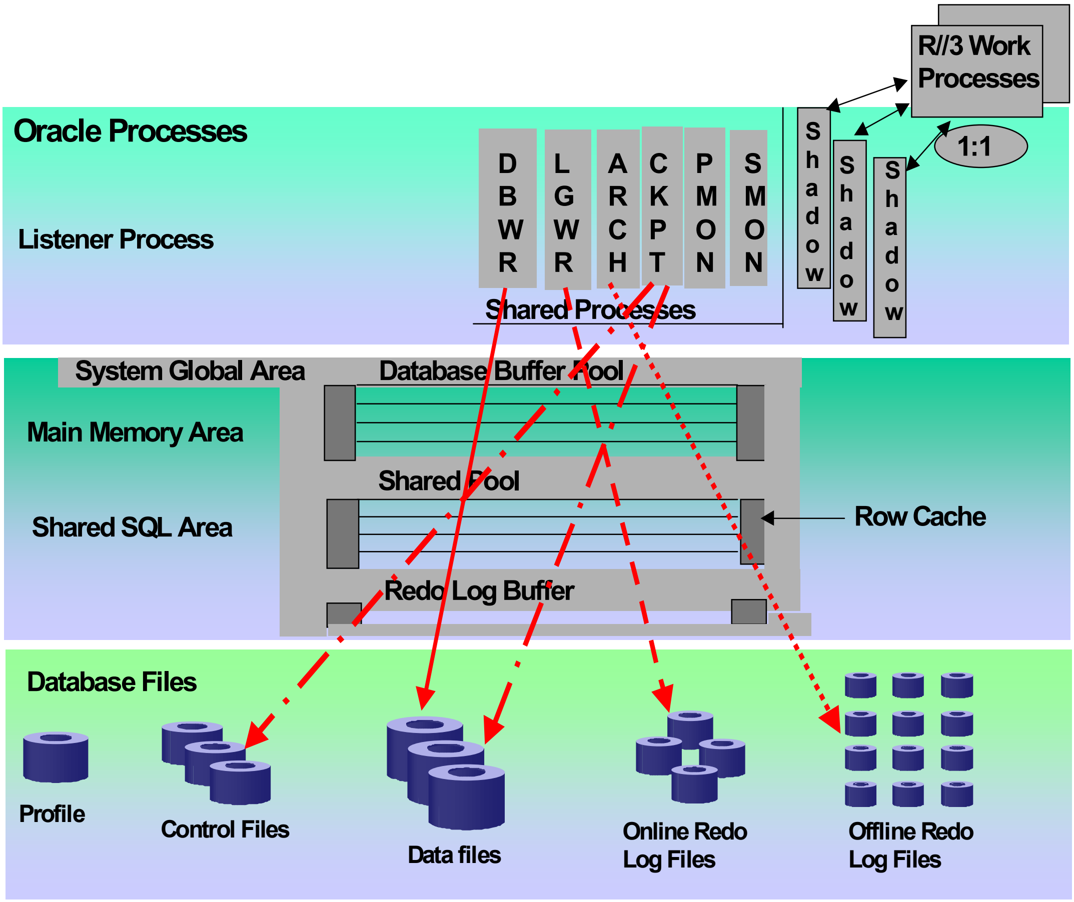


Figure 10: Overview of Oracle 'write' process to Disks

B) Oracle Database Growth and Impact on Storage

The Oracle database uses tablespaces, which hold the database objects, such as tables and indexes. On disk, a tablespace consists of one or more data files. Adding datafiles to it can increase the capacity of a tablespace. The R/3 naming convention for tablespace names is PSAP<tablespace_name><extension>. The abbreviations in the tablespace name are part of the data file name.

Depending on the size of each data record, several data records can be stored in an Oracle data buffer. Within a tablespace, the extents belonging to different data objects compete for storage space. SAP R/3 for Oracle on non-RAID5 storage systems separates data and index objects into different tablespaces.

In non-RAID5 storage architecture implementations, the emphasis is to distribute data and index tablespaces to different physical devices, thus improving database performance. With ESS's robust RAID5 architecture and its use of striping different types of data on all RAID arrays, the performance hot spots are eliminated while I/O performance improves through parallel access to all members of the array.

Oracle stores tables and indexes in individual data blocks. When new storage space is required for a table or an index, one or more contiguous data blocks of a data file are allocated to form an extent. Each table and index is assigned to a tablespace, which consists of one or more data files at the operating system level. All table and index data is stored in the data files of the tablespace.

Several storage parameters influence the growth of Oracle data objects. During installation of an R/3 system, while creating an SAP table or index, the default storage parameters - INITIAL, NEXT and MAXEXTENT values are used. The first extent (INITIAL EXTENT) should be large enough for the expected table or index size. If an extent of a data object becomes full during an insert or update operation, the Oracle storage management system attempts to allocate another extent in the tablespace.

The growth of an object, such as a table, index, or rollback segment, is determined according to the size specified by the NEXT parameter.

Oracle allocates extents to an object up to the limit of the MAXEXTENTS parameter. When the maximum number of extents for an object is reached, the MAXEXTENTS parameter can be increased after checking the size of the table's NEXT parameter and setting the AUTOEXTEND parameter to true. Under special circumstances, depending on the nature of the tables (static or dynamic) three additional storage parameters PCTFREE, PCTUSED, and PCTINCREASE are applied.

Tablespace Reorganization addresses the problem of fragmentation to optimize the amount of storage space. Extents are merged together to reduce the number of extents in the database and some data files are merged together to reduce the number of data files in the database.

Online Reorganization should be done during light transaction load on the tables to be reorganized. However, very large tables should be reorganized using export / import functionalities of Oracle. Block usage for the reorganized object will be optimal after a table or index is reorganized. Additional storage space is required to perform a re-organization and to store intermediate data, in database or in /oracle/SSD/sapreorg directory.

Performance Considerations in Database Layout

Customer experiences have shown that poor database performances at SAP installations often result from high I/O wait times for database access. I/O contention in a non-RAID5 environment can be traced to:

- Inefficient application design (expensive, unnecessary, and poorly qualified statements).
- Data not evenly distributed across many disk cylinders.
- Heavily accessed tables or indexes not distributed or striped across many disks.
- Oracle database and file objects (tablespaces and directories) not optimally laid out, thus causing numerous Oracle threads or processes to access the same disks as shown in
  Figure 11.

In the past, the database layouts in non-RAID5 environments, have often tried to separate tablespaces onto different devices, e.g., separating data and index tablespaces and putting each data tablespace on different drives. The consequence was that the database often ended in hotspots on devices that carried a single tablespace. Manual intervention was necessary in each case to alleviate the problem. Additional performance degradations occur over time due to changing application profiles and data growth leading to more hotspots and restructuring demands.

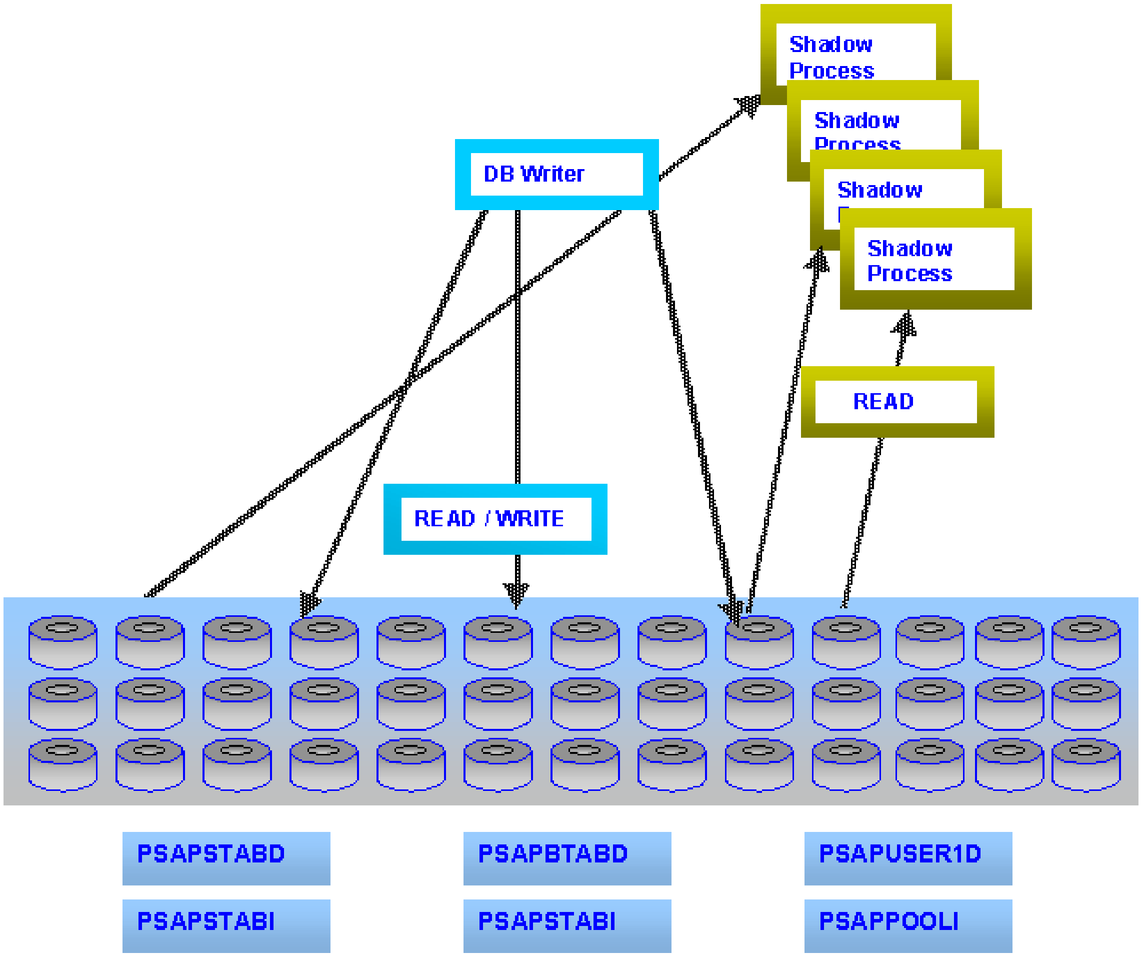


Figure 11: SAP R/3 and Oracle Read/Write Access - Impact on Disk Arrays

With advances in storage subsystem technologies and in particular advanced RAID5 implementations such as in the ESS, applications can now experience significant performance improvements, high availability and recoverability.

C) Oracle Memory Management

As shown in Figure 12, Oracle's basic memory structure includes:

- The System Global Area (SGA) which contains:
    - The Database buffer cache (enabling logical reads)
    - The Redo log buffer
- The Shared pool which contains:
    - Shared SQL Area where parsed SQL statements are cached for shared access by shadow processes
    - Row cache which holds the Oracle Data Dictionary information
    - Program Global Area (PGA) and Sort Areas

A user call refers to a shadow process which accesses a shared SQL statement. A recursive call refers to the Row Cache making a physical read to load Oracle Data Dictionary objects from the system tablespace. To increase the entire shared pool (Row Cache + Shared SQL area), the parameter SHARED_POOL_SIZE in INIT.ORA is tuned.

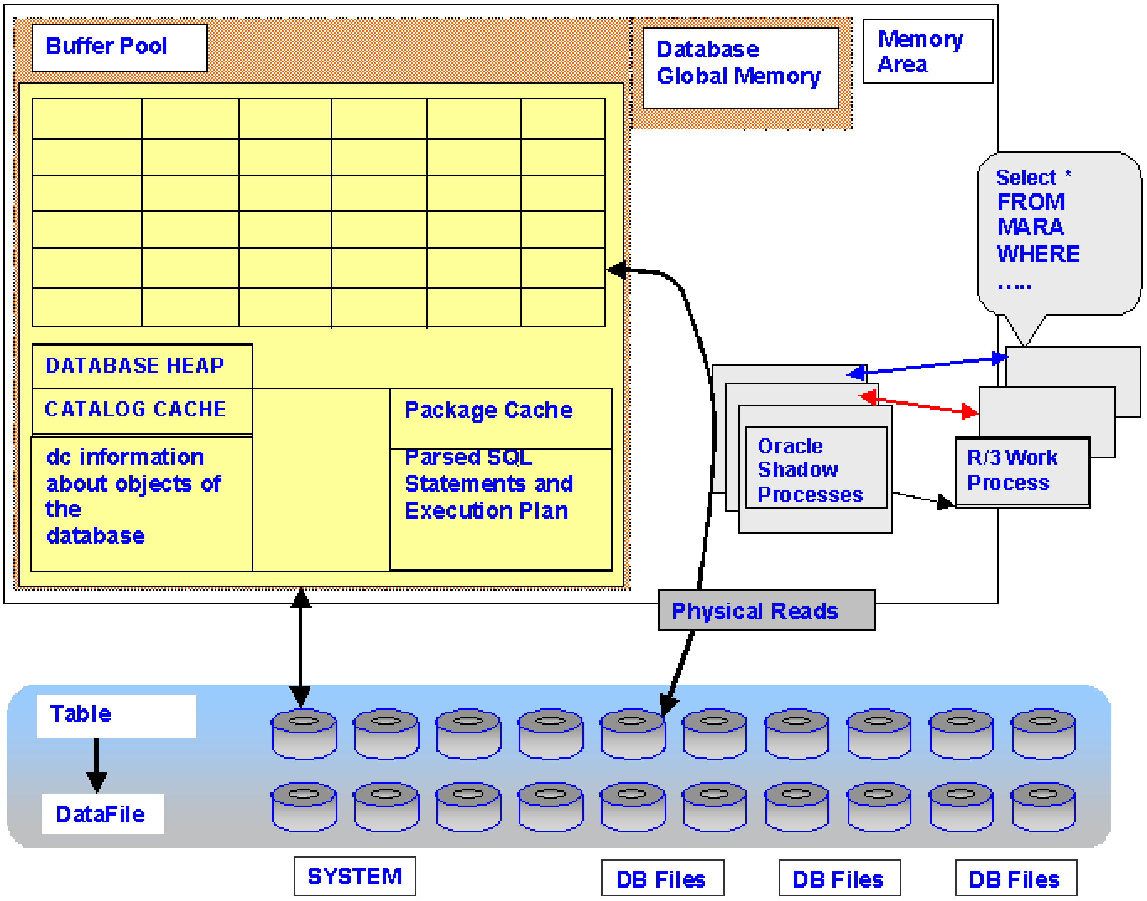


Figure 12: Oracle Memory Management for each SAP work Process

D) Backup and Recovery Management for Oracle in SAP R/3 Environments

The type and frequency of backups determine the speed and success of the recovery process. Various backup methods exist today. The DBA needs to determine the kind of backup procedures that are required for a particular implementation. This section gives an overview of various backup types commonly used for SAP R/3 databases.

Backups can be broadly categorized into physical backups and logical backups. A physical backup is a backup where the actual physical database files are copied from one location to the other (usually from disk to tape). Operating system backups, backups using Recovery Manager, cold backups, and hot backups are examples of a physical backup. Logical backups use Oracle's EXPORT/IMPORT functionality.

SAP R/3's kernel executables - BRBACKUP and BRRESTORE can be used to manage database backups and log file backups. During a database backup, the data files, the online redo log files, profiles, and control file are backed up.

Internal Operation of Hot Backup [10]:

During an Online backup, the Oracle database and the R/3 System remain available. Oracle automatically writes to the next online REDO log file in a round robin fashion while the HOT BACKUP is in progress. Owing to continuous Data Manipulation Langurage (DML) activity, Online backups as referred to in Section 4A and as detailed below, cause a reduction in system performance.

At the ALTER TABLESPACE BEGIN BACKUP command, the file header's SCN is advanced to the SCN number captured when this command is issued. The checkpoint SCN in the backup files must be the same as when the backup started. After the initial checkpoint, succeeding checkpoints cease to update the file headers when in HOT BACKUP mode.

When an ALTER TABLESPACE BEGIN BACKUP command is issued, the data files that belong to the tablespaces get flagged as HOT-BACKUP-IN-PROGRESS. This command checkpoints all the data files that are in HOT BACKUP mode. The checkpointing process that runs during the execution of the ALTER TABLESPACE BEGIN BACKUP command flushes all the dirty buffers that belong to the data files in the tablespaces and ensures that only blocks that are changed during the HOT BACKUP are written to Redo Log file.

The above activity also points to the important fact that depending on the I/O activity, excessive redo logs can be generated for data files during HOT BACKUP mode as Oracle records the activities in the REDO logs.

The phenomenon of SPLIT BLOCKS arises during the online backup. Depending upon how the operating system (AIX) copies blocks, it is possible for a HOT BACKUP to contain an inconsistent version of a given data block. This inconsistency arises due to the difference in the operating system and Oracle block sizes. By checkpointing and logging the before image of a data block to the redo log file before the first change, it can be used later to reconstruct a fractured block during recovery.

To verify the consistency of the block before recovery, Oracle compares the version number at the beginning of the block to the version number at the end of the block to determine whether the block has been split during a hot backup. The before image of the block in the redo log is copied to disk before applying the redo changes.

In order to make a consistent database copy as mentioned in section 4A, Oracle is put in HOT BACKUP mode using the ALTER TABLESPACE BEGIN BACKUP command on all the tablespaces on the primary system. In this mode, Oracle does not stop writing to the datafiles (it actually allows continued operation of the database almost exactly as during normal operation).

The database files are read by server processes and written by the database writer (DBWR) throughout the backup, as they are when a backup is not taking place. The only difference manifested in the open database files is the freezing of the checkpoint SCN, and the incrementing of the HOT-BACKUP SCN.

The following actions take place:

1. Oracle checkpoints the tablespace, flushing all changes from shared memory to disk while the SCN markers for each datafile in that tablespace are "frozen" at their current values.

2. While further updates will be sent to the datafiles, the SCN markers will not be updated until the tablespace is taken out of backup mode. Oracle switches to logging full images of changed database blocks to the redo logs.

3. To avoid inconsistencies in block images as mentioned before about SPLIT BLOCK technique, it would log the entire image of the block after the change instead of recording the change vector that could show how it changed a particular block.

4. The SMBR works through this datafile, backing it up at volume / LUN level. After all the tablespaces have been put in HOT BACKUP mode, certain CLI commands are executed to activate the ESS's advanced functions such as FlashCopy, a near-instant local volume copy function. Upon the execution of this command, Oracle's HOT BACKUP mode is terminated by the execution of the END BACKUP command, which returns database to normal state.

5. After an online backup is finished, if using the SAP backup management tool BRBACKUP, a log switch is initiated This creates a new online redo log file and the Oracle Archiver process in turn copies the previously active online redo log file directly into /oracle/<SID>/saparch directory. The entire log information generated during the online backup is contained in the offline redo log files.

6. When an attempt to start the SAP/Oracle instance is made on the "decme" host, Oracle looks at the data file and sees an old SCN value - the SCN marker it had before the hot backup began. When a "recover" command is applied, Oracle begins to apply redo logs against this data file. Since the redo logs contain a complete image of every block that changed during the backup, Oracle can rebuild this file to a consistent state.

Upon the execution of a full Online backup, the following files are backed up:

- The data files of all tablespaces belonging to the Oracle database
- The control file
- The profiles (init<SID>.ora, init<SID>.sap, and init<SID>.dba)
- Archive log files
- Contents of /oracle/SSD/sapreorg, /oracle/SSD/sapbackup directories

Logical Backups

During a logical backup of an Oracle database, using Oracle EXPORT utility, the data in the database is copied but the location of the data is not recorded. This utility copies the data and the database definitions or the Meta data and saves them to a binary OS file in Oracle internal format. While a logical backup takes longer to complete than a physical backup, it is used to detect data block corruptions, restore tables accidentally lost by users and reduce database fragmentation.

Using Oracle's IMPORT utility, the data is restored back from the binary Oracle internal format into the Oracle database format. Thus Logical backups provide a very flexible tool to migrate an SAP / Oracle database from one machine / platform to another. Logical Backups using EXPORT / IMPORT features are widely used by SAP R/3 to create homogeneous system copies.

SAP R/3 Backup Objects

SAP delivers the following tools for performing database backups:

- The program BRBACKUP backs up the data files, the control file, and the database redo log files where necessary.
- The program BRARCHIVE backs up the offline redo log files of the database.
- Both BRBACKUP and BRARCHIVE record the actions performed in log files.
- These log files can be used in case of a database restore, and can be analyzed by the program BRRESTORE. This program can restore all files belonging to the database system from the backups
- The database backup tools support standard backups (disk and tape).

External data management systems like Tivoli Storage Manager's Tivoli Data Protector for R/3 or Legato's Networker enable backup and recovery procedures to be integrated with the existing operations of an SAP R/3 landscape. These systems improve backup throughput and allow easy management of large number of tapes.

Overview of Recovery Issues in SAP / Oracle Environments

Every recovery option for SAP R/3 using Oracle is very important and has its own use, and it is crucial that DBA's understand how each recovery option works. Recovery processes vary, depending on the type of failure that has occurred, the structures that have been affected, and the type of recovery that is required.

On the recovery side, some recovery procedures require DBA intervention, whereas other internal recovery mechanisms are transparent to the DBA.

Block-level recovery is the simplest type of recovery, and is automatically done by SAP / Oracle. It is done when a process dies just as it is changing a buffer. The online redo logs for the current thread are used to reconstruct the buffer and then write to disk. If a process dies abnormally while modifying a block, SAP / Oracle will do a block-level recovery, which is automatic and doesn't require human intervention.

On the other hand, if a data file has been lost, recovery requires additional steps. Some of the common errors or failures include the following [10]:

- User error
- Statement failure
- Process failure
- Network failure
- Instance failure
- Media failure

A user deleting a row or dropping a table is a typical example of **user error**. The recovery procedure might be as simple as importing from a logical backup, or might involve a more complicated procedure such as doing point-in-time recovery from a physical backup (DB_D) on the backup host, exporting the table, and, finally, importing it into the production database.

A **statement failure** occurs when SAP/Oracle statements are selecting from a table that doesn't exist and trying to do an insert resulting in the statement's failure due to unavailable space in the table. Recovery from such failures is automatic. Upon detection, SAP / Oracle usually will roll back the statement, returning control to the user or user program. The user can simply re-execute the statement after correcting the problem.

A **process failure** due to abnormal termination of a process caused by SAP / Oracle itself (such as when a user performs a ^C from SQL*PLUS during non standard SAP loads at Oracle level or during admin operations). If the process that is terminated is a user process, a server process, or an application process, the Process Monitor (PMON) performs process recovery. PMON cleans up the cache and frees up resources that the process was using.

PMON resets the status of the transaction table in the rollback segment for that transaction, releasing the locks or latches acquired by the terminated process. It then removes the process ID from the list of active processes.

PMON doesn't clean up the processes that have been killed by Oracle. During an abnormal background process termination, Oracle must be shut down and restarted.Oracle uses SMON to automatically perform roll forward, and Oracle's transaction recovery which will roll back any uncommitted transactions.

Network failures can occur in a SAP three-tier configuration. In such cases, PMON will roll back the uncommitted Logical Unit of work (LUW) of each of the processes.

An **instance failure** can be caused by a physical (hardware) or a design (software) problem—for example, when one of the database background processes (DBWR) detects that there is a problem on the non-RAID5 disk and can't write to it.

In such a situation, an error message is written to a log file and the background process terminates. Crash recovery or instance recovery is automatically done depending on the I/O activity at the failure time, database instance recovery might take a long time.

Crash recovery has to roll forward and then transaction recovery has to roll the transaction back, which might take a long period of time.

Thread recovery is done automatically by SAP / Oracle when it discovers that an instance died leaving a thread open. Thread recovery is performed as part of either crash recovery or instance recovery. If the database has a single instance, then crash recovery is performed. This requires the DBA to simply start up the database, and crash recovery is automatically performed by Oracle.

If multiple instances are accessing the database and if one of the instances crashes, the second instance automatically performs instance recovery to recover the first thread. Either way, the goal of thread recovery is to restore the data block changes that were in the cache of the instance that died, and to close the thread that was left open. Thread recovery always uses the online redo log files of the thread it is recovering.

Media recovery, the last type of recovery, is executed in response to a recovery command. It is used to make a backup data file become current, or to restore changes that were lost when a data file went offline without a checkpoint. During media recovery, archived logs as well as online log files can be applied.

In an **ESS RAID5** environment these types of failures are minimized owing to the subsystem's fault tolerant design.

Media failures are the most dangerous failures in non-RAID5 environments resulting in potential loss of data if proper backup procedures are not followed, and could involve more time to recover in comparison with other kinds of failures mentioned above. A typical example of a media failure is a non-RAID5 disk controller failure or a disk head crash, which causes all Oracle database files residing on that disk (or disks) to be lost.

However, with ESS's fault tolerant RAID5 architecture, the impact of media failures is greatly minimized.

SAP R/3 data layout

SAP R/3 and the importance of Database Layout in Oracle Environments [7]:

With the progressive complexity in application system designs, combined with rapid explosion in database sizes, it is necessary to distribute data in disk systems in a way that allows the fastest possible access to it. While only 1% of the tablespaces in the SAP R/3 system are characterized by high growth and thus high I/O impact, these tablespaces essentially dictate the overall size and performance of the database. Storage architectures that provide the capability to distribute these tablespaces across as many disks as possible are able to deliver superior performance in this demanding OLTP environment**.** These architectures include the ability to reassign very high growth objects to new tablespaces.

In preparation for SAP R/3 Oracle data layout for the ESS, the following sequence of activities needs to be considered:

- Find Number and types of SAP systems to be installed in the system landscape that typically includes Development, Quality Assurance and Productive system instances.
- Start with the Planning of an SAP landscape topology with estimation of hardware requirements for storage, network and server machines. This creates a hardware requirements list: ESS storage subsystems, number of SCSI / FC connections, server machines, network connectivity etc.
- Sizing done by HW partners to establish requirements for:
    - CPU
    - Storage space, bandwidth
    - Network bandwidth
    - System infrastructure / landscape (inter system communication)
- SAP installation guide
- Assignment of SAP systems to hosts
- Estimations for storage needs of individual database systems / tablespaces

- Collect space requirements from SAP applications, their databases (tablespaces, logging), 3rd party software (Bolt-ons)
  - Consider ESS / Host / DB / Application restrictions
  - Map requirements to AIX logical volumes
  - Group the logical volumes to volume groups
  - Design ESS configuration and perform configuration steps with ESS Specialist (creating LUN's as shown in Figure 16) to create AIX volume groups and logical volumes

SAP R/3 Backup Objects and File Definitions for Oracle

All objects in the R/3 environment need to be backed up. These objects are:

1. R/3 data objects:
   a. R/3 archiving objects
   b. R/3 Interfaces
   c. SAP Executables
2. Computing center data such as:
   a. The Operating System
   b. Third Party Programs connected to R/3
3. Database objects

Figure 13 depicts the SAP Backup Objects and Figure 4 depicts the Oracle File Systems for SAP.

Oracle Backup Objects for SAP

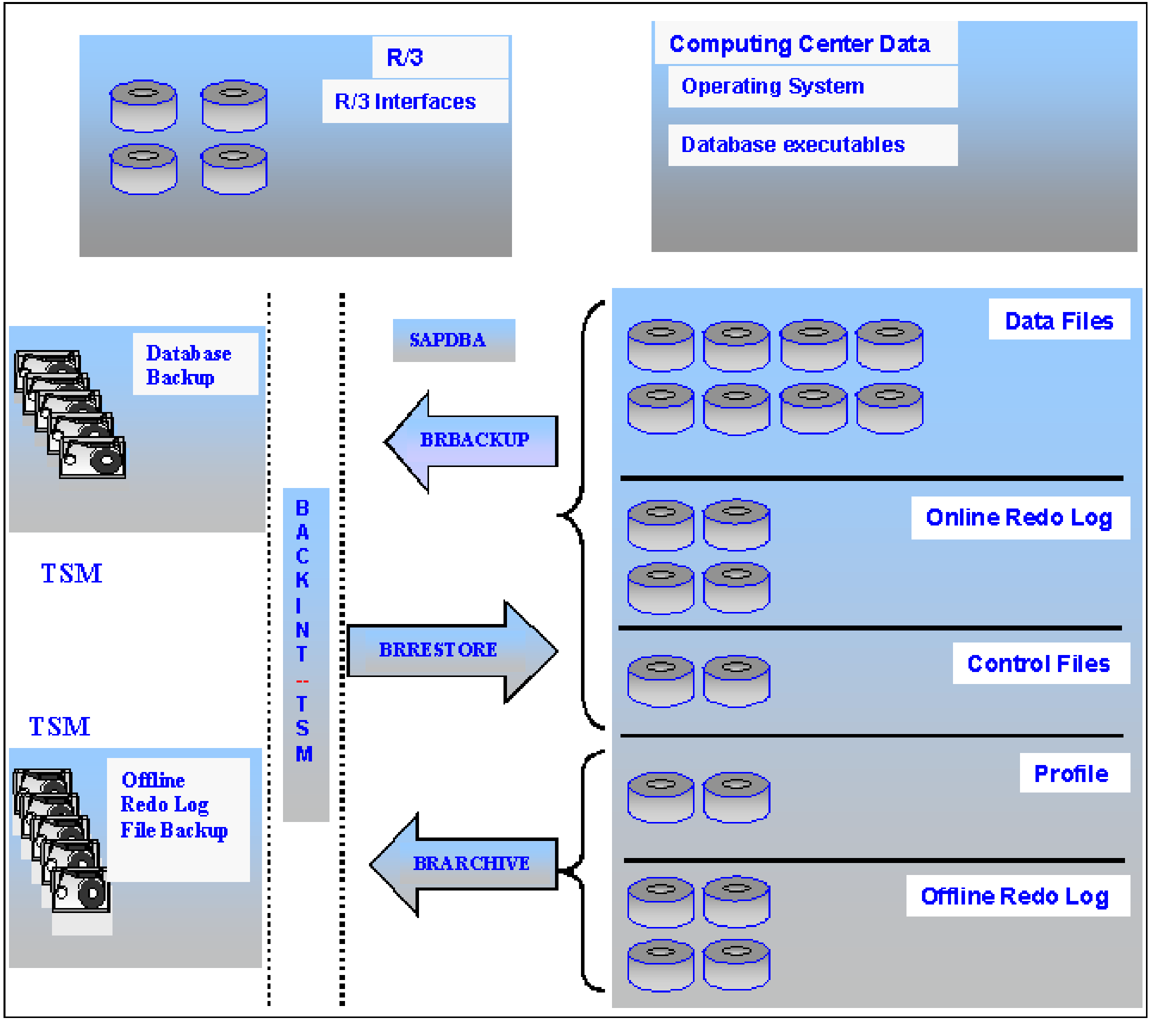


Figure 13: Oracle Backup Objects for SAP

As depicted in Figure 13, SAP R/3 database backup is accomplished through the SAPDBA tool that invokes SAP's BRBACKUP, BRRESTORE and BRARCHIVE utilities. These utilities, in turn, are integrated into TSM using SAP approved BACKINT interface also called TDP for R/3. Computing Center data and R/3 data are backed up using Operating System utilities (cpio, tar) or by using products such as TSM, Legato Networker, and Veritas Net Backup etc.

As shown in Oracle File System for SAP in Figure 4, the volume groups LSS10data, LSS12data, LSS14data and LSS16data are 64GB each in size. The other volume groups are sapvg comprising of 1GB LUNs as shown in Figure 7. The volume group sapvg is used for the SAP and Oracle binaries, and logvg holds the online and archive logs.

SAP R/3 uses standardized Directory and File names in the Oracle environment. As shown in Figure 13, the SAP R/3 objects – Oracle control files, Oracle data files, offline redo log files & online redo log files are required for a consistent database backup and restore/recovery.

The Oracle binaries are located in the subdirectory 'bin' of the ORACLE_HOME directory. The ORACLE_HOME directory is also required on each server with a database client. The environment variables SAPDATA_HOME and SAPDATA<1, 2, 3 …6> point to the directories containing database-specific files, such as data files, online redo log files, and offline redo log files. The operating system users <SID>adm (System ID) and ora<SID> (on the database server) require the following environment variable:

ORACLE_SID = <SID> (System ID on the database server, SSD in Figure 4)

E) ESS Raid5 Disk Layout Considerations for SAP R/3 Environments

The RAID5 arrays / ranks are connected via device adapters to the ESS fault tolerant control unit that contains, cache and NVRAM as shown in the Figure 14 below. The control unit attaches to host systems via SCSI or fiber channel adapters.

With proper distribution of all kinds of data across all available devices using state of the art storage systems like ESS, the whole I/O subsystem is made available to each individual database and file objects (tablespace, directory, table, index).

As hotspots tend to occur on a subset of the tablespaces or even on a single tablespace, by striping these tablespaces across all devices, the cumulative performance of all devices and device adapters is made available to each individual tablespace to improve performance.

ESS's RAID5 architecture, with maximum width striping of all Oracle tablespaces, relieves the DBA from hotspot identification and tracking which might arise from changing application profiles, shifting workloads and rapid data growths in typical non-RAID5 storage environments.

The RAID5 striping of ESS accommodates single tablespace growth while totally avoiding any additional cost of striping tablespaces across all devices.

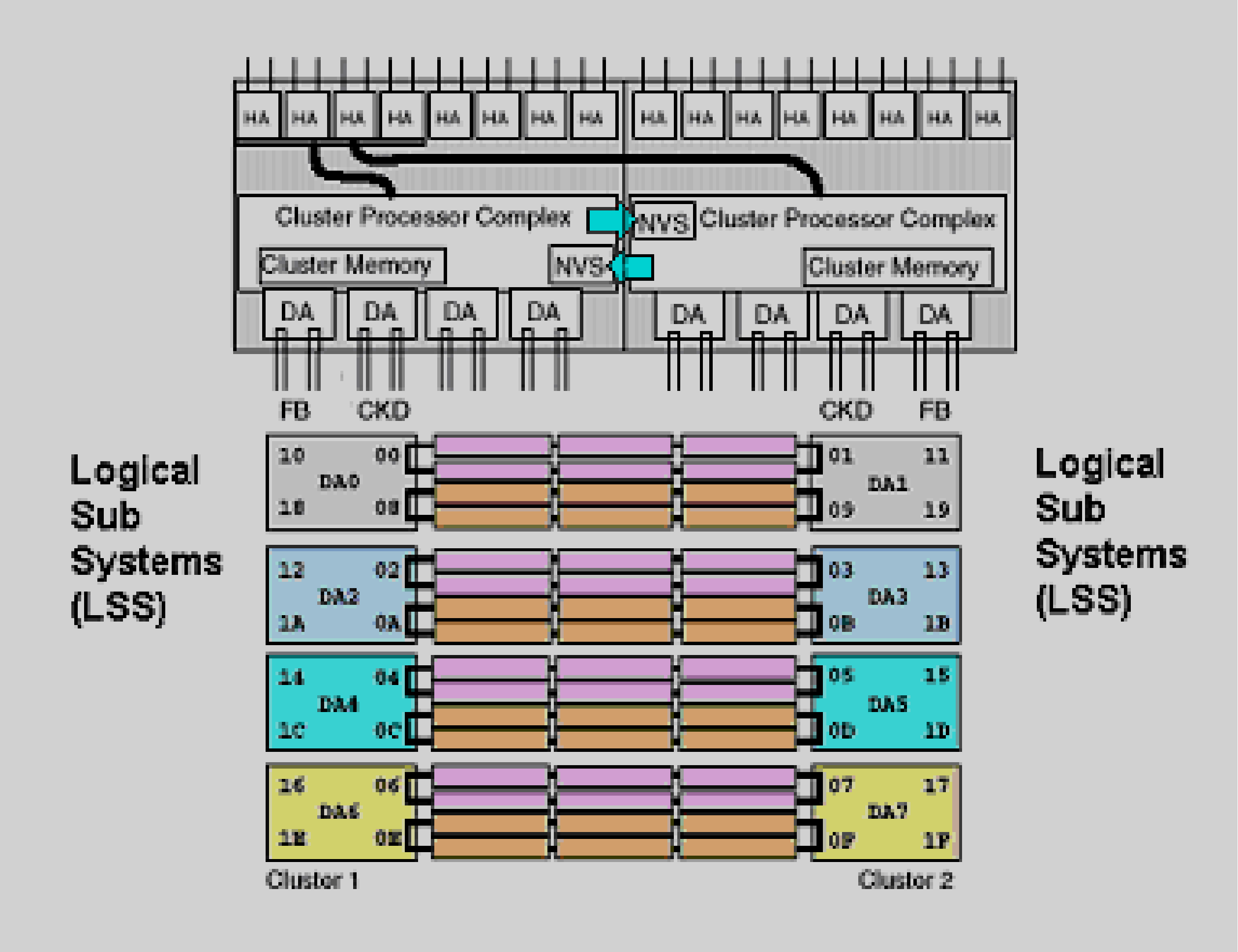


Figure 14: ESS's Fault Tolerant Architecture and Logical Sub Systems Design

Each Oracle tablespace and file system is striped across all RAID5 volumes. This is the optimum configuration since the ESS's built-in I/O capabilities are automatically and dynamically "allocated" on demand to all the tablespaces.

Spreading a logical volume across separate hdisks on the same array would provide no performance benefit, because all of the activity is still going to a single RAID5 array. The best technique for maximizing throughput of ESS is to balance workload across as many RAID5 arrays as possible.

ESS's configuration for performance may need to be balanced with availability considerations. If all databases sharing the ESS are required to be highly available, then the databases will have to be spread over the ESS clusters and Device Adapter (DA) pairs to ensure that there are alternate paths to the same data. This may lead to data files of multiple databases residing on the same array/s. Such configurations should be laid out with due attention to the I/O load that a single array can handle and possible I/O contention. Keeping in mind that it is the array that determines the throughput, not the number of logical disks, consideration for LUN sizes for the ESS should be based on the application and customer requirements. For handling high sequential read/write I/O as in Business Warehouse workloads, ensure that the tablespace layout makes use of all the arrays in an ESS.

Typically, one would allocate 8 or 16 GB LUNs over the available storage and then distribute the data across the RAID5 arrays to get optimum performance from the ESS. Further distribution of I/O is achieved by creating AIX logical volumes using the ESS logical volumes defined in multiple RAID arrays. Monitoring the I/O access pattern can then be used for further tuning and re-distributing heavily accessed database files across the arrays.

F) Sub System Device Driver (SDD)

ESS's SDD, a High Availability Automatic I/O Path Failover Function in the ESS, enhances data availability. It takes advantage of multiple active paths, distributes the I/O workload and supports more than one path from the host to the shared LUNs. If a failure occurs in the data path between the Host System and the ESS, SDD automatically switches the I/O to another path. It will also move the failed path back online after a repair is made. A single LUN can appear as 2 to

16 LUNs. This avoids I/O bottlenecks that could occur when too many I/O operations are directed to a common device using the same I/O path.

The following Figure 15 highlights the necessary steps in preparing the layout for an ESS based SAP R/3 Oracle system using fixed block (FB) for Open systems format.

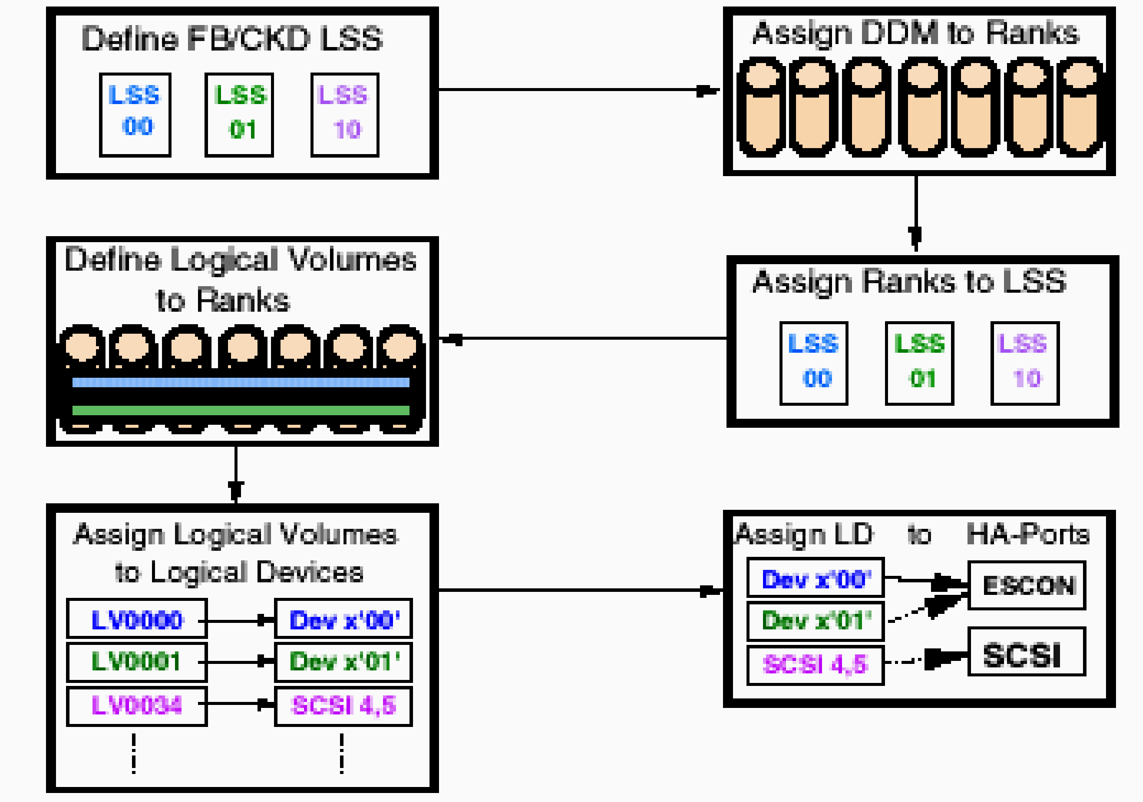


Figure 15: Steps in preparing the ESS for data layout and connectivity to hosts

In AIX, the Logical Volume Manager uses the 8GB LUNs, defined in ESS, by assigning them to volume groups as hdisk*s*. Each ESS logical disk translates to AIX like an hdisk as shown in Figure 16. With each hdisk in AIX, the host limits the number of concurrent requests to a value of '20' that it will send to ESS.

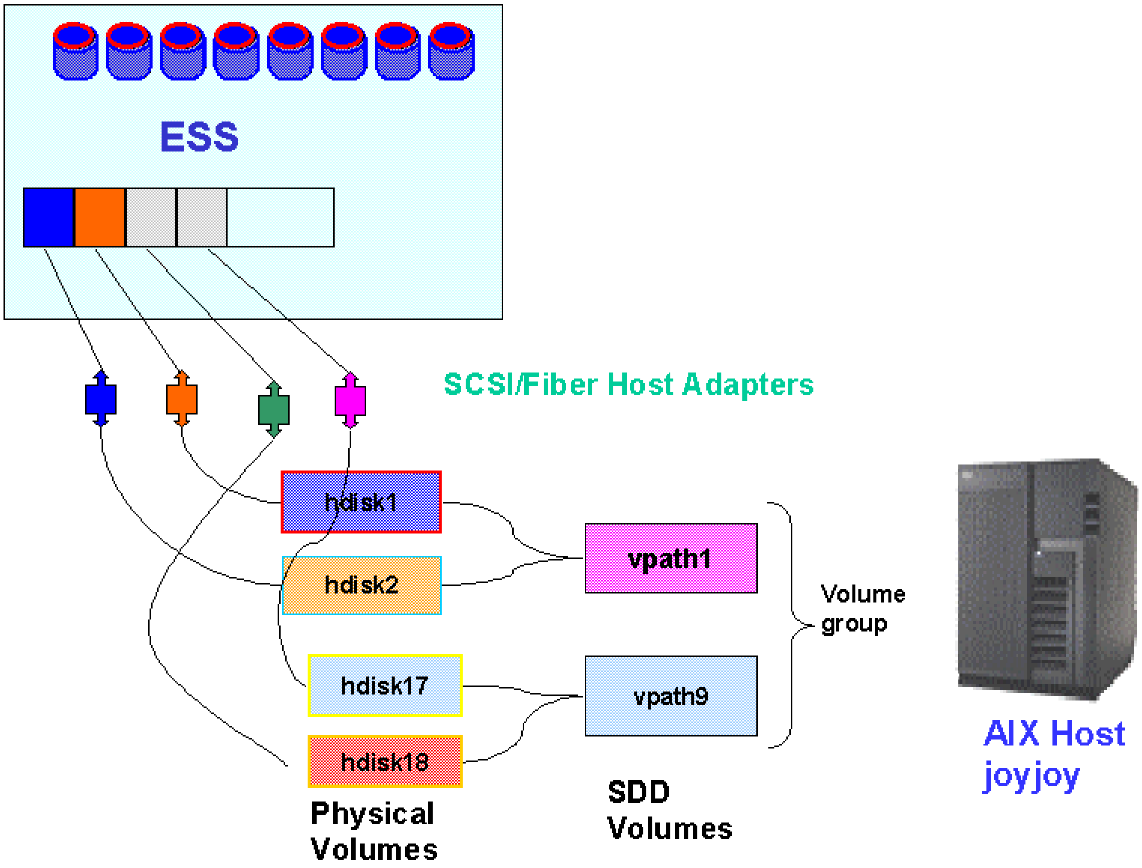


Figure 16: Vpath to Hdisk Mapping by SDD

G) Volume Layout for Oracle

The following table, Figure 17, captures the volume layout for SAP R/3 in ORACLE / ESS environment with a mapping of vpaths to hdisks using Subsystem Device Driver (SDD).

Vpath Device	Serial No	Hdisk mapping				DA loop	Volume Group
Vpath1	60112005	Hdisk6	Hdisk46	Hdisk90	hdisk126	Loop A	LSS16data
Vpath0	60012005	Hdisk5	Hdisk45	hdisk89	hdisk125	Loop A	LSS16data
Vpath3	60312005	Hdisk8	Hdisk48	Hdisk92	hdisk128	Loop A	LSS16data
Vpath4	60C12005	Hdisk9	Hdisk49	hdisk85	hdisk129	Loop B	LSS16data
Vpath5	60D12005	Hdisk10	Hdisk50	Hdisk86	hdisk130	Loop B	LSS16data
Vpath6	60E12005	Hdisk11	Hdisk51	hdisk87	hdisk131	Loop B	LSS16data
Vpath7	60F12005	Hdisk12	Hdisk52	hdisk111	hdisk151	Loop B	LSS16data
Vpath8	61812005	Hdisk13	Hdisk53	hdisk94	hdisk133	Loop A	Logvg
Vpath9	61912005	Hdisk14	Hdisk54	hdisk93	hdisk134	Loop B	Logvg
Vpath40	21912005	Hdisk33	Hdisk74	hdisk114	hdisk153	Loop B	Sapvg
Vpath41	61B12005	Hdisk166	Hdisk174	hdisk182	hdisk190	Loop B	Sapvg
Vpath10	40012005	Hdisk170	Hdisk178	hdisk186	hdisk194	Loop B	Sapvg
Vpath11	40112005	Hdisk16	Hdisk56	hdisk96	hdisk136	Loop A	LSS14data
Vpath12	40212005	Hdisk17	Hdisk57	Hdisk97	hdisk137	Loop A	LSS14data
Vpath1	60112005	Hdisk6	Hdisk46	Hdisk90	hdisk126	Loop A	LSS16data
Vpath15	40D12005	Hdisk20	Hdisk60	Hdisk100	hdisk140	Loop B	LSS14data
Vpath16	40E12005	Hdisk21	Hdisk61	hdisk101	hdisk141	Loop B	LSS14data
Vpath17	40F12005	Hdisk22	Hdisk62	hdisk102	hdisk142	Loop B	LSS14data
Vpath18	41812005	Hdisk23	Hdisk63	hdisk103	hdisk143	Loop A	Logvg
Vpath42	41A12005	Hdisk167	Hdisk175	hdisk183	hdisk191	Loop A	Sapvg
Vpath43	41B12005	Hdisk168	Hdisk176	hdisk184	hdisk192	Loop B	Sapvg
Vpath44	01A12005	Hdisk171	Hdisk179	hdisk187	hdisk195	Loop A	Sapvg
Vpath47	01B12005	Hdisk172	Hdisk180	hdisk188	Hdisk196	Loop B	Sapvg

Figure 17: ESS Volume Mapping & Layout

H) The Role of the ESS Specialist

With ESS Specialist, the WebGUI interface for the Enterprise Server, all the tasks needed for setting up arrays, logical subsystems, logical unit numberings, host definitions for connectivity via SCSI / Fiber, PPRC path definitions and Copy services tasks can be administered and managed. This is shown in Figure 18.

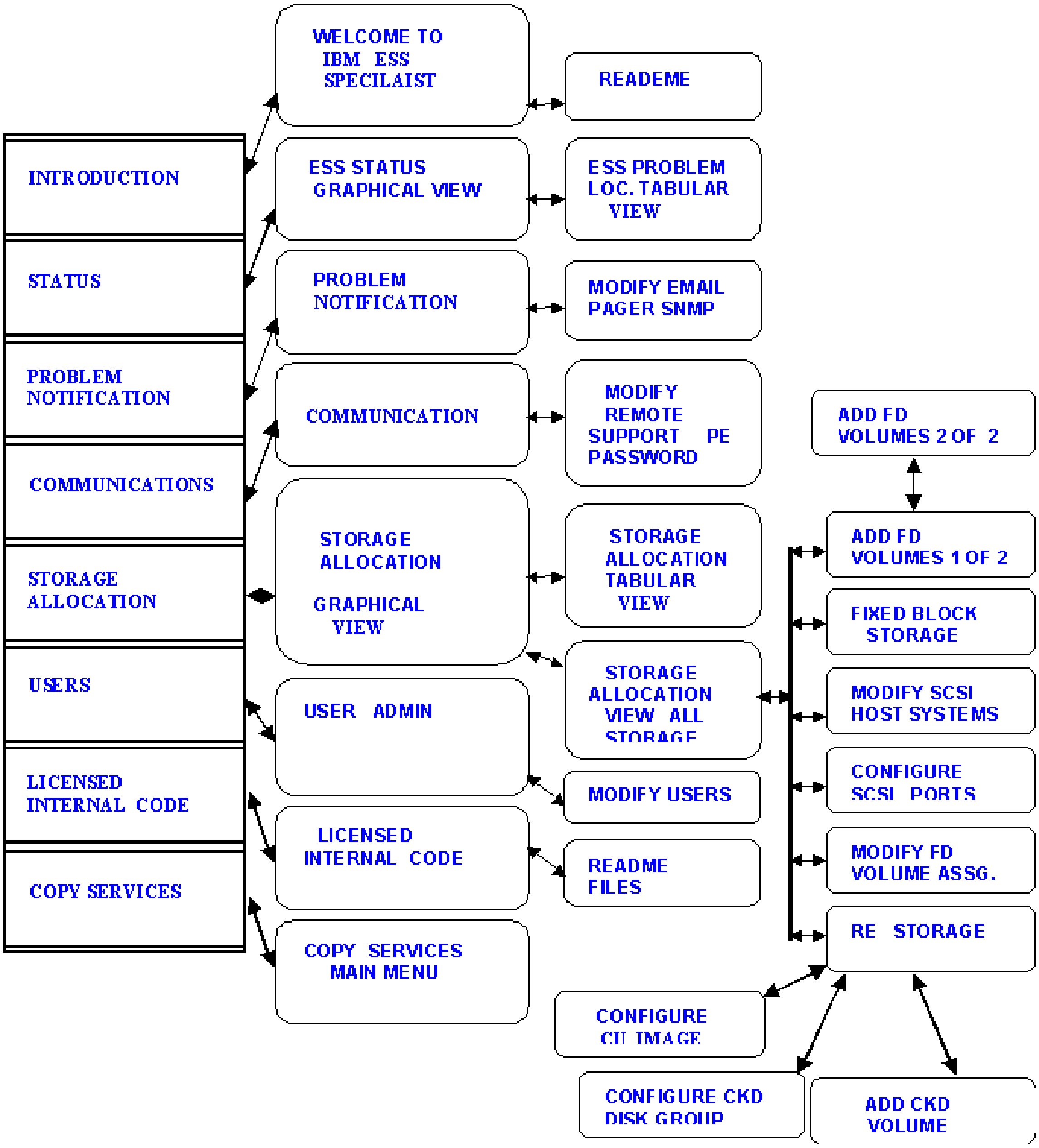


Figure 18: ESS Specialist Views and Functions

References

1) IBM Enterprise Storage Server - IBM Document Number: SG24-5465-00
http://www.redbooks.ibm.com/

2) Implementing the Enterprise Storage Server in your Environment -IBM Document Number SG24-5420-00 http://www.redbooks.ibm.com/

3) IBM Enterprise Storage Server Performance Monitoring and Tuning Guide - IBM Document Number: SG24-5656-00

4) ESS Layout Considerations for a Heterogeneous System Landscape - Dr. Jens Claussen, SAP Advanced Technology Group, December 2000

5) SAP R/3 Storage Management - Split Mirror Backup Recovery on IBM's Enterprise Storage Server (ESS) - Siegfried Schmidt, SAP AG, AdvancedTechnology Group, February 2000, http://service.sap.com/atg or
http://www.storage.ibm.com/hardsoft/products/ess/whitepaper.htm

6) Database Layout for SAP Installations with DB2 UDB for UNIX and Windows - Dr. Jens Claussen, SAP AG, Advanced Technology Group, February 2001,
http://service.sap.com/atg

7) Database Layout for R/3 Installations under ORACLE, Terabyte Project, SAP AG
http://service.sap.com/atg or
http://www.storage.ibm.com/hardsoft/products/ess/whitepaper.htm

8) DB2 Internals for Administrators: With V7 Updates May 16 & 18, 2000 - Matt Huras, IBM Toronto Labs

9) Fundamentals of Database Layout SAP AG, Version 1.0

http://service.sap.com/atg , March 2000

10) Oracle 8i Backup & Recovery Handbook - Rama Velpuri and Anand Adkoli, Oracle Press, Osborne McGraw Hill, 2001

11) Storage Management for SAP and DB2 UDB: Split Mirror Backup / Recovery With IBM's Enterprise Storage Server (ESS) - Sanjoy Das, Siegfried Schmidt, Jens Claussen, BalaSanni Godavari, August 2001,

http://service.sap.com/split-mirror or

http://www.storage.ibm.com/hardsoft/diskdrls/technology.htm

12) SSQJ Documentation Version 9.B – SAP Advanced Technology Group,
http://service.sap.com/atg