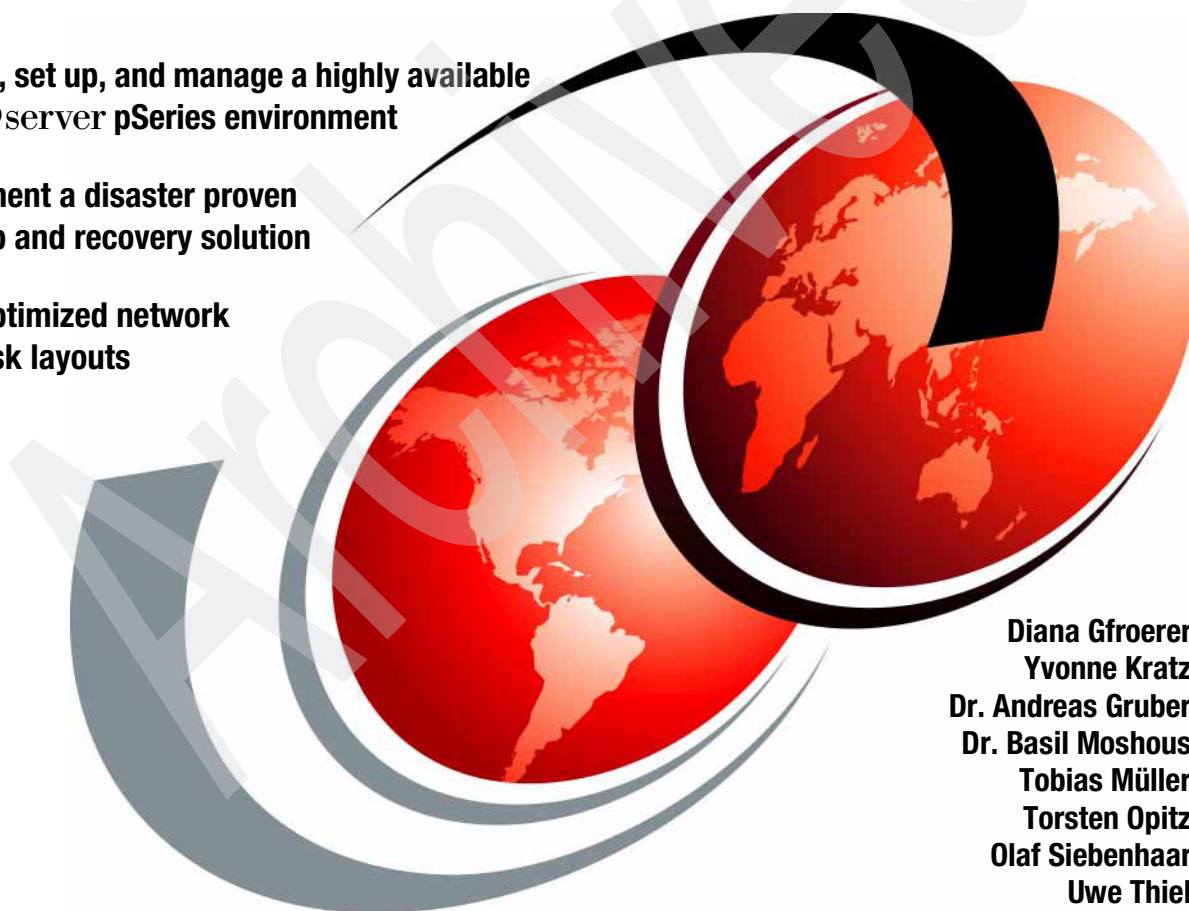


A Holistic Approach to a Reliable Infrastructure for SAP R/3 on AIX

Design, set up, and manage a highly available IBM @server pSeries environment

Implement a disaster proven backup and recovery solution

Plan optimized network and disk layouts



Diana Gfroerer
Yvonne Kratz
Dr. Andreas Gruber
Dr. Basil Moshous
Tobias Müller
Torsten Opitz
Olaf Siebenhaar
Uwe Thiel



International Technical Support Organization

**A Holistic Approach to a Reliable Infrastructure for
SAP R/3 on AIX**

October 2001

Archived

Take Note! Before using this information and the product it supports, be sure to read the general information in “Special notices” on page 485.

First Edition (October 2001)

This edition applies to SAP R/3 Release 4.6C, Oracle 8.1.7, DB2 UDB Version 6.1 and 7.1, HACMP Version 4, and TSM Version 4, for use with the AIX Version 4.3.3 ML8 Operating System.

Comments may be addressed to:
IBM Corporation, International Technical Support Organization
Dept. JN9B Building 003 Internal Zip 2834
11400 Burnet Road
Austin, Texas 78758-3493

When you send information to IBM, you grant IBM a non-exclusive right to use or distribute the information in any way it believes appropriate without incurring any obligation to you.

© Copyright International Business Machines Corporation 2001. All rights reserved.

Note to U.S Government Users – Documentation related to restricted rights – Use, duplication or disclosure is subject to restrictions set forth in GSA ADP Schedule Contract with IBM Corp.

Contents

Figures	13
Tables	17
Preface	19
The team that wrote this redbook	20
Special notice	22
IBM trademarks	23
Comments welcome	24
Chapter 1. Introduction	1
1.1 A holistic approach	2
Part 1. Concepts	5
Chapter 2. Requirements for a reliable SAP R/3 environment	7
2.1 Characteristics of an SAP R/3 environment	8
2.1.1 SAP R/3 architecture	8
2.1.2 SAP R/3 system landscapes	10
2.1.3 Service layer model	11
2.2 Definition of requirements	12
2.2.1 Run-time requirements	12
2.2.2 Non-run-time requirements	13
2.2.3 Definition of high availability and downtime	13
2.3 Constraints for real environments	16
2.4 Guidelines for the implementation	17
2.5 Organizational aspects for SAP R/3 operations	21
2.5.1 Roles in a productive environment	21
2.5.2 Interactions between roles in an SAP R/3 environment	22
2.5.3 Service provider model	24
2.6 Service level agreement in an SAP R/3 environment	24
2.6.1 Definition of service level agreement	24
2.6.2 General criteria for service level agreements	25
2.6.3 SAP R/3 specific criteria	27
2.7 Guidelines for operating an SAP R/3 environment	30
2.7.1 Definition of general rules for interaction	31
2.7.2 Landscape-wide operational processes	32
2.7.3 Single system operational processes	33

Chapter 3. Architecture	35
3.1 Components of the infrastructure	36
3.1.1 Structure of an SAP system	36
3.1.2 IBM @server pSeries and RS/6000	42
3.1.3 Storage subsystems	46
3.1.4 Local Area Network (LAN)	55
3.1.5 Software components	58
3.2 Building a basic model	65
3.2.1 Server for the database and central instance (DB/CI)	66
3.2.2 Application server (AP)	67
3.2.3 The storage subsystem (disk storage)	69
3.2.4 The networks (switch)	70
3.2.5 The backup subsystem	71
3.2.6 Administration server (Admin)	73
3.2.7 Requirements met by the basic model	76
3.3 Building a fault-tolerant model	77
3.3.1 Identifying single points of failure in the basic model	77
3.3.2 Cluster solution for SAP R/3	80
3.3.3 Storage subsystem considerations	82
3.3.4 Network considerations	83
3.3.5 Distribution of SAP R/3 services to application servers	84
3.3.6 Performance considerations	85
3.4 Building a disaster-tolerant model	86
3.4.1 Additional requirements for disaster-tolerant environments	86
3.4.2 Additional storage subsystem	88
3.4.3 Dual library and TSM server	89
3.4.4 Shadow database	90
Part 2. Implementation	93
Chapter 4. Disk storage	95
4.1 Basic understanding of disk mechanics	97
4.2 Redundant array of independent disks (RAID)	99
4.3 Disk interconnection architectures	101
4.4 Disk subsystems based on RAID 5 implementations	105
4.4.1 General idea of improving RAID 5 performance	105
4.4.2 IBM Enterprise Storage Server	106
4.5 Disk layout for an SAP R/3 system	109
4.5.1 Different file types in an SAP R/3 installation	110
4.5.2 Layout criteria	111
4.5.3 SAP R/3 specific considerations	112
4.5.4 Example for a disk layout	113
4.5.5 Guidelines for tablespaces	118

4.6 Logical Volume Manager considerations	120
4.6.1 Journaled file systems versus raw logical volumes.	120
4.6.2 Mirror Write Consistency and databases	123
4.7 General requirements.	125
4.7.1 Operating system requirements	125
4.7.2 Application requirements.	126
4.8 Basic requirements.	128
4.8.1 A basic SSA configuration.	128
4.8.2 A basic ESS configuration.	129
4.9 Fault-tolerance requirements	131
4.9.1 A fault-tolerant SSA configuration.	131
4.9.2 A fault-tolerant ESS configuration.	132
4.10 Disaster-tolerance requirements.	133
4.10.1 A disaster-tolerant SSA configuration	133
4.10.2 A disaster-tolerant ESS configuration	134
4.10.3 A disaster-tolerant ESS configuration based on SAN switches	136
Chapter 5. Storage Area Networks	137
5.1 Introduction to Storage Area Networks (SAN).	138
5.2 Storage Area Networks based on Fibre Channel	140
5.2.1 Fibre Channel protocol layers	140
5.2.2 Fibre Channel technology	141
5.3 SAN products	142
5.3.1 IBM SAN Data Gateway	142
5.3.2 IBM SAN Fibre Channel Switches	142
5.3.3 McDATA ES-3016 and ES-3032 Switches	143
5.3.4 McDATA ED-6064 Enterprise Fibre Channel Director	144
5.4 Building a Storage Area Network	144
5.4.1 SAN connections and services	145
5.4.2 Design considerations for a SAN	146
Chapter 6. Network	149
6.1 Network requirements	150
6.1.1 Latency requirements of SAP R/3 Release 4.6.	151
6.1.2 Bandwidth requirements of SAP R/3 Release 4.6	151
6.1.3 Bandwidth requirements for backup and recovery tasks	154
6.2 Network technologies for IBM @server pSeries	155
6.2.1 Characteristics of different network technologies	155
6.2.2 Characteristics of switches	157
6.2.3 Bandwidth of different network technologies.	159
6.2.4 Characteristics of Inter Switch Links	160
6.3 Definition of the different networks	161
6.3.1 The front-end network.	161

6.3.2	The backup network	165
6.3.3	The control network	166
6.4	Network layout for different configurations	167
6.4.1	Network layout for the basic configuration	167
6.4.2	Network layout for the fault-tolerant configuration	169
6.4.3	Network layout for the disaster-tolerant configuration	172
6.5	Configuration of switches and network adapters	174
6.5.1	Configuration of switches	174
6.5.2	Configuration of network adapters and TCP/IP options	175
Chapter 7.	Backup and recovery	185
7.1	Introduction	186
7.1.1	Terminology	187
7.1.2	General aspects	189
7.2	Backup objects and methods	191
7.2.1	AIX operating system	191
7.2.2	SAP R/3 database including log files	192
7.2.3	Archived log files of the SAP R/3 database	193
7.2.4	SAP R/3 and database executables	194
7.2.5	SAP R/3 Transport Management System files	194
7.2.6	Interfaces of the SAP R/3 system	195
7.2.7	TSM database and recovery log	195
7.2.8	LVM configuration	196
7.2.9	Summary	196
7.3	TSM concept	197
7.3.1	Virtual nodes	197
7.3.2	Storage pools	199
7.3.3	Domain, management classes and copy groups	201
7.3.4	Devices	204
7.3.5	Scheduling	205
7.3.6	TSM database and recovery log	207
7.3.7	Using NIM as part of the TSM concept	209
7.4	Preparing for disaster	210
7.4.1	Vaulting	211
7.4.2	Dual library and TSM server/library sharing	213
7.5	Test and operation of the implemented solution	220
Chapter 8.	High availability	221
8.1	Introduction	222
8.2	Manual and automatic takeover	224
8.3	HACMP basics	226
8.3.1	Standard features	226
8.3.2	Required customization	227

8.3.3 Application monitoring with HACMP/ES	227
8.4 Implementation of HACMP for SAP R/3	228
8.4.1 Topology	229
8.4.2 Resource groups	231
8.4.3 Application start and stop handling	241
8.4.4 Odds and ends	244
8.5 Testing of the cluster	248
8.5.1 Cluster diagram (state-transition diagram)	248
8.6 Extensions for the disaster-tolerant model	252
8.7 Maintenance of clusters	256
8.7.1 Install programs and patches	256
8.7.2 File systems	258
Chapter 9. Shadow database	259
9.1 Business needs for a shadow database	260
9.2 Definition of a shadow database	262
9.3 Benefits of a shadow database	263
9.3.1 Logical database errors	263
9.3.2 Physical database errors	264
9.3.3 Database consistency checks	264
9.3.4 Independent backup	265
9.3.5 Geographical disaster-tolerance	265
9.3.6 Upgrade improvement	265
9.4 Implications of a shadow database	266
9.4.1 Technical implications	266
9.4.2 Organizational implications	267
9.5 Implementation of a shadow database	268
9.5.1 General considerations	268
9.5.2 Large shadow database server	269
9.5.3 Small shadow database server	269
9.5.4 Considerations for a disaster-tolerant scenario	271
9.5.5 Performance enhancements	271
9.6 Comparison with FlashCopy	272
9.6.1 Time delay	272
9.6.2 Hardware separation	273
9.6.3 Consistency check	273
9.6.4 Backup from copy	273
9.6.5 Creation of test systems	273
9.6.6 Summary	274
Chapter 10. Hints and tips	275
10.1 Principles of administration	276
10.1.1 Naming concept	277

10.1.2	Importance of synchronizing	279
10.1.3	Easy interface for administration	279
10.1.4	Principles of the SP system	280
10.1.5	System documentation	282
10.2	AIX implementation	283
10.2.1	Network Installation Manager (NIM)	283
10.2.2	Time synchronization	284
10.2.3	File system for logs and documentation	285
10.2.4	File systems	285
10.2.5	Redirect the console	286
10.2.6	Clever tools	287
10.2.7	Remote commands	287
10.2.8	Configuration files	289
10.2.9	Printing time out	294
10.2.10	Address Resolution Protocol cache (ARP cache)	294
10.3	SAP R/3 implementation	294
10.3.1	Multiple SAP R/3 systems on one server	295
10.3.2	Paging space	296
10.3.3	User limits	296
10.3.4	Easy access to installation media of SAP R/3	297
10.3.5	Oracle recommendations	298
10.3.6	Number range buffering	299
10.3.7	Clean up jobs	300
10.3.8	Update V3	300
10.4	Operation of the SAP R/3 infrastructure	301
10.4.1	Server journal	301
10.4.2	Starting point for profiles	301
10.4.3	Time differences and daylight saving time	302
10.4.4	Background processing	302
10.4.5	Avoid client copies	303
	Chapter 11. Performance	305
11.1	AIX shared memory management	306
11.1.1	Virtual memory	307
11.1.2	Segment layout of processes	307
11.1.3	Sharing segments between processes	309
11.1.4	Shared memory regions	310
11.2	SAP R/3 instance buffers	312
11.2.1	SAP R/3 buffer types	312
11.2.2	Checking SAP R/3 buffer configuration	313
11.2.3	Arrangements of SAP R/3 buffers in pools	314
11.2.4	Implications of the AIX version for SAP R/3 buffers	315
11.2.5	Buffer tuning	317

11.3 Models for SAP R/3 Extended Memory	318
11.3.1 A definition of SAP R/3 Extended Memory	318
11.3.2 Standard Extended Memory configuration (32-bit)	321
11.3.3 AIX ES shared memory implementation (32-bit)	323
11.3.4 AIX ES shared memory implementation (64-bit)	325
11.3.5 AIX ES Extended memory configuration recommendations	327
11.4 AIX concepts	329
11.4.1 CPU	330
11.4.2 The Virtual Memory Manager	332
11.4.3 Asynchronous disk I/O	338
11.4.4 Disk I/O pacing	342
11.5 Tools to monitor AIX	345
11.5.1 The vmstat command	345
11.5.2 The vmtune command	347
11.5.3 The iostat command	348
11.5.4 The filemon command	349
11.5.5 The netstat -m command	351
11.5.6 The topas command	353
11.5.7 The monitor tool	355
11.5.8 The SAP R/3 transaction ST06	357
11.6 AIX performance hints	357
11.6.1 High value of disk I/O wait	359
11.6.2 High disk activity	359
11.6.3 Paging	360
11.6.4 The run queue	361
11.6.5 Long running batch jobs	361
11.6.6 Related information	361
Part 3. Operation	363
Chapter 12. SAP R/3 system copy	365
12.1 Introduction	366
12.1.1 Reasons for copying an SAP R/3 system	367
12.1.2 Terminology	367
12.2 Methods for SAP R/3 system copy	369
12.2.1 R3load procedure	370
12.2.2 Backup and restore procedures	370
12.2.3 Export and import procedures	371
12.3 Organizational preparations	372
12.4 Technical preparations	372
12.4.1 Preparing the source system	373
12.4.2 Preparing the target system	376
12.5 Performing an SAP R/3 database copy	378

12.5.1	Example of backup and restore procedure of a DB2 database . . .	379
12.5.2	Example for backup/restore procedure of an Oracle database. . .	382
12.6	Subsequent technical actions	387
12.6.1	Actions on operating system level.	387
12.6.2	Actions on database level.	390
12.6.3	Actions on SAP R/3 system level	393
12.7	Special treatment for the development system	402
12.7.1	Version history	403
12.7.2	Repair and transport requests.	403
12.7.3	Registration of developers and objects	404
12.7.4	IMG projects	404
12.7.5	SAP R/3 user	405
12.8	Reference information	405
Chapter 13.	Daily tasks to prevent error situations	407
13.1	Monitoring methods	408
13.2	AIX operating system	409
13.2.1	Checking the AIX error log	409
13.2.2	Checking file systems	410
13.2.3	Checking mirrors of logical volumes (LV)	411
13.2.4	Checking mailboxes	411
13.2.5	Checking printer queues	412
13.2.6	Checking network routes.	412
13.2.7	Checking the AIX system console log.	412
13.2.8	Differences for an SP environment	413
13.2.9	Monitoring HACMP	415
13.2.10	Summary for AIX, SP, and HACMP daily checks	415
13.3	Database system	416
13.3.1	Checking database free space	416
13.3.2	Update optimizer statistics	417
13.3.3	Running Oracle database system check.	418
13.3.4	Checking for database errors	419
13.3.5	Searching for missing indices	419
13.3.6	Monitoring performance	419
13.3.7	Reorganization	420
13.3.8	Summary for DB regular checks and tasks.	421
13.4	SAP R/3	422
13.4.1	Housekeeping batch jobs	422
13.4.2	Checking SAP R/3 instances and application servers	423
13.4.3	Checking the SAP R/3 system log	423
13.4.4	Checking background jobs	424
13.4.5	Checking update records	424
13.4.6	Checking ABAP dumps.	425

13.4.7	Checking TemSe consistency	425
13.4.8	Checking for lock entries	425
13.4.9	Monitoring performance and workload	426
13.4.10	Analyzing trace files of work processes	427
13.4.11	Summary for SAP R/3 checks	427
13.4.12	Transactions for administration assistance	428
13.5	Backup	429
13.5.1	Checking the TSM client schedules	429
13.5.2	Checking the TSM administrative schedules	431
13.5.3	Checking the TSM server activity log	432
13.5.4	Checking TSM scheduler daemons	433
13.5.5	Checking SP CWS mksysb	433
13.5.6	Summary for backup operations daily checks	433
13.6	Cleaning the log files	434
Chapter 14.	Troubleshooting and first aid after system failure	437
14.1	Steps to locate and analyze a faulty component	438
14.2	Transactions	443
14.2.1	Transactions AL08 and OS07	443
14.2.2	Transaction SM21	443
14.2.3	Transaction SM51	444
14.2.4	Transactions ST02 and ST03	445
14.2.5	Transaction STMS	448
14.3	Tasks	448
14.3.1	Archive file system full	448
14.3.2	Asynchronous I/O failed (Oracle)	451
14.3.3	Database processes	452
14.3.4	Event 'Checkpoint not complete' (Oracle)	454
14.3.5	Free space problem	455
14.3.6	Instance file system full	459
14.3.7	Report RDDIMPDP	459
14.3.8	SAP R/3 processes	460
14.3.9	Shared memory segment too large (Oracle only)	462
14.3.10	startsap fails (Oracle only)	463
14.3.11	Transport directory	464
14.4	Tools	465
14.4.1	Active listener process [lsnrctl] (Oracle only)	465
14.4.2	Allocatable memory [memlimits]	466
14.4.3	Available logon groups [lgst]	467
14.4.4	Check SAP profiles and logs [sappfpar]	467
14.4.5	Cleanup shared memory [showipc and cleanipc]	468
14.4.6	SAP R/3 connection to database [R3trans]	469
14.4.7	Online help	471

14.4.8 Problems with brarchive and brbackup	473
14.5 AIX commands	474
14.5.1 Connectivity check [ping]	475
14.5.2 Login to a host.	476
14.5.3 Performance problems [vmstat]	476
Related publications	479
IBM Redbooks	479
Other resources	480
Referenced Web sites	482
How to get IBM Redbooks	483
IBM Redbooks collections.	483
Special notices	485
Abbreviations and acronyms	487
Index	491

Figures

2-1	Three-tier SAP R/3 system architecture	9
2-2	The common SAP R/3 system landscape	10
2-3	Service delivery model	11
2-4	Responsibilities and roles in an SAP R/3 environment	21
3-1	Architecture of SAP R/3 process communication	39
3-2	A basic infrastructure model	65
3-3	A model for a fault-tolerant cluster	81
3-4	A disaster-tolerant infrastructure model	87
4-1	The disk storage	96
4-2	Data transfer in an SSA loop	103
4-3	Supported connections of SSA Optical Extenders	104
4-4	Term definition in RAID 5	106
4-5	Schematic ESS overview	107
4-6	Overview of the three layers	109
4-7	Influences on the disk layout	111
4-8	Access activity on the SAP R/3 standard tablespaces	112
4-9	Disjointed disk areas	116
4-10	Sample partitioning of an ESS with eight ranks	118
4-11	Overview of I/O processing in AIX	122
4-12	Basic SSA configuration	128
4-13	Basic ESS cabling configuration	130
4-14	Fault-tolerant SSA configuration	131
4-15	Two hosts connected with four access paths to the ESS	132
4-16	Disaster-tolerant SSA configuration	134
4-17	Disaster-tolerant disk subsystem connection	135
4-18	ESS in a SAN configuration	136
5-1	The Storage Area Network	138
5-2	Layers of the Fibre Channel architecture	140
5-3	SAN scenario in an SAP R/3 environment	146
6-1	The network	150
6-2	Illustration of the front-end network and access network	162
6-3	Introduction of the server network	163
6-4	Network topology for the basic configuration	167
6-5	Physical layout of network connections from servers to switches	168
6-6	Connection with a single network adapter to the front-end network	169
6-7	Connection of two network adapters to the front-end network	170
6-8	Network topology for the fault-tolerant configuration	171
6-9	Network topology for the disaster-tolerant configuration	173

7-1	Backup and recovery	186
7-2	Normal database operation	187
7-3	Backup of a database	188
7-4	Restore and recovery of a database	189
7-5	Backup objects, management classes, and storage pools	204
7-6	NIM solution for restore of system backups (mksysb)	209
7-7	Dual library and TSM server	213
7-8	Distribution of storage pools to shared libraries	215
7-9	Schematic picture of backup paths to the libraries lowlib and hilib	216
7-10	Primary and secondary NIM server	218
7-11	NIM client server relationships	219
8-1	High availability	222
8-2	IP labels and subnets	223
8-3	A resource group can be taken over manually or automatically	226
8-4	Scope of the high available environment	229
8-5	Concentrated versus separated SPoFs	232
8-6	Service, boot, and standby IP label in an SAP R/3 cluster	233
8-7	Handling of NFS mounts in the highly available state	239
8-8	Handling of NFS mounts in the takeover scenarios	240
8-9	State-transition diagram for automatic operation	250
8-10	Legend for the state-transition diagram	251
8-11	State-transition diagram for manual operation mode	252
8-12	Disaster-tolerant model	253
8-13	Adapter groups	254
9-1	The shadow database	260
9-2	Main reasons for system outages in database environments	261
9-3	Basic principle of a shadow database	263
9-4	Database swap in case of an error in the production database	270
9-5	Disk assignment for operation after a swap	271
10-1	Hints and tips	276
10-2	LPP source handling	284
11-1	Performance	306
11-2	Segments of a process	308
11-3	Two processes sharing segments	310
11-4	Memory types of SAP R/3 work processes	319
11-5	Standard Extended memory configuration (32-bit)	322
11-6	Alternative AIX ES Extended Memory configuration (32-bit)	324
11-7	Alternative AIX ES Extended Memory configuration (64-bit)	326
12-1	System copy	366
12-2	Sample SAP R/3 system landscape	378
13-1	Daily tasks to prevent error situations	408
14-1	Troubleshooting and first aid after system failure	438
14-2	Determine a faulty component if SAP R/3 transactions abort / dump	441

14-3 Determine a faulty component if a user cannot connect to SAP R/3 . . 442

Archived

Tables

2-1	Landscape wide operational processes	32
2-2	Operational processes related to single systems	33
3-1	Number of processes per SAP R/3 instance and system	41
3-2	IBM @server pSeries and RS/6000 server types	42
3-3	RS/6000 SP classic node types	43
3-4	Work process configuration for a minimal central instance	67
3-5	Active application components for several cluster states	82
3-6	Distribution of SAP R/3 services	84
3-7	Application performance degradation for different failure scenarios	85
4-1	Throughput and I/O per second of a single disk	98
4-2	I/Os per second in RAID 1 and RAID 5 arrays for random access	100
4-3	I/O price in standard RAID 1 and RAID 5 arrays	101
4-4	Bandwidth and throughput of interconnection technologies	102
4-5	Overview of file types	110
6-1	Network bandwidth between the application and presentation layers	153
6-2	Required sustained bandwidth for the backup network	154
6-3	Features of common network adapters for IBM @server pSeries	155
6-4	Bandwidth of common network adapters for IBM @server pSeries	159
6-5	Required bandwidth of the front-end network for interactive work	164
6-6	Required bandwidth for the backup network and recommendations	165
6-7	Settings for Ethernet ports	175
6-8	Recommended values of the attributes for the TCP/IP protocol	183
7-1	Backup objects and methods	196
7-2	Virtual Nodes for an SAP R/3 environment	198
7-3	Storage pools	201
7-4	Management classes and copy groups	202
7-5	Client schedules	205
7-6	Administrative schedules	206
8-1	Contents of the HACMP resource groups	235
8-2	HACMP public network configuration	236
8-3	Modified SAP R/3 profiles	237
9-1	Protection against failures	274
10-1	Example for files to be distributed	291
11-1	Summary of shared memory management for 32 bit processes	311
11-2	Parameter recommendations for AIX ES shared memory	327
11-3	Example for a starting point for disk I/O pacing	344
13-1	Important SP specific daemons for nodes and CWS	413
13-2	AIX, SP and HACMP checks	415

13-3	Regular DB checks and tasks	421
13-4	Suggested house keeping jobs	422
13-5	SAP R/3 regular checks	427
13-6	Backup log files and protocols	430
13-7	Backup operations daily checks	433
13-8	Log files	434

Preface

This redbook gives a broad understanding on how to plan, set up, and maintain a reliable infrastructure for SAP R/3 on AIX. It is a reflection of the real-life experience gathered and documented by a team of SAP R/3 Basis Consultants that planned, installed, set up, and maintained complex SAP R/3 landscapes in IBM @server pSeries environments for many years. This redbook is an invaluable source of information to consultants, IT specialists and architects, system administrators, and sales representatives that work with SAP R/3 environments.

This redbook is organized into three major parts.

Part 1, “Concepts” on page 5 contains the following chapters:

- ▶ Chapter 2, “Requirements for a reliable SAP R/3 environment” on page 7 provides a high level overview of the SAP R/3 architecture and describes the organizational aspects that are important for the operation of SAP R/3 systems. It enables you to define the requirements for your SAP R/3 environment, including the definition of service level agreements, roles, responsibilities, and processes.
- ▶ Chapter 3, “Architecture” on page 35 provides information on the basic building blocks for an SAP R/3 environment. It introduces three different solution models for SAP R/3 landscapes that are used as reference models throughout this redbook. It is therefore mandatory to read this chapter in order to be able to understand the contents of this redbook.

Part 2, “Implementation” on page 93 contains the following chapters:

- ▶ Chapter 4, “Disk storage” on page 95 provides recommendations for the design of the disk subsystem in SAP R/3 environments. It helps to choose the technology and layout of the disk subsystem, looking at connection mechanisms, fault- and disaster-tolerant solutions, and performance and sizing requirements.
- ▶ Chapter 5, “Storage Area Networks” on page 137 provides an introduction to Storage Area Networks (SANs) including information on technology, protocols and products. It gives practical hints and tips for the design and implementation of a SAN.
- ▶ Chapter 6, “Network” on page 149 provides information on the network layout in an SAP R/3 environment in regards to performance, reliability, availability, scalability, security, and manageability.

- ▶ Chapter 7, “Backup and recovery” on page 185 discusses data backup and recovery concepts of an SAP R/3 environment based on Tivoli Storage Manager (TSM) and Tivoli Data Protection (TDP).
- ▶ Chapter 8, “High availability” on page 221 provides information on design, implementation, and configuration of highly available SAP R/3 environments, based on High Availability Cluster Multi Processing (HACMP) for AIX.
- ▶ Chapter 9, “Shadow database” on page 259 provides information on the principles and benefits of a shadow database, including a discussion of commercial products that help to implement a shadow database. It also compares shadow databases with other techniques that provide database availability.
- ▶ Chapter 10, “Hints and tips” on page 275 discusses practical hints and tips to implement and run a reliable SAP R/3 infrastructure.
- ▶ Chapter 11, “Performance” on page 305 provides information to set up a well performing SAP R/3 system in terms of memory management and I/O throughput. It is not intended as a troubleshooting guide.

Part 3, “Operation” on page 363 contains the following chapters:

- ▶ Chapter 12, “SAP R/3 system copy” on page 365 provides information on setting up and performing an SAP R/3 system copy. It is meant to be used as a supplement to existing SAP R/3 documentation.
- ▶ Chapter 13, “Daily tasks to prevent error situations” on page 407 describes tasks that can be used on a regular basis to detect and prevent error situations in an SAP R/3 landscape.
- ▶ Chapter 14, “Troubleshooting and first aid after system failure” on page 437 outlines a structured path for problem determination in an SAP R/3 infrastructure and describes problem determination activities on operating system, database, and SAP R/3 levels.

The team that wrote this redbook

This redbook was produced by a team of specialists from Germany working at the International Technical Support Organization, Austin Center.

Diana Gfroerer is an International Technical Support Specialist for IBM @server pSeries and AIX Performance at the International Technical Support Organization, Austin Center. She writes extensively and teaches IBM classes worldwide on all areas of AIX, with a focus on performance and tuning. Before joining the ITSO in 1999, Diana Gfroerer worked in AIX pre-sales Technical Support in Munich, Germany, and led the Region Central, EMEA, and World Wide Technical Skill Communities for AIX and PC Interoperability.

Dr. Andreas Gruber is a Senior IT Specialist at IBM Global Services in Munich, Germany. He holds a Ph.D. in Physics from the University of Munich. He has ten years of experience in high performance UNIX computing. He has worked at IBM for five years. His areas of expertise include AIX, RS/6000, HACMP, and SAP R/3. Dr. Andreas Gruber is a Certified SAP R/3 Basis Consultant.

Yvonne Kratz is a Senior IT Specialist from IBM Germany. She has eight years of IT experience, including three years of implementation experience in SAP R/3 environments. She holds a Bachelor's Degree in Technical Information Systems from the University of Applied Studies in Mannheim. Her areas of expertise include AIX, relational database management systems, and TSM. Yvonne Kratz is a Certified SAP R/3 Basis Consultant.

Dr. Basil Moshous is a Senior IT Specialist at IBM Global Services in Munich, Germany. He holds a Ph.D. in Physics from the Technische Universität München. He has six years of experience in scientific computing on UNIX. He has worked at IBM for two years in the field of SAP R/3 infrastructure design and implementation. His areas of expertise include AIX, Linux, and SAP R/3. Dr. Basil Moshous is a Certified SAP R/3 Basis Consultant.

Tobias Mueller is a Senior IT Specialist in Germany. He has five years of experience in the AIX and SAP R/3 infrastructure field. He holds a master degree in computer science from Technische Universität München. His areas of expertise include AIX, High Availability, and relational database management systems. Tobias Müller is a Certified SAP R/3 Basis Consultant.

Torsten Opitz is a Senior IT Specialist at IBM Global Services in IBM Germany. He has worked at IBM for nine years and has eight years of experience in the AIX and SAP R/3 field. His areas of expertise include HACMP, SAP R/3, and SAP R/3 Migrations. He holds a degree in Computer Science from Technical University Dresden. Torsten Opitz is a Certified SAP R/3 Basis Consultant.

Olaf Siebenhaar is a Certified Consulting IT Specialist at IBM Global Services in IBM Germany. He has worked at IBM for nine years and has eight years of experience in the AIX and SAP R/3 field. His areas of expertise include HACMP, SAP R/3, and SAP R/3 Migrations. He holds a degree in Computer Science from Technical University Dresden. Olaf Siebenhaar is a Certified SAP R/3 Basis Consultant.

Uwe Thiel is a Senior IT Specialist at IBM Global Services in Munich, Germany. He has nine years of IT experience, including six years of infrastructure implementation experience in SAP R/3 environments. He holds a Bachelor's Degree in Technical Information Systems from the University of Applied Studies in Stuttgart. His areas of expertise include AIX, TSM, and technical infrastructure. Uwe Thiel is a Certified SAP R/3 Basis Consultant.

Thanks to the following people for their invaluable contributions to this project:

International Tech Support Organization, Austin Center

Budi Darmawan and Wade Wallace

IBM Austin

Matthew Accapadi and Richard Cutler

IBM Chicago

David Sacks

IBM Dallas

Dan Braden

IBM Foster City

Walter Orb

IBM Germany

Herbert Diether, Justus Reich, Carsten Weinelt

Special thanks to the following people without whom this project could not have been realized:

IBM Germany

Dr. Antonio Palacin, Gerd Rechmann, Uwe Schütt, Nurcan Tezulas, Fred Wenzel

Finally, we would like to thank our customers that have supported our ideas to implement new solutions.

Jürgen Schlagenhauser's leading edge technology environment motivated us to reach the ultimate goal of superior SAP R/3 performance.

Alex Hufnagl supported our team with invaluable material and information for creating this book.



Special notice

This publication is intended for all people who are involved in the design, implementation, and operation of an SAP R/3 infrastructure, including system programmers, system administrators, IT specialists, IT managers, consultants, sales representatives, and project managers. The information in this publication

is not intended as the specification of any programming interfaces that are provided by AIX Version 4.3.3 or SAP R/3. See the PUBLICATIONS section of the IBM Programming Announcement for AIX Version 4.3.3 for more information about what publications are considered to be product documentation.

IBM trademarks

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

AIX®	AIX 5L™
Approach®	Balance®
Chipkill™	Cross-Site®
CUA®	DB2®
DB2 Universal Database™	Domino™
e (logo)® 	Enterprise Storage Server™
ESCON®	FICON™
FlashCopy™	IBM ®
Informix™	iSeries™
Lotus®	Lotus Notes®
Magstar®	Manage. Anything. Anywhere®
NetView®	Notes®
OS/2®	PBT®
Perform™	Planet Tivoli®
Predictive Failure Analysis®	pSeries™
Redbooks™	Redbooks Logo 
RISC System/6000®	RS/6000®
S/390®	SAA®
Sequent®	SP™
SP1®	StorWatch™
Tivoli®	Tivoli Enterprise™
TME®	Ultrastar™
xSeries™	zSeries™

Comments welcome

Your comments are important to us!

We want our IBM Redbooks to be as helpful as possible. Send us your comments about this or other Redbooks in one of the following ways:

- ▶ Use the online **Contact us** review redbook form found at:

ibm.com/redbooks

- ▶ Send your comments in an Internet note to:

redbook@us.ibm.com

- ▶ Mail your comments to the address on page ii.

Introduction

In a nutshell

This IBM Redbook:

- ▶ Defines three different models of reliable SAP R/3 infrastructure.
- ▶ Discusses all important aspects of the major components of these models.
- ▶ Gives some guidance for operation and maintenance of the SAP R/3 environment.

An SAP R/3 environment is reliable, if it has a design that prevents errors, reduces downtime, and optimizes the handling. This redbook describes and discusses all aspects of the life cycle of such a reliable SAP R/3 infrastructure. The redbook includes a variety of information that is important from the requirement analysis up to the operation of an SAP R/3 environment, including activities such as instance creation, operation, and maintenance.

1.1 A holistic approach

This redbook is intended for all people who are involved in the design, implementation, and operation of an SAP R/3 infrastructure. It gives some inspiration and comprehensive guidance to system programmers, administrators, IT managers, sales representatives, and project managers.

Every chapter addresses a certain audience and demands different levels of experiences and skills. In general, the first chapters discuss concepts while the last chapters contain technical details. You should be familiar with the fundamentals of the topics discussed in each chapter, because we only explain the principals in a short review.

The book is split into three parts, which contain aspects of the concept, the implementation and the operation of an SAP R/3 infrastructure.

For readers that are interested in a complete overview, we recommend you read the book from the beginning to the end. All other readers should read at least Chapter 3, "Architecture" on page 35, because there we define three different models of SAP R/3 environments, which are used throughout the whole redbook as a basis for discussions.

A suitable selection of chapters to read depends on the project phase to which you refer. For example, you should read different selections of chapters while searching information for bidding, proposal, implementation, or operation and maintenance. The following list provides a guidance through the chapters:

Part 1

In Chapters 2 and 3, we discuss requirements of a reliable SAP R/3 infrastructure and we derive three different implementation models, which are called *basic*, *fault-tolerant*, and *disaster-tolerant* model. This part is especially valuable for solution architects, sales representatives and IT managers.

Part 2

In Chapters 4 to 9, we discuss special aspects of components on which our models are based. These chapters are especially valuable for system programmers and administrators who are responsible for implementing an SAP R/3 infrastructure.

In chapters 10 and 11, we discuss specific tasks that arise during a project's life cycle. These chapters are especially valuable for system programmers and administrators who are responsible for these tasks.

Part 3

In chapters 12 to 14, we discuss the aspects of operation and maintenance of a reliable SAP R/3 infrastructure. This part is especially valuable for system administrators.

This redbook is based on up to eight years of individual experience with SAP R/3, collected by a team of seven Certified SAP Basis Consultants who gathered a lot of practical experience in numerous projects. The three different models of SAP R/3 environments represent the essence of the authoring team's personal experiences. Currently available components are used in these models.

This redbook does not contain universally valid statements that can be used in each and every situation. We try to give you the means to enable you to set up and maintain a well working SAP R/3 environment.

The principle of this redbook's content is to implement a reliable SAP R/3 infrastructure. We concentrate on the reliability of the SAP R/3 infrastructure, but also tune the SAP R/3 systems to get good performance values. However, an SAP R/3 system that is optimized for highest performance may interfere with the premises we use for our design of a reliable SAP R/3 environment. There is a trade-off between reliability and performance, depending on the business needs of the SAP R/3 system.

In the last few years, SAP has developed many new products with different flavors. Examples are the New Dimension products (like SAP CRM or SAP APO). Some of these products are based on a technical base similar to SAP R/3. For example, the SAP Business Warehouse has a strong affinity to a standard SAP R/3 system. Therefore, most of the concepts described in this book are also applicable to these products.

This book focuses on AIX 4.3.3, SAP R/3 Release 4.6C, TSM Version 4, HACMP Version 4, and the database management systems DB2 Version 6 and 7, as well as on Oracle Version 8.1.

Text boxes have been placed at the left hand side of the text throughout this redbook in order to highlight important sections.

Pitfall ahead!

A shaded text box with the words "Pitfall ahead!" is a marker for important information on certain problems that can arise within the section's scope.

Bright idea!

A shaded text box with the words "Bright idea!" emphasizes paragraphs in which the authoring team's solutions and experiences are described.



Part 1

Concepts

Requirements for a reliable SAP R/3 environment

In a nutshell:

- ▶ Define the requirements for your SAP R/3 environment.
- ▶ Create a service level agreement, especially for internal operations.
- ▶ Define roles, responsibilities, and processes for your SAP R/3 landscape.
- ▶ Adopt the guidelines for implementing and operating SAP R/3 systems.

This chapter provides a high level overview of the SAP R/3 architecture. It shows an SAP R/3 landscape and the systems that are relevant in such an environment.

In this chapter, we describe various requirements of a information system and define the key concepts of availability and reliability.

We give advice on guidelines that are important for the implementation of an SAP R/3 environment. The constraints for the design and the implementation are covered as well.

This chapter also describes the organizational aspects that are important for the operation of SAP R/3 systems. We present a model for the interactions in an SAP R/3 environment. The requirements inside the model and the implications for a service level agreement are discussed.

2.1 Characteristics of an SAP R/3 environment

This section describes, in a nutshell, the architecture of a single SAP R/3 system, a typical system landscape consisting of several SAP R/3 systems and a service layer model for SAP R/3.

2.1.1 SAP R/3 architecture

Generally, an SAP R/3 system is a distributed system with a three-tier client/server architecture. All SAP R/3 data is contained in a relational database management system, building the database layer. The whole processing of the data is done at the application layer, providing the business logic. Users access the system via a graphical front-end application (SAP GUI) running on their PCs. The set of all SAP GUI instances constitute the presentation layer. The software components of the three layers (database layer, application layer, and presentation layer) communicate with each other over networks. Figure 2-1 on page 9 illustrates the three-tier SAP R/3 architecture.

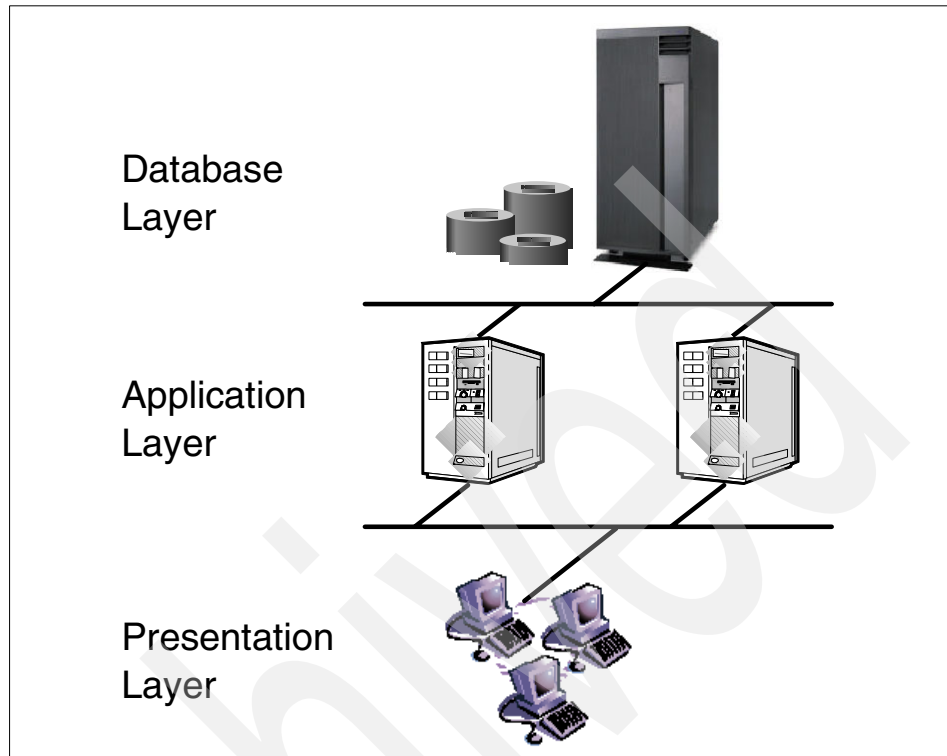


Figure 2-1 Three-tier SAP R/3 system architecture

There is exactly one database server per SAP R/3 system. An *application server instance* is a set of processes providing the business logic. The application layer is built up by one or more application servers instances. One instance is the so called *central instance* carrying special services for the communication within an SAP R/3 system. It is not mandatory that all layers are distributed to different hosts. For example, in a so called *central system*, the database server and central instance are combined on a single machine.

SAP GUI processes are clients of an application server instance. All application server instances are clients of the database. Furthermore, there can be external systems that exchange data with SAP R/3 via interfaces like Electronic Data Interchange (EDI) or that use Remote Function Calls (RFCs) for triggering actions inside SAP R/3.

SAP R/3 is mainly an Online Transaction Processing (OLTP) system where many users access business data in an interactive mode. This imposes far more read than write activity on the database. There is also batch processing for calculating business reports or processing data from external systems, for example.

Bright idea!

Batch processing should be limited to times with little interactive usage, because it can have a significant performance impact on the interactive users. It usually generates heavy writes to the database, thus invalidating any buffered or cached data.

The business functionality of SAP R/3 is grouped into modules, such as financial accounting (FI), material management (MM), and sales and distribution (SD). Although the installation of an SAP R/3 system makes all modules available, you have to customize the modules you intend to use according to your business needs.

2.1.2 SAP R/3 system landscapes

An SAP R/3 system landscape usually consists of at least three separate SAP R/3 systems used for development, quality assurance, and production. There can be even further systems for training or demonstration purposes, for example. Figure 2-2 shows the common SAP R/3 system landscape.



Figure 2-2 The common SAP R/3 system landscape

Of course, there are different requirements for aspects, such as performance or availability for each of the SAP R/3 systems in a landscape. A development system usually has a much smaller database than the production system because no transaction data is stored in it. Because the number of developers is normally only a fraction of the number of productive users, the development

system can run on a less powerful machine. Obviously, the production system has the highest demand on availability of the systems in an SAP R/3 landscape, so you probably want a fault-tolerant or even disaster-tolerant installation for production. Other implications arise when looking at storage.

Pitfall ahead!

A common request is to copy the production system onto the quality assurance system, because you want to test with current production data. Therefore, you certainly have to have the same disk space available for both systems.

2.1.3 Service layer model

The purpose of every information technology (IT) system is the provision of one or more services to the user with a reasonable performance and preferably without any unplanned downtime. The stacked layers of an SAP R/3 system, each delivering services to the next higher layer, are shown in Figure 2-3.

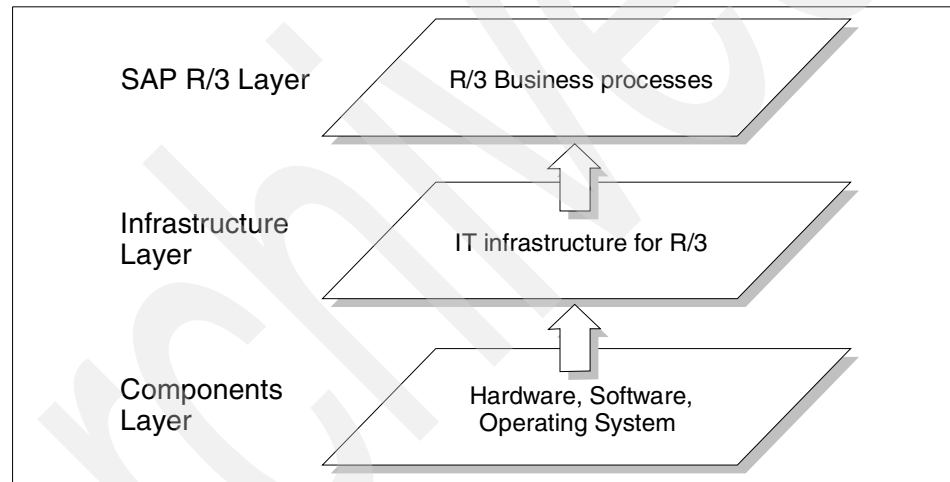


Figure 2-3 Service delivery model

An SAP R/3 system delivers business processes to the end user at the SAP R/3 layer. Below lies the infrastructure layer. The IT infrastructure for SAP R/3 is the assembled implementation of all technical components that are necessary for the operation. Substantially, these are the servers running the database and application code. But communication networks, storage subsystems, and facilities for backup and recovery also belong to the infrastructure layer. The infrastructure layer delivers the environment where the SAP R/3 application processes run. At the bottom layer reside single components of hardware and operating systems. These components that are part of the overall infrastructure deliver single services, such as storage space or a communication link to the distributed environment.

Different parties (Section 2.5, “Organizational aspects for SAP R/3 operations” on page 21) responsible for these three layers have to collaborate to provide an overall good and reliable SAP R/3 service for the end user. The users are not aware or even interested in the kind of hardware the database is running on or to which application server they are logged onto. They are only interested in the continuity of the service and the offered functionality.

If business requirements demand an availability of the SAP R/3 system beyond a certain percentage, the system has to be implemented upon a highly available infrastructure. In such an environment, every single point of failure (SPoF) has to be eliminated. This can be achieved through redundancy in components and cluster technologies.

Bright idea!

It is the responsibility of SAP R/3 basis administrators and IT system administrators to proactively monitor the system and to use proper system management techniques to reduce planned and unplanned downtime. Furthermore, there has to be an effective and tested backup and recovery concept at the infrastructure level to prevent data loss.

At the bottom line, fault resilient and robust hardware components, as well as stable operating systems, contribute to a reliable and highly available infrastructure.

2.2 Definition of requirements

In this section, we give some definitions of requirements that are essential for services provided by an information system. We distinguish between run-time and non-run-time requirements and have a closer look to the term downtime, because availability is one of the most important requirements for today's mission critical information systems.

2.2.1 Run-time requirements

The following list gives definitions for run-time requirements:

Performance	Given a specific workload, the performance of an information system can be measured in terms of response time, throughput, and costs.
Reliability	The degree of stability for the service provision of an information system.
Availability	The percentage of a time interval in which an information system was able to provide its service.

Scalability	The capability of an information system to function within a given quality of service when faced with varying complex workloads at run time.
Security	A set of personnel, data, operational, software, and hardware safeguards and procedures designed to prevent unauthorized use of a service or stored information.
Manageability	A degree of complexity of an information system that still allows trained system administrators to keep the system operational and to respond to error situations in an orderly manner.

2.2.2 Non-run-time requirements

The following list gives definitions for non-run-time requirements:

Efficiency	Effective provision of a service as measured by a comparison of profit with cost.
Scalability	The ability to enhance the performance of an information system by adding additional processing units, storage space, or network bandwidth without changing the system's overall structure.
Maintainability	The degree of ease of performing administration tasks or restoring an information system or components of it to a working state, if an error or a breakdown occurs.
Safety	Limitations of physical access to the hardware components of an information system and all precautions and installations to protect the system from damage caused by fire, water, or accidental human interventions.

2.2.3 Definition of high availability and downtime

In this section, we define terms regarding high availability and system outages.

A *fault-tolerant* system is capable of recovering from a software error or a failure in a hardware component. The design has to cover all possible single failures but cannot usually cover all double or multiple failures occurring at the same time. Thus, single points of failure have to be eliminated through redundancy in all critical components. Additional cluster software enables an automated recovery in the case of an error. At best, a user does not even notice an interruption of service. More serious failures, such as the breakdown of a central server, require a new logon to the system after a standby server has taken over.

A *disaster-tolerant system* is a fault-tolerant system that can provide its service even if a whole computer center is lost in case of, for example, a fire. A second computer center in another location has to be in place that can take over the operation. All data has to be mirrored between the locations. Usually a disaster will reduce the quality of service, but business critical data processing has to be held up.

There are two kinds of downtime, *planned* and *unplanned*, that you have to be aware of when implementing and operating a reliable, fault-tolerant or even disaster-tolerant system.

Availability is defined as the time where a system was able to provide its service reduced by the outages based on a given time interval. The outage time t_{out} is defined as the sum of planned and unplanned downtime:

$$t_{out} = (t_{out_planned} + t_{out_unplanned})$$

Planned downtime ($t_{out_planned}$)

There are maintenance tasks for an information system that require the system to be brought down. The following list shows common causes for planned downtime ordered by the frequency of occurrence:

- ▶ Routine maintenance windows
- ▶ Offline backups
- ▶ Changes to parameters that require a restart of the application
- ▶ Upgrades of the application, database, or operating system software
- ▶ Maintenance on hardware components
- ▶ Power outages

Bright idea!

Careful planning of maintenance tasks is essential to minimize the planned downtime. The goal is to combine as many single actions as possible into one maintenance window.

Unplanned downtime ($t_{out_unplanned}$)

The unplanned downtime is the time where a system is not available because of an error situation. Errors can be classified into hardware, software, and human failures. Unplanned downtime is defined as the time interval between breakdown and completed error recovery and can be further detailed as:

$$t_{out_unplanned} = t_{recognize} + t_{analysis} + t_{reconst} + t_{recall} + t_{restore} + t_{recovery} + t_{start}$$

Where:

$t_{\text{recognize}}$

Period of time that elapses between the occurrence of the error and recognition by a system administrator.

It can be minimized by preventive monitoring, possibly supported by a system management tool.

t_{analysis}

Period of time that is needed to analyze the error and to decide whether the problem can be fixed and if so how or whether the system has to be restored.

It can be minimized if there is a prepared list of known error situations and their solution (see Chapter 14, "Troubleshooting and first aid after system failure" on page 437).

t_{reconst}

Period of time that is needed (in the case of a disaster) for the reconstruction of the absolutely necessary infrastructure components, such as the assembly of replacement hardware and the installation of the operating systems.

It can be minimized if there are procedures prepared for reconstruction and if installation documentation and *up-to-date* system documentation is available (see Section 10.1.5, "System documentation" on page 282).

t_{recall}

Necessary amount of time to locate all needed backup medias and to bring them online.

It can be minimized through a good documentation of the location of backup medias that have been checked out, or through the use of a backup system with automatic media management and retrieval (see Chapter 7, "Backup and recovery" on page 185).

t_{restore}

Period of time that the technical restore of all lost information takes.

It can be minimized through a well implemented recovery concept based on a parallel restore from disk or fast tapes.

t_{recovery}

Period of time for the forward recovery of a database to a certain point in time.

It can be minimized through appropriately powerful

hardware and if available parallel recovery (see Chapter 11, “Performance” on page 305).

t_{start} Period of time for starting the system after completed recovery.

The overall goal for implementing and operating a reliable system is to minimize the unplanned downtime. Fault-tolerant or disaster-tolerant systems cannot reduce the probability for unplanned downtime, but they can vastly reduce the length of it. Consider all the given factors and the maximum amount of downtime that is acceptable for your business. Based on that, constitute the requirements for your system in a service level agreement. Refer also to Section 2.6, “Service level agreement in an SAP R/3 environment” on page 24 for more details.

2.3 Constraints for real environments

Even if it is possible to plan the implementation of an information system from scratch, there are constraints that can negatively affect the design of an optimal infrastructure. In this section, we divide the constraints into several categories and give examples that, however, do not represent a complete list.

- ▶ Technical
 - The reuse of existing server machines
 - The integration with already installed network technologies, such as token-ring
 - The requirement to integrate legacy systems into a new environment
 - The need to interoperate at certain internal or external interfaces with defined standards or protocols
- ▶ Personal
 - Existing skills of staff imply the choice of operating or database management systems
- ▶ Business
 - Project plans and deadlines
 - Budget limitations
 - Corporate implementation standards or guidelines
- ▶ Organizational
 - Centralized versus decentralized processing and storage of data for several company locations
 - Inhouse operation versus outsourcing of an information system

- Topology and environmental
 - Geographical location and number of the computer centers (important for disaster-tolerant environments)
 - Available floor space and weight limits for the raised floor
 - Capacity of available air conditioning and chillers

2.4 Guidelines for the implementation

In this section, we provide guidelines for implementing a reliable infrastructure for SAP R/3. These concepts are universally valid and do not depend of specific hardware components or software products that we use for our infrastructure models in Section 3.1, “Components of the infrastructure” on page 36.

Bright idea!

Observing these guidelines for the implementation in this chapter leads not only to a more reliable and stable system environment, but also helps to ease the operation of SAP R/3, and thus reduce the total costs of ownership (TCO).

Obey the SAP standards

For several reasons, it is always a good idea to stick very closely to the concepts and directions given by SAP in their installation manuals, operating system requirements lists, and design White Papers for system management and high availability. These documents are available from the SAP Service marketplace at:¹

<http://service.sap.com/instguides>

or

<http://service.sap.com/systemmanagement>

In the first place, it is easier to get support from SAP in case of problems, if your SAP R/3 system is installed according to SAP standards. Also, the lookup of possible failure causes in the SAP Notes database is more efficient. You can find all the SAP Notes referred to in this redbook at:

<http://service.sap.com/notes>

If you cannot find a solution for your problem in that database, you can create a customer message at SAP asking for help. Often, SAP technical service employees will then request access to your system for troubleshooting. They may refuse service if they detect that your system is not installed according to SAP guidelines.

¹ You need an SAPNet user ID for access to the SAP Service marketplace.

Pitfall ahead!

Even if you find that a nonstandard implementation or a certain workaround is functioning (given a particular release of SAP R/3), you cannot be sure that it will still work in a future release. In addition, if you make changes to SAP R/3 application start and stop scripts or environment definitions, you will have to reapply them after a release upgrade, because these scripts are frequently replaced during the upgrade procedure.

Use reasonable concepts for naming

Consistent naming of all infrastructure elements that exist company-wide eases documentation, inventory, and troubleshooting. The goal is to introduce names that are non-ambiguous over the whole system environment. Important examples are:

- ▶ Host names and labels for network adapters (IP labels)
- ▶ Names for storage objects, such as volume groups, logical volumes, file systems, storage pools, and so on
- ▶ Printer names
- ▶ User and group names and identifiers
- ▶ Cable labels

Be careful when coding locations into names. This makes locating components very easy but complicates change management if a component is moved. For example, the effort for propagating a name change of a printer within the system is usually higher than for moving the printer into another room.

Concepts for naming conventions are presented in Section 10.1.1, “Naming concept” on page 277.

Use unification

It is essential for a complex system environment to keep all parts as uniform as possible. This is true for hardware components, as dealing with several different device drivers often leads to a higher effort for testing.

Unification is also important for software components. Keeping operating system versions for different machines or database releases for different SAP R/3 systems at the same level simplifies system management, because usually every release level of software has its own problems and bugs. Program temporary fixes (PTF) should be applied to all instances of the program within a system environment. Of course, every PTF or bugfix should be tested before applying it to a production system.

Bright idea!

Generally, unification makes troubleshooting more easy and allows analogy conclusions. If a system shows a certain behavior, then an identically configured system is very likely to show the same behavior.

Implement a high degree of automation

Preferably all routine administration tasks should be automated. This significantly reduces the daily workload for system administrators and operating staff and simplifies administration. Furthermore, automation prevents that tasks are missed for single systems where required, or performed for wrong systems by accident. Important fields for automation are:

- ▶ Starting and stopping of SAP R/3 systems including all required auxiliary processes
- ▶ Archiving database redo log files and creating database backups
- ▶ Creating system images for backup
- ▶ Collecting up-to-date system documentation
- ▶ Scanning log files for errors
- ▶ Distributing software packages and configuration files

Manage global configuration information in a single place

Configuration information that is identical for a set of machines should be edited in a single place. This concept prevents the existence of outdated versions on different hosts.

Examples are directories for the lookup of host names and services or printer configurations. After changing these files on a master system, there should be an automatic distribution to all client systems.

Introduce administration domains

Concepts, such as unification and automation, enable the implementation of administration domains. Administration domains can specify the set of machines to which software packages and configuration files are distributed. You can schedule automated tasks for all members of an administration domain. This concept permits complex administration problems to be solved by simply assigning components to or removing them from an administrative domain.

The keynote is to group systems into domains that should be treated in the same way regarding administration. You can limit automated tasks or administration scripts to certain domains to reduce the danger of executing actions on wrong machines.

Observe obviousness and avoid side effects

You should be able to reproduce all changes to a system due to single implementation steps or administrative tasks. This can be made sure through the effective logging of all single steps or documenting of all affected components.

Avoid side effects or hidden actions whose impacts can complicate troubleshooting to a large extent.

Introduce housekeeping

At the time of implementing a system, you should pay attention that there is an effective housekeeping of temporary resources in place. You run into the risk of needing more and more disk space, because after a certain period of time, it is hardly possible to clean up the system manually.

Many components of a system create huge amounts of only temporarily needed information, such as log files, error reports, or dumps. In addition, users and administrators create temporary files needed for data exchange between systems, batch input, or administration tasks. You run into the risk of needing more and more disk space, because after a certain period of time, it is hardly possible to clean up the system manually. Some important areas for preferably automated housekeeping are:

- ▶ Remove core dumps from systems.
- ▶ Prune log files and keep only a few generations.
- ▶ Archive and delete temporarily needed information from interfaces and data exchange directories.
- ▶ Be sure that automated administrative tasks do not leave behind temporary files.

Create documentation and keep it up-to-date

As already mentioned, an accurate and current system documentation is essential for troubleshooting or for reconstruction after loss of parts of your environment in case of a disaster. Do not forget to update your documentation after applying changes to the system.

You should, preferably, implement mechanisms for automatically creating as-is configuration snapshots of all components in regular intervals and store this information electronically on an external server, keeping at least two generations.

2.5 Organizational aspects for SAP R/3 operations

SAP R/3 is a system that is used for Enterprise Resource Planning (ERP). It covers many aspects operating a company, such as accounting, controlling, materials management, production planning, or marketing. Therefore, the availability and performance of SAP R/3 affects many different departments in a company. In order to implement and run an SAP R/3 system in a reliable and stable manner, many different teams have to work together. In this chapter, we describe the organizational aspects that have to be considered for the operation of the whole system.

2.5.1 Roles in a productive environment

There are many people involved in the operation of an SAP R/3 environment, mainly because the architecture of the software is divided into many layers and because the scope of the business related tasks is very wide.

In Figure 2-4, we show the typical areas of responsibility and some important roles that constitute a productive environment.

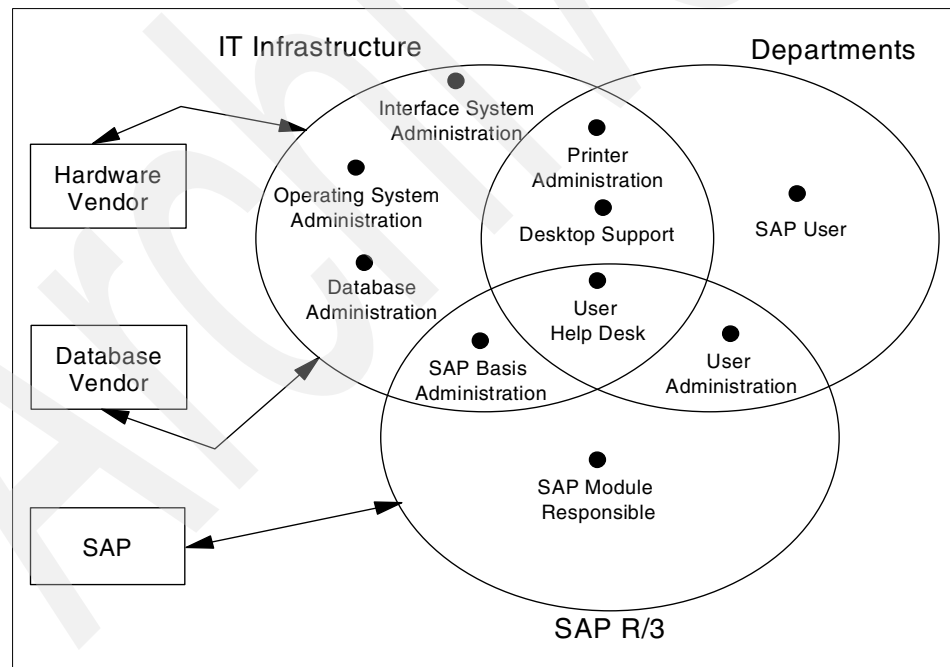


Figure 2-4 Responsibilities and roles in an SAP R/3 environment

There are three main areas of responsibility:

- ▶ IT Infrastructure
- ▶ SAP R/3
- ▶ Departments

The *IT Infrastructure* area contains the roles that are responsible for the hardware, operating systems, databases, middleware, communication, and external interfaces.

The *SAP R/3* area contains roles that are dealing with the SAP R/3 modules, including the SAP R/3 basis module and the implementation of business processes.

The *Department* area contains roles that have direct responsibilities for work in the departments, like the SAP R/3 users, the User Help Desk, or printer administration.

Each of the roles belongs to at least one area, but there are also roles that lie in the intersection of two or three areas. Thus, there are no strict allocation of roles to the areas. This reflects the way how companies work and how the responsibilities are distributed in organizations. The printer administration is an example for this, because the printing software is running centrally, but the printer hardware is very often purchased and run locally in the departments.

In every SAP R/3 installation, there are also vendors of software and hardware components involved that have a relationship to different roles inside the SAP R/3 production environment.

2.5.2 Interactions between roles in an SAP R/3 environment

There is a very complex relationship between the roles in an SAP R/3 environment. In order to illustrate this, we give some examples of typical processes:

Creating a new user

If there is a new employee in the department, this employee has to get an SAP R/3 user account. First, the new employee contacts the User Help Desk and asks for the new account. The User Help Desk checks with the SAP R/3 Module Responsible to see if the user will be granted an account and what kind of authorizations are necessary for this employee. The SAP R/3 Module Responsible instructs the User Administrator to create the user account and to assign the necessary authorization profiles.

Printing problems

An SAP R/3 user encounters a problem with printing and calls the User Help Desk. They check with the SAP R/3 Basis Administration and see that the print spool system is working correctly. The SAP R/3 Basis Administrators contact the Database Administrators because there seems to be a problem with the database. The Database Administrators find that there is not enough space in the database and ask the Operating System Administrators to increase the size of a file system. The SAP R/3 Basis Administrator tells the Printer Administrator to restart the print queues and to repeat the printing.

Upgrade of SAP R/3

The SAP R/3 Module Responsibles need new functionality, which is part of a new SAP R/3 release. The SAP R/3 Basis Administrators check, with the Database Administrators and the Operating System Administrators, the compatibility of the new release with the database and the operating system version. If there are dependencies, an action plan is developed that contains upgrades of the operating system or the database. They also contact the Interface System Administrator to make sure that there are no incompatibilities with the external systems. The SAP R/3 Module Responsibles ask the SAP R/3 users and check the interface systems when a downtime for an upgrade of the system is acceptable and schedule a date for the implementation. The Desktop Support is contacted to check whether a new SAP GUI is needed before the upgrade and whether the desktop hardware meets the requirements for a new SAP GUI release.

Bright idea!

As can be seen from these examples, the interaction between different roles in a productive SAP R/3 environment is very complex and there has to be very intense communication between the roles to fulfill the tasks. It is important to define standard processes for many tasks.

But there will always be situations where the standard processes will not be sufficient. This is when the informal network becomes important.

The informal network is a description of a way of communication between employees that know each other beyond the scope of their defined roles. This enables communication and information exchange between departments and employees that normally do not work directly with each other. This can be of invaluable help when there are problems that go beyond the boundaries of organizations. This informal network enables the solution of these problems by skipping intermediate levels of support and by thinking globally instead of being limited to small aspects.

2.5.3 Service provider model

The interactions between people in different roles can be modeled as a process where one party offers a service and the other party receives a service. Thus, the relationship between them can be described as that of a service provider and a service recipient.

This is a very common structure in the strategic outsourcing business, but can also be adopted to inhouse projects and system operations. The service provider offers a certain service and also offers a quality for this service. The service recipient can accept the offering or ask for a modification of the service or the quality of this service.

Pitfall ahead!

What you have to keep in mind is that every service has a certain cost structure, depending on the quality of the service. This is an obvious statement in an outsourcing environment, where everybody agrees that a higher performance of a system also costs more money. But this also proves to be true in internal projects, where it is more difficult to negotiate prices.

It is essential that all the parties are involved while discussing the service levels, because it may otherwise lead to unrealistic expectations or demands from one of the service recipients. Only a service provider can give reasonable numbers for the cost of a certain service.

2.6 Service level agreement in an SAP R/3 environment

Bright idea!

In a complex environment like SAP R/3, it is a reasonable approach to model the relationships of the people as an exchange of services between a provider and a recipient. In this section, we describe general criteria that should be kept in mind when setting up and operating a complex system. We also mention important criteria that are especially suitable for SAP R/3 environments.

2.6.1 Definition of service level agreement

A service level agreement is a contractual arrangement that defines services and measurements for these services. It describes service level performance objectives with consequences for not meeting those objectives.

The main reasons for creating a service level agreement are:

- ▶ Management of the relationship between service provider and recipient
- ▶ Definition of the roles and responsibilities
- ▶ Definition of the services that are involved

- ▶ Definition of measurements for the services

It helps to define different types of measurements for structuring the service level agreement. The following types may be used:

- ▶ Response time
- ▶ Completion time
- ▶ On time delivery
- ▶ Availability
- ▶ Capacity

These are the measurable quantities that can be used in contracts or agreements.

2.6.2 General criteria for service level agreements

There are many different areas of service that can be measured and may be used in a service level agreement, depending on the environment. In the following sections, we describe a selection of important service areas and what they cover.

Server systems

In the server area, the main quantities for measuring services are:

- ▶ Time of planned maintenance outage per month
- ▶ Time of unplanned operating system outage per month
- ▶ Hardware outage times per month
- ▶ Peak and average CPU utilization of servers
- ▶ Peak and average memory utilization of servers
- ▶ Peak and average I/O utilization of servers

Storage

In the storage area, there are many different categories that can be measured and used in a service level agreement:

- ▶ Tape or optical storage capacity
- ▶ Tape mounts per month
- ▶ Simultaneously used tape drives
- ▶ Disk storage capacity
- ▶ Maximum throughput per server

- ▶ Disk hardware outage times per month
- ▶ Tape hardware outage times per month

Network

In the network area, the availability and capacity services that can be measured are:

- ▶ Network bandwidth
- ▶ Network response time
- ▶ LAN utilization
- ▶ Network availability per month

Application

In the application area context, the following service criteria are measurable:

- ▶ Response time of transactions
- ▶ Time of planned application maintenance per month
- ▶ Time of unplanned application outage per month
- ▶ Batch processing completion on time

Backup and recovery

In the backup and recovery area, the following major services are available:

- ▶ Backed up files per month
- ▶ Backup volume per month
- ▶ Restored files per month
- ▶ Restored volume per month
- ▶ Restore speed
- ▶ Number of successful recovery tests

Printing and output distribution

In the printing area, the quantities that can be used as criteria in a service level agreement are:

- ▶ Number of printers
- ▶ Number of pages per month
- ▶ On time delivery of output

Customer service center

The customer service centers may be measured by the following services criteria:

- ▶ Calls per month
- ▶ Average call service time
- ▶ First call resolution rate
- ▶ Speed to answer calls

Problem management

Problem management is closely related to the customer service center and has the following measurable services:

- ▶ Problems per month
- ▶ Reaction times
- ▶ Resolution times

Change management

Change management, in terms of software development or documentation, may be covered by the following criteria:

- ▶ Change requests per month
- ▶ Changes that have completed on time

Asset management

The asset management contains the management of orders and assets. The following service level measurements exist:

- ▶ Orders per month
- ▶ Order placement time
- ▶ Number of tracked assets

2.6.3 SAP R/3 specific criteria

Bright idea!

In the previous section, we have given a list of service areas that are not dependent on a specific application. This list is useful for defining service levels for single system components, as it is often used in outsourcing engagements. In the context of an SAP R/3 environment, it is useful to have a more holistic approach for the definition of service levels.

We split the specific criteria into two categories, representing different groups of general service level criteria. The SAP R/3 specific criteria are using a business process oriented approach, and each of them consists of a combination of several general criteria.

Performance and capacity related criteria

These criteria provide an end to end view instead of giving an isolated measurement that is derived from the performance or availability of a single subsystem. They comprise the items *server system*, *storage*, *network*, and *application*.

► Response Time

The response time is defined as the elapsed time between the end of an inquiry or demand on a computer system and the beginning of a response. SAP R/3 measures the response time of every transaction. In a real environment, there are transactions that are used more often and others that are run only once per month. There might also be transactions that are very time critical because other systems rely on a response in a given time.

Therefore, it is important to define service levels for each transaction or for a group of transactions.

Because of statistical fluctuations, even the response time for a specific transaction is always distributed. Therefore, it makes sense to define a percentage of the number transactions that should not exceed a certain threshold.

► Run time

The run time is a parameter that is similar to the response time. While the term response time is used for the description of elapsed times in interactive applications, the term run time is used to describe the time for long-running applications, also known as batch jobs.

Due to the use of non-standard programs for batch jobs, the service level objectives for run time should be defined separately for each job.

► Number of users

In SAP R/3 environments, this term is used in different contexts. One definition describes the number of *named users*, that is, the number of accounts with login names and passwords.

More important in terms of service levels is the number of *concurrent users* that are logged on to the system and active at the same time, because only these are consuming resources of the system.

Pitfall ahead!

- Number of batch queues

The number of batch queues can be defined in the SAP R/3 configuration files. They determine the maximum number of batch processes that can run simultaneously. It is important to choose this number depending on the available hardware resources. Due to the scalable architecture of SAP R/3, it is possible to add resources dynamically. This scalability is only limited by the performance of the database server.

Availability related criteria

These criteria give service levels for the availability of SAP R/3 system components, but they are not directly related to performance. They cover aspects of the general criteria for *customer service center*, *problem management*, *change management*, *printing and output distribution*, and *backup and recovery*.

Pitfall ahead!

- Service hours

Every business unit must define the service hours during which the SAP R/3 system has to be available and the kind of support that has to be available. Depending on the time of the day and day of the week, there can be different service levels in place. This helps reduce costs because there does not have to be user support during night times or weekends.

- Planned downtimes

SAP R/3 environments still require downtimes in case of maintenance of operating system, database, or application upgrades. It is advisable to define time windows in advance for this type of regular work.

- Print management

Printing is an extremely sensitive area in SAP R/3 environments because many business processes rely on documents. In most companies today, the print systems are decentralized, which means that there are many printers distributed in departments and offices.

For high volume print services, it is important to have a reliable print subsystem in the operating system or a specialized printing software. It is also necessary to have on-site support for the various printer related problems.

- Database growth

While operating an SAP R/3 system, there is a constant growth of data inside the database. The increase in volume depends on the business type and the SAP R/3 modules that each company deploys. The growth has to be monitored and projected into the future.

You have to specify certain size limits when appropriate actions should take place, for example, the archiving of data, the reorganization of database contents, or the acquisition of additional disk space.

- ▶ Interface availability

An SAP R/3 system is often the central source for information, and exchanges data with a lot of external systems, like planning systems, automatic storage and retrieval systems, or points of sale. These external interfaces are business critical, and the availability and throughput of the communication paths is an important service.

- ▶ Transport management system

The software development in SAP R/3 is managed and controlled by the Transport Management System. It is a vital component in the development and test cycle and the availability of this component is mandatory for fixing software defects. It is very often coupled or integrated with a workflow system that manages the authorizations for software or component releases.

- ▶ Problem resolution

In case of problems, there has to be a defined process for resolving them. There might be several owners that are involved in problem determination and solving; the developer of the software component, the database administrator, or the operating system administrator.

It is not reasonable to define a maximum time for the resolution, but it makes sense to define a maximum time until there is a response to a problem and how the problem resolution can be managed.

- ▶ Restore time

The design of the system should take into account aspects of business continuation in case of disasters, if there is a business requirement. But even if there is a disaster-tolerant architecture in place, it will be necessary to recover the damaged data center after a disaster to restore the full performance. In these cases or for logical errors, it is vital to restore the data from long-term storage.

The time until the SAP R/3 system is operational again or the time until the full performance has been restored is an important criterion.

2.7 Guidelines for operating an SAP R/3 environment

Operating a productive SAP R/3 environment is a complex task that involves many people. There are some general guidelines that should be obeyed in order to make the operations as smooth as possible. In this section, we describe some general rules and guidelines that help to organize the work.

2.7.1 Definition of general rules for interaction

In Section 2.5, “Organizational aspects for SAP R/3 operations” on page 21, we show some of the key roles in a typical SAP R/3 environment. We now define guidelines for their interactions that we have drawn from our experience with many different SAP R/3 environments.

Bright idea!

It is not possible to define exactly which department or employee has to take on which role because this would limit the freedom of management for organizing their companies. Nevertheless, it is possible to assign tasks and work objects to roles that are responsible for them.

In order to write about rules for interaction, we define new terms that facilitate the description of the operational processes:

Object	A general term for describing a piece of software, hardware, or area of responsibility.
Owner	A role that is responsible for an object.
Requestor	A role that is requesting a certain service.
Executor	A role for the implementation of a service task.
Recipient	A role that is receiving the service.
Time scale	Amount of time that is usually needed to implement the service task.

A prerequisite for the operation of the system is the definition of a service level agreement between the owner of objects and the requestors or recipients of the services. Besides the measurable quantities and criteria there should also be documents that define contact persons, action items in case of non-fulfilment of services, and reconciliation boards in case of critical situations.

The typical process flow for an operation is as follows:

- ▶ The requestor is asking for a modification of an object.
- ▶ The owner is checking dependencies of the modification with other objects and coordinates the modification with other owners.
- ▶ The owner of the object assigns the task to the executor.
- ▶ The recipient is receiving information upon completion of the task.

During this process, all activities should be recorded automatically and documented as detailed as possible. To make sure that the interdependencies between objects are known, every owner is responsible for the documentation of critical processes where their objects are involved. There should be test scenarios for the critical processes, so that regression tests can be done.

2.7.2 Landscape-wide operational processes

Every SAP R/3 landscape consists of several systems, with a development and a production system as the minimum configuration. Some processes affect the whole landscape and are described in this section. As an example we show the main operational processes in Table 2-1. Table 2-1 and Table 2-2 on page 33 should be augmented by the columns for *Requestor*, *Executor* and *Recipient* in a productive installation. We do not want to dictate a certain organizational structure and thus leave these columns out.

Table 2-1 Landscape wide operational processes

Owner	Object	Task	Time scale
SAP R/3 administrator	SAP system	Installation System copy Releasing system Release upgrade Hot Packages Deletion of system	Week Week Hour Week Hour Hour
	Customizing	Creation Test Release	Day Day Minute
	Programs	Creation Test Release	Day Day Minute
	Transports	Creation Packaging Release Export Import	Minute Minute Minute Minute Minute

2.7.3 Single system operational processes

The processes that have to be defined on each SAP R/3 system are described in Table 2-2. There are, of course, processes that are always affecting more than one system, for example, the downtime of the master system for the transport management. All objects that influence other objects on different systems have to be analyzed by the owners as mentioned before.

Table 2-2 Operational processes related to single systems

Owner	Object	Task	Time scale
SAP R/3 basis administrator	Instance	Creation Configuration Monitoring Deletion	Day Hour Minute Hour
	SAP client	Copy Release Deletion	Day Hour Hour
	SAPNet connection	Unlocking Locking	Minute Minute
	Batch jobs	Creation Planning Monitoring Recovery	Hour Minute Minute Minute
User administrator	Authorization profile	Creation Test Release	Hour Hour Hour
	SAP user	Creation Deletion	Minute Minute
Interface system administrator	Interface	Development Test Release Monitoring	Day Day Hour Minute
	Data transfer	Setup Monitoring Deletion	Hour Minute Hour
Database administrator	Database	Reorganization Monitoring Backup and Recovery	Hour Minute Hour

Owner	Object	Task	Time scale
Operating system administrator	High availability	Ensure availability	Hour
	Operating system and scripts	Monitoring backup, recovery upgrades, and fixes	Minute Hour Hour
	Hardware	Procurement Integration Monitoring	Week Day Hour
	Network	Setup Monitoring	Hour Hour
Printer administrator	Printing list	Creation Management	Minute Minute
	Printer definitions for SAP+operating system	Creation Monitoring Deletion	Minute Minute Minute
Desktop support	SAP GUI SAPHelp	Installation Upgrade	Hour Hour
User help desk	SAP R/3 user problems	Receive calls Resolution Escalation	Minute Minute Hour
	SAP R/3 technical problems	Receive calls Resolution Escalation	Minute Minute Hour
	SAP R/3 infrastructure problems	Receive calls Resolution Escalation	Minute Minute Hour

Architecture

In a nutshell:

- ▶ Tailor the solution according to requirements in a service level agreement.
- ▶ Use only reliable hardware and software components for building systems.
- ▶ Identify and assess single points of failure in your system design.
- ▶ Maintain a holistic approach when integrating the components.

This chapter provides information on the basic building blocks for an SAP R/3 environment. We show the software and hardware components that are needed and suitable for a reliable implementation.

The main part describes three different solution models that cover three different sets of requirements:

- ▶ The basic model is a reference model, which covers the minimal needs of every SAP R/3 implementation.
- ▶ The fault-tolerant model shows the extensions that have to be made if the requirements ask for a higher availability.
- ▶ The disaster-tolerant model describes a scenario that allows business continuation in case of a disaster.

These models are referenced in the following chapters where single implementation details are given.

3.1 Components of the infrastructure

In this section, we explain the structure of an SAP R/3 environment and describe the elements of the SAP R/3 software architecture. We describe the components that an SAP R/3 system landscape is made of. They are the elements we choose from when creating a complete operating environment for SAP R/3.

3.1.1 Structure of an SAP system

We briefly introduce the different types of SAP R/3 systems that are used in productive environments and explain the internal architecture of SAP R/3.

System landscape

SAP has adopted a common software development methodology, which consists of the typical software development life cycle phases:

- ▶ Development
- ▶ Quality assurance
- ▶ Production

There should be a separate SAP R/3 system for each of these phases.

Development system

The development system is used to customize the SAP R/3 software according to the needs and specification of the company it is used for. If the standard features of the SAP R/3 software are not sufficient, it is possible to add functionality through programming inside the SAP R/3 framework.

The information about customization settings and software developments are automatically recorded by the SAP R/3 software. This information is stored inside the development system and can be exported to a shared file system, so that other systems can import this information. This mechanism of transportation is used by the Transport Management System (TMS) of SAP R/3.

The development system is used by SAP R/3 specialists, such as consultants and developers. The database size is usually small because there is no master data stored in the system. There is also no extensive growth of data because there are no business transactions running on this system.

Quality assurance system

The quality assurance system is used to make tests of the customization settings and the new or modified programs on larger and realistic data sets. The database size is identical to the size of the production system and will grow accordingly.

In order to have a test environment of the quality assurance system that is mostly identical to the production system, it is recommended that you perform regular copies of the production database. The quality assurance system is also mainly used by SAP R/3 specialists that run integration tests on this system. Due to the fact that the work on this system very often consists of mass tests, the load on the hardware can be very high.

Production system

The production system is used for the business transactions of a company. Customizations and programs that have passed the tests at the quality assurance system are transported into the production system. The production system is used by the SAP R/3 users and by external systems for data interchange. The database contains master data and transaction data and usually grows constantly at varying speed.

SAP recommends a minimal system landscape that consists of a development system and a production system, which should both run on their own physical hardware.

SAP offers the choice to run the quality assurance tests in a separate client inside the SAP R/3 development system. We do not recommend this approach because it is a tedious and unreliable task to copy the whole master and transactional data from the production system into the development system. SAP also does not officially recommend that you copy the production clients, due to their size and the resulting fragmentation of tables in the target system.

The implementation of the three SAP R/3 systems on their own hardware prevent any performance impacts of tests done at the development or quality assurance system on the production system.

In many SAP R/3 environments, there are the following additional systems involved:

- ▶ Training
- ▶ International Demo and Education System (IDES)
- ▶ Migration
- ▶ Pre-production

The training system is used for the education of the SAP R/3 end users. In companies with several hundred or thousand employees that have to be trained in the use of SAP R/3, it makes sense to run a separate training system. These systems usually operate on a copy of the production system or a subset thereof.

The IDES system is a predefined system from SAP that contains a virtual company with a typical customization. It can be used for the education of SAP R/3 users and developers.

The migration system is typically used for testing SAP R/3 release upgrades or upgrades of the database and operating system.

The pre-production system is used in highly sensitive companies like banks or insurance companies. For security reasons, they do not copy the production databases onto the quality assurance systems but keep only a subset of master and transactional data there. The pre-production system is very often an exact copy of the production system in terms of software and hardware. The final test for all modifications, including hardware components, operating system, database and other software components, is done in the pre-production system.

Database server

One of the core components of an SAP R/3 system is the Relational Database Management System (RDBMS). It contains data structures and the source code for the implementation of the business processes. The SAP R/3 system supports products from a number of different RDBMS vendors.

The database is a single entity in the SAP R/3 architecture, which means that it is not, in general, easily scalable by adding more instances of it. Although there is a parallel RDBMS implementation based on DB2 for S/390, this is out of the scope of this book. The only other product on the market that is supported by SAP is Oracle Parallel Server, but SAP does not encourage the use of it either as a high availability solution or as an option to boost performance.

Pitfall ahead!

It is important to choose a powerful and scalable platform for the database server so that the growth of performance demands can be satisfied without changing the system structure or the hardware.

Due to this property of the database server being a single instance, it is also important to deploy a reliable hardware platform or even increase the reliability by the implementation of a high availability hard- and software solution.

The RDBMS is providing access to the business data and is responsible for storing the transactional data. Thus, it needs a powerful I/O subsystem and I/O connections with a high bandwidth.

Application server

An application server is a physical system that is running one or several SAP R/3 *application server instances*. The application server instance is a collection of processes and memory areas that are responsible for executing programs and offering services.

The business logic of SAP R/3 is implemented with the 4th Generation Advanced Business Application Programming language (ABAP/4). It is an interpreted programming language that is executed by the SAP R/3 processes. This allows you to keep the source code for the business transactions platform independent. Only a fraction of about ten percent of the SAP R/3 code has to be applied to the different operating system platforms.

In Figure 3-1 we depict the processes that are communicating in an SAP R/3 system. We show the different types of services that are necessary for the correct operation of the SAP R/3 system. All the processes are communicating with the database, which we do not show in the figure for simplicity's sake. The SAP R/3 users connect to the SAP R/3 system from their SAP GUIs through the Dispatchers or indirectly through the Message service.

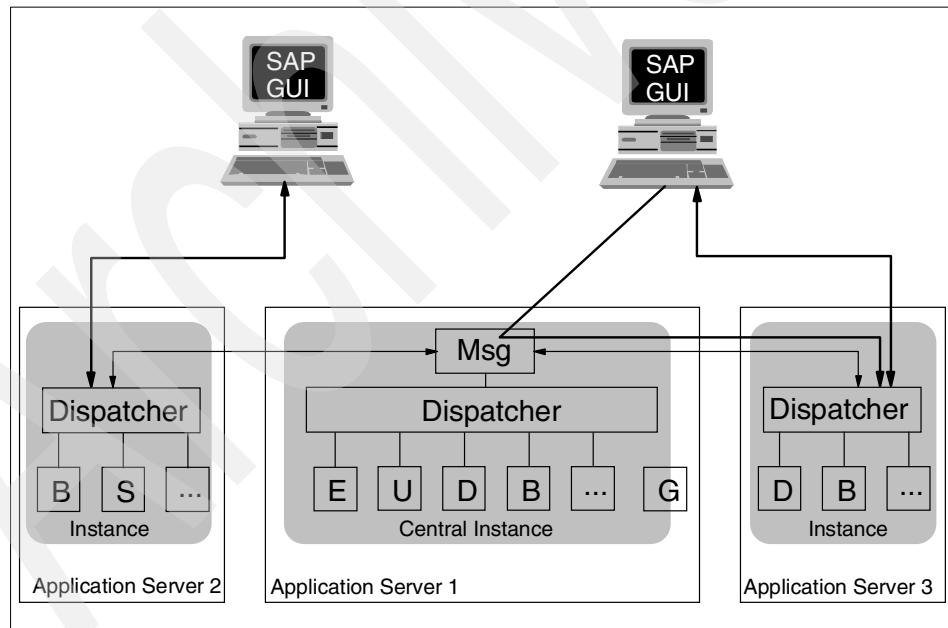


Figure 3-1 Architecture of SAP R/3 process communication

The following services are implemented as processes in a UNIX environment and have the following main objectives:

Message (Msg)	This service coordinates the communication between different instances of a single SAP R/3 system. It is also used for logon purposes, where a load balancing algorithm is implemented.
Dispatcher	The dispatcher service receives the requests from the SAP GUI clients, selects one of the dialog processes for execution, and redirects the requests to a free process.
Dialog (D)	The dialog service is interpreting the ABAP/4 code and is executing the business logic. This and the following types of processes are also called <i>work processes</i> .
Batch (B)	The batch service is processing programs that do not need interaction with SAP R/3 users.
Enqueue (E)	The enqueue service is responsible for the locking of business objects that are used in critical transactions.
Update (U)	The update service works on changes to the database that are requested asynchronously. There are two different types of update work processes: the V1 and the V2 services. The fundamental difference between these two types is the handling of SAP R/3 locks (enqueues). A V1 update work process will process all V1 modules of an update request and then release the SAP R/3 locks. If V2 update work processes exist and there are V2 update modules used, the work process will then pass the update request on to a V2 update work process.
Spool (S)	The spool service is creating spool requests and it is formatting the output for print requests.
Gateway (G)	The gateway service is responsible for filtering client requests and for the connection to external systems. It can run as a standalone process on a separate system and thus is shown in the figure as belonging only partially to the instance.

The different services can be combined in various ways but there are rules for the minimum and maximum number of processes, both regarding an instance and the whole SAP R/3 system. These numbers are summarized in Table 3-1.

Table 3-1 Number of processes per SAP R/3 instance and system

Service	Number of processes per SAP R/3 instance	Number of processes per SAP R/3 system
Message	0 or 1	1
Dispatcher	1	Number of instances
Dialog	≥ 2	≥ 2
Batch	≥ 0	≥ 1
Enqueue	0 or 1	1
Update	≥ 0	≥ 1
Spool	≥ 0	≥ 0
Gateway	1	Number of instances

Checking this table shows that there are two services in an SAP R/3 system that are potential single points of failure (SPoF): the message service and the enqueue service. Due to this property, the instance that is hosting these two services is also called the *Central Instance (CI)*.

The work processes of an SAP R/3 instance are using a common area of memory for the interprocess communication and the buffering of user data and object code. The available physical and addressable virtual memory are important factors for the performance of the SAP R/3 system.

Pitfall ahead!

The maximum amount of memory that can be used for SAP R/3 is usually limited by the operating system. The use of a 32-bit operating system results in a theoretical limit of 4 GB virtual memory. The realistic limits are below that. For AIX, it is about 2.7 GB per process. The use of 64-bit hardware and operating systems lifts these limitations to values that are no longer a restriction for the SAP R/3 system.

3.1.2 IBM @server pSeries and RS/6000

The IBM @server pSeries and IBM RS/6000 servers are designed for a broad range of applications serving small, medium, and large businesses. They are symmetric multiprocessor (SMP) machines that are well suited for mission-critical commercial, large e-business, or Enterprise Resource Planning (ERP), Supply Chain Management (SCM), and Customer Relationship Management (CRM) environments. The systems are all based on 64-bit processor and memory architectures for fast access to large amounts of data.

Types of servers

In this section, we describe IBM @server pSeries and RS/6000 types that are suitable for an SAP R/3 platform. For a description of all types and models, please refer to the *RS/6000 Systems Handbook 2000 Edition*, SG24-5120 and the World Wide Web at <http://www.ibm.com/servers/eserver/pseries>

Bright idea!

In order to satisfy the requirements for reliability, manageability, and scalability, we only consider rack servers and enterprise servers. Tower servers do not fulfill the needs of a data center environment where SAP R/3 systems have to run. The models in Table 3-2 are considered as possible platforms in the following scenarios.

Table 3-2 IBM @server pSeries and RS/6000 server types

Type	Number of CPUs	Clock rates in MHz	Memory	PCI slots
p640	1-4	375	256 MB - 16 GB	5
p660	1,2,4,6	600,668	512 MB - 32 GB	28
M80	2,4,6,8	500	1 - 32 GB	56
p680	6,12,18,24	600	4 - 96 GB	53

The models p660 and p680 are also available at lower clock rates of 450 MHz. Only the model p660 with six CPUs is only available at a clock rate of 668 MHz.

Reliability, Availability, and Serviceability (RAS)

The IBM @server pSeries and RS/6000 have been designed to provide inherent RAS features. The following features are found in all the sever types of Table 3-2:

- ▶ Error Checking and Correction (ECC) protection on main memory, L1 and L2 caches, and internal processors
- ▶ Predictive failure analysis on processors, memory, I/O components, and disks
- ▶ Fault tolerance with N+1 redundancy and concurrent maintenance for power and cooling

- ▶ Processor run-time and boot-time deallocation based on run-time errors
- ▶ Hot swappable disks
- ▶ Service processor
- ▶ Highly reliable components

Some of the servers even offer more advanced features:

- ▶ Chipkill memory for eliminating memory errors (p660 and p680)
- ▶ Hot-plug PCI slots (p660 and M80)

These features have been included to ensure that the servers operate when required, perform reliably, and efficiently handle infrequent failures in a non-disruptive fashion. They also provide the possibility to repair the system concurrently or on a deferred basis, minimizing downtime, or deferring it to a more convenient point in time.

More details on RAS features and implementation details can be found in the document *The IBM @server pSeries 680 Reliability, Availability, Serviceability*, found at:

http://www.ibm.com/servers/eserver/pseries/hardware/whitepapers/p680_reliability.html

RS/6000 SP

The SP system is basically a group of RS/6000 or IBM @server pSeries servers that can work collaboratively and are controlled by the Control Workstation (CWS). It acts as a single point of control and manages the whole group of RS/6000 or IBM @server pSeries servers, also called *nodes*, as a unique system. The CWS has two connections to each node: an SP-Ethernet network for the system administration and an RS-232-connection for hardware control. The node types shown in Table 3-3 are available for SP frames.

Table 3-3 RS/6000 SP classic node types

Type	Number of CPUs	Clock rates in MHz	Memory	PCI slots
Power3 Thin Node	2,4	375	256 MB - 16 GB	2
Power3 Wide Node	2,4	375	256 MB - 16 GB	10
Power3 High Node	4,8,12,16	375	256 MB - 64 GB	53

Three different types of SP systems are possible:

- ▶ The classic SP with the thin, wide, and high SP nodes, which only fit in a special rack, the SP frame. These systems can scale up to 128 nodes.
- ▶ SP systems with SP nodes in the SP frame with SP-attached servers. Up to 16 p680 servers, or up to 32 midrange RS/6000 M80 or p660 rack-mounted servers can be integrated in these SP-clusters.
- ▶ Clustered Enterprise Servers, a cluster of non-SP building blocks (RS/6000 M80 and IBM @server pSeries p660 and p680), which has the same building rules and limitations as the SP attachment of these servers.

Nodes or attached enterprise servers can be interconnected by an optional high bandwidth, low-latency switch for high-performance internode communications. The following switch options are available:

- ▶ The SP Switch offers a bandwidth of up to 150 MB/s for all node types.
- ▶ The SP Switch2 provides a bandwidth of up to 500 MB/s only between high nodes.

The concept of the RS/6000 SP offers a new level of manageability to a loosely coupled cluster of RS/6000 and IBM @server pSeries servers. The most important idea is the centralized system management performed by the Control Workstation. The CWS is responsible for installation and software maintenance of all cluster nodes. The configuration information of all nodes is stored in a central database on the CWS. Furthermore, the CWS is the central point of dynamic hardware and software monitoring, a functionality which can be regarded as the built-in monitoring of the whole SP cluster environment. With this approach, a higher availability and a better cost efficiency of system services can be achieved.

With the introduction of the enterprise server into the SP cluster concept, an extremely high scalability is possible, because now the horizontal scalability of the RS/6000 SP is combined with the vertical scalability of the enterprise server.

Performance classifications

The RS/6000 and IBM @server pSeries servers have a long history of outstanding performance and scalability. For the deployment of SAP R/3 on the different servers, it is essential to know the relative performance of the servers in comparison to each other, but also their performance in terms of SAP R/3 requirements. Performance numbers of RS/6000 and IBM @server pSeries server can be obtained from the following Web site:

http://www-1.ibm.com/servers/eserver/pseries/hardware/system_perf.html

The most important number is the Relative OLTP (ROLTP). It is an estimate of commercial processing performance derived from an IBM analytical model. The model simulates some of the system's operations, such as CPU, cache, and memory. However, the model does not simulate disk or network I/O operations. An IBM RS/6000 Model 250 is the baseline reference system and has a value of 1.0.

The most important figure in an SAP R/3 environment is SAPS, which is a definition of throughput coined by SAP capacity planning and performance testing personnel. 100 SAPS are defined as 2,000 fully business processed order line items per hour in the standard SAP SD application benchmark. The latest measured numbers and estimates for many types of server systems can be obtained from the infoservice at the IBM SAP International Competence Center (ISICC) which can be contacted at isicc@de.ibm.com or the benchmark page of SAP at <http://www.sap.com/benchmark>.

Thoughts on server selection

The most important step before the implementation of an SAP R/3 environment is the selection of a suitable server. All the types of servers shown in the previous section fulfill the requirements that have been specified in Chapter 2, "Requirements for a reliable SAP R/3 environment" on page 7 in respect to RAS features. However, there are more features that come into play when implementing SAP R/3. More input is needed for the proper selection of hardware components, for example:

- ▶ Number of SAP R/3 users and their activity profile
- ▶ Number and types of SAP R/3 modules
- ▶ Initial size of the database
- ▶ Growth of the database
- ▶ Activity of the database and the resulting I/O
- ▶ Number of interface systems
- ▶ Type of batch jobs and the affordable time window
- ▶ Backup time window
- ▶ High availability requirements

The process of gathering the input and doing some calculations based on experience with SAP R/3 systems is called *sizing* and should be done by a team of qualified sales and technical personnel.

Bright idea!

The sizing process will determine the requirements for the number of CPUs, the size of the memory, the number of network cards, the number of I/O interfaces, and the internal bandwidth of the system. Based on the results of the sizing process only certain types of servers will be matching the requirements.

3.1.3 Storage subsystems

Today's business relies heavily on the instant availability of large amounts of data. In the past few years, the ever-increasing demand on storage capacity has driven the development of new technologies for storage products. Spending for disk storage capacity has constantly increased due to the high demand of applications, although disk storage per GB has become less expensive. The increased storage volume also triggered the development of new tape technologies in order to provide long term storage for archiving and for the backup of the vast amount of data.

Properties

Storage subsystems have to fulfill a number of requirements that are unique to this kind of hardware. The main difference between typical electronic IT equipment and storage components is the fact that in storage technology, there are many mechanical components involved that are stressed heavily during the lifetime of the product. These mechanical parts have to be engineered very carefully, and there is a need for high quality in manufacturing. The following objectives have to be met:

- ▶ Reliability and robustness of disk and tape drive mechanics
- ▶ Reliability and durability of the storage media
- ▶ Reliability and robustness of media changer robotics
- ▶ Affordability of the components

Apart from mechanical requirements, there are also requirements that are related to software, physical, and logical design:

- ▶ Scalability of a subsystem
 - Number of physical connections
 - Number of media in a tape library
 - Number and size of disks in a disk subsystem
- ▶ Maintainability of a subsystem
 - Concurrent software and hardware maintenance
- ▶ Compatibility with different platforms
 - Number of supported platforms

- Types of connections
- ▶ Interoperability of tape drives and tape media
- ▶ Security regarding physical and logical access
- ▶ Redundancy of components
- ▶ Manageability of the subsystem
- ▶ Performance of the subsystem
 - Throughput
 - Tape mounts
 - Average access times

Bright idea!

All the criteria mentioned above have to be taken into account when selecting a tape storage subsystem for an SAP R/3 system. The sizing of these components has to be done according to requirements that have been stated in the service level agreement.

Disk storage products

In an SAP R/3 environment, the disk technology is an important factor for performance and reliability of the whole system. There are many different technologies on the market that we can choose from. They usually do not differ in the underlying technology, which is based on standard hard disk drives. The main difference comes from the attachment technology and the communication protocol that is used for the transmission of data. Details on the different technologies and protocols are given in Chapter 4, “Disk storage” on page 95. In this section, we give an overview of the available products and discuss their deployment in SAP R/3 environments.

Small Computer System Interconnect (SCSI)

IBM offers a SCSI-based storage system, the 2104 Expandable Storage Plus. It is a flexible, scalable, and low-cost disk storage for RS/6000 and pSeries servers in a compact package, available both as a standalone and a rack-mounted version. The rack-mounted model DU3 drawer can reside in a variety of 19-inch racks. It can be populated with up to fourteen IBM Ultrastar disk drives, available as 9.1 GB, 18.2 GB, and 36.4 GB models. Drive capacities can be intermixed, and drives can be added in any increment, to a maximum capacity of 509 GB.

The Expandable Storage Plus can attach to RS/6000 servers by using SCSI-2 Fast and Wide, Ultra SCSI, Ultra2 SCSI, and Ultra 3 SCSI. For the highest performance and availability, the enclosures can be combined with the IBM PCI 4-channel Ultra3 SCSI RAID Adapter (FC 2498), enabling 160 MB/s throughput performance and multiple RAID options. Distances up to 20 meters are supported between disk enclosures and pSeries or RS/6000 servers using Ultra3 SCSI adapters.

More information on this product can be found at the IBM Storage site:

<http://www.storage.ibm.com/hardsoft/products/expplus/expplus-spec.htm>

Serial Storage Architecture (SSA)

IBM has implemented a powerful industry-standard serial storage technology with the Serial Disk System 7133. Unlike SCSI bus configurations, SSA devices, such as the 7133 Serial Disk System, are configured in loops and allow multiple concurrent operations to occur in separate sections of these loops. It provides outstanding performance, availability, and attachability:

- ▶ Performance with advanced SSA bandwidth of 160 MB/s
- ▶ Multiple system attachment and subsystem partitioning for distributed systems
- ▶ Provides high availability with redundant data paths, redundant cooling units, and two power supplies
- ▶ Facilitates remote mirroring
- ▶ Up to 10 km connection distances with the Advanced SSA Optical Extender

The rack-mountable 7133 Advanced Model D40 drawer is designed for integration into a supported 19-inch rack. The 7133 Advanced Model T40 is a free-standing desktide tower unit. Both models feature high-performance 36.4 GB, 18.2 GB, and 9.1 GB disk drives with 10000 rpm, which provide a capacity of up to 582 GB per tower or drawer and 3.5 TB per host adapter

IBM SSA technology features significant availability advantages over SCSI-based technology. If a cable failure occurs on the loop, the SSA adapter automatically continues accessing disks through an alternate path. If a disk failure occurs, the hot-swappable drives can be removed and replaced without disrupting the communication between the adapter and other disks on the loop. To further increase availability, the 7133 Advanced Models monitor and provide detailed information on the status of power, cooling, and disk drives.

More information on this product can be found at the IBM Storage site:

<http://www.storage.ibm.com/hardsoft/products/7133/7133-spec.htm>

Network Attached Storage (NAS)

Network Attached Storage devices are high-performance storage appliances that provide shared data to clients and other servers on a Local Area Network (LAN). IBM offers two products with this technology.

The IBM Network Attached Storage 200 series products is well suited for small SAP R/3 environments in a rack configuration. Combined with their ease of installation and easy to use features, the low cost per megabyte makes them an excellent fit for environments where large amounts of inexpensive storage are required. The system provides the following features:

- ▶ High reliability via redundant, hot swap power supplies and hard disk drives, so business operations can continue in the event of a sub-system failure.
- ▶ Snapshot capability with 250 persistent True Image data views for quick data protection, enables backups without slowing the system, and allows files to be restored quickly and accurately.
- ▶ Fully integrated, pre-loaded software suite allows a minimum amount of dedicated IT resource for setup due to simplified installation and integration into the IP network via an easy-to-use Web browser.
- ▶ On-board Advanced System Management processor with Light Path Diagnostics, Predictive Failure Analysis and Remote Connect capabilities helps to ensure system uptime.
- ▶ Scalability from 216 GB to 1.74 TB.

The IBM Network Attached Storage 300 is designed to meet storage requirements for more demanding environments. It offers the same features and benefits as the 200 series products, but provides a second high-performance server with these additional advantages:

- ▶ Increased levels of system reliability via a second engine, for mission critical processes
- ▶ Enhanced performance, so additional users in a department or small enterprise can maintain high productivity levels
- ▶ Highly scalable storage capacity, ranging from 360 GB to 3.24 TB, providing flexibility and control over the initial purchase investment

More information on this product can be found at the IBM Storage site:

<http://www.storage.ibm.com/snetwork/nas/>

Fibre Channel (FC)

The Fibre Channel protocol and the technology has been developed to satisfy the storage requirements of data centers and high performance applications. It offers both a high bandwidth for data transfers and the capability to bridge long distances. The storage subsystem that fits perfectly into this scenario is the IBM Enterprise Storage Server (ESS).

The IBM ESS is a high-performance disk storage solution ideally suited for storage consolidation purposes and for the attachment of performance critical systems across the enterprise. It offers the following features:

- ▶ Provides superior storage sharing for UNIX, Windows NT, Windows 2000, Novell NetWare, and all IBM @server series
- ▶ Provides high performance with two powerful four-way RISC SMP processors, large cache, and serial disk attachment
- ▶ Features industry-standard, state-of-the-art copy services, including FlashCopy, Peer-to-Peer Remote Copy, and Extended Remote Copy, for rapid backup and disaster recovery
- ▶ Uses redundant hardware and RAID 5 disk arrays to provide high availability for mission-critical business applications
- ▶ Enables enterprises with multiple heterogeneous hosts to scale up to 13.9 TB while maintaining excellent performance
- ▶ Provides fast data transfer rates with attached hosts through Fibre Channel, UltraSCSI, ESCON, and FICON interfaces
- ▶ Increases administrative productivity by centralizing operations management and providing users with a single interface via a Web browser

Support for 24x7 operations is built into the IBM Enterprise Storage Server. RAID 5 disk arrays help provide data protection while remote copy technologies allow fast data backup and disaster recovery. The IBM Enterprise Storage Server is a high-performance RAID 5 storage server featuring dual active processing clusters with fail-over switching, hot spares, hot-swappable disk drives, nonvolatile fast write cache, and redundant power and cooling.

The IBM Enterprise Storage Server also contains integrated functions to help prevent storage server downtime by constantly monitoring system functions. If a potential problem is detected, the IBM Enterprise Storage Server automatically "calls home" to report the problem. A technician can be dispatched to make repairs, often before the problem is noticed by data center staff. Maintenance, including licensed internal code revisions, can typically be performed without interrupting operations.

The IBM StorWatch Enterprise Storage Server (ESS) Specialist helps storage administrators control and manage storage assets for the IBM Enterprise Storage Server. With a browser interface, they can access the ESS Specialist from work, home or on the road through a secure network connection.

The IBM StorWatch Enterprise Storage Server (ESS) Expert helps storage administrators monitor the performance of all connected IBM Enterprise Storage Servers in the enterprise. This innovative software tool provides performance statistics, flexible asset management, and tracks a variety of capacity information through a common available browser interface. As such, this tool enables the administrators to centrally manage all Enterprise Storage Servers located anywhere in the enterprise. This is a fee licensed feature.

More information on this product can be found at the IBM Storage site:

<http://www.storage.ibm.com/hardsoft/products/ess/ess.htm>

Discussion of the disk technologies and products

Bright idea!

The disk storage subsystems and the connection technologies described in the previous sections all provide reliable storage space for an SAP R/3 environment. Nevertheless, it is important to point out a few features that are different between these solutions.

The SCSI solution is working for small SAP R/3 systems with local storage requirements. Due to the distance limitations of SCSI, it is not possible to build disaster-tolerant solutions. The restrictions on cable lengths, electrical termination of the SCSI bus, and device addresses on a SCSI bus are reasons why fault-tolerant configurations with shared SCSI devices are not easy to set up and manage.

The serial storage architecture offers far easier management of devices due to the fact that it is not a bus system but an architecture based on a loop topology. In addition, the option of using Advanced SSA Optical Extenders enables long distance connection for implementing disaster-tolerant solutions, which is extremely important for business critical productive SAP R/3 systems.

Inherent in the SSA architecture is the possibility to have up to eight host systems and 48 devices in a loop, which can all be configured dynamically into and out of the loop. This facilitates the dynamic reconfiguration and the expansion of disk storage, which is a common task in an SAP R/3 environment. It also makes a fault-tolerant configuration far easier where two or more hosts need access to the same disk devices.

The NAS technology is ideally suited for small to medium SAP R/3 systems because it enables the quick availability of database file space via the Network File System (NFS) protocol from the NAS server. Depending on the network technology, it is easy to scale the bandwidth from 100 Mb/s Ethernet to Gigabit Ethernet, for example. Unfortunately, in case of even higher performance requirements it is necessary to split up the networks to reach a higher throughput. This is where the disadvantages of NAS show, because the management advantages disappear in high performance scenarios. Adding fault tolerance criteria to the requirements produces a very complex network structure that is difficult to set up and manage.

Fibre Channel technology combines the strength of the previously mentioned technologies. It supports a virtually unlimited number of devices, the distance limitations are in the range of kilometers and the throughput of the adapters and the fiber medium is very high and will even be increased in the next level of the Fibre Channel specifications. In case of an increased bandwidth need, and for a fault-tolerant configuration, there are software enhancements that provide multi-path I/O for Fibre Channel connections. All these features make the Fibre Channel technology ideally suited to build infrastructures for large high performance SAP R/3 production systems.

There is more information on disk technology and implementation in Chapter 4, “Disk storage” on page 95.

Tape storage products

Tape backups cannot be replaced by features like snapshot technologies and RAID protection for disk storage subsystems that have become standard features. Administrators of data centers still rely on the security and safety added by moving the business critical data to long term storage. It is the only reliable solution for the recovery of some disaster scenarios and logical errors in data sets.

Bright idea!

In an SAP R/3 environment, there are many different objects that have to be stored on tape. All the backup tasks have to be performed regularly and should be automated as much as possible to reduce the risk of human error. Also, the restore of the data should run without manual intervention after being triggered by an administrator. These requirements ask for a tape storage management system, consisting of software and hardware components.

The hardware solution must contain a tape library that is able to automatically select and move data cartridges from their storage cells to the tape drives. The size of the library should be sufficient to store the data capacity of all the SAP R/3 databases, the operating systems, interface data, and the application software of the SAP R/3 system. This capacity depends on the service level agreement, where it is specified how many versions of the data have to be stored over which period of time.

The number and the speed of the drives essentially determine the time window that is needed to make backups and, what is even more important, the time for restores. They also have to be chosen according to the requirements that have been specified in the service level agreement. In the following sections, we describe two technologies that are suitable for SAP R/3 environments in terms of reliability, performance, and scalability.

IBM Ultrium LTO technology

The Linear Tape Open (LTO) technology brings unprecedented levels of reliability, capacity, and performance to scalable automation for open systems tape backup. Developed jointly by IBM, Hewlett-Packard, and Seagate, LTO's open technology is ideal for a wide-range of open systems streaming data environments, such as backup or archive.

The IBM 3580 Ultrium Tape Drive is the building block of the new family of scalable, flexible tape solutions. It offers a capacity of 100 GB of uncompressed data per cartridge and has a sustained data transfer rate of up to 30 MB/s using the hardware compression of the drives. The IBM 3580 features Ultra2/Wide SCSI Low Voltage Differential (LVD) or Ultra/Wide SCSI High Voltage Differential (HVD) interfaces and a Fibre Channel Arbitrated Loop attachment.

The IBM 3584 UltraScalable Tape Library starts with an entry base frame and can be scaled by adding up to five expansion frames. You can choose a media/drive mix based on particular application needs (more drives and fewer media slots or more media slots and fewer drives):

- ▶ Each frame can contain up to 12 drives, with a maximum of 72 drives per library.
- ▶ Available drive types IBM LTO Ultrium or DLT 8000.
- ▶ Expandable to handle from 87 up to 2,481 tape cartridges.
- ▶ Storage capacity of up to 496 TB (with 2:1 compression).
- ▶ Supports IBM @server xSeries, pSeries, iSeries, HP, Sun, and Windows NT platforms, even simultaneously.

More information on this product can be found at the IBM Storage site:

<http://www.storage.ibm.com/hardsoft/tape/3584/index.html>

IBM Magstar technology

The Magstar 3590 Model E Tape Subsystem is a powerful integrated storage solution that provides the highest levels of capacity, performance, and data reliability. The drives have the following properties:

- ▶ Native drive data rate up to 14 MB/s
- ▶ Sustained data rates of 42 MB/s with Fibre Channel and 34 MB/s with Ultra SCSI attachment (with 3:1 compression)
- ▶ 40 GB native cartridge capacity with the Extended High Performance Cartridge Tape (120 GB with 3:1 compression)
- ▶ Choice of Dual Ultra SCSI or Dual Fibre Channel ports for enhanced sharing and availability options

The Magstar 3494 Tape Library, a tape automation system, consists of individual frame units for modular expansion that provides a wide range of configurations. This flexibility enables organizations to start small and grow in an affordable and incremental manner.

The basic building block of the Magstar 3494 is the Lxx Control Unit Frame, which contains a library manager, a cartridge accessor, up to two tape drives, and slots for the storage of tape cartridges. To the Lxx frame, you can add drive frames and storage unit frames, and a high-availability model to create a maximum configuration of 16 frames. A choice of two optional convenience I/O stations provides the capability to add or remove up to 30 cartridges at a time without stopping the operation of the cartridge accessor. The following features are provided by the Magstar 3494:

- ▶ Connection of up to 32 SCSI or Fibre Channel drives
- ▶ Mounting and dismounting up to 610 cartridge per hour with the Dual Active Accessor option
- ▶ Design for high reliability and the elimination of key single points of failure
- ▶ Expandable to handle from 160 up to 6,240 tape cartridges
- ▶ Storage capacity of up to 748 TB (with 3:1 compression)
- ▶ Supports IBM @server zSeries, pSeries, iSeries, HP, Sun, and Windows NT platforms, even simultaneously

More information on this product can be found at the IBM Storage site:

<http://www.storage.ibm.com/hardsoft/tape/3494/index.html>

Discussion of the tape technologies and products

Both tape technologies are very well suited for the storage of business critical data, such as SAP R/3 databases. The experience and reliability of the Magstar drive and tape technology has entered into the design of the Linear Tape Open technology. As such, LTO achieves similar performance numbers, although offering it for a lower price.

Pitfall ahead!

The main differentiator between both technologies is the availability of two Fibre Channel or SCSI ports on a single Magstar drive. In combination with the recoverable path feature of the underlying device driver, this enables a high available attachment of the tape drives as needed for reliable data center operations.

3.1.4 Local Area Network (LAN)

SAP R/3 is based on a client/server architecture. This design requires a reliable and performance optimized network connection between the servers and the clients. Designing a network consists of certain aspects, such as bandwidth, latency, availability, topology, redundancy, and security. All of these topics have an influence on the operation of an SAP R/3 environment. We do not go into the details of global network design, which usually must take the whole infrastructure of a company into account. We only concentrate on the requirements imposed by an SAP R/3 environment. In this section, we present possible building blocks that can be used for the construction of a network. More details on protocols and SAP R/3 specific requirements can be found in Chapter 6, “Network” on page 149.

Host adapters

The following adapters are available for connecting RS/6000 and IBM @server pSeries to local area networks. We have only selected Ethernet and ATM adapters, because they are the only types that offer a sufficient bandwidth for the communication of SAP R/3 systems. Information on the features of the SP Switch adapters can be found in the redbook *RS/6000 SP and Clustered @server pSeries Systems Handbook*, SG24-5596.

10/100 Mbps Ethernet PCI Adapter

This Fast Ethernet adapter, Feature Code (FC) #2968, offers PCI-based RS/6000 and pSeries users an easy migration path from 10 Mb/s to 100 Mb/s LANs without requiring a change of adapter. It is compatible with IEEE 802.3 and 802.3u specifications. The adapter has one RJ-45 connection that supports connections to 100BaseTx and 10BaseT networks.

- ▶ Supports auto-negotiation of media speed and duplex operation
- ▶ Supports both full and half duplex operation over 10BaseT networks using the RJ-45 connector

- ▶ Supports 100 Mb/s full duplex using Category 5 Unshielded Twisted Pair (UTP) cable

4-Port 10/100 Base-TX Ethernet

This adapter, FC #4951, provides the same functionality as four standard Ethernet Adapters (FC #2968) while taking up only one PCI slot. The features of each port are:

- ▶ Supports four RJ-45 connectors for UTP or STP cabling
- ▶ Fits in full-sized PCI slots
- ▶ Supports 32/64-bit PCI data width
- ▶ Operates at PCI bus speed of 33 MHz
- ▶ Supports half or full duplex operation
- ▶ Supports installations via Network Installation Manager (NIM)
- ▶ Auto-negotiation for detecting speed and duplex capability across each port

Gigabit Ethernet - SX PCI Adapter

This adapter, FC #2969, is a 1000 Mb/s PCI Ethernet adapter that is compatible with IEEE 802.3z specifications. The adapter has one external fiber connection that attaches to 1000BaseSX networks using 50 and 62.5 μ m multimode cables with SC connectors. This adapter will perform best in a 64-bit 50 MHz or 66 MHz slot, but will also function in a 32-bit 33 MHz slot. It offers the following features:

- ▶ Full duplex operation
- ▶ Supports jumbo frames
- ▶ Supports installations via Network Installation Manager (NIM)

IBM 10/100/1000 Base-T Ethernet PCI adapter

This adapter, FC #2975, provides one Gigabit Ethernet connection via Category 5 Unshielded Twisted Pair (UTP) for selected RS/6000 and pSeries systems. It offers the following features:

- ▶ 10/100/1000 Mb/s support
- ▶ Full duplex operation at 1 Gigabit
- ▶ Full or half duplex at lower speeds

Turboways 622 Mbps PCI Multimode Fibre ATM Adapter

This adapter, FC #2946, provides access to an ATM switch at a speed of 622 Mb/s from RS/6000 and pSeries PCI systems. The following features are included:

- ▶ Up to 1024 virtual connections

- ▶ Permanent Virtual Circuits (PVC) and Switched Virtual Circuits (SVC) supporting the ATM Forum specifications for User Network Interface (UNI) 3.0 and 3.1 signalling
- ▶ Uses 62.5 µm multimode fibre
- ▶ Supports classical IP and ATRP over ATM according to Request For Comments (RFC) 1577
- ▶ Supports LAN emulation

Turboways 155 ATM PCI Multimode Fibre Adapter

This adapter, FC #2988, provides access to an ATM switch at a speed of 155 Mb/s from RS/6000 and pSeries PCI systems. It has the same characteristics as the adapter with feature code #2946, except for the data rate.

Turboways 155 ATM PCI Unshielded Twisted Pair Adapter

This adapter, FC #2963, provides access to an ATM switch at a speed of 155 Mb/s from RS/6000 and pSeries PCI systems. It has the same characteristics as the previous adapter with feature code #2988, except for the attachment. It attaches to Category 5 Unshielded Twisted Pair (UTP CAT-5) wiring.

More information on RS/6000 and communications adapters can be found in the redbook *RS/6000 Systems Handbook 2000 Edition*, SG24-5120 and in *RS/6000 Adapters, Devices, and Cable Information for Multiple Bus Systems*, SA38-0516. The detailed adapter installation and user guides can be found on the Web:

http://www.ibm.com/servers/eserver/pseries/library/hardware_docs/options.html

LAN switching products

IBM and Cisco agreed on a strategic alliance to combine the proven attributes of IBM's server platforms, application enabling software, and commitment to open standards with Cisco's Internet Operating System (IOS) software, routers, and switches. The companies are improving interoperability and providing the most available, scalable, manageable, and secure solutions in the industry.

Cisco offers a variety of switching products that scale to fit the needs of any network campus. Technology solutions supported include Ethernet, Fast Ethernet, Gigabit Ethernet, Asynchronous Transfer Mode (ATM), and Fibre Distributed Data Interface (FDDI). The following families of products are suitable for attaching servers in an SAP R/3 environment.

Catalyst 4000 Family

Offers superior performance, value, and optimized total cost of ownership for 10/100/1000 Mb/s Ethernet switching. This switch provides 24 Gb/s of switching bandwidth and provides expansion to 96 ports of 10/100 Mb/s Ethernet or 36 ports of Gigabit Ethernet.

Catalyst 5000 Family

Including the Catalyst 5000 and 5500 Series, this family of products offers the highest port densities (up to 256 ports) with flexible uplinks, ATM, FDDI, 10/100 and 1000 Mb/s, and CiscoAssure intelligent network features

Catalyst 6000 Family

The Catalyst 6000 Family, consisting of the Catalyst 6500 Series and the Catalyst 6000 Series, delivers high performance switching solution designed to address increased requirements for gigabit scalability, high-availability, and multilayer switching in backbone/distribution and server aggregation environments

The following Redbooks give additional information and help for implementing complex networks:

- ▶ *IP Network Design Guide*, SG24-2580
- ▶ *RS/6000 ATM Cookbook*, SG24-5525

3.1.5 Software components

The software components of the infrastructure are equally important as the hardware components. The most reliable hardware is not working satisfactory if the software or the operating system crashes the machines constantly or requires a reboot of the system due to minor configuration changes. In this section, we describe the software components that are selected as the basis of our infrastructure for SAP R/3 systems.

Operating system

The AIX operating system runs across the entire range of RS/6000 and IBM @server pSeries systems, from entry-level servers and workstations to powerful supercomputers like the RS/6000 SP. Not only does AIX scale across systems of different sizes, it scales across different technologies and hardware platforms by delivering binary compatibility.

The following features are key components of AIX:

- ▶ Dynamic kernel

The kernel of AIX is dynamic, which means that it is possible to add devices during the run time of the system and to load kernel extension without the necessity to reboot the system. There are also no static kernel parameters in the AIX kernel that require a recompile of the kernel. The parameters can be adjusted during run time.

► 64-bit

The AIX kernel is enabled for the processing of 64-bit programs while being able to concurrently run 32-bit and 64-bit executables. Even for 32-bit programs, the kernel allows the addressing of large real memory areas for file caching and memory mapped files.

► Logical Volume Manager

One of the masterpieces of the operating system is the Logical Volume Manager (LVM). It facilitates the management of disk storage in a very flexible way, enabling features, such as the import and export of a complete set of disks, the extension of file systems during run time, mirroring, and striping of data.

► System Management Interface Tool (SMIT)

IBM is a long-time leader in UNIX systems management. AIX offered the first UNIX point-and-click management tool which transforms a complex process into a series of simple tasks, the System Management Interface Tool (SMIT). It has been extended to a complete Web-based System Manager, which runs on any JAVA enabled platform.

► Software installation

The installation of software and fixes, as well as the version control for software components, is completely managed by the operating system. The software management tools take dependencies and prerequisites into account and automate many installation tasks. For remote installations the Network Installation Manager (NIM) is used, which is a product that is an inherent part of the operating system. Also, the complete backup of the operating system image and restore capabilities from tape are core features of AIX.

► Error logging

AIX has implemented an error logging and reporting facility that complements the syslog mechanisms of standard UNIX systems. All hardware and software subsystems of RS/6000 and IBM @server pSeries use this mechanism to enable preventive error notification and analysis.

► Workload Manager

AIX Workload Manager (WLM) is an operating system feature introduced in AIX Version 4.3.3. It is part of the operating system kernel at no additional charge. AIX WLM delivers the basic ability to give system administrators more control over how scheduler, Virtual Memory Manager (VMM), and device driver calls allocate CPU, physical memory, and I/O bandwidth to classes of processes, based on user, group, application path, process type, or application tag. WLM is ideally suited to balance the demands or requests of competing workloads when one or more resources are constrained.

Bright idea!

For a complete list of the operating system features, refer to the IBM AIX Web site:

<http://www.ibm.com/servers/aix/products/aixos/specs/index.html>

The following Redbooks contain more information on AIX and the features of the latest versions of the operating system:

- ▶ *AIX Version 4.3 Differences Guide*, SG24-2014
- ▶ *AIX 5L Differences Guide*, SG24-5765
- ▶ *AIX 5L Workload Manager (WLM)*, SG24-5977

Middleware

In many environments there are software components that are not appreciated as being important; very often, they are summarized as middleware. Nevertheless, we want to stress the importance of these components that are essential for a reliable operation of the systems.

Parallel System Support Program (PSSP)

The SP Cluster Software Parallel System Support Programs for AIX (PSSP) is a collection of administrative and operational software applications that run on each node of an SP system and on the CWS. Built upon the system management tools and commands of the AIX operating system, PSSP enables system administrators and operators to better manage SP systems and their environments.

Sets of software tools and related utilities, including application programming interfaces (APIs), have been grouped together to offer easier administration of installation, configuration, device management, security administration, error logging, system recovery, and resource accounting in the SP environment.

The System Data Repository (SDR) is a central repository that contains the specific SP configuration information and operational information. It only resides on the CWS. The PSSP software component for system administration and operation contains all the tools required for entering and changing configuration information, such as:

- ▶ Listing and modifying configuration properties of the nodes in the SDR and distributing the information to the nodes
- ▶ Parallel system management tools and commands for performing management functions concurrently across multiple SP nodes
- ▶ File collections for managing files and directories on multiple nodes and keeping them in sync
- ▶ Login control for blocking an unauthorized user or group access to a specific SP node or a set of nodes

- ▶ Consolidated accounting for centralizing records at the node level (for tracking use by wall clock time rather than processor time) and gathering statistics on parallel jobs

A consolidated system graphical user interface, RS/6000 SP Perspectives, provides a common launch pad for PSSP system management applications through direct manipulation of system objects represented as icons. This interface is tightly integrated with the problem management infrastructure. It allows users to easily create and monitor system events and provide notification when events occur. The interface is highly scalable for large systems, and can be easily customized to accommodate varying environments.

High Availability Cluster Multi Processing (HACMP)

Clustering is the linking of two or more computers or nodes into a single, unified resource. High availability clusters enable the parallel access of data, redundancy, and fault resilience required for business-critical applications such as SAP R/3.

HACMP for AIX is a control application that links RS/6000 and IBM @server pSeries servers into highly available clusters. There are two different choices for HACMP, the classic version and the HACMP Enhanced Scalability (ES) version. HACMP automatically detects system or network failures and eliminates single points of failure by managing failover to a recovery resource. HACMP/ES offers additional availability benefits through the use of the Reliable Scalable Cluster Technology (RSCT) function of AIX, which allows to monitor all kinds of system resources.

The following features are important properties of HACMP:

- ▶ Multiple availability configurations
 - Scalability from two-node to 32-node clusters
 - Flexibility through customization of the configuration according to business needs
 - Cluster snapshot utility can save and restore configuration and configuration changes
 - Facilitates cloning of additional clusters
 - Allows maintenance of multiple cluster configurations
 - Allows quick backout of configuration changes
- ▶ Integration into system management interface
 - Interfaces to install and configure highly available cluster systems as well as maintain them on the network
 - Monitoring utilities to manage and tune clusters

- Visualization of relationships between cluster hardware and software resources
- Automatic discovery of the connections between clusters
- ▶ Cluster Single Point of Control (C-SPOC)
 - Allows cluster management from a single system console for the cluster
 - Enhances systems management by helping to reduce system administration errors
 - Enables system tasks to be performed once for all cluster nodes
- ▶ Dynamic reconfiguration
 - Enables continuous system maintenance without disruption
 - Allows cluster resources to be added or removed without disruption

These features of HACMP are perfectly suited to integrate a set of software products into a high available overall solution. This is the main task when building an infrastructure for SAP R/3 environments.

Databases

SAP has decided from the beginning that SAP R/3 should be based on open system standards. One of the main components of an SAP R/3 system is the database, which holds the data and the code for the business logic. SAP has implemented interfaces to the major relational database management systems on the market, depending on the choice of the platform.

The standard language for manipulating data in RDBMS environments is the Structured Query Language (SQL). Although SQL is standardized to a high degree, there are still varying syntax and semantics for the different RDBMS products. SAP has implemented some special functions in the database interface to adopt to these different implementations.

This adoption has been carried out for database systems that were able to satisfy the high performance requirements of SAP R/3. The following databases are supported by SAP for UNIX platforms:

- ▶ DB2 Universal Database
- ▶ Informix
- ▶ Oracle
- ▶ SAP DB

In this redbook, we only cover the database products DB2 Universal Database and Oracle, because they are the most wide-spread products in the UNIX market segment for SAP R/3. These two databases run on IBM hardware as well as on non-IBM machines, such as Sun and Hewlett-Packard. In an SAP R/3 environment, they are supported for various operating systems, such as AIX, Windows, Linux, and Sun's Solaris.

Pitfall ahead!

DB2 and Oracle databases have a long history and are deployed in thousands of installations for SAP R/3 systems and other applications. They fulfill the basic requirements of SAP and are reliable and high performance RDBMS systems. Oracle database administration is integrated with the SAP R/3 database administration tool *SAP DBA*. The administration of DB2 in SAP R/3 environments is performed with a plug-in for the DB2 *Control Center*.

Storage management

In an SAP R/3 environment, the backup of the business critical data is an extremely important area. The protection of the data through archiving and regular backups is the only safeguard in case of user errors or disaster scenarios. The required hardware components have been described in the section "Tape storage products" on page 52. The same focus should be put on the software solution, which is shown in this section.

Tivoli Storage Manager (TSM)

The Tivoli Storage Manager utilizes high performance patented technologies to protect and manage mission-critical business information in an enterprise-wide environment. Tivoli Storage Manager offers optional products that provide an end-to-end scalable backup and restore solution which spans from desktop systems to mainframes on over 35 platforms. The main features include:

- ▶ Centralized storage management
- ▶ Central administration provides server management from any TSM client platform via robust administrator capabilities
- ▶ LAN-free backup and tape sharing
- ▶ Automated network incremental and sub-file backup, archive, and retrieval
- ▶ Broadest range of supported storage devices, including over 250 disk, removable cartridge drives, and tape systems
- ▶ Server-to-server communication enables support for objects to be sent to or received from another TSM server directly
- ▶ Space management file migration
- ▶ High-speed, policy-based disaster recovery
- ▶ Advanced tape management capability maximizes tape usage to minimize number of tapes needed

- ▶ Optional compression can reduce network traffic and transmission time
- ▶ Optional data protection for messaging software, databases, and applications

Tivoli Storage Manager is the product of choice in more than one million systems worldwide, including more than 80 of the Fortune 100 companies.

Tivoli Data Protection for R/3

Tivoli Storage Manager offers integration with other Tivoli products to enable a scalable, customizable solution that grows with your organizational needs. One of the important components is the Tivoli Data Protection (TDP) for R/3, which has formerly been known as *BACKINT/ADSM*.

TDP for R/3 is SAP certified and builds upon SAP DBA, a set of database administration functions integrated with R/3, for database control and administration. Tivoli Data Protection for R/3 seamlessly connects SAP DBA with the Tivoli Storage Manager to support large volume backup, recovery, data cloning, and disaster recovery of multiple SAP R/3 systems in heterogeneous environments. It offers the following features:

- ▶ Parallel backup and restore
Using native SAP utilities, TDP for R/3 backs up and restores database objects in parallel with the specified number of multiple sessions.
- ▶ Parallel backup paths
In order to reduce network-induced bottlenecks, TDP for R/3 can simultaneously use multiple communication paths for data transfer with its Tivoli Storage Manager server.
- ▶ Parallel backup servers
Multiple Tivoli Storage Manager servers can also be used for backup and restore simultaneously to eliminate bottlenecks in server capacity. Each server is defined to TDP for R/3 via a "server statement" in the profile.
- ▶ Multiple redo log copies
For availability and security reasons, TDP for R/3 is able to store multiple copies of the same redo log file on different physical volumes.
- ▶ Alternate backup paths
In order to improve the availability of the communication network between the R/3 database server and backup server, TDP for R/3 can use alternate communication paths for data transfer with its Tivoli Storage Manager server.
- ▶ Alternate backup servers
If one server is down, backups and restores can be transferred to and from an alternate backup server with Tivoli Data Protection for SAP R/3.

These features make TDP for R/3 an indispensable product that is highly recommended to use in an SAP R/3 environment.

3.2 Building a basic model

Now that we have introduced all necessary infrastructure components for an SAP R/3 system landscape, we can use these building blocks to construct real environments for different requirement classes.

Bright idea!

This section covers our suggested basic infrastructure model for SAP R/3 on IBM @server pSeries. This model does not claim to be a reference architecture, but meets basic requirements, such as reliability, scalability, and manageability, as defined in Chapter 2, “Requirements for a reliable SAP R/3 environment” on page 7.

The basic model is suitable for SAP R/3 systems where availability is not a critical demand. Usually this applies to development or quality assurance systems and to production systems whose demanded availability is below 98 percent.

Figure 3-2 on page 65 shows the components of the basic model that are described.

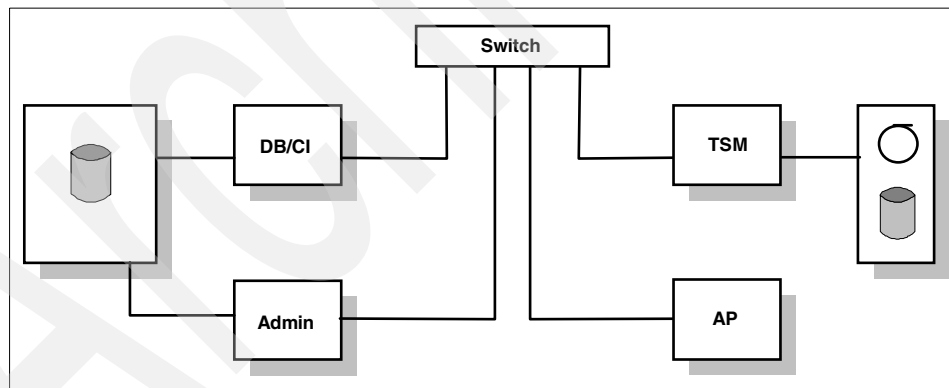


Figure 3-2 A basic infrastructure model

3.2.1 Server for the database and central instance (DB/CI)

Servers for SAP R/3 development and quality assurance systems are normally configured as central systems because their number of users usually do not require additional dialog instances and hence application servers. This is a two-tier implementation of an SAP R/3 system, because a database and an application server instance are combined on a single machine. Choose an IBM @server pSeries model providing a sufficient SAPS value for a central system according to your estimated workload.

The current minimal recommended configuration for SAP R/3 release 4.6x is a central system IBM @server pSeries Model 620 1-way, with 450 MHz with 1 GB memory and 20 GB internal disk space

For productive systems with an 32-bit SAP R/3 kernel, an internal disk space for paging of at least three times the amount of main memory is proposed. For better reliability, you should mirror the volume group containing AIX (rootvg) and the paging space to different internal disks.

Pitfall ahead!

For production systems that do not have special availability requirements, we recommend a configuration where the database server carries the database system and a *minimal central instance*. A minimal central instance provides the *message* and *enqueue* service for the entire SAP R/3 system, and dialog and spool service only for administration purposes. One or more application servers (refer to Section 3.2.2, “Application server (AP)” on page 67) cover the workload from interactive users (dialog and update work processes), batch processing (batch and update work processes), and printing (spool work processes). User access to a minimal central instance should be restricted to administrators. Normal users that logon to the system via a logon group are distributed only to the application server(s). Thus, the minimal central instance is not part of a production logon group. This approach has the following advantages:

- ▶ The critical services of an SAP R/3 system (database, message, and enqueue) are combined on a single machine. This adds to a higher reliability and simplifies the handling when starting and stopping the system, for example.
- ▶ For administration purposes, the system can be run without further application server instances. In this case, end users cannot log on to the system and accidentally disturb maintenance work.
- ▶ For performance reasons, enqueue and message service should reside on the same machine, because the enqueue service communicates *only* with the message server. There is no resource usage conflict between these CPU intensive service processes and the (mainly) I/O bound database.

Of course, a minimal central instance must have a few dialog work processes and a spool work process for administration purposes. As a rule of thumb you can use the work process distribution shown in Table 3-4 for a minimal central instance.

Table 3-4 Work process configuration for a minimal central instance

Dialog	Update	Enqueue	Batch	Message	Spool
2	2	1	0	1	1

64-bit considerations

AIX supports 64-bit since Version 4.3. The AIX kernel allows you to run 32-bit and 64-bit applications simultaneously. The theoretical maximum for an application virtual address space is 4 GB within a 32-bit architecture. In practice, a single 32-bit SAP R/3 instance on AIX cannot use more than 2.7 GB main memory with the standard memory management model. Single 32-bit SAP R/3 user contexts are even limited to a maximum of approximately 1.4 GB.

Pitfall ahead!

For IBM @server pSeries servers with 2 GB and more main memory, it is therefore recommended that you install the 64-bit SAP R/3 kernel. Furthermore, the memory management for SAP R/3 is simplified to a great extent with the 64-bit kernel (see Chapter 11, “Performance” on page 305).

You can find the latest compatibility matrix for SAP R/3 kernel release, database versions, and AIX at the SAP platform site:

<http://service.sap.com/platforms>

The 64-bit SAP R/3 kernel 4.6D supports DB2 as of Version 7.1 and Oracle as of Version 8.0.5.

SAP recommends a minimum paging space for productive 64-bit systems of 20 GB, or at least the amount of main memory, if it exceeds 20 GB.

3.2.2 Application server (AP)

The basic infrastructure model in Figure 3-2 on page 65 shows exactly one application server (AP), but the required numbers of application servers depend on different circumstances, which are described in the following sections. As already mentioned, central systems do not have an application server at all. From an administrative point of view, central systems are the easiest to manage SAP R/3 systems, because a distributed system consisting of several servers is always harder to manage.

Sizing

If the supposed workload of the system requires a 3-tier solution, you have to be aware of the ratio that specifies how many application servers are able to saturate a database server. The IBM SAP International Competence Center (ISICC) evaluates the performance of IBM @server pSeries servers in terms of the SAPS count. The capacities of a server for the role as a central system or a database server are derived from the standard SAP SD benchmark. As there are no certified application server benchmarks, the SAPS capacity of an IBM @server pSeries server in the role of an application server is calculated by the saturation ratio that was derived from feedback of real productive environments.

The average ratio for SAP R/3 release 4.6 was found to be 1:6. This means that six application servers are able to saturate an equally powerful database server in terms of CPU usage.

SAP R/3 instance buffers

Pitfall ahead!

Every SAP R/3 application instance buffers data (such as compiled ABAP programs and buffered tables) during operation to reduce the requests to the database. The quality, that is, the hit rate, of these caches is extremely important for the performance of an SAP R/3 system. Particularly interactive users suffer from a bad response time if the quality of the buffers is bad. The buffered data is module dependent.

If your business uses several modules, such as FI, CO, and MM, then users of different modules interfere at the instance buffers, producing an overall bad buffer hit rate and worse response times. The solution is to separate users of different modules via logon groups and distribute them to different application servers where they do not have to compete for buffer space. This could increase the number of required application servers. Another approach is to install two or more instances onto a single application server (as long as the machine is powerful enough).

A similar problem arises if an instance serves both interactive users and batch processing. If it is possible to separate interactive usage and batch usage to different times, you can adapt the instance to the different types of workload using operation modes.

With operation modes, it is possible to modify the work process distribution of a given instance for different time intervals. It is not possible to alter the total number of work processes of an instance. A common configuration consists of an operation mode during business hours for interactive usage with a large number of dialog work processes and only few batch work processes, and one operation mode during night times for batch processing with a diametrical distribution.

Pitfall ahead!

For systems where a separation of interactive and batch usage is not possible, it is recommended that you have dedicated application servers.

Special purpose instances

There are further reasons to have instances dedicated to a special purpose running on different application servers (if necessary). Some systems may require a dedicated spool application server for mass printing. There are constellations that require an additional instance if the system supports several codepages, as one instance can handle only certain language combinations.

Availability considerations

With more than one application server in place, availability is another factor that must be considered. While the database and the enqueue and message service are single points of failure in an SAP R/3 system, all other services (dialog, batch, update, gateway, and spool) can be distributed to several application server instances. The message server has a global view of the available services and their locations within an SAP R/3 system. The SAP R/3 logon load balancing mechanism, based on logon groups that include several instances, distributes interactive users to another available instance if an application server breaks down. Equally batch jobs are sent to an available batch work process by the message server.

Bright idea!

Avoid assignment of jobs to a dedicated batch server (if not highly available).

Of course, the work process configurations of the instances have to be configured accordingly, which could be contrary to the considerations described in “SAP R/3 instance buffers” on page 68. There is often a trade-off between performance and availability issues.

3.2.3 The storage subsystem (disk storage)

Disk space stores the data of an SAP R/3 system. In an IBM @server pSeries environment storage subsystems such as SSA drawers or ESS are usually used instead of internal hard disks for performance, availability, and scalability reasons. SSA, SCSI, or fibre channel (FC-AL) adapters connect the servers to the storage subsystem. Yet the trend is to store the electronic data of the whole enterprise in a single SAN.

The data of an SAP R/3 system can roughly be divided into the classes operating system files, database and application executables, database files (table spaces and redo logs), and instance and interface data. Operating system files and the paging space reside on the internal disks of the server. All database and

application specific data is stored within the storage subsystem. There are different requirements in terms of performance and availability for the mentioned data classes that has to be taken into account when designing a disk layout for the system.

For the SAP R/3 servers of our model, it is sufficient to connect only the DB/CI server to the storage subsystem, because application servers mount the required files via NFS from the DB/CI server.

Chapter 4, “Disk storage” on page 95 covers all important aspects of storage for SAP R/3 in detail.

3.2.4 The networks (switch)

As described in Section 2.1, “Characteristics of an SAP R/3 environment” on page 8, SAP R/3 has a three-tier client/server architecture. All three layers are connected to each other over communication networks.

Bright idea!

The model proposes a switched networking infrastructure. Although Figure 3-2 on page 65 shows only single lines connecting the servers with the switch, that does not mean that the servers are equipped only with a single network adapter. Thus, the connections in the figure are a simplified representation of a server connected to a switch over one or more network adapters.

We distinguish three different networks: front end, backup, and control network. They do not have to be necessarily different physical networks, but can also be *virtual networks* sharing the same network infrastructure.

Front-end network

The communication of SAP R/3 application servers with the database server is called *server communication*. The server communication is of great importance to the performance of an SAP R/3 system. Therefore, the network connecting the servers within the computer center (*front-end network*) should be built up onto a reliable network technology, providing high throughput of data with a minimal latency.

Pitfall ahead!

The front-end network should only include the servers of the SAP R/3 system and should not be used for any non-SAP data traffic (for example, backup data) for performance reasons.

The *access network* includes all network segments that SAP GUIs use to access the SAP R/3 system and consists of corporate LANs and WAN connections outside of the computer center. The switch connects the access network to the front-end network but separates the traffic between the servers from the access network. The access network is not part of this redbook’s scope.

Backup network

The *backup network* carries the traffic for backing up or restoring databases or file systems. The servers have separate network adapters for the backup network in order to avoid an impact to the server communication and a degradation of response times for interactive users. The backup network connects the servers to the backup subsystem and also needs to be a high bandwidth network, because it is the goal to have as short times for backup or recovery as possible.

Control network

We propose that you connect all servers of the SAP R/3 environment through a further *control network* for administrative purposes only. The control network does not need to be a high speed network, but should be a private physical network independent of other corporate networks (if possible). It should grant access to the servers for administrators only and should be secured from any unauthorized access. Remote and parallel shell command execution for the superuser root simplifies administration to a great extent. If this is limited to the control network, it is not even a security problem. Other infrastructure components that have network interfaces for administration, such as switches, disk subsystems (ESS), or tape libraries, should also be connected to the control network.

Chapter 6, “Network” on page 149 covers all important aspects of networks for SAP R/3 in detail.

3.2.5 The backup subsystem

This section describes the components for a complete backup and recovery solution for SAP R/3 and discusses aspects of locating the components for disaster safety of data.

Components and assembly

We recommend a solution for backup and recovery of the SAP R/3 system environment consisting of several hardware and software components.

The basic model in Figure 3-2 on page 65 includes a dedicated IBM @server pSeries server (TSM) running the server part of the Tivoli Storage Manager (TSM) (further referred to as the *TSM server*).

Pitfall ahead!

It is not recommended that you use one of the SAP R/3 servers (DB/CI or AP) simultaneously as a TSM server because of performance and availability reasons. Be sure to size the TSM server sufficiently in terms of I/O bandwidth according to the amount of created backup data and demanded backup and restore times.

The TSM server is connected to the high bandwidth backup network and should have a connection to the control network. Attached to the TSM server is a tape library. The tape drives of the library are connected to the TSM server via SCSI adapters or, for higher demands, in terms of throughput or distance, via Fibre Channel adapters. The tape library can also be part of a SAN and therefore be shared with other environments, such as a host system, for example.

Pitfall ahead!

In order to implement a reliable and automated backup and recovery solution, a tape library, including a tape robot for media retrieval, is highly recommended.

The TSM server has also a connection to a storage subsystem because disk storage pools improve the backup and restore times for small files (to a large extent). The TSM database should be stored on internal disks or on a storage device that is independent of the storage subsystem that holds the application data.

The TSM client software is installed on all servers that back up files on the TSM server. All operating system files, application, and database executables, and all interface data files are incrementally backed up in regular intervals by the TSM client. The Tivoli Data Protection for R/3 software enhances the TSM client software by providing an API for the standard SAP backup and recovery commands (**brarchive**, **brbackup**, and **brrestore**). It is used for high performance backups of the SAP R/3 database tablespace files and for archiving database redo logs into the TSM server.

Disaster safety for data

Even if your business does not require that you implement a fault-tolerant or even disaster-tolerant infrastructure for your whole environment, it is intolerable to lose the data of the SAP R/3 system in a disaster. So even for the basic model, we present two approaches that allow you to restore your data after a disaster.

Vaulting

Vaulting means storing your backup media in a safe place outside the computing center. After the creation of a backup of the production systems *and* of the TSM database, you have to check out the relevant tapes from your library and store them in another location, for example, in a safe in another fire compartment. This solution is cheaper but imposes a higher administration effort and exposes your data to human errors.

Remote library

The second solution is to install the tape library and, optionally, the TSM server in a remote location. If you lose your computing center, the library, including the backup media that contains system and database backups, is still available.

Pitfall ahead!

Depending on the location of the TSM server, you have to consider the following:

- ▶ TSM server that remains in computer center
 - Do not forget to back up the TSM database to specially labelled tapes into the remote library. Otherwise, you will not be able to restore data from the tapes if you lose the TSM database.
 - Plan to attach the tape drives to the TSM server with fiber optics. SCSI attachments are usually limited to 20 m. Technical solutions to overcome this constraint are expensive and limited in performance.
- ▶ TSM server that is installed together with the library in a remote location
 - Be sure that the link between the backup network switch inside the computing center and the remote TSM server provides the same bandwidth as your backup network technology thus not becoming a bottleneck.
 - You have to have a disk subsystem that is only used by the TSM server, or a long distance attachment to the storage subsystem, or SAN of the computing center. In the latter case, the loss of the disk storage pools must not limit the restore capability. The TSM database should reside in the same location as the TSM server.

Chapter 7, “Backup and recovery” on page 185 covers all important aspects of implementing a reliable backup and recovery solution for SAP R/3.

3.2.6 Administration server (Admin)

Every SAP R/3 environment should include a further server for administration purposes (*Admin*). A suitable machine could be an entry server or a workstation. It should be equipped with a graphics adapter and a monitor to provide a graphical system administration platform. The Admin server should be connected to the control and backup network. It may be attached to the storage subsystem or has to have sufficient internal disk space for its server purposes.

Bright idea!

Besides being an administration platform, the Admin server should provide the following services.

Software distribution

AIX contains the product NIM for software distribution. You should use NIM to guarantee a consistent software level throughout the servers of your environment.

The Admin server should be configured as a NIM server and should be the master for the following NIM resources:

- ▶ Licensed Program Product (LPP) source

The LPP source is a file system of the NIM server that contains all base level filesets of the installed LPPs in the environment. Additional filesets can be installed to a subset or to all servers of the environment from the NIM server via the control network.

- ▶ Program Temporary Fix (PTF) source

The PTF source is a file system of the NIM server that contains available PTFs or AIX maintenance levels for the base level filesets of the LPP source. PTFs can be installed to a subset or to all servers of the environment from the NIM server over the control network.

- ▶ Shared Product Object Tree (SPOT)

It is possible to boot an IBM @server pSeries server from the control network using NIM. Using the bootp protocol and tftp, the client retrieves its AIX boot image from the NIM server. During the network boot, the NIM client can fetch further required device drivers and LPPs normally stored locally under /usr/lpp from the SPOT resource of the NIM server.

- ▶ System image (mksysb) source

The backup and recovery concept is described in detail in Chapter 7, “Backup and recovery” on page 185. Backup demands that system images (mksysb) of all servers of the environment are stored regularly on the NIM server. Thus, the NIM server allows a network boot and a restore of the whole base operating system, including the contents of the root volume group of any server of the environment over the network.

Documentation Web server

The Admin server should be configured as a Web server for the following documentation:

- ▶ System configuration

An invaluable help for troubleshooting and reconstruction of a server is to have an up-to-date system documentation of all servers of your environment in place. There are tools to regularly collect all configuration details of the base operating system of a server and store the information in HTML format on the Admin server. Then it is possible to look at the system configuration of any server with a Web browser, even if the relevant server is down. System documentation should include the detailed LVM configuration (volume groups, logical volumes, and file systems), paging spaces, detailed configuration of every network adapter and related TCP/IP network device, printing subsystem configuration, the contents of all important configuration

files included in the directory /etc (environment, passwd, groups, hosts, and so on), and the contents of the AIX error report. It is recommended that you keep at least two generations of system documentation created in a reasonable time interval.

- ▶ TSM server configuration

As with the system configuration, it is useful to have the documentation server provide the TSM server configuration in HTML format.

- ▶ AIX documentation

The AIX documentation (commands, manuals, messages, and guides), including links to online AIX documentation resources, should be provided by the documentation server.

- ▶ Database documentation

The documentation of the SAP R/3 database system (installation, SQL, and administration commands reference) should be provided by the documentation server.

- ▶ TSM documentation

The TSM documentation (commands, manuals, messages, and guides), including links to online TSM documentation resources, should be provided by the documentation server.

- ▶ SAP R/3 Library

The Admin server can also be a Web server for the SAP R/3 Online Documentation (SAP Library). If also end users use this server to access the SAP Library, the Admin server also has to be connected to the front-end network. This might be a security concern.

Master for distribution of global configuration files

As described in Section 2.4, “Guidelines for the implementation” on page 17, configuration files that should be kept consistent within the whole server environment or within an administration domain (/etc/hosts, /etc/services, /etc/passwd, /etc/group, and /etc/environment, for example) should be edited only on the Admin server. After editing, these files are distributed to the relevant hosts.

Refer to Chapter 10, “Hints and tips” on page 275 for more information concerning this topic.

3.2.7 Requirements met by the basic model

In Chapter 2, “Requirements for a reliable SAP R/3 environment” on page 7, we introduced requirements for an SAP R/3 system infrastructure. In this section, we discuss which of these are met by the proposed basic model.

Reliability	<p>The model is assembled out of reliable components. IBM @server pSeries servers and their operating system AIX provide leading-edge RAS features. All data of the systems is stored redundantly (RAID 1 / RAID 5) within the storage subsystem. An effective backup/recovery solution that can easily be enhanced to a disaster-tolerant solution is in place.</p>
Performance	<p>The IBM @server pSeries servers are available for even highest performance requirements. High performance storage subsystems based on SSA or ESS technologies and efficient distribution of data guarantees highest I/O throughput. Separate high speed networks for application and backup traffic meeting the required bandwidth and latency are in place.</p>
Scalability	<p>The basic model is highly scalable both in its components and in its whole architecture. IBM @server pSeries servers provide CPU capacity upgrades on demand. The main memory can be extended to as much as 96 GB for the IBM @server pSeries model p680. 64-bit AIX allows applications to access large amounts of virtual address space. The possibility to add I/O drawers lets you upgrade some servers with additional high speed adapters for higher throughput to storage and backup subsystems. The ESS scales to 11 TB disk storage (maximum). Additional tape drives let the backup and recovery solution scale because TSM and Tivoli Data Protection for SAP R/3 support parallel back up and restore from several tape drives.</p> <p>The SAP R/3 architecture implicitly provides scalability, as the performance of the system can be enhanced through additional application servers.</p>
Manageability	<p>IBM @server pSeries servers, in combination with AIX, provide outstanding system management facilities, such as service processors, SMIT management tool and NIM for software distribution.</p>

The dedicated Admin server and workstation of the basic

	model provides an effective management platform for even a large environment consisting of many servers.
Availability	The basic model provides a reliable, but not fault-tolerant, infrastructure, as it contains several single points of failure. Therefore, without enhancements, it is not suitable for mission critical applications that have to be available 7x24 hours.

In the following sections, we show how the basic model can be enhanced with further components to achieve fault and disaster tolerance for an SAP R/3 system environment.

3.3 Building a fault-tolerant model

The proposed basic model in Figure 3-2 on page 65 is not a suitable productive SAP R/3 system where availability is a critical requirement. For such systems, the basic model has to be enhanced in order to eliminate all single points of failure.

In the following sections, we identify the single points of failure and introduce additional components to gain redundancy for a fault-tolerant model.

3.3.1 Identifying single points of failure in the basic model

We distinguish between single points of failure of the application design on one hand and of infrastructure components on the other hand.

Single points of failure of the application design

The critical services inside the SAP R/3 application architecture are the database system and the message and enqueue service. A failure of any of these services stops the function of the whole SAP R/3 system.

Single points of failure of the infrastructure

A failure of one of the following components of the basic model leads to a service interruption of the SAP R/3 system.

Global power supply of the computer center

A failure in the global power supply of the computer center leads to an instant shutdown of all servers. If there is a battery buffered UPS in place, the capacity usually allows only a coordinated shutdown of the servers, but cannot ensure the provision of the service for a longer time. Diesel driven emergency power supplies may be used to protect from long term power outages.

Server DB/CI

The server DB/CI runs the database and the central instance of the SAP R/3 system. The critical message and enqueue services of SAP R/3 are part of the central instance. The following situations lead to a service interruption:

- ▶ Crash of the central instance
- ▶ Crash of the database system
- ▶ Crash of the operating system AIX
- ▶ Hardware failure: CPU, memory, and system planar
Failures of these components normally crash the machine, producing a system dump.
- ▶ Adapter or cabling failure: connection to storage subsystem
If there is a failure of an adapter or a cable that connects the DB/CI server to the storage subsystem, access to the instance or database file systems is interrupted. The system may not crash, but will not be able to read or write any data.
- ▶ Adapter or cabling failure: network connection to the front-end network
If the adapter or a cable that connects the server DB/CI to the front-end network fails, both communication between DB/CI and the application server AP and between DB/CI and the SAP GUIs is interrupted.

Server AP

If there is only one SAP R/3 application server, and the central instance is configured as a minimal central instance (refer to Section 3.2.1, “Server for the database and central instance (DB/CI)” on page 66), the application server exclusively provides the dialog, batch, and spool service of the SAP R/3 system. In this case, the following situations lead to a service interruption:

- ▶ Crash of the application instance
- ▶ Crash of the operating system AIX
- ▶ Hardware failure: CPU, memory, and system planar
Failures of these components normally crash the machine producing a system dump.
- ▶ Adapter or cable failure: network connection to the front-end network
If the adapter or cable fails that connects the application server AP to the front-end network fails, both communication between DB/CI and AP and between AP and the SAP GUIs is interrupted. In addition, the access to the global file systems that are mounted via NFS from DB/CI is lost.

Storage subsystem

In case of a breakdown of the whole storage subsystem, the access to the database and instance file systems is lost.

Switch

A failure of the switch interrupts any communication in the front-end network. In this case, both the connection between the application servers and the database or central instance and the connection between front-end and access network is lost.

Failure of the backup and recovery solution

A failure of a component of the backup solution (TSM server or tape drive, for example) can lead to one of the following severe situations:

- ▶ **Archiver stuck**

If you are not able to archive the redo log files of the database to tape, sooner or later the file system will run out of free space. The database stops at once, interrupting the SAP R/3 service. In this case, you have to move the redo log files to another location until the failing component is repaired.

- ▶ **Loss of the ability to restore backup data**

It is not a huge problem if it is a temporary problem (because of a hardware failure of the TSM server or a tape drive, for example) that can be easily repaired. A defective backup media is more serious, because the errors might not be detected. The worst case is a fundamental implementation failure that is not recognized.

Important: Test the successful restore and recovery of your database and the restore of file systems before going live with the backup and recovery solution and after any major change of a component. Otherwise, you risk data loss.

Failure of other components

Failures of other components may complicate the operation or administration of the SAP R/3 system but do not cause an interruption of the service itself. This can easily be seen on the Admin server.

3.3.2 Cluster solution for SAP R/3

The only possibility to achieve fault-tolerance for the database and for the central instance containing the critical services message and enqueue of an SAP R/3 system is the implementation of a cluster, as these components cannot be distributed redundantly to multiple servers. Within a cluster, the server DB/CI is protected through a backup server that can take over in case the server DB/CI breaks down.

The cluster software of choice within an IBM @server pSeries environment is HACMP. HACMP monitors the correct function of cluster nodes through permanent exchange of heartbeat packets between the nodes on multiple point to point connections over several network adapters. Thus, HACMP is able to distinguish between single network adapter failures, the breakdown of a whole network, and the crash of a cluster node. According to the event, automatic recovery actions are executed by event scripts, such as the swap of a service address of an adapter with a standby address or the complete takeover of a resource group to a backup node. A resource group can include both hardware and software resources, such as volume groups, logical volumes, file systems, and service addresses of network adapters and applications.

Figure 3-3 on page 81 shows our proposal for a fault-tolerant infrastructure model. The cluster is part of the model and consists of the servers DB/CI HA and APserv HA.

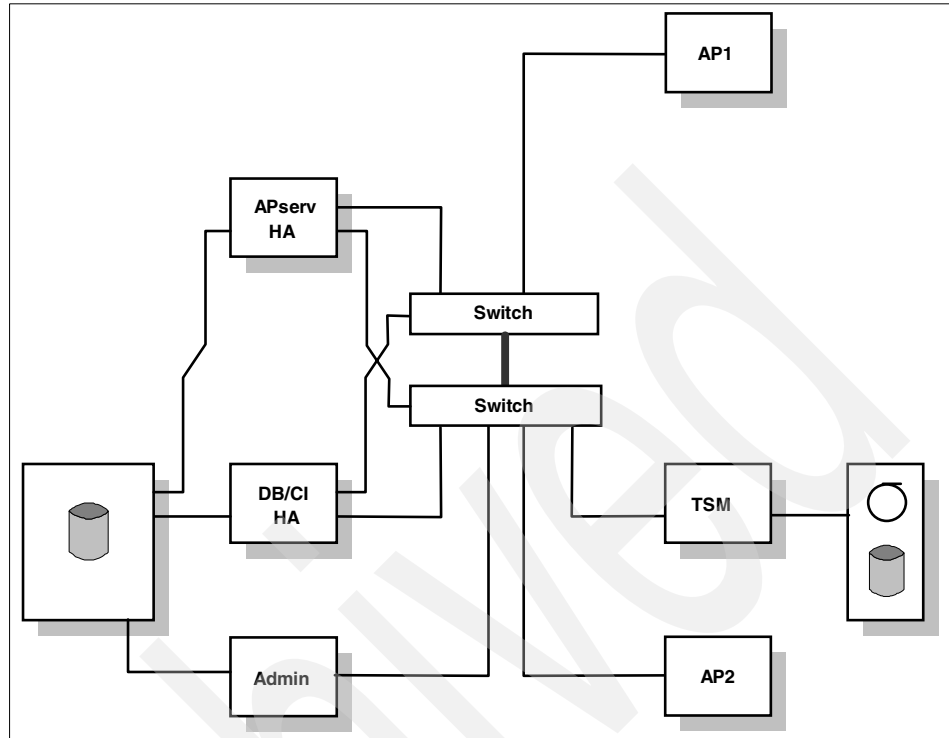


Figure 3-3 A model for a fault-tolerant cluster

The server APserv HA is an additional server in comparison to the basic model. It should be the same model and equally equipped as the server DB/CI HA in order to be able to deal with the workload after a takeover.

Bright idea!

As you do not want to configure the backup server as an idle hot standby (because of cost reasons), we recommend you use the backup server as a powerful application server for interactive users during normal operation. The following list briefly describes the characteristics of the cluster:

- ▶ Cluster nodes: DB/CI HA and APserv HA
- ▶ Cascading resource groups configured for two-way takeover
- ▶ Resource group DB/CI: database and minimal central instance
- ▶ Resource group AP: one or more dialog instances
- ▶ Both cluster nodes attached to storage subsystem
- ▶ Both cluster nodes attached to two separate switches for the front-end network

The following sections discuss the cluster characteristics in detail.

Takeover relationship

The nodes DB/CI HA and APserv HA build a two-node cluster. Each of the cluster nodes is configured as a backup node for the other. We use a cascading takeover relationship for both resource groups. A cascade defines a priority list for a resource group that specifies which of the active cluster nodes acquires the resource group when starting the cluster or in case of a failover. Simply worded, for our cluster, there is a cascade for the DB/CI resource group that specifies DB/CI HA as the primary server and APserv HA as the backup server and vice versa for the other resource group AP. After a takeover, both resource groups are active on the remaining node. Table 3-5 shows the active application components of SAP R/3 and their distribution for the cases of normal operation and the crash of one node.

Table 3-5 Active application components for several cluster states

State	Node DB/CI HA	Node APserv HA
Normal operation	Database Central Instance (minimal)	Dialog Instance
Crash of DB/CI HA	-	Database Central Instance (minimal) Dialog Instance
Crash of APserv HA	Database Central Instance (minimal) Dialog Instance	-

Refer to Chapter 8, “High availability” on page 221 for a detailed description of the contents of the resource groups and for implementations aspects.

3.3.3 Storage subsystem considerations

In Section 3.3.1, “Identifying single points of failure in the basic model” on page 77, we identified the storage subsystem as a single point of failure in case of a complete breakdown. The fault-tolerant model connects both cluster nodes to the storage subsystem, but does not introduce further redundancy for it. Redundancy of the storage subsystem highly depends on the used technology.

Pitfall ahead!

For our model, we simply demand fault-tolerance of the storage subsystem. Solutions to achieve fault-tolerance for the storage subsystem can be found in Chapter 4, “Disk storage” on page 95.

3.3.4 Network considerations

In comparison to the basic model, the fault-tolerant model introduces a second switch and an *Inter Switch Link* to eliminate the switch as a single point of failure. In this section, we take a closer look at the different networks, as Figure 3-3 on page 81 was a simplification and did not show different networks.

Front-end network

The front-end network is the only critical network regarding the provision of the SAP R/3 service. For that reason, the fault-tolerant model introduces a second switch. The cluster nodes have two network adapters connected to the front-end network. For each node, there is a standby adapter, which has to be attached to the second switch. In case of a failure of a switch, exactly one of the front-end network adapters of both cluster nodes is still connected to a working switch. These scenarios are covered in detail in Chapter 6, “Network” on page 149.

Pitfall ahead!

The most important thing to consider is the fact that the switch of the basic model connects the front-end network inside the computer center to the access network outside. For fault-tolerance, you therefore have to connect the access network to both switches. Otherwise, at least a part of the SAP GUIs will not have access to the SAP R/3 system in case of a switch failure.

Backup network

The fault-tolerant model does not propose a redundant configuration for the backup network. A failure of the backup network is not considered a critical situation. Even if you equip each of the cluster nodes with two adapters in the backup network and configure them, in a similar way, to the front-end network, the TSM server itself would remain a single point of failure. In case of failure of the switch (where the TSM server is attached to), you should have a spare port in the other switch in place. Then it is possible to attach the TSM server manually to the other switch. In case of failure of the adapter for the backup network, the resource group DB/CI can be transferred to the other cluster node for backups.

Control network

A failure of the control network only complicates the administration of the SAP R/3 environment. We recommend you use a physically independent network for the control network. Therefore, the control network should not be connected to the same switches as the front-end network and the backup network.

3.3.5 Distribution of SAP R/3 services to application servers

Similar to a central SAP R/3 system is the basic model; of course, it is possible to run a fault-tolerant clustered SAP R/3 system without any additional application servers. However, if larger workloads demand further application servers, their front-end network adapters should be uniformly attached to both front-end network switches. Be aware that a failure of one switch will disconnect half of the additional application servers, thus significantly reducing the overall performance of the system. The fault-tolerant model shows two further application servers, AP1 and AP2.

Bright idea!

Table 3-6 shows how SAP R/3 services can be distributed in the fault-tolerant model.

Table 3-6 Distribution of SAP R/3 services

Service	DB/CI	APserv	AP1	AP2
Message	1	-	-	-
Enqueue	1	-	-	-
Dialog	2 ^a	0..n	0..n	0..n
Update (v2)	2 (1)	0..n (m)	0..n (m)	0..n (m)
Batch	0	0..n	0..n	0..n
Spool	1 ^b	1..n ^c	0	0
Saprouter	1	-	-	-
Special purposes	-	x	-	-

a. Only for administration, not part of a logon group

b. Only for administration, not for productive use

c. As of release 4.0b, more than one spool work process can be configured

Message and enqueue service are part of the minimal central instance. Also, dialog and spool work processes of the central instance should be restricted to administrative purposes.

Dialog, update, and batch processes can be distributed to all application servers APserv, AP1, and AP2. Refer to the considerations in “SAP R/3 instance buffers” on page 68 for interferences between dialog usage and batch processing.

Use the highly available application server APserv as the spool server for productive printers. As output devices (printers) are generally tied to a single spool server, you would have to reassign output devices after the failure of a spool server. There is also a possibility to define an alternative spool server for an output device within SAP R/3.

If you use saprouter for access to SAP GUIs over WAN connections or for access to the SAPNet, it should run on the DB/CI server to be fault-tolerant.

If your systems run special purpose instances (refer to “Special purpose instances” on page 69) or if there are external systems that use hardcoded references to a certain application server (RFC destination, for example), these services should be provided by the highly available application server APserv.

3.3.6 Performance considerations

The last section dealt with the functional distribution of SAP R/3 services to achieve fault tolerance. In this section, we give an example for the fault-tolerant model on how the overall system performance in terms of SAPS is affected in case of a breakdown of a server or a switch. Of course, this is a simplified example, because the performance degradation of a certain service can differ from the degradation of the overall performance, because it depends on the actual distribution of the services. This kind of performance calculation can be easily adapted to other environments. The given SAPS counts of the example are based on a real IBM @server pSeries system environment. They are scaled to smaller units for simpler calculation.

We suppose that the server DB/CI has a relative SAPS count of 9 when used as a database server with a minimal central instance. As explained in Section 3.2.2, “Application server (AP)” on page 67, the performance ratio between the role as database server and the role as application server for a given server is 1:6. Thus, the application server APserv has a relative SAPS count of $9 / 6 = 1.5$. We suppose that the application servers AP1 and AP2 are less powerful servers and have a SAPS count of 1. The overall application server performance is the sum of the single application server SAPS counts. For normal operation, this is $1.5 + 1 + 1 = 3.5$, corresponding to 100 percent. Table 3-7 shows the overall performance degradation of the SAP R/3 system for different failure scenarios.

Table 3-7 Application performance degradation for different failure scenarios

Scenario	Database relative SAPS count	Application relative SAPS count	Overall application server performance
Normal operation	9	3.5	100 percent
Loss of DB/CI	4.5	2.75	78 percent

Scenario	Database relative SAPS count	Application relative SAPS count	Overall application server performance
Loss of APserv	4.5	2.75	78 percent
Loss of AP1 or AP2 Loss of a switch	9	2.5	71 percent

Bright idea!

It is assumed that, in the case of a breakdown of a cluster node, after the takeover of the resource group, the database and the application server share the performance of the remaining node in a ratio of 1:1. Therefore, both database SAPS count and application server SAPS count are cut in half. It is possible to guarantee this share ratio of 1:1 by the use of the AIX Workload Manager.

3.4 Building a disaster-tolerant model

Section 3.3, “Building a fault-tolerant model” on page 77 introduced a fault-tolerant infrastructure model where all single points of failure of an SAP R/3 system environment were eliminated. In this section, we describe additional demands that are posed to disaster-tolerant solutions.

Bright idea!

We enhance the fault-tolerant model in order to achieve disaster tolerance. In this context, we also discuss the capabilities of a *shadow database* system.

3.4.1 Additional requirements for disaster-tolerant environments

If your vital business relies on a running IT system you have to consider implementing a disaster-tolerant system to achieve business process continuity. Of course, a disaster-tolerant implementation includes fault tolerance. But in the case of a disaster (where you probably lose the whole computer center), it is critical to get the system up and running again in a short period of time. In this section, we only cover the technical and topological requirements for an environment, to reduce the unplanned downtime to a minimum in case of a disaster. Organizational aspects of system change management (to reduce the planned downtime) are covered in Chapter 2, “Requirements for a reliable SAP R/3 environment” on page 7.

We define a disaster as an incident where all components providing the service of the SAP R/3 system are cut off suddenly and cannot be recovered within an acceptable time. This does not necessarily mean that the components are destroyed. A possible *soft disaster* would be a complete loss of electrical power within the computer center, for example. On the contrary, a fire that destroys the whole building with the computer center could be called a *hard disaster*.

Pitfall ahead!

For both disaster types, there should be plans for reconstruction and restart in place.

Our approach towards a disaster-tolerant solution for a SAP R/3 system is to distribute the components of the fault-tolerant model to two separate computer centers in different fire compartments. It is assumed that only one computer center is lost even in the case of a hard disaster. Furthermore, the complete loss of one computer center must not affect the operation of any component of the remaining computer center.

Two separate computer centers

Figure 3-4 shows a proposal for a disaster-tolerant infrastructure model for SAP R/3. All components above the dashed horizontal line reside in the computer center *High*; all components below reside in the computer center *Low*.

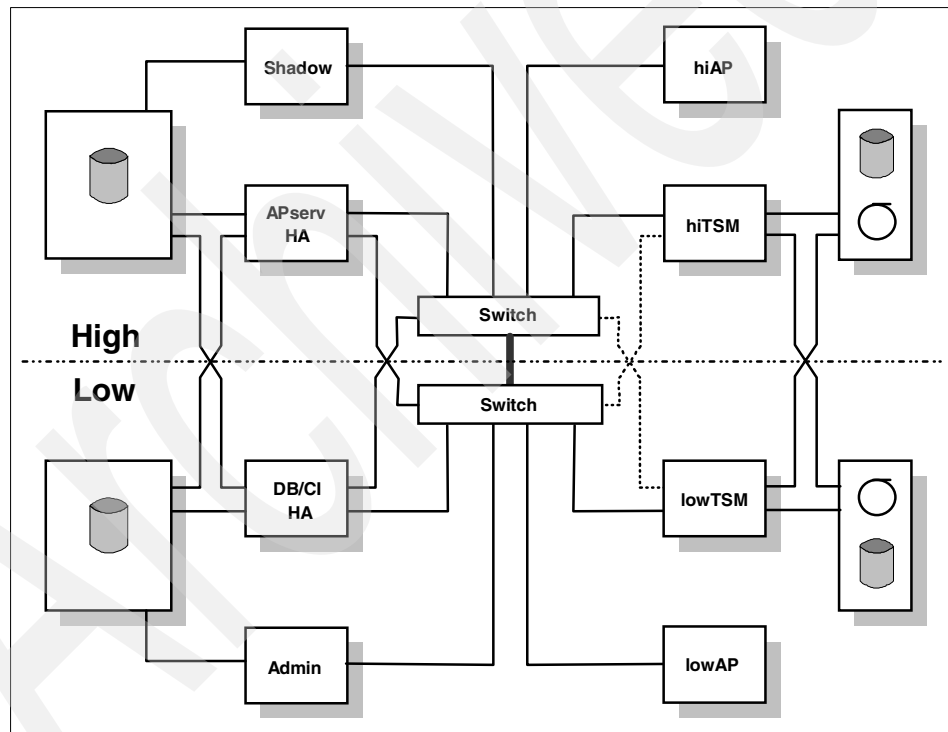


Figure 3-4 A disaster-tolerant infrastructure model

Arrangement of the components

Each of the cluster nodes (DB/CI HA and APserv HA) is placed into one computer center. Also, the additional application server (hiAP and lowAP) has to be distributed. The highly available Inter-Switch Link between the switches now connects all three networks (front end, backup, and control) of both computer centers. As with the fault-tolerant model, the access network has to be connected to both switches.

The disaster-tolerant model requires further components in comparison to the fault-tolerant model. An additional storage subsystem has to be installed in the second computer center. Because the reconstruction of a computer center after a disaster will probably last a long time the backup solution is doubled also. Therefore, a second tape library and TSM server is required. The server *Shadow* is a further optional component and hosts a shadow database system. The importance of the additional components are discussed in the following sections.

Pitfall ahead!

In case of a disaster that includes the loss of one computer center, the cluster performs an automatic failure recovery. Business process continuity is guaranteed, because the business critical SAP R/3 system is immediately up and running again. Of course, there is a performance degradation, because half of the number of additional application servers are not available. The degradation can be estimated with a calculation similar to the example we provided in Section 3.3.6, “Performance considerations” on page 85 and amounts 50 percent.

3.4.2 Additional storage subsystem

In case of a disaster, a single storage subsystem is an unacceptable single point of failure. All data of the SAP R/3 system must therefore be mirrored to one storage subsystem in each computer center using AIX mirroring. Both cluster nodes have to be attached to both storage subsystems, which requires twice the number of adapters within the servers. To overcome the distance between a cluster node and the remote storage subsystem, you will likely have to use fiber optical connections. Samples of disaster-tolerant configurations are covered in detail in Section 4.10, “Disaster-tolerance requirements” on page 133.

Important: Check regularly during normal operation that both mirrors of your data do not have stale partitions. Otherwise, you risk losing the implemented disaster tolerance.

There are further checks that should be performed, from time to time, during routine maintenance windows for standby connections to the storage subsystem that are only used in case of a takeover. These checks should ensure that access to disk resources is possible in a takeover case, because the loss of this possibility may not be detected during normal operation of the cluster.

3.4.3 Dual library and TSM server

The first step to achieve a disaster-tolerant backup and recovery solution is to have another tape library in the second computer center. For the libraries, a fiber attachment or a SAN has to be realized assuming a distance of more than 20 m between the computer centers.

Achieving disaster tolerance for the backup solution

Let us assume that the TSM server we proposed for the basic model (refer to Section 3.2.5, “The backup subsystem” on page 71) is now called *lowTSM* in the disaster-tolerant model. The additional tape library is placed in computer center High and is also attached to the TSM server *lowTSM*.

From now on, backups of the production database can be alternately stored on tapes in both locations. Archived redo log files are stored simultaneously on one tape in both locations. This lets you restore and recover the database independently from tapes of both locations.

Of course, these enhancements are not sufficient to cover a disaster in computer center Low. In this case, there would be a tape library but no TSM server in computer center High. Thus, we introduce an identically equipped IBM *@server* pSeries server *hiTSM* and connect it to the backup network (upper switch) and to both tape libraries.

In the next step, the TSM server instance on machine *lowTSM* is prepared in such a way that it can be taken over manually to machine *hiTSM*. For this to work, it is required that all data belonging to the TSM server instance is kept in a volume group that is mirrored on two external storage subsystems that can be accessed by both servers *lowTSM* and *hiTSM*.

Whether you use the same storage subsystems as for the SAP R/3 system or separate ones is a trade-off between independency from the backup data and consolidation of external disk space. This configuration builds a disaster-tolerant, manual cluster for the TSM server *lowTSM*. In this case, the machine *hiTSM* would be a hot standby node for the TSM server *lowTSM*.

A second TSM server instance for scalability

In order to use the node hiTSM in a reasonable way during normal operation, we suggest you have a separate TSM server instance running on node hiTSM. There are many reasons for having two separate TSM servers. The most important one is scalability. Usually TSM is used in large companies as the backup solution for the whole IT infrastructure and not only for the SAP R/3 systems. Often, there are many non-productive systems to back up.

Bright idea!

Company-wide file servers with millions of files to back up impose a high demand on a TSM server. As both backup time windows and the size of a single TSM database are limited in practice, it is reasonable to split backups to two or more TSM servers.

For this disaster-tolerant model, it is assumed that both nodes lowTSM and hiTSM run a separate TSM server instance. Both TSM instances can be taken over manually to the other node. Thus, both TSM server databases are mirrored to two external storage subsystems. Both TSM servers are configured to access all tape drives of both libraries, apart from tape drives that are used exclusively for other environments in a shared library environment. This builds a cluster of two TSM server instances that can back up and restore data in a disaster-tolerant way.

For operations details of this disaster-tolerant backup and recovery solution, refer to Chapter 7, “Backup and recovery” on page 185.

3.4.4 Shadow database

The disaster-tolerant model shows a shadow database server *Shadow*. The proposed solution is disaster-tolerant even without the shadow database. However, this optional component provides further valuable features to the SAP R/3 infrastructure.

A *shadow database* is a separate database management system that works on a copy of the production database. The shadow database is mounted, but not opened, and is constantly in recovery mode. The archived redo log files of the production database are copied to the shadow database server. The shadow database applies these redo log files to its copy of the database, with a certain time delay.

These are the most important features and advantages of a shadow database system:

- Recovery from logical errors. If a logical failure, for example, a user error that corrupts the production database is detected within the time delay for applying the redo log files, then the shadow database represents still a

consistent error-free state. The shadow database can be copied back, in this case, to the production server for a fast recovery.

- ▶ The consistency of the redo log files is guaranteed, because the shadow database instantly discovers an error when it applies a redo log file.
- ▶ Backups of the production database can be performed from the shadow database system, thus reducing I/O workload on the production server or eliminating backup outages.

A shadow database provides the largest benefit if there is a possibility to swap the productive database with the shadow database copy in a very short time. Downtime due to a recovery is reduced to a minimum if there is neither a restore of a database backup from tape nor a copy from the shadow database back to the production discs necessary. Chapter 9, “Shadow database” on page 259 describes such a configuration in detail.



Part 2

Implementation

Archived

Disk storage

In a nutshell:

- ▶ Fibre Channel technology offers the best performance and flexibility.
- ▶ Use the AIX Journaled File System (JFS) and the Logical Volume Manager (LVM) with Mirror Write Consistency (MWC) for business critical data.
- ▶ Use intelligent file placement, even with high performance storage systems.
- ▶ Use the Subsystem Device Driver (SDD) to optimize throughput and availability.

In this chapter, we describe the fundamental principles of disk drive technology. We explain different ways of distributing data and redundancy information on several disks and visualize the implications of these arrangements on disk subsystems that are built from these components.

A discussion of various disk storage connection mechanisms follows, focusing on advanced optical attachments that are deployed in modern storage products.

The impact of the SAP R/3 database design on the requirements for the file and disk layout is shown. Practical examples for the extraction of sizing and file placement information from an SAP R/3 system are given for DB2 and Oracle databases.

Based on the information about performance and sizing requirements, rules for placing database objects are given. They cover the whole scope of the AIX Logical Volume Manager, such as file level, file system, logical volume, volume group, and physical disk.

We also explain Mirror Write Consistency and discuss the AIX Journaled File System in SAP R/3 environments and give recommendations.

Finally, there are examples for the correct attachment of disk subsystems, taking into account such criteria as fault and disaster tolerance.

This chapter covers the highlighted area in Figure 4-1, which is derived from the SAP R/3 infrastructure model that is used throughout the redbook.

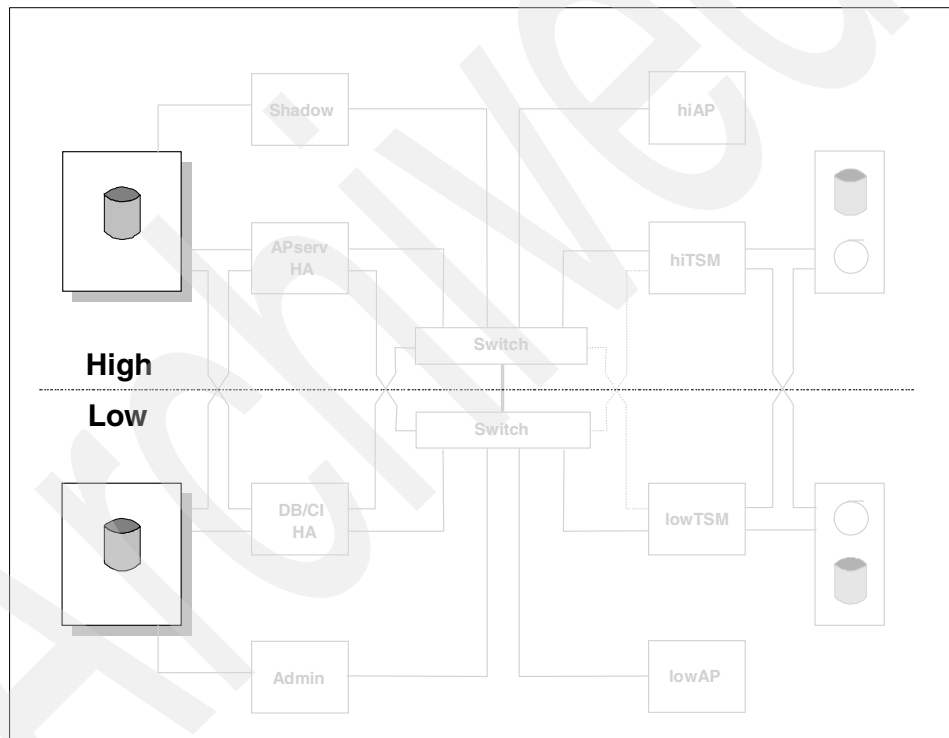


Figure 4-1 The disk storage

4.1 Basic understanding of disk mechanics

A basic knowledge of hard disk technologies is a prerequisite for the following sections. Even though sophisticated disk subsystems become more and more important, they still rely on individual disks. The mechanics of one hard disk in a complex storage subsystem still determines the performance of the whole system. Neglecting this fact may lead to performance bottlenecks.

In a simplistic model, a hard disk drive basically consists of a moveable magnetic head and a spinning platter below this head. The hard disk can read and write data very fast in a sequential way, which means without too many seeking moves of the head over the data area. This sequential mode determines the maximum transfer rate a disk can reach.

If the access pattern to data on disk is random, the average throughput that can be achieved is far less. This is due to the fact that there are seeks of the head involved, which consume a considerable amount of time. An additional performance penalty is the rotational latency, which is when the head has to wait for the platter to turn until the correct sector is positioned under the head for the next read or write access. This latency is determined by the rotational speed.

Due to the physical construction of a disk, we are capable of calculating the basic mechanical limits of a disk. The time for a single I/O request, independent of the fact whether it is a read or write operation, is the sum of all the above mentioned times and can be calculated with the following formula:

$$\text{I/O time} = \text{seek} + \text{latency} + \text{data transfer}$$

The disk manufacturer provides the information for seek, latency, and data transfer times in their technical specifications. Other internal disk factors, connection latencies, and the queuing time until the request is processed by the disk are ignored for simplicity.

The I/O time of a single disk with random 4 KB I/O operations can be determined, for example, with an assumed data transfer time of 0.3 ms for a 4 KB I/O, a seek time of 4 ms, and a latency of 5.7 ms.

$$\text{I/O time} = 4 \text{ ms} + 5.7 \text{ ms} + 0.3 \text{ ms} = 10.0 \text{ ms}$$

In one second, which is equal to 1000 ms, the maximum number of random operations for a single disk can be calculated as:

$$\text{max I/O per second} = 1 \text{ second} / \text{I/O time} = 1000 \text{ ms} / 10.0 \text{ ms} = 100$$

About 100 random 4 KB I/O operations per second result in a throughput of $100 * 4 \text{ KB/s} = 400 \text{ KB/s}$, which is significantly less than the maximum transfer rate of 9 MB/s that can be measured in sequential mode. It is possible to calculate an artificial number of I/O requests for sequential operations by dividing the transfer rate of 9 MB/s by the size of a 4 KB I/O request, which results in 2250 I/O requests per second.

A summary for the resulting numbers is given in Table 4-1.

Table 4-1 Throughput and I/O per second of a single disk

	Throughput (MB/s)	I/O per second
Sequential I/O	9.0	2250
Random I/O	0.4	100
Note: These values are hypothetical and do not reflect the performance of a particular disk model.		

The results clearly show that we can achieve higher data throughput rates with sequential disk access, that is, by fewer head movements. With many random accesses to the disk, the throughput is reduced substantially.

Pitfall ahead!

The primary goal for an efficient disk layout is to reduce the head movements of a single disk head to gain more data throughput. Besides better performance, there is less wear on the mechanical parts of the disk, and so a longer fault free operation can be achieved.

The secondary goal is to satisfy a larger I/O request by using as many heads as possible. If one disk has a sequential throughput of 9 MB/s, then ten disks will have a throughput of 90 MB/s if the data is spread over all disks.

Bright idea!

To reach our goals, we have to distribute the data evenly on many disks to balance the I/O of different groups of files. Each type of file has different requirements that impact the choice of disks, the connection between disks, the loops, and the adapters.

A disk sizing based solely on the database size may lead to a performance bottleneck in most situations. A database with a size of 70 GB fits easily on two 36.4 GB disk drives. Databases operate most of the time in random access mode. Assuming the numbers given in the previous example calculations, two disks support only 200 random I/O operations per second in total, whereas eight disks with a capacity of 9.1 GB provide the same disk space but four times the number of I/O operations per second, that is, 800 I/Os per second, and therefore a much higher throughput.

The distribution of all performance critical data to improve the throughput has to be managed carefully. We show the strategy and an example of a disk layout in “Disk layout for an SAP R/3 system” on page 109.

Employees of IBM can obtain more information by looking at *Disk Sizing, Data Layout and Tuning for AIX - Presentations*, by Dan Braden, found at:

<http://dscrs6k.aix.dfw.ibm.com>

Further performance aspects of an SAP R/3 system are covered in detail in Chapter 11, “Performance” on page 305.

4.2 Redundant array of independent disks (RAID)

To spread data over several disks, redundant arrays of independent disks can be used, also known as RAID. The following descriptions give short definitions of the several RAID levels:

- RAID 0** Data is spread over more than one disk (N disks) (this is called striping). There is no redundancy for data security in case of a defect. It has the lowest cost of all RAID implementations. Striping increases the throughput of sequential I/O, because the transfer rate is multiplied by the number of used disks.
- RAID 1** Data is mirrored over separate disks (2*N disks). It offers high availability (up to half of the disk can fail without data loss). It doubles the cost in comparison to RAID 0, but also doubles the bandwidth for reads if the operating system supports it.
- RAID 4** Data is on N disks, one disk contains checksum information called parity (N+1 disks). The parity disk is the hot spot of the disk subsystem and limits the total bandwidth. It offers the possibility of extending the data area by one disk without affecting other disks. RAID 4 is used by some vendors of Network Attached Storage (NAS) for easier administration, but pay for the cost of performance.
- RAID 5** Data is spread over several disks (N+1 disks) together with parity checksums. In contrast to RAID 4, the parity is not stored on a dedicated disk, but is also spread in a round-robin fashion across all disks in the disk group. It offers higher availability than RAID 0 (one disk can fail without data loss). The performance is decreased because one I/O write request requires four I/O operations on the disks.
- RAID 10** Mixture between RAID 0 and RAID 1, that is, data is mirrored and striped over several disks to increase bandwidth.

The most widely used implementations in productive environments are RAID 1 and RAID 5. The following sample calculations illustrate the bandwidth of these technologies for a better comparison.

We assume a configuration requiring 91 GB of protected storage using 9.1 GB drives. For a RAID 1 implementation, we need 20 disk drives; for a RAID 5 implementation, we need 11 disks, the capacity of 10 disks for data and one for parity. We further assume that the I/O load is random and that a disk can perform 100 random 4 KB I/O operations.

Write operations in RAID 5 implementations are subject to a performance penalty, because four I/O operations are needed to satisfy one write request. The following tasks have to be performed by the RAID 5 implementation of most RAID controllers:

- ▶ Read the block that needs modification (1st I/O).
- ▶ Read the related parity block (2nd I/O).
- ▶ Remove the knowledge of the overwritten block from parity.
- ▶ Calculate the new parity, including the new data block.
- ▶ Write the data block (3rd I/O).
- ▶ Write the related parity block (4th I/O).

One logical write operation on a RAID 1 system has to perform two physical I/O operations, one write on each copy of the data, before it is complete.

Table 4-2 summarizes the values for reads and writes in the two scenarios.

Table 4-2 I/Os per second in RAID 1 and RAID 5 arrays for random access

	Read	Write
RAID 1 (20 drives)	$20 * 100 \text{ I/Ops} = 2000 \text{ I/Ops}$	$2000 \text{ I/Ops} / 2 = 1000 \text{ I/Ops}$
RAID 5 (11 drives)	$11 * 100 \text{ I/Ops} = 1100 \text{ I/Ops}$	$1100 \text{ I/Ops} / 4 = 275 \text{ I/Ops}$
Note: These values are hypothetical and do not reflect the performance of a particular RAID system.		

The results indicate that a RAID 5 implementation offers less performance than a RAID 1 implementation. The arguments in favor of RAID 5 are usually not based on performance, but on price. A RAID 5 configuration costs less than RAID 1, because it needs fewer disks to provide the same usable capacity and the costs scale with the number of disks.

For a fair comparison of costs, it is necessary to calculate the costs per performance unit, which can be expressed as price per I/O operations that can be completed in a given time interval.

For the calculation of the price that we have to pay to get an I/O done, we use the following formula:

$$\text{disk number (N)} * \text{disk price (D\$)} / \text{I/O per second} = \text{price per I/O}$$

Table 4-3 summarizes the I/O price in standard RAID 1 and RAID 5 arrays.

Table 4-3 I/O price in standard RAID 1 and RAID 5 arrays

	Read	Write
RAID 1 (20 drives)	20 D\$ / 2000 I/Ops = 0.01 D\$ / I/Ops	20 D\$ / 1000 I/Ops = 0.02 D\$ / I/Ops
RAID 5 (11 drives)	11 D\$ / 1100 I/Ops = 0.01 D\$ / I/Ops	11 D\$ / 275 I/O = 0.04 D\$ / I/Ops
Note: The values given in this table are hypothetical and do not reflect the performance of a particular RAID system.		

Pitfall ahead!

From this price calculation, we can draw the conclusion that the price per performance ratio of RAID 1 when performing read requests is the same as for RAID 5. Considering write operations, the value for RAID 1 is twice as good as the value for a RAID 5 configuration.

These are theoretical values and apply to RAID 5 implementations without any involved write caching. “Disk subsystems based on RAID 5 implementations” on page 105 shows how the RAID 5 performance penalties can be overcome and how intelligent processing of the data can turn a modified RAID 5 implementation into a performance optimized data handling. The IBM Enterprise Storage Server removes the RAID 5 penalties over RAID 1 and therefore offers an enormous price advantage compared to disk subsystem solutions based on RAID 1.

4.3 Disk interconnection architectures

We have taken a closer look on the disk mechanics and some possibilities of grouping disks together. The interconnection technology between individual disks is an important matter as well, because the overall throughput depends on this technology.

The industry standard for disk connection technology is the SCSI-3 standard architecture. Based on this standard several flavors of SCSI-3 have been implemented in the past, for example, the SCSI Parallel Interface, the Fibre Channel (FC), and the Serial Storage Architecture (SSA).

Table 4-4 gives a rough overview of the different SCSI implementations, neglecting the impact of other active communication partners on the bus and the subsequent throughput reduction.

Table 4-4 Bandwidth and throughput of interconnection technologies

Technology	Theoretical bandwidth (MB/s)	Maximum throughput (MB/s)
SCSI 1	4	3
SCSI 2	10	8
SCSI 2 F/W	20	17
Ultra SCSI	40	32
Ultra2 SCSI	80	72
Ultra3 SCSI	160	128 (estimated)
SSA 80 (per loop)	80	35
SSA 160 (per loop)	160	90
FC	100	90

The SCSI parallel interface is the most mature technology, which offers various types of interfaces, cables, and connectors. It, however, has certain limitations concerning maximum number of devices, maximum throughput, and connection distance. As a connection interface for internal media (disk devices, tape devices, and CD-ROM) or for the usage in small workgroup or departmental servers, the SCSI parallel interface is the first choice. However, if a large amount of disks has to be connected, the parallel approach reaches its limits. The reason is the increase of bus contention, also called arbitration, which can easily consume 40-50 percent or more of the stated capacity of the bus.

The remaining options for interconnections between drives and servers are Fibre Channel (FC) and the Serial Storage Architecture (SSA). Both allow elongation with fiber optics and offer the possibility to connect communication partners over much longer distances than that of the SCSI parallel interface (25 m).

Fibre Channel is a further stage in the development of SCSI and uses serial technology. Fibre Channel Arbitrated Loop (FC-AL) still has the constraints of arbitration, which may cut down the throughput seriously. Fibre Channel Switched Fabric, however, removes the arbitration overhead.

SSA is also a serial bus protocol and offers the option of connecting a large number of disks without the disadvantage of a high communication overhead, due to arbitration. SSA offers a high throughput and the flexibility to expand the storage by as many disks as you want without server downtime until the maximum of disks (48 disks per loop and 96 disks per adapter) is reached. Due to the ring topology of SSA, it is possible to open a loop, to exchange or add some devices or even whole drawers and close the loop again without any service interruption. So a smooth upgrade to more devices is possible.

Another aspect of flexibility in an SSA implementation is the option of having several host adapters in one loop. For highly available configurations, up to eight different servers can be connected to the same loop, in order to increase the number of access paths for availability.

Figure 4-2 shows the SSA loop principle of eight disks connected to an adapter.

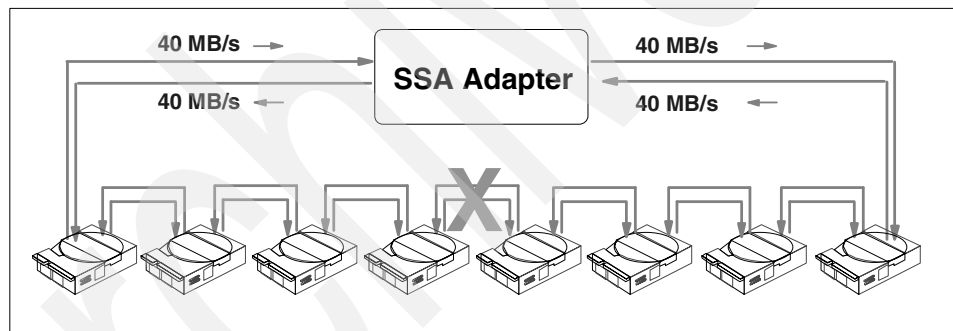


Figure 4-2 Data transfer in an SSA loop

The following are some of the SSA features:

- ▶ Ring topology, no single point of path failure (see large X in Figure 4-2)
- ▶ Link error recovery procedures
- ▶ Automatic selection of alternate path
- ▶ Automatic loop rebuild after repair
- ▶ Hot swappable cables and disk drive modules

For elongation of the distance between the server and the disks or between the servers, you can choose SSA Optical Extenders. The IBM SSA Advanced Fibre Optical Extender (FC 8851) allows distances up to 10 km with single-mode fibers.

Pitfall ahead!

There are some general rules for using Optical Extenders:

- ▶ The older black colored Optical Extender, FC #5500, works with multi-mode fiber cables. The maximum allowed distance is 2.4 km.
- ▶ The newer blue colored Advanced Optical Extender, FC #8851, works only with single-mode cables. The maximum allowed distance is 10 km.
- ▶ When the Advanced Optical Extender is used with multi-mode fibers, then the Mode Conditioning Patch Cord, FC #8852 for 50 μm fibers and FC #8853 for 62.5 μm fibers, must be installed. The resulting maximum distances are reduced to 2.4 km.
- ▶ An old Optical Extender at one end and a new one at the other end of one link is not allowed.
- ▶ A copper cable attached to port A1 (B1) and an Optical Extender attached to port A2 (B2) on the same adapter, and vice versa, is not supported.
- ▶ The maximum allowed light attenuation with FC #5500 is 3 dB; with FC #8851, it is 8 dB. These values must not be exceeded.

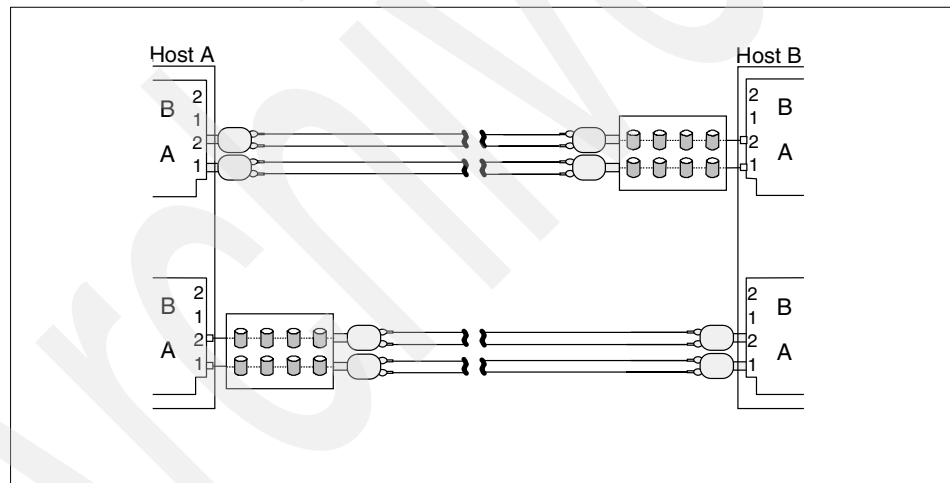


Figure 4-3 Supported connections of SSA Optical Extenders

Bright idea!

The following recommendations are important for correctly implementing the Optical Extenders:

- ▶ If the loop is connected between two hosts, as shown in Figure 4-3, the loop is logically divided into two parts, one connecting the SSA adapter ports #2 and one connecting the SSA adapter ports #1. Both parts of the loop should contain the same number of disks.
- ▶ It is important to keep the access delay times from both ports of the SSA adapter to all disks nearly identical. The delay times are dominated by the

fiber paths. Therefore, the Optical Extenders have to be directly attached to both ports of an SSA adapter, as shown in Figure 4-3.

- ▶ It is advisable to measure the total path light loss during installation of the system (to ensure that it is within specification) before using the fiber optic links.
- ▶ Fiber optic cables are aging and may increase the attenuation by 0.5 dB in a year as a function of the environmental conditions.
- ▶ The usage of 10 km of fiber will introduce 100 μ s of delay and bring down the maximum data rate per I/O channel from 40 MB/s to 1 MB/s.

Further information about SSA can be found at:

<http://www.hursley.ibm.com/ssa>

You can also find out more about SSA by using the redbook *Understanding SSA Subsystems in Your Environment*, SG24-5750.

4.4 Disk subsystems based on RAID 5 implementations

In Section 4.2, “Redundant array of independent disks (RAID)” on page 99, we learned the reason why RAID 5 has the reputation of being slow, which is the write penalty for I/O requests that consists of reading and writing the parity information. There are some disk subsystems that are based on RAID 5 technology, but allow very fast access nevertheless.

4.4.1 General idea of improving RAID 5 performance

RAID 5 implementations can avoid some of the write penalty by performing full stripe writes or partial stripe writes. In RAID 5, the data is spread in so-called data stripes on several disks, as shown in Figure 4-4. All these data stripes belong to one data stripe. One parity checksum information belongs to one data stripe.

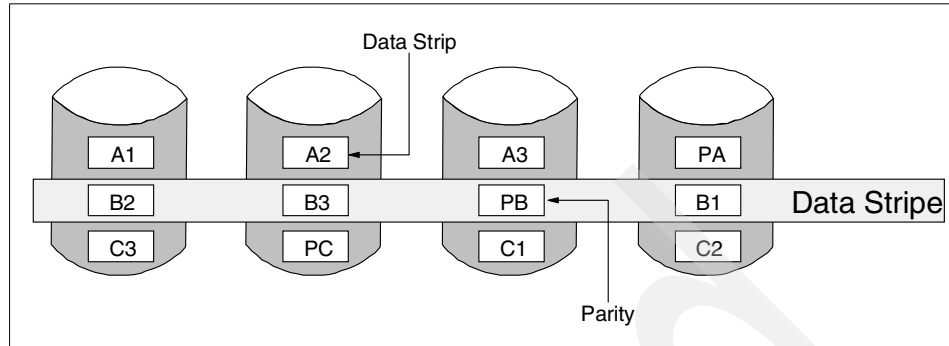


Figure 4-4 Term definition in RAID 5

The marked data stripe in Figure 4-4 on page 106 consists of the data stripes B1, B2, and B3, and the parity information PB.

When all data of a data stripe is changed at once and staged for writing, the old parity information is no longer needed and does not have to be read from disk for recalculating the new parity. The new parity can be calculated on the fly from the new data and can be written with the same I/O access. This is called a *full stripe write* and is characterized by the fact that the read of the previous stripes and of the previous parity can be omitted completely. Generally, this is seen when performing sequential writes.

Random writes usually will not alter the whole stripe, but only parts of it. Even then an optimization is possible if the parity information is cached. This means that parity is read only once for several data stripes and kept in memory and when a write occurs, only the new calculated parity has to be written. Thus, multiple reads of parity data can be avoided.

In order to make efficient use of these optimization techniques, the RAID 5 system has to gather information on the incoming data and has to recognize sequential access patterns. Modern disk subsystems deploy efficient processing technology and large caches in order to offer tremendous performance.

4.4.2 IBM Enterprise Storage Server

A disk subsystem like the IBM Enterprise Storage Server (ESS) offers a great bandwidth, gives you the opportunity for storage consolidation in a single system, and reduces the administration effort. Instead of handling hundreds of single disks, large data volumes can be managed literally with some mouse clicks.

The ESS consists of standard components, which are shown in the schematic overview in Figure 4-5.

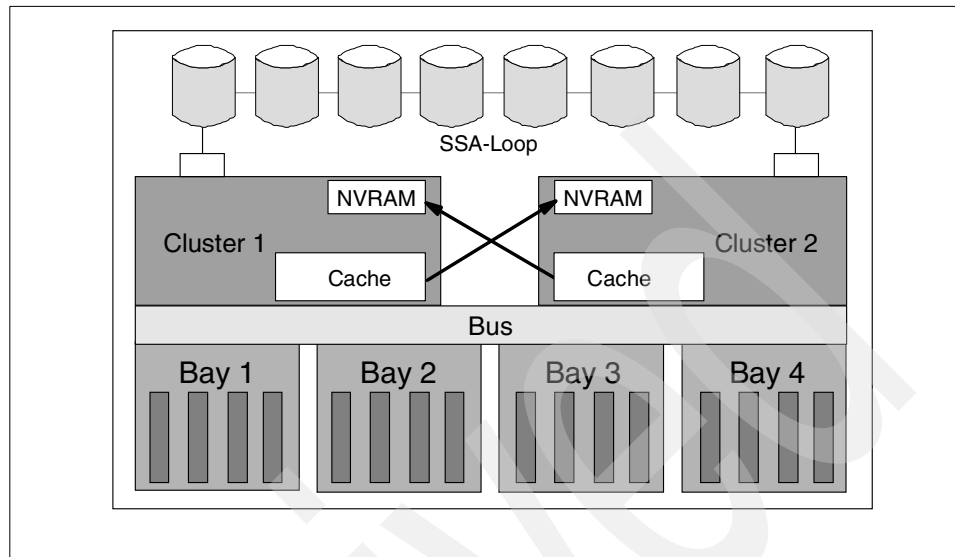


Figure 4-5 Schematic ESS overview

The core of an ESS is a highly available cluster of powerful servers. These servers are based on RS/6000 and IBM @server pSeries technology. Multiple gigabytes of memory for caching purposes can be configured for these cluster servers. Most of this memory is used as a read-only cache. Both cluster servers mirror 192 MB of memory to the second cluster, which is a battery backed Non Volatile Random Access Memory (NVRAM). This amount of memory (384 MB) is used as a read/write cache, and the NVRAM provides the required safety for the data that has not been written to the disks yet.

Pitfall ahead!

The ESS write cache is the main reason for avoiding the RAID 5 write penalty, as a write request is immediately stored into the cache and an I/O complete indication is quickly returned to the host. With sequential writes, the full stripe writes come into play, to further reduce I/O operations. For random writes, additional I/O operations to maintain the parity information may be necessary. They are handled internally and asynchronously by the ESS.

A large part of the ESS is equipped with SSA disks. These disks are connected to the cluster servers with SSA loops and adapters. Therefore, both internal cluster servers have full access to all disks in case of a failure of the other. Furthermore, disk interconnection based on SSA technology offers maximum performance and throughput for the disk access. SSA offers the highest performance in the marketplace.

The SSA disks are grouped together in ranks. A rank is a JBOD or RAID 5 formatted disk pack, which comprises eight volumes. Therefore, it is also called an eight-pack. This rank can be formatted as a RAID 5 array and, depending on the SSA loop configuration, it is defined as a 7+P array or as a 6+P+S array. This means that in the second case, a hot spare disk (S) is reserved. This reduces the net capacity, but increases availability, and enables deferred maintenance work.

The ESS contains the advantages of SSA inside and offers the advantages of Fibre Channel outside of the system. The high data throughput of SSA in combination with many disks and the easy handling of FC attached storage make the ESS an ideal solution for high performance storage requirements.

Servers attach to the ESS over the connectors in one of the four bays. One bay holds up to four host bay adapters. There are three different types of adapters:

- ▶ SCSI with two connectors
- ▶ Fibre Channel with one connector
- ▶ ESCON with two connectors

This adds up to a maximum of 32 different connectors. This area is discussed in depth in Section 5.2, “Storage Area Networks based on Fibre Channel” on page 140.

The ESS provides an outstanding read and write performance both with sequential and random I/O. The main reasons for this behavior are:

- ▶ High *read performance* is offered due to a multi-gigabyte memory cache, which satisfies many read requests and reduces the number of physical disk I/O for cache hits. Up to 64 internal high-speed SSA data paths to disks provide extensive I/O parallelism.
- ▶ *The fast write cache* returns an I/O complete indication quickly and the write on the disks is handled asynchronously in the background.
- ▶ *Data striping* naturally tends to balance I/O activity evenly across a set of disks (improving average response time) and allows multiple requests to the same logical volume to proceed in parallel if the data resides on multiple disks.
- ▶ Due to *write preempts*, it is possible that data is written to cache and updated there again by the application before it is written to disk. So the write and additional penalties are further reduced.
- ▶ Every SSA adapter in the ESS has its own *disk adapter cache* for the disks it manages. This cache holds frequently referenced data and parity information. This can potentially eliminate reading them from the disk in order to update the parity information when changed data is destaged to disk.

- *Full stripe destage* occurs, when the ESS recognizes that multiple application I/O operations comprise an entire stripe. In that case, it will create the parity information on the fly and write data and parity to multiple disks in parallel in a single I/O operation. This is especially useful for online redo logs in database environments, which are written in a sequential way.

More features of the ESS like FlashCopy are covered in Chapter 9, “Shadow database” on page 259, and further information to these topics can be found at <http://www.storage.ibm.com/storage>

4.5 Disk layout for an SAP R/3 system

The disk layout for an SAP R/3 system consists of several steps. They involve all the layers from the application down to the physical disk.

Database layer	The database layer is built up by the database (DB) subsystem and comprises tables, tablespaces, and containers (DB2) or data files (Oracle).
Logical layer	The logical layer consists of the components that belong to the Logical Volume Manager, that is, file systems, logical volumes, and volume groups.
Physical layer	The physical layer is represented by the hardware, such as disks, loops, adapters, drawers, and racks.

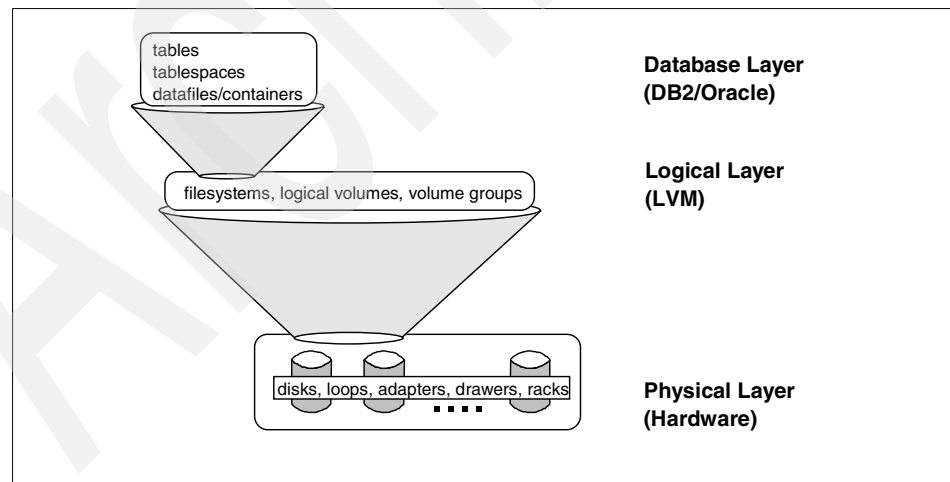


Figure 4-6 Overview of the three layers

The relationship between the layers is depicted in Figure 4-6. Our considerations for the disk layout have an impact on these three layers, the three layers influence the design in return.

For example, if we plan a disaster-tolerant configuration, the hardware has to be placed in two locations, and this implies a mirroring of the components in the logical layer. The interdependencies between the layers are the central theme for the following sections.

4.5.1 Different file types in an SAP R/3 installation

An SAP R/3 system consists of several types of files and storage spaces, which should be handled in a different way, considering the different requirements for the data.

Table 4-5 gives you a rough overview of the file types.

Table 4-5 Overview of file types

File types and storage space	Performance critical	Access method
Operating system files	no	random
Operating system paging space	yes	random
SAP R/3 executable	no	random
Database executables	no	random
Database redologs	yes	sequential
DB data (containers / data files)	yes	random

We have to take care of three performance critical objects:

- ▶ The performance of the paging space should not be an issue, because it should not be used in the first place. As soon as paging space is needed, the system will become very slow, because all data manipulation in memory will require disk activity in order to swap pages in or out. In this case, paging space is performance critical, but then increasing the physical memory is a much more efficient way to tune system performance.
- ▶ All changes to the database are first written to the database redo logs, then written to the data files, and finally confirmed to the application. Therefore, the database redo logs are very performance critical. They are written in a strict sequential way. Due to the totally different access characteristics, they should not be mixed with other performance critical file types in the same disk area.
- ▶ The database (DB) data has the largest volume of the installation. Data changes in this area are written asynchronously in a random fashion to the

disk. Most I/O operations are observed here. Therefore, we put our emphasis on the design of the database layout.

The following sections cover mainly the distribution of the database data.

4.5.2 Layout criteria

The basis for a good disk layout are the following design criteria:

- ▶ Data security
- ▶ Good performance
- ▶ Easy administration
- ▶ Clarity in design
- ▶ Flexibility for growth

The design criteria should always be taken into account when a new layout is being designed or an existing layout is being extended. This could originate from database growth or high-level requirements for the availability of the SAP R/3 system.

In Section 4.1, “Basic understanding of disk mechanics” on page 97, we have discussed the influence of disk numbers on the throughput. Therefore, the two influencing factors, sizing and performance, should always be kept in mind throughout the layout process (see Figure 4-7).

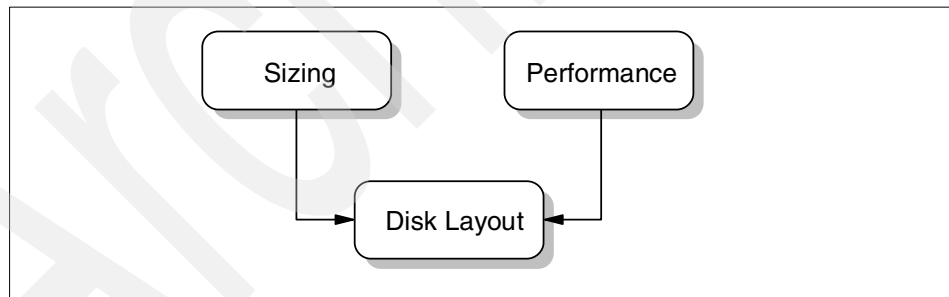


Figure 4-7 Influences on the disk layout

There are several possibilities for designing the layout; it depends, for example, on the data size, the transaction rates, the user numbers, or a particular application requirement. Another layout technique is based on I/O transactions. SAP R/3 accesses the database system with workloads typical for traditional online transactions, online analysis, decision support, and batch reporting system. So it is possible to model the I/O behavior of an SAP R/3 system based on performance numbers of existing systems of these types.

4.5.3 SAP R/3 specific considerations

Based on a forecast for the growth and the access of several SAP R/3 tables, the logical layer, that is, the distribution of data files on the disk, can be drafted. For a first estimate, the most active tablespaces in an average system can be consulted. They are shown in Figure 4-8, separately for DB2 and Oracle.

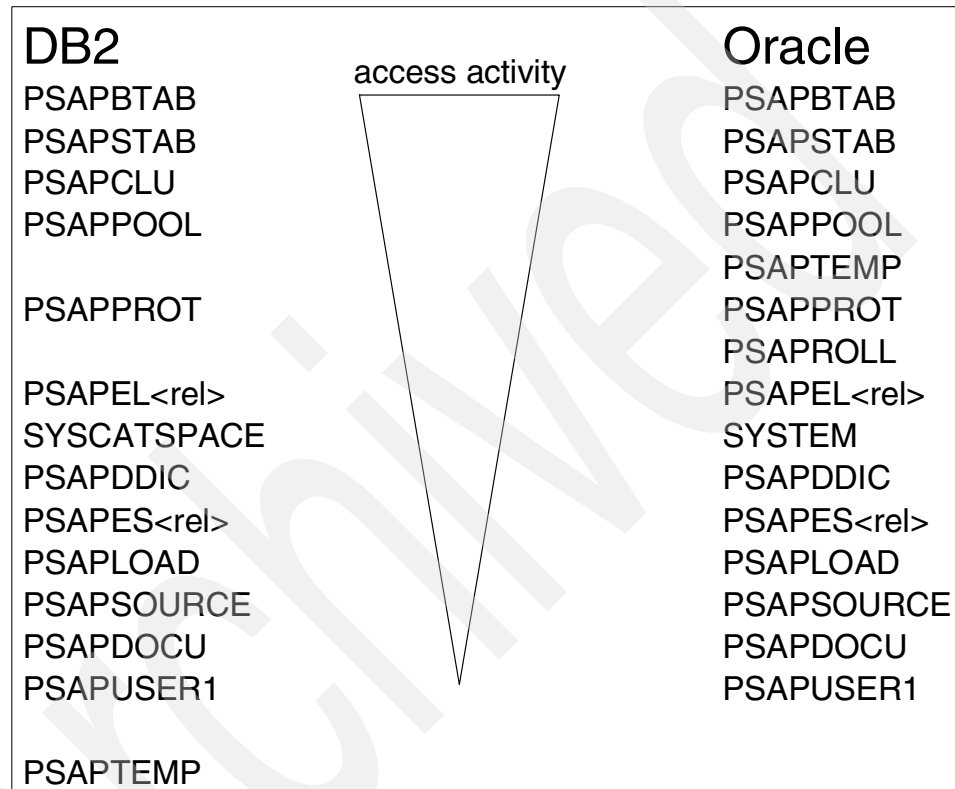


Figure 4-8 Access activity on the SAP R/3 standard tablespaces

The first main difference between DB2 and Oracle is the access to the tablespace PSAPTEMP. In Oracle, there is much activity on PSAPTEMP for the sorting of the data and SQL operations, such as group by statements. Thus, the PSAPTEMP tablespace has to be considered as performance critical. In DB2, the physical tablespace PSAPTEMP is used differently from Oracle, because of the handling of sorting in memory and the special treatment as a system managed tablespace. In the DB2 implementation of SAP R/3, the PSAPTEMP tablespace exists for compatibility with some SAP R/3 tools and transactions. The second difference is the absence of the tablespace PSAPROLL in DB2, which contains the rollback segments in Oracle. This mechanism is not used in DB2.

Bright idea!

There are two scenarios for more detailed planning:

- If a production system does not yet exist, you can base your layout on the data, which can be extracted from a test system. With the experience from a similar workload that was recorded during the running of test cases, you can estimate more accurately the access pattern and the growth of the tablespaces. In this case, the disk layout has to be reviewed a few weeks or months after the start of the production system.
- If there is already a production system, you can create a very detailed design. For example, if you plan to upgrade your server or the storage hardware, it is practical to check the layout for bottlenecks. The upgrade gives you the chance of adjusting the design to better fit your requirements.

SAP R/3 stores all access and performance statistics in tables inside the database. The transaction code (TX) in SAP R/3 that leads you to the database performance data is ST04:

DB2

TX ST04 leads you directly to TX DB6COCKPIT. The tab Tablespaces contains the read and write requests to the physical disks.

Oracle

In TX ST04 you chose Detailed Analysis-File system Requests and get the physical I/O requests per data file.

Bright idea!

You may export the performance data into a spreadsheet, for example, Lotus 1-2-3 or Microsoft Excel, for sorting and calculating sums. There you can aggregate the access values and balance them between several disk areas. You should concentrate on the high activity workloads and keep in mind that it is impossible to optimize all accesses of all tablespaces.

The result so far is a listing of all tablespaces with all sizes and their access activity. If there are tablespaces specific to your environment, these should be considered as well. The next section shows an example of this design approach.

4.5.4 Example for a disk layout

Bright idea!

The operating system should be separated from the SAP R/3 files to simplify system management tasks, for example, backup and recovery, and for system upgrades.

A first approach for a database file layout is to consolidate data files that are not performance critical on one disk and store each of the critical files on a separate disk. But the result could be a database that is distributed over hundreds of disks and could not be handled anymore. Therefore, a different technique has to be found.

Analyzing disk access

We provide you with an example based on an Oracle database. When using Oracle, it is necessary to make the disk layout based on data files. When using DB2, the layout has to be made based on tablespaces. The load balancing within a tablespace, that is, between the several containers that belong to a tablespace, is done by DB2 itself, and, therefore, the layout process in DB2 is much easier.

Example 4-1 is an excerpt from a productive SAP R/3 system, which lists the physical data file accesses for reads and writes on the contained tables.

Example 4-1 Output of transaction ST04

Filename	Reads	Writes	Blk Reads	Blk Writes
apqdd.data1	13922	23297	13922	23297
apqdi.data1	13204	31836	13204	31836
bkpfd.data1	11412	548	11412	548
bkpfd.data2	82297	1603	82297	1603
bkpfi.data1	737	660	737	660
bkpfi.data2	627	310	627	310
bsidd.data1	176941	3518	176941	3518
bsidd.data2	198145	3493	198145	3493
bsidi.data1	17542	2637	17542	2637
bsidi.data2	52691	8222	52691	8222
btabd.data1	157034	35538	164873	35538
btabd.data2	54319	19661	71091	19661
[...]				
Sum	29567514	13035133	34029768	13314417

The numbers in the second and third column represent the number of physical I/O operations on disk; the fourth and fifth columns represent the number of 8 KB database blocks that are read from or written to disk. In most cases, the reads are correlated with the block reads and the writes are correlated with the block writes and are of the same order of magnitude. For this reason, it is usually enough to consider only one of the two columns.

Distributing disk workload

With a spreadsheet tool, we are able to total and sort the accesses for the tablespaces, which are shown in Example 4-2. When using DB2, the process starts here.

Example 4-2 Cumulated and sorted output of transaction ST04

Tablespace	Reads	Writes	Blk Reads	Blk Writes
PSAPROLL	471674	4075818	471674	4075818
PSAPBTABD	6679177	2572343	9893607	2572343
PSAPBTABI	4764893	1911307	4846368	1911307

PSAPPROTD	46331	462439	49176	462439
PSAPSTABI	2992097	425992	3259881	425992
PSAPSTABD	3377007	418441	3621519	418441
PSAPPROTI	91458	359492	92353	359492
PSAPCLUD	87676	220801	87676	220801
PSAPAPQDD	110574	187724	110574	187724
[...]				
Sum	29567514	13035133	34029768	13314417

In this example, we show only the most active tablespaces. The first two tablespaces, PSAPROLL and PSAPBTABD, comprise 60 percent of all write accesses. The first eight tablespaces, PSAPROLL to PSAPCLUD, comprise 80 percent of all write accesses.

Pitfall ahead!

It is evident that most of the considerations should be made with respect to these heavily used tablespaces. In most cases, 10 to 20 percent of the tablespaces obtain 70 to 90 percent of the workload.

We put an emphasis on these 10 percent of tablespaces for distributing the data and position these on distinct disk areas. The other tablespaces are merged with the heavily used tablespaces in order to balance between hot and cold spots on the disk, which means areas with a very high access and areas with little access.

Bright idea!

Therefore, we put the eight most heavily used tablespaces on separate disk areas. For simplicity's sake, we assume eight disk areas and merge the remaining tablespaces in a round-robin fashion to these areas.

Furthermore, you have to take into account that the tablespaces have different dynamic growth characteristics. This balancing of the sizes of the disk areas can also be done in the spreadsheet.

Bright idea!

When sizing tablespaces, you have to take into account that there is a four percent overhead for the file system directory structure information (inodes). A spreadsheet helps to calculate and round the resulting physical partition (PP) quantities automatically to decrease the possibilities of arithmetical errors.

For example, if we need a file system with eight 2 GB data files and have a PP-size in the volume group of 32 MB, the logical volume size is:

$$\text{PP number for the logical volume} = 8 * 2 \text{ GB} / 32 \text{ MB} * 1.04 = 532.48$$

This value has to be rounded to 533 PPs. Using this size as a minimum, the logical volume can be created, which holds the file system for the eight data files.

Upon completion, the information in the spreadsheet can be exported and used to create shell scripts (with **sed** and **awk** or within vi) for creating logical volumes and journaled file systems, for moving or renaming data files and for remote copies (refer to Chapter 12, “SAP R/3 system copy” on page 365). It is also useful to document of file system growths in order to forecast the required disk space in the future.

With the information of the spreadsheet, the database layer is mapped onto the logical layer, which means on the data files, which will reside within file systems according to the existing physical disks (physical layer). The physical layer is represented by the hardware resources, such as racks, drawers, adapters, and disks, and is planned according to the disk layout criteria mentioned in Section 4.7, “General requirements” on page 125.

Disk grouping with SSA

For a native SSA configuration, a recommended technique is to group four or eight disks to one disk area and put one logical volume on it either by striping on stripe size level or on a physical partition level (the so-called *poor man's striping*).

Figure 4-9 gives you an overview of this technique. We have two disjointed areas named Volume A and B with their mirrors A' and B'. The volumes are striped over four disks in a round-robin fashion on 16 disks.

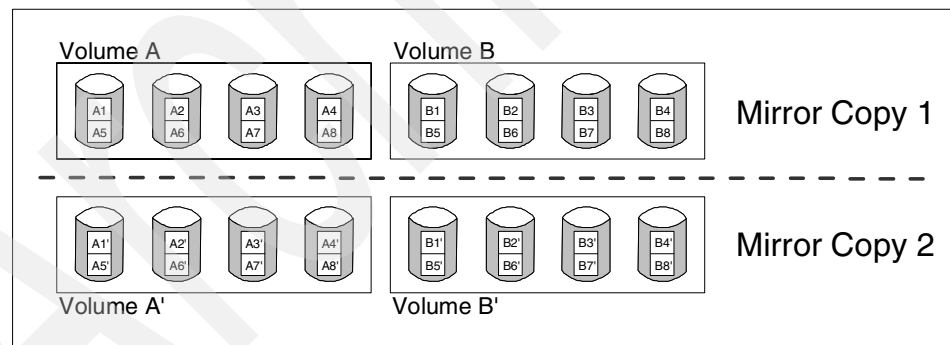


Figure 4-9 Disjointed disk areas

Pitfall ahead!

The use of larger logical volumes offers easier administration through a smaller number of volumes, and a better performance, because of parallel data access to four or eight independent disks.

Tip: A new location attribute known as *Super Strict* has been introduced in AIX 4.3.3 and reduces the possibility of data loss due to a single disk failure by disallowing a second or third mirrored partition to share the same disks as another copy. This is useful for mirroring and striping (RAID 10) with SSA on the Logical Volume Manager level.

Logical Volume sizing with ESS

In an ESS, a logical volume is defined as a RAID 5 stripe set spread over seven or eight disks in a rank. It has a correlation to a native SSA configuration (from a systems management point of view). The disjointed disk areas are represented through the large volumes on the different ranks.

There are three approaches to configuring the ranks of an ESS and splitting them into useful volume sizes:

- Configure dedicated volumes for the explicit needed size.

This seems to be a good approach in the beginning, because you create the volumes as large as needed without leaving empty space. However, after a few weeks of usage, you might need larger volumes or a different partitioning. There is no possibility of deleting volumes in the ESS, so this approach should be avoided.

- Create all volumes in all ranks of one size (for example, 8 GB volumes).

In this case, you have enough flexibility to respond to changing demands by assigning the needed storage space to a connected server. The disadvantage is the large number of volumes, which would cause an administrative overhead. If we assume an ESS with a capacity of 1470 GB, the number of 8 GB volumes is about 180. If we create larger volumes, for example, 16 GB volumes, this logical volume size may be too large for some requirements and too small for other.

- Partition your ranks with useful sizes in an easy to understand pattern.

A rank in the size of 105 GB could, for example, be split up in three volumes with: 40 GB, 40 GB, and 25 GB in size. If you configure all your ranks of your ESS in this way, you keep the number of volumes and the administration overhead small and you have the possibility to react to different requirements with different sizes. And you have the chance to use the whole rank without having remainders that are not usable. If you need smaller volumes, you can choose a pattern like 40 GB, 40 GB, 15 GB, or 10 GB.

Bright idea!

Figure 4-10 on page 118 shows you a sample partitioning of an ESS and the assignment to the connected systems.

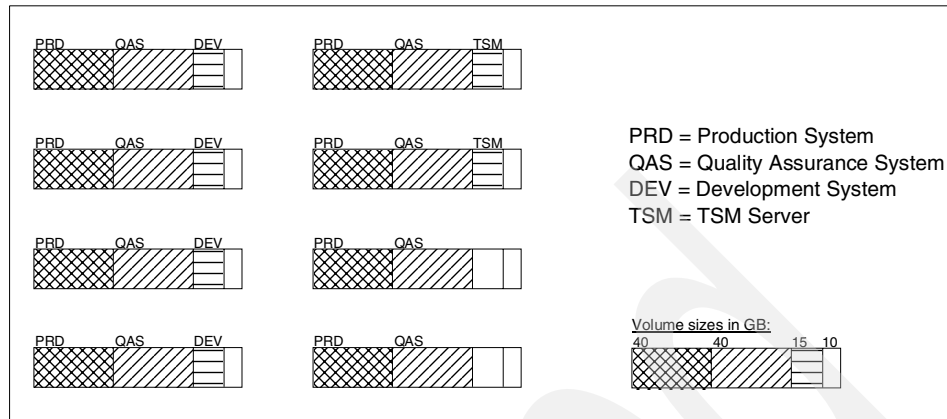


Figure 4-10 Sample partitioning of an ESS with eight ranks

Bright idea!

It is possible to create file systems with very large sizes. For administrative purposes, large file systems maybe considered convenient. But you should not exceed certain sizes, depending on your requirements.

With FlashCopy, a file system check may be necessary. A file system with 150 GB needs about 15 minutes to be checked. A 600 GB file system needs one hour. But four file systems with 150 GB each can be checked in parallel and need a check time of 15 min. So a compromise between easy administration and other important factors has to be found.

4.5.5 Guidelines for tablespaces

Today's relational database management systems (RDBMS) offer a lot of features for managing large databases. In the mid-1990s, a database with a size of over one terabyte was hardly manageable. The basic object for organizing tables is a tablespace. Database tuning based on parameters for tables and tablespaces is an important issue for database administrators, but you cannot expect that it will drastically improve the response time of queries that are badly formulated.

If a query to the database requests one million rows, the RDBMS has to read them and deliver them to the query. For example, a query like "give me all invoices stored in the database" could cause the reading of millions of data rows and will take a considerable amount of time. Even worse is that the result is not very useful, because the user wants to check only few of these invoices.

If the query is changed to “give me all invoices from this company code, with this supplier code, and from this month, which have not been paid yet,” the RDBMS has the possibility to look for fast access paths provided by indexes and may deliver the result after a few I/O requests in a considerably shorter time.

It is evident that it is impossible to make the execution of one million I/O request as fast as the execution of one hundred I/O requests by changing the number of disks or the layout on the disk, but with an optimized disk layout, we may improve the response time from one second to a tenth of a second. This is the improvement that is achievable by a performance oriented disk layout.

In Section 4.5.4, “Example for a disk layout” on page 113, we have seen the frequently accessed tables. It is not useful to put every heavily accessed table in a dedicated tablespace. This will lead to a large number of tablespaces widely scattered across the disks. As a result, the administration gets more complicated. SAP recommends you extract tables with a large growth rate and a high access rate from their default tablespaces and to put them together in a tablespace called PSAPCUSTOM. Then this tablespace is spread over as many disks as possible to balance the I/O activity. Otherwise, the concentration of these tables would lead to hot spots on the disks and would require a restructuring.

Bright idea!

There are some practical guidelines for the administration of tablespaces:

- ▶ Put the data and the index tablespaces on separate disk areas.
- ▶ The containers/data files should not exceed the size of 2 GB, in order to offer the possibility of parallel access in backup and restore operations. There will also be fewer problems with third party applications, such as backup tools.
- ▶ The tablespaces should be filled to a maximum of 70 percent in order to decrease the administration overhead that is caused by checking the free space of the tablespaces continuously. This also avoids fragmentation and offers a better performance.
- ▶ Extending the database is only possible by allocating a complete stripe set in a new disjointed disk area. This can be realized only by extra hardware or an unused rank in an ESS.

Special considerations for DB2:

Bright idea!

- ▶ The tablespace containers should all be equal in size. DB2 puts the next extent of a table in a round robin fashion in the next container, so the containers should have equal sizes to avoid hot spots on the larger ones.
- ▶ Refer to the white paper *Database Layout for SAP Installations with DB2 UDB for Unix and Windows*, found at: <http://service.sap.com>, for a more general approach. In this white paper, topics such as maximum tablespace sizes in DB2 are also covered.

Special considerations for Oracle:

Bright idea!

- ▶ The data files should be all equal in size. Oracle puts the next extent of a table in a round robin fashion in the next data file, so the data files should have equal sizes to avoid hot spots on the larger ones.
- ▶ The table parameters INITIAL EXTENT and MAX EXTENT should be the same size.
- ▶ The size of INITIAL and MAX EXTENT should be considerably large. For example, a value between 4 MB and 128 MB is recommended by SAP.
- ▶ Refer to the white paper *Database Layout for R/3, Installations under ORACLE, 50038473*, found at <http://service.sap.com>, for more details to the approach with the tablespace PSAPCUSTOM.

4.6 Logical Volume Manager considerations

The Logical Volume Manager (LVM) is a core component of AIX for the administration of file systems and logical volumes. LVM offers a lot of features for making administration tasks easier. The following two sections take a deeper look into some of these features.

4.6.1 Journaled file systems versus raw logical volumes

Both databases that are covered in this book offer the possibility to use journaled file systems and/or raw logical volumes for their data. In AIX, there is the possibility to use raw logical volumes, which are controlled by the Logical Volume Manager. Any direct access to raw disks is not supported in AIX. Please refer to Figure 4-11 on page 122 for an overview.

Until now, we have not discussed whether the implementation is based on the AIX Journaled File System (JFS) or on raw logical volumes (raw LV). This question was not an issue in the preceding sections.

There has been an ongoing debate on the use of raw logical volumes (raw devices) versus journaled file systems, especially in database environments. Advocates of raw logical volumes stress the performance gains that can be realized through their use, while JFS supporters emphasize the ease of use and manageability features of file systems. As with many other aspects of system design, there is a trade off between performance and manageability.

In order to better understand the performance advantages associated with raw logical volumes, it is helpful to have an understanding of the impact of the journaled file system cache. Most UNIX file systems set aside an area of memory to hold recently accessed file data, thereby allowing a physical I/O request to be

satisfied from memory instead of from disk. In AIX, this area of memory is known as the file system buffer cache. If an application requests data that is not already in memory, AIX will read the data from disk into the buffer cache so that it can be used by the application. In addition, AIX journaled file systems deploy a sequential read-ahead mechanism to pre-fetch data into memory when it determines that a file is being accessed sequentially.

In non-database environments, the AIX buffer cache can significantly reduce I/O wait time for heavily accessed files. However, the performance benefits of file system caching in database environments are not so clear. This is due to the fact that most modern Relational Database Management Systems (RDBMS) also allocate a region of memory for caching frequently accessed data. The end result, when using journaled file systems, is that the data is buffered twice: once in the file system buffer cache and once in the RDBMS cache. In most cases, the extra memory used by the file system buffer cache could be used more efficiently by the database buffers.

This buffering of the file system cache can be disabled in AIX. The file system buffer can be decreased, for example, to a minimum of three percent of real memory and a maximum of five percent of real memory by issuing the following command:

```
/usr/samples/kernel/vmtune -p 3 -P 5
```

The **vmtune** command is delivered in the fileset bos.adt.samples. For details, see Section 11.5.2, “The vmtune command” on page 347.

The memory saved by reducing the file system cache can then be allocated to the database to increase the size of the data buffers. Refer to Chapter 11, “Performance” on page 305 for more information. Figure 4-11 on page 122 shows the I/O processing in AIX.

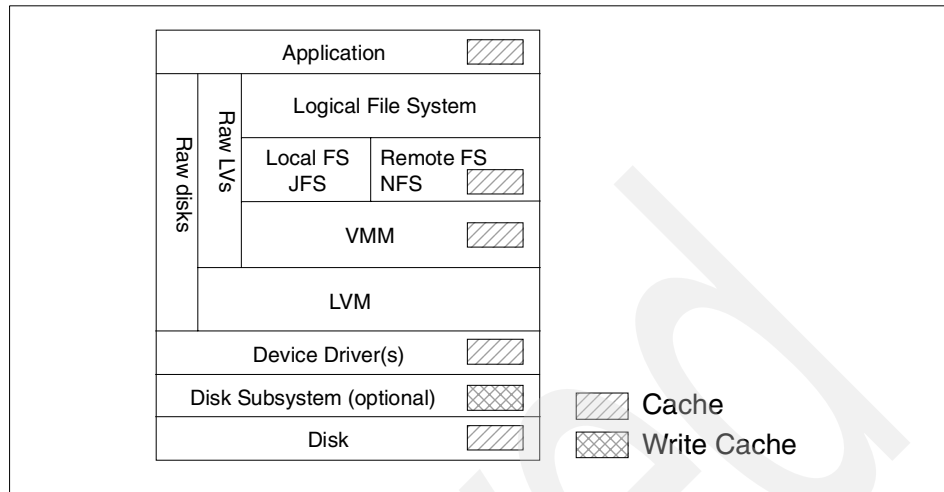


Figure 4-11 Overview of I/O processing in AIX

Raw LVs bypass the AIX file system buffer cache entirely by directly accessing the underlying device. The benefit of raw logical volumes is that there is no inode management overhead, as opposed to JFS, where the inode is locked when the file is accessed. The main drawback of using raw logical volumes lies in the increased administration efforts associated with their use.

Since raw logical volumes do not exist as files at the UNIX level, many of the traditional tools and utilities for managing data (for example, `cp`, `mv`, or `tar`) will not work. Backup and recovery operations can be especially difficult when using raw logical volumes. Many third party vendor backup applications cannot directly read raw logical volumes and must rely on the UNIX `dd` command to copy the raw data to a UNIX file system prior to backing up the data. For the daily backup, self-written scripts have to be created and tested.

Pitfall ahead!

Restores are more complicated, as the data must first be restored to a UNIX file system and then copied to the raw logical volume. If this approach is used, additional disk space will be required for the journaled file systems used to temporarily hold the data contained in the raw logical volume.

Pitfall ahead!

If the raw logical volumes can be backed up directly with the TSM client (since Version 3.7) or directly to a locally attached tape drive using the `dd` command, this will not be an issue. However, partial restores (like with single database files) for media recovery are not possible. You must restore a complete raw logical volume, which could take much longer and increases the downtime of the system

in case of a failure. Refer to “Unplanned downtime (tout_unplanned)” on page 14 for the calculation of downtimes. The administrator’s knowledge about dealing with files in a standard file system is much higher than the knowledge of raw logical volumes in case of a failure, and should not be neglected.

Many raw logical volume benchmarks point to an overall disk I/O throughput gain of 0-20 percent when compared to journaled file systems in a random workload scenario. However, the actual performance gains that can be realized in a typical database environment will vary depending on the I/O workload mix of the application. Applications that perform a large amount of random I/O operations, such as Online Transaction Processing (OLTP) systems, benefit the most from the use of raw logical volumes. Applications that perform a large amount of sequential I/O operations, such as Decision Support Systems (DSS), benefit from the sequential read ahead feature of the journaled file systems. The same observation can be made for (sequential) background processing and for (random) online workload in an SAP R/3 system.

The difference in performance results from the different locking implementations when accessing disk blocks. RDBMS vendors recommend raw devices for faster disk I/O.

Bright idea!

For a easily manageable system, only the journaled file system approach is recommended. The easier administration, the faster backup and recovery, and the usage of tools for SAP R/3 compensate the possible performance constraints due to inode lock contention. Our conclusion is that a potentially increased database performance of 0-20 percent in online processing does not justify the daily overhead in administration of raw LVs.

4.6.2 Mirror Write Consistency and databases

For mirrored logical volumes, AIX provides the option of switching Mirror Write Consistency (MWC) on or off. The default of setting it on is an often questioned issue. When we turn it on, AIX reserves some space on the outer edge of every disk for the consistency cache record, where the consistency information is written. Upon a single write operation, a flag is written first in this consistency cache and then the data is written into the data area of a disk. After a successful write of the data to both disks, the consistency information is updated in an asynchronous way. In normal operation, we get a performance penalty due to this behavior because the head of the disk has to move to the cache area first and then moves to the data area to write the data record. The cache information is stored for every 128 KB record of data, so a sequential write of 100 KB needs only one access to the mirror write consistency cache.

In case of a system crash, the two mirrored data areas may have different data stored on the disks. Because of the sudden crash of the machine, the operating system cannot decide, after the reboot, if one disk contains good or bad data and which one is the right one, so the copy to read from is randomly chosen. The mirror write consistency cache helps to find the inconsistent data areas fast without running a full synchronization of the whole volume group. Then the data of both areas are synchronized to contain the same information.

When the database system is started, it performs a crash recovery and compares the redo logs and the checksums stored there with the checksums of the data stored in the data files. In case of a mismatch, the data is rewritten to the data file and everything is consistent.

Pitfall ahead!

You could tune up your database write access times by switching the MWC feature off, but you may lose the integrity for your data.

The reason is that if the MWC is turned off, nobody cares about the synchronization of your data. After a crash, the database system runs a crash recovery as well and compares the redo log data with the data file. However, the database system may read the good copy of the data and marks everything as fine. The bad copy is not checked at all. Two reads of the same logical data block may lead to two different results, because they access two different physical data blocks. This inconsistency or corruption of your database may remain undetected.

In a high performance environment, you may switch off the Mirror Write Consistency, but you have to take care of the consistency on your own. This can be managed by a complete and forced synchronization of the logical volumes after a system failure or by removing the copies of the logical volume before mounting the file systems. In case of a high available environment, this is not advisable at all, so the automatic handling of consistent logical volumes is highly recommended.

Bright idea!

Our conclusion for a *reliable* environment is to switch the MWC on!

IBMers can find details about this in the IBM intranet in the white paper *AIX Mirror Write Consistency with Oracle Databases*, by Walter Orb, found at:

<http://dscrs6k.aix.dfw.ibm.com>

The effects are almost the same for DB2 databases.

4.7 General requirements

The software fundament for every system is the operating system. In Section 3.2.6, “Administration server (Admin)” on page 73, we already discussed the usage of NIM to have identical operating system levels on all machines for less administrative work. Other aspects for easier administration are considered in this chapter.

4.7.1 Operating system requirements

The following recommendations apply to the operating system disks that constitute the rootvg volume group:

Bright idea!

- ▶ The rootvg comprises the operating system related files, the default logical volumes, and the paging space only. This keeps the backup of AIX (mksysb) small and makes the move to another hardware or operating system easier.
- ▶ The sizing of the rootvg should provide comfortable space for the /, /tmp, and /var file systems, as these file systems may have a serious impact on an application if they are full.

The daily or hourly effort of monitoring these file systems can be reduced drastically by extending the file systems to reasonable sizes. For example, a full /tmp can also cause pending print jobs, backup failures, or a failing **installp** command. As a result, overhasty actions to correct the failure may imply other mistakes. Assuming a price of \$5000 for a 36 GB disk; the cost for a 100 MB /tmp directory is about \$15. This assumes less than five minutes of unplanned downtime, which can be caused by a full /tmp directory or an accidental mistake when removing temporary files.

Bright idea!

- ▶ For a reliable environment, the rootvg has to be mirrored in a manner similar to application data volume groups. The mirroring should include the paging space; otherwise, a failing rootvg disk, including paging space, will cause a system crash.

If we assume two existing disks (hdisk0/hdisk1) in the rootvg and two added disks (hdisk2/hdisk3), the mirroring of a rootvg is managed by these steps:

1. Place all the volumes, including the paging spaces, on the first two disks (hdisk0/hdisk1) according to your requirements.
2. Create the logical volume copies with an exact identical mapping using the command:

```
/usr/sbin/mirrorvg -m rootvg hdisk2 hdisk3
```

3. Find out the disks on which the boot logical volume (blv=hd5) copies reside with the command:

```
/usr/sbin/lslv -l hd5
```

We assume that `hdisk0/hdisk2` is the result.

4. Check the contiguity of the `hd5` volume. The `hd5` volume must reside within the first 4 GB of the disk. You can check this by issuing the command:

```
/usr/sbin/lslv -m hd5
```

The PP number multiplied with the PP size must be below 4 GB.

5. Write the boot record and the boot logical volume on the new disk using the command:

```
/usr/bin/bosboot -a -d /dev/hdisk2
```

6. Add the new disk to the bootlist using the command:

```
/usr/bin/bootlist -m normal -o hdisk0 hdisk2
```

7. Reboot the system to activate the changed quorum settings.

4.7.2 Application requirements

Bright idea!

These recommendations apply to application volume groups:

- ▶ Additional volume groups should be defined for the other disks attached to the system. The maximum number of disks in a volume group should be 16 for performance and handling reasons. This fits well to the size of an SSA drawer 7133. Furthermore, using the **chvg -t** command, it is possible to increase the number of partitions supported per disk if a higher capacity disk is being used in a volume group.
- ▶ Do not use 32 disks in a single volume group. You cannot migrate the data of a single defective disk to another new one, as you are not able to extend the volume group beyond 32 disks.
- ▶ Do not use the big volume group feature, which allows more than 32 disks in a volume group. This will decrease performance significantly when handling file systems or logical volumes. With 64 disks, the creation of a logical volume can take hours.
- ▶ It is very useful to keep one spare disk somewhere connected and available to the system. In case of a disk failure, you may migrate the contents of the failing disk temporarily to the spare disk. After the exchange of the failing disk, you can migrate back and no further action is required. You can migrate the data manually (**extendvg/migratepv**) or you can use the AIX command **replacepv** to assist you with these tasks. Starting with AIX 5L, a hot spare disk can be defined in a volume group.
- ▶ Logical volumes should have a meaningful name, for example, `lvsapmntPRD` for the `/sapmnt/PRD` file system, because you might migrate a volume group

from a different system and the logical volume names have to be unique. Example 4-3 shows a possible nomenclature.

Example 4-3 Logical volume names and correlating file systems

df -k					
Filesystem	1024-blocks	Used	Free	%Used	Mounted on
/dev/hd4	196608	32648	163960	17%	/
/dev/hd2	983040	929088	53952	95%	/usr
/dev/hd9var	524288	127744	396544	25%	/var
/dev/hd3	262144	38612	223532	15%	/tmp
/dev/hd1	131072	65508	65564	50%	/home
/dev/lvusrsapPRD	1048576	630228	418348	61%	/usr/sap/PRD/DVEBMGS00
/dev/lvextdataPRD	15728640	10067048	5661592	65%	/global/extdataPRD
/dev/lvhomePRD	262144	9616	252528	4%	/global/homePRD
/dev/lvsapmntPRD	1703936	1164584	539352	69%	/global/sapmntPRD
/dev/lvsapspoolPRD	10485760	992996	9492764	10%	/global/sapspoolPRD
/dev/lvsaptrans	5242880	4388264	854616	84%	/global/saptrans
/dev/lvoraclePRD	13107200	10275408	2831792	79%	/oracle/PRDDB
/dev/lvdataPRD.01	104005632	82078404	21927228	79%	/oracle/PRDDB/sapdata1
/dev/lvdataPRD.02	156172288	108095148	48077140	70%	/oracle/PRDDB/sapdata2
/dev/lvdataPRD.03	156172288	124835432	31336856	80%	/oracle/PRDDB/sapdata3
/dev/lvdataPRD.04	156172288	105202884	50969404	68%	/oracle/PRDDB/sapdata4

- ▶ Logs of the journaled file systems (jfslog) should be renamed to a meaningful and unique name (for example, lvjfslogSID) and should be mirrored. It is advised that you define the mirror of the jfslog to a different disk in a nonsymmetric way. For example, if disk A is mirrored to disk A' and disk B is mirrored to B', we would place the jfslog on A and the mirror on B'. If there are many accesses to the volumes and to the jfslog, and if the jfslog is put on A and A', it is likely that A and A' fail in the same time due to the identical workload. This effect can be avoided through nonsymmetric mirroring.
- ▶ If the jfslog puts a very high workload on the disk, it is recommended that you put the jfslog on a separate disk to improve performance.
- ▶ The general rule of thumb for the jfslog size is to have 4 MB of jfslog for each 2 GB of file system space. This is often overlooked, but if you conform to this recommendation, it keeps a comfortable safety clearance.
- ▶ Write and run a script to create the logical volumes and the journaled file systems. Scripting makes you think and plan ahead and makes sure everything is done in the right order.
- ▶ If possible, predefine logical volumes at their maximum expected sizes. This maximizes the probability that performance critical logical volumes will be contiguous and in the desired location.

4.8 Basic requirements

In the following sections, we discuss the storage layouts that meet the requirements for a basic, for a fault-tolerant, and for a disaster-tolerant configuration. First, we give an example for an SSA configuration and then for an ESS configuration.

4.8.1 A basic SSA configuration

For a typical server, the recommendations given in Figure 4-12 can be made for a valid SSA configuration.

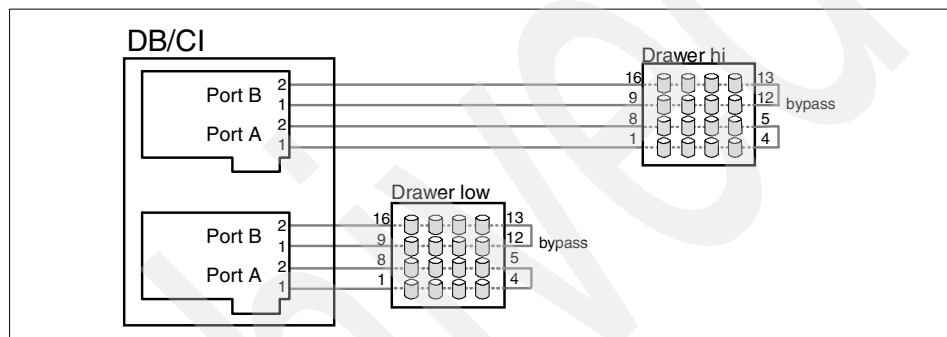


Figure 4-12 Basic SSA configuration

If you want to provide a reliable server environment, try to avoid the occurrences of single points of failure (SPOF). A single adapter, a single loop, a single drawer, a single disk, and a single power-supply are examples of single points of failure.

In Figure 4-12, we show two adapters, four loops, and two independent drawers with doubled power-supplies. The data is striped on the disks (RAID 0) of one drawer and is mirrored (RAID 1) to the other drawer, so that we have a RAID 10 configuration. A lot of components may fail in this configuration without resulting in loss of access to the server's data on the disks.

A RAID 5 configuration, based on a native SSA concept without large caches (like in the ESS), has performance limitations compared to a RAID 1 configuration, as discussed in Section 4.1, "Basic understanding of disk mechanics" on page 97. Additionally, using only one drawer is a potential single point of failure in a RAID 5 configuration.

Pitfall ahead!

If the SSA cabling is done according to Figure 4-12 on page 128, you make use of the bypass card, which is an integrated component in newer SSA Drawers (7133-020/600 and 7133-D40/T40). Keep in mind that an integration of an additional server in a loop in this cabling scheme is not possible during run time. The cabling has to be changed according to Figure 4-14 on page 131 to avoid an invalid loop configuration in case of a shutdown of one server.

A special treatment is required for the quorum of the application volume groups. If one of the two drawers fails completely, then the volume group loses its majority and the volume group with the file systems will be closed. This behavior can only be deactivated by disabling the quorum checking. In this case, the volume group stays online even if half of the disks or less is available. The AIX command for switching off the quorum for the volume group `appl_vg` is:

```
chvg -Qn appl_vg
```

Pitfall ahead!

A problem can arise if one disk fails and remains undetected. In case of a varyoff of the volume group or a reboot, the volume group cannot be varied on again. The `varyon` has to be executed with the force flag (`varyonvg -f appl_vg`) to remove the missing disk out of the volume group definition and to vary on the volume group again.

Bright idea!

To prevent this manual handling effort, the disks have to be monitored to recognize a disk failure and to exchange the disk before the next reboot cycle.

The manual effort can be omitted by introducing a quorum buster into the loops. Refer to Section 4.10, "Disaster-tolerance requirements" on page 133 for more information on this topic.

4.8.2 A basic ESS configuration

Figure 4-13 on page 130 shows the cabling in an ESS configuration. In this case, the quorum checking is not an issue, because we do not mirror the logical volumes. The reliability is completely based on RAID 5 and other availability features of the subsystem. If any of the logical disks in the ESS is not available, the application cannot deliver its service anymore.

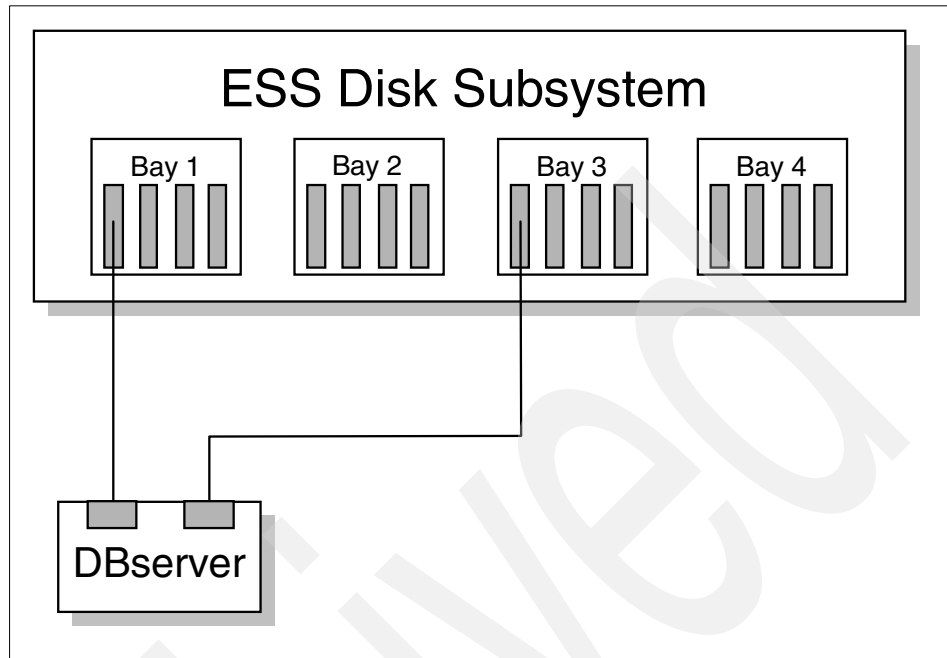


Figure 4-13 Basic ESS cabling configuration

Pitfall ahead!

Whereas a single point of failure is eliminated inside the ESS by the design, the connected host must have at least two connections to two different bays to exclude a single point of failure in the disk storage access path.

The reason for attaching the DBserver to Bay 1 and Bay 3 is that Bay 1 and 2 are internally closer attached to cluster node 1 of the ESS, and Bay 3 and 4 are closer related to cluster node 2 of the ESS. So if there is an internal failure of one of the cluster nodes and an internal fail over occurs, the external access to only one of the bays might be affected.

Bright idea!

During a concurrent microcode update on the ESS, every bay is restarted one after the other, so a connected host must be attached to two different bays to maintain accessibility to the disks. In order to make the multiple paths transparent to the operating system and the application, a multi-path I/O capable software must be used. In AIX, the Subsystem Device Driver (SDD) provides the ability to access the disk storage over two or more paths and thus eliminates outages due to a failing link.

For more information on the SDD refer to the document *IBM Subsystem Device Driver / Data Path Optimizer on an ESS*, by Jesse I. Adams III, found at:

<ftp://ftp.software.ibm.com/storage/subsystem/tools>

4.9 Fault-tolerance requirements

A fault-tolerant configuration requires further elimination of single points of failure. The previous configuration was based on a single server and this single server represents the single point of failure. The new configuration takes two servers into account. In this case, the server itself may fail and the service that the server provided can be taken over by the other machine.

4.9.1 A fault-tolerant SSA configuration

By introducing a second server, the cabling for SSA has to be changed. A failing adapter or host leads to an open loop, but not to an invalid configuration of the loop. If the A and B port of one adapter would be offline and the configuration would be as in Figure 4-12 on page 128, the bypass card would establish a connection between connector 1 and 16 of the SSA drawer, which would lead to a shortcut between the A- and the B-Loop. This automatic reconfiguration would produce an invalid configuration. There is no way to configure two adapters (hosts) with two loops and one drawer so that both loops remain intact when one adapter is removed.

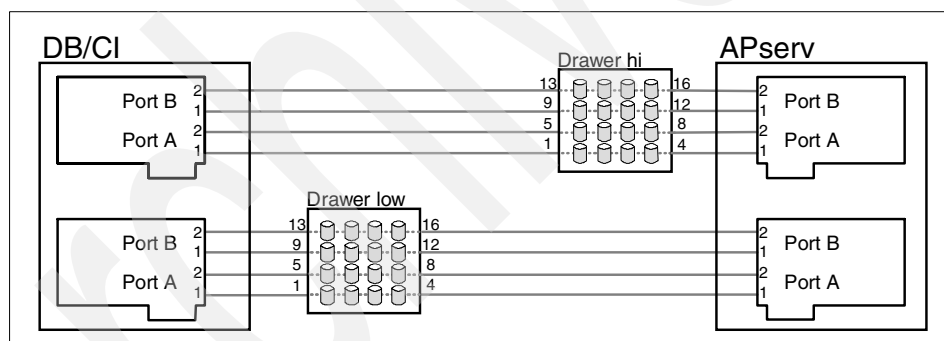


Figure 4-14 Fault-tolerant SSA configuration

This configuration allows both servers to access all disks. Even if one server fails, the other has full access to the disks. Based on this configuration, a manual switch over, to provide a fault-tolerant service, can be implemented.

The software implementation of the cluster software, which is necessary to provide such an automatic fail over in a fault-tolerant configuration, is discussed in Chapter 8, "High availability" on page 221.

The same special quorum treatment has to be taken into account, as discussed in Section 4.8.1, "A basic SSA configuration" on page 128.

4.9.2 A fault-tolerant ESS configuration

With an ESS, we have to connect the second server (APserv) as well as the first one (DB/CI) to the ESS disk subsystem. Both servers connect to the ESS with two data links in minimum, and the volumes have to be assigned to both servers in the internal ESS configuration.

The connection between the host systems and the ESS can be a copper cable for SCSI or a fiber optic cable for Fibre Channel. Figure 4-15 gives an overview of this configuration.

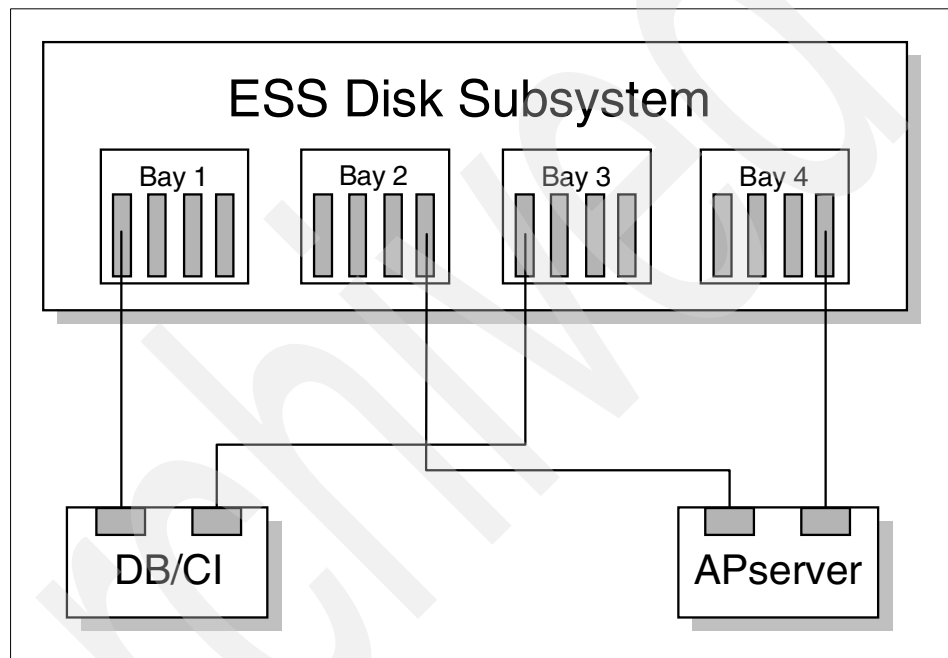


Figure 4-15 Two hosts connected with four access paths to the ESS

In case of failure of one host, the other host has full access to the disks and is able to take over the applications that were formerly running on the other server in order to provide its service. A manual switch over, to provide fault-tolerant service, can be implemented.

The software implementation of the cluster software, which is necessary to provide an automatic fail over in a fault-tolerant configuration, will be discussed in Chapter 8, "High availability" on page 221.

4.10 Disaster-tolerance requirements

A disaster-tolerant solution may be required for business critical data. In this case, two computing centers in different fire cells have to be provided, into which the available hardware can be distributed. The following sections show examples for such disaster-tolerant configurations.

4.10.1 A disaster-tolerant SSA configuration

Basically the layout stays the same as in Section 4.9.1, “A fault-tolerant SSA configuration” on page 131. The difference comes through the introduction of SSA Optical Extenders to increase the distance between the components for better disaster protection.

Through the separation of the servers and disks, a loss of 50 percent of the disks may occur because of a blown fuse. Therefore, we have to change our quorum checking to off. Otherwise, the volume group closes and the application fails (if 50 percent of the disks are not available anymore). As previously mentioned, manual intervention is required to force the volume group active again after a close (if one disk is missing).

In disaster-tolerant environments, it is unacceptable that a blown fuse requires manual handling for a switch over. The technical solution for the problem of losing the majority of disks is the introduction of a quorum buster. A quorum buster consists of a disk that is placed in a third location. If one computing center fails because of a disaster, the majority of disks (50 percent plus one disk) are still available.

For a quorum buster, one disk per volume group is needed. This disk is included in the volume group in a third location. Absolutely no data is allowed on this quorum disk.

Tip: Create a logical volume with type quorum on the quorum buster disk, so nobody can use the space unintentionally for a logical volume on this disk or by creating a journaled file system in it.

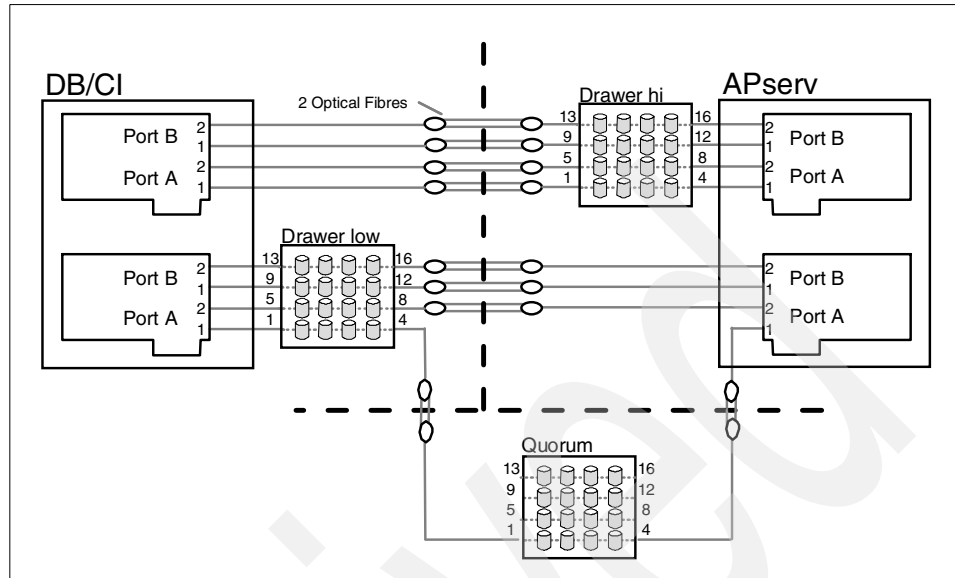


Figure 4-16 Disaster-tolerant SSA configuration

For a disaster-tolerant solution, a cluster software that executes all the necessary actions for a switch over is absolutely required. Refer to Chapter 8, “High availability” on page 221 for more details.

4.10.2 A disaster-tolerant ESS configuration

The ESS is a highly available disk subsystem and offers redundancy for every component. The ESS itself represents a single point of failure in the case of a disaster. Therefore, the ESS has to be duplicated as well to a second location if a disaster solution is required. The connected servers need two connections to both ESS systems to satisfy the requirements of eliminating single points of failure.

For larger distances, the use of fiber optic cables is necessary, as shown in Figure 4-17 on page 135.

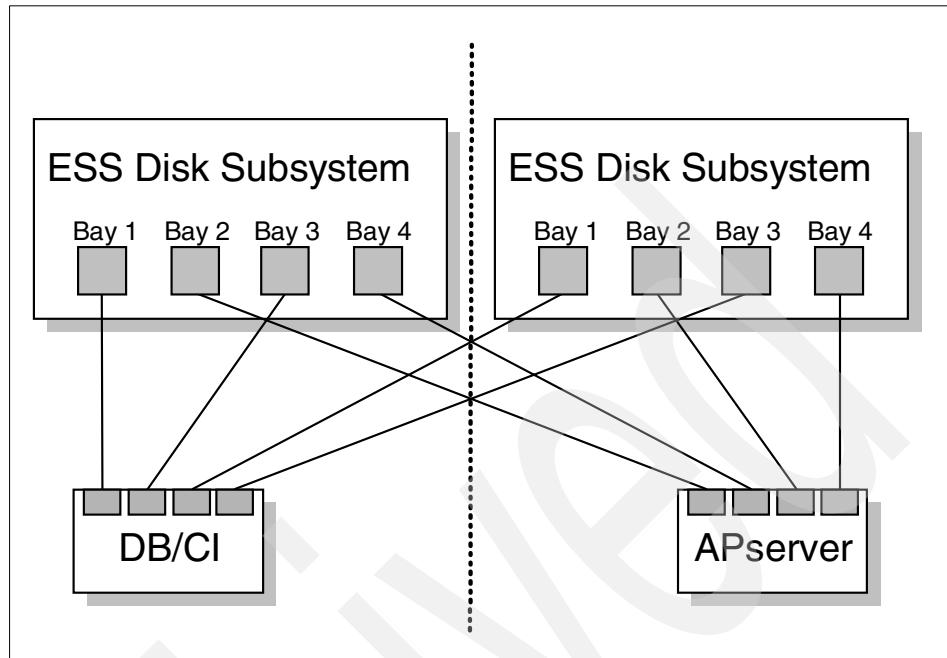


Figure 4-17 Disaster-tolerant disk subsystem connection

Bright idea!

From the operating system's point of view, we create mirrors of the logical volumes on the second ESS. In this mirrored configuration, we have to consider the constraints of the quorum checking of the volume group, as discussed in Section 4.8.1, "A basic SSA configuration" on page 128.

Within the ESS, we use RAID 5 as a performance optimized and safe implementation of data striping. On top of that, we mirror on Logical Volume Manager level. Besides the disaster safety, we achieve a higher throughput with load balancing between the two ESSs.

In the previous SSA section, we discussed the constraints with quorum checking. With no quorum buster, we have to switch the quorum off. In contrast to a standard SSA configuration, a single missing disk cannot fail and interfere with the quorum checking, because we have RAID 5 secured volumes. Either the ESS is up and running and 100 percent of the volumes are available or it is down and 0 percent disks are available. But it is not possible that 75 percent of the total number of volumes are available in a mirrored ESS environment.

In order to increase the availability (in case of disaster) and to manage a switch over between the systems in the different computing centers automatically, the forced activation of the volume groups has to be switched on.

See “Configuration with no quorum buster” on page 255 for more details.

4.10.3 A disaster-tolerant ESS configuration based on SAN switches

The number of connections required for availability reasons is very high. Even if the throughput of four Fibre Channel connections is not needed, they are necessary for a disaster-tolerant environment. In order to decrease the number of needed adapters, Fibre Channel links, and connections, we introduce SAN components into the configuration.

In Chapter 5, “Storage Area Networks” on page 137, the components of a SAN based connectivity are explained in detail. Refer to this chapter for more information. The two switches in Figure 4-18 build the core of the SAN implementation and are connected to each other. Every link in the SAN exists twice, so if one link fails, there is always another link that the servers can rely on.

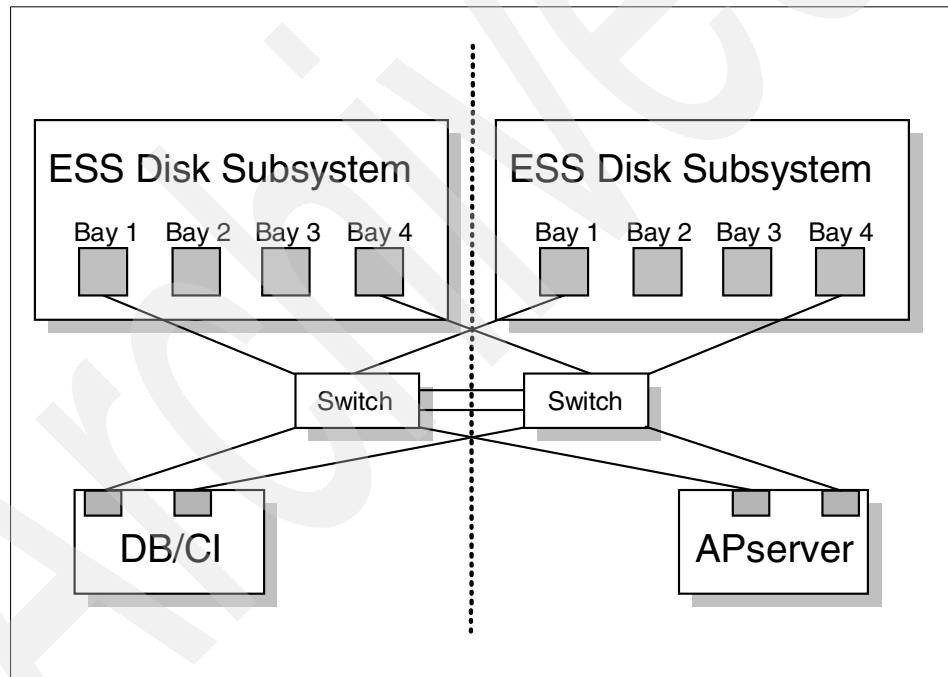


Figure 4-18 ESS in a SAN configuration

In the following chapters, we discuss the network layouts and the requirements of backup and recovery. With that knowledge, we discuss the implementation of fault- and disaster-tolerant configurations in Chapter 8, “High availability” on page 221.

Storage Area Networks

In a nutshell:

- ▶ Consider reliability and availability for the design of your SAN.
- ▶ Use products from one manufacturer to avoid interoperability problems.
- ▶ Use single-mode fibers if distances between components are nearly 500 m.
- ▶ Use multi path I/O for all storage devices when possible.

In this chapter, we give a short introduction to Storage Area Networks (SAN), explain where they are used, and describe the benefits of this new paradigm.

We describe the underlying physical technology and show the protocols that can be used within a SAN. We give an overview of the most important products that are available for building a Storage Area Network and describe their features and advantages.

We show an example for building a Storage Area Network and give practical hints and tips for the design and the implementation of the components in disk and tape environments.

This chapter covers the highlighted area in Figure 5-1 on page 138, which is derived from the SAP R/3 infrastructure model that is used throughout the book.

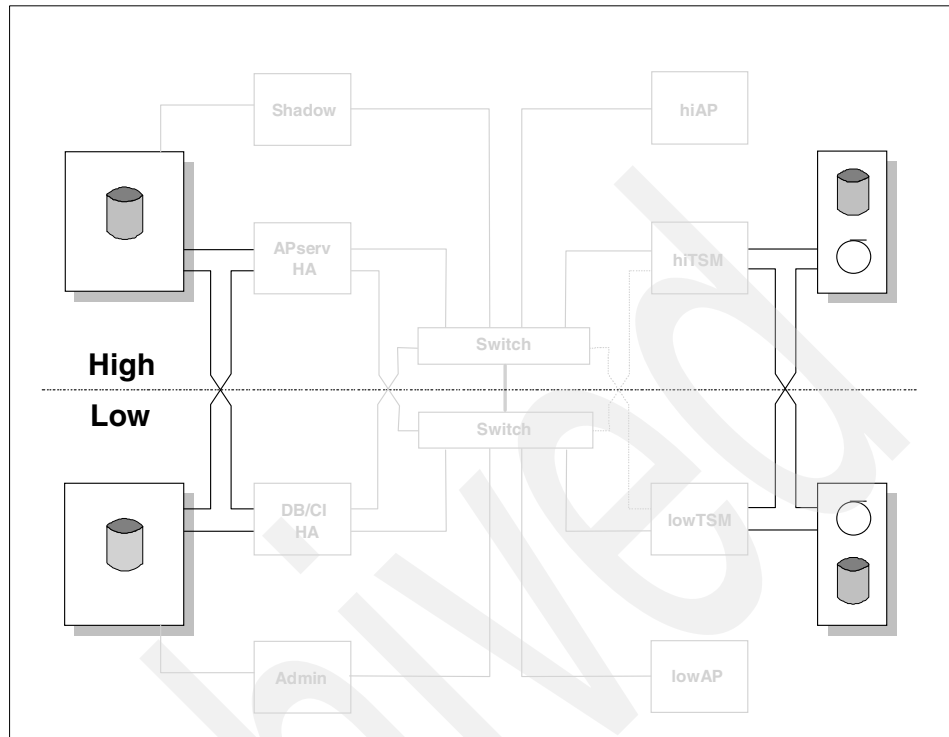


Figure 5-1 The Storage Area Network

5.1 Introduction to Storage Area Networks (SAN)

Companies today are using electronic data processing for almost every aspect of daily business. This has the effect of a dramatically increased data volume in recent years. The move from mainframe based centralized systems to decentralized client/server computing in a downsizing process has led to an environment of highly scattered data storage. The combination of these trends results in the requirement for access to large storage spaces from different locations and different platforms.

There are two different options to manage these requirements. One is the reintegration of the distributed client/server systems into data centers; the other is the consolidation of the storage space and making it centrally available on an as-needed basis.

In the second case, it is necessary to transport the data between the storage systems and the server systems. There are two different approaches:

- ▶ Transport the data via traditional communication networks, using file serving features such as the Network File System (NFS) and Server Message Block (SMB).
- ▶ Build a new specialized storage network architecture parallel to the existing communication network.

The first approach is called Network Attached Storage (NAS); the second approach is called Storage Area Network (SAN).

A SAN shows the following benefits:

- ▶ Facilitates universal access and sharing of resources
- ▶ Supports unpredictable, explosive information technology growth
- ▶ Provides affordable 24 x 365 availability
- ▶ Simplifies and centralizes resource management
- ▶ Improves information protection and disaster tolerance
- ▶ Enhances security and data integrity of new computing architectures

Fibre Channel (FC) is the underlying technology for SANs. It is a highly-reliable, gigabit interconnect technology that allows concurrent communications between servers and storage subsystems using different protocols.

Similar to standard communications networks, Fibre Channel networks use the same types of components, such as gateways for protocol conversions, hubs, and switches. Details on the architecture of the Fibre Channel protocol and technologies can be found in Section 5.2.1, “Fibre Channel protocol layers” on page 140. In the next part of this section, we show products that can be used as building blocks for the implementation of a SAN.

There are two Redbooks available that give an introduction and important criteria for the design of a Storage Area Network:

- ▶ *Introduction to Storage Area Network, SAN, SG24-5470*
- ▶ *Designing an IBM Storage Area Network, SG24-5758*

5.2 Storage Area Networks based on Fibre Channel

The technology for the connection of storage subsystems to servers has undergone revolutionary change in the past few years. As described in “Disk storage products” on page 47, there are three major technologies available: SCSI, SSA, and Fibre Channel. While SSA has already overcome many of the restrictions of SCSI, the technology has been leveraged with the introduction of Fibre Channel. In this section, we describe the basic principles of Fibre Channel technology and how Fibre Channel enables the building of Storage Area Networks.

5.2.1 Fibre Channel protocol layers

Fibre Channel is an industry standard that is based on serial data transfer via copper or optical links. It supports a bandwidth of up to 1 GB/s with a proposed increase of up to 4 GB/s. It supports the physical connection of devices in point-to-point or switched topologies.

The specifications shown in Figure 5-2 have been defined in a layered approach in order to facilitate the implementation of Fibre Channel based products and to ensure a framework for protocol extensions.

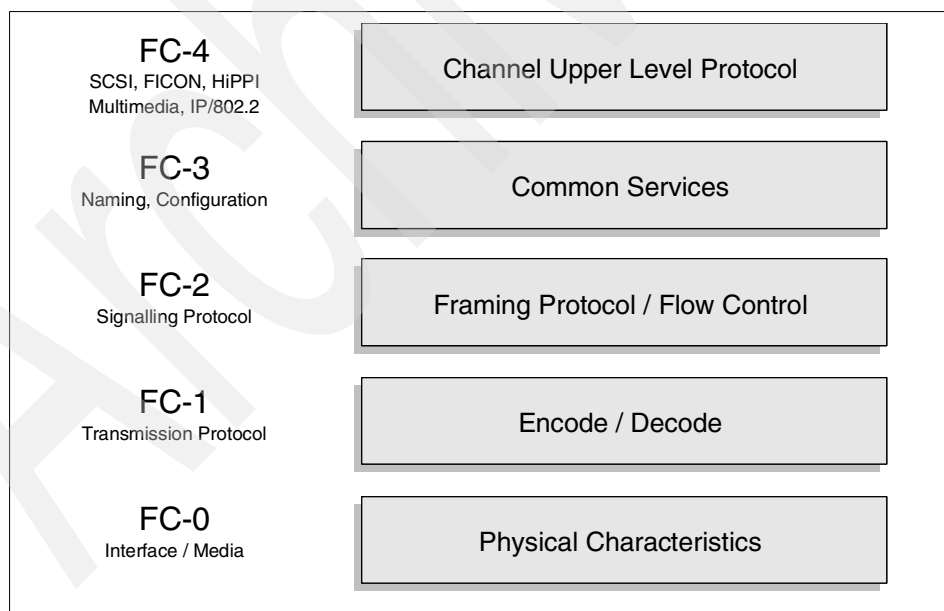


Figure 5-2 Layers of the Fibre Channel architecture

The layers have the following purpose:

FC-0	The lowest layer on Fibre Channel transport represents the physical media: copper, single-mode, or multi-mode fibres.
FC-1	This layer contains the 8b/10b encoding scheme.
FC-2	This layer handles framing and protocol, frame format, sequence/exchange management, and ordered set usage.
FC-3	This layer contains common services used for naming and configuring of multiple FC nodes.
FC-4	This layer handles standards and profiles for mapping upper-level protocols, such as SCSI and FICON, onto the Fibre Channel protocol.

This layered structure enables the implementation of different upper level protocols, for example, the High-performance Parallel Interface (HiPPI), an 800 Mbps interface normally used in supercomputer environments. Also, the Fibre Channel-Arbitrated Loop protocol has been implemented, which represents a logical loop in a point-to-point topology (for compatibility reasons) with the SCSI protocol with regard to arbitration in storage environments.

5.2.2 Fibre Channel technology

The layers FC-0, FC-1, and FC-2 are also described as FC-PH, the Fibre Channel physical and signaling standards layer, which indicates signaling used for cable plants, media types, and transmission speeds. The most important definitions of the physical layer are:

- ▶ Fibre optic cables
 - Multi-mode: Short-wave, 62.5 and 50 μm , up to 500 m
 - Single-mode: Long-wave, 9 μm , up to 10 km
- ▶ Host adapter connectors
 - Fibre optic cable duplex SC connector
 - Small Form Factor (SFF) duplex LC connector
- ▶ Network component connectors
 - Gigabit Interface Converters (GBIC)
 - SFF optical transceivers

The network component connectors convert copper interfaces to fibre optical interfaces.

5.3 SAN products

In this section, we describe SAN components that can be used as the building blocks of departmental or enterprise-wide SAN solutions. For the integration of SCSI based technologies into a Fibre Channel based SAN, there are gateways that convert the different physical layer protocols. The connection between all FC devices and host adapters is performed by fully interconnected switches. In data center environments, these switches are also called *Directors*, as they usually offer the option for Fiber Connection (FICON) or ATM connections.

5.3.1 IBM SAN Data Gateway

The IBM SAN Data Gateway, model 2108-G07, is a hardware solution that enables the attachment of SCSI storage systems to Fibre Channel adapters on specific Intel-based servers running Windows NT and UNIX-based servers from IBM.

The SAN Data Gateway provides two short-wave Fibre Channel ports and four Ultra SCSI Differential ports to attach disk or tape storage devices. This can be expanded to a maximum six short-wave ports, or two short-wave and two long-wave ports. It supports the following devices:

- ▶ IBM Enterprise Storage Server
- ▶ IBM Magstar 3590 Tape Subsystem in stand-alone, Magstar 3494 Tape Library, and Magstar 3590 Silo Compatible Tape Subsystem environments
- ▶ IBM Magstar MP 3570 Tape Subsystem or Magstar MP 3575 Tape Library Dataserver

The SAN Data Gateway is available as a rack-mounted unit or as a stand-alone tabletop unit.

5.3.2 IBM SAN Fibre Channel Switches

The IBM SAN Fibre Channel Switches, models 2109-S08 and S16, which are OEM products from the Brocade SilkWorm family, provides Fibre Channel connectivity to many different server and storage systems. It supports the following features:

- ▶ Enables connectivity to Fibre Channel-attached disk subsystems, such as the IBM Enterprise Storage Server and the IBM Modular Storage Server.
- ▶ Enables connectivity to Fibre Channel-attached tape subsystems, such as the IBM Magstar 3590 Fibre Channel drives and LTO Ultrium Fibre Channel drives.

- ▶ Supports the interconnection of up to 252 IBM SAN Fibre Channel Switches to provide enterprise-level scalability.
- ▶ Each port delivers up to 100 MB/s, full-duplex data transfer.
- ▶ Offers non-blocking eight-port (S08) and sixteen-port (S16) models with short or long wave fiber optic connections.

The IBM StorWatch SAN Fibre Channel Switch Specialist provides a comprehensive set of management tools for the switch. It supports a Web browser interface for flexible, easy-to-use integration into existing enterprise storage management structures. The Specialist provides security and data integrity by limiting host system attachment to specific storage systems and devices, also known as *zoning*.

5.3.3 McDATA ES-3016 and ES-3032 Switches

McDATA Fabric Switches are offered in a sixteen port model and a thirty-two port model, also known as IBM 2031-016 and 2031-032. Each model includes sixteen short-wave transceivers for device interconnection at a maximum distance of 500 meters. A mixture of short-wave and long-wave (20 km) ports can be configured by adding transceivers.

- ▶ Provides Fibre Channel connectivity for UNIX-based servers and Intel-based servers running Windows NT and 2000.
- ▶ Enables connectivity to Fibre Channel-attached disk storage and to McDATA ES-1000 Switch for Fibre Channel-attached tape storage.
- ▶ Utilizes IBM SAN Data Gateway for Ultra SCSI-attached IBM tape storage.
- ▶ Provides non-blocking fabric switching for scalable departmental SANs.
- ▶ Offers rack space saving with new Small Form Factor (SFF) LC transceivers.
- ▶ Provides high availability with redundant, hot-swappable fans, power supplies, optics, and concurrent firmware activation.
- ▶ Multiple management options include enterprise-to-edge SAN management.

Each port delivers up to 100 MB/s full-duplex data transfer. Full throughput at extended distances (up to 100 kilometers) is enabled with long wave optics.

McDATA Fabric Switches may be shipped as standalone units or they may be configured into a McDATA FC-512 Cabinet, an IBM 2101 or 7014 Rack, or an industry standard 19" rack.

5.3.4 McDATA ED-6064 Enterprise Fibre Channel Director

The McDATA Enterprise Fibre Channel Director, offered as IBM 2032-064, provides the scalability demanded by rapidly growing mission-critical applications. It provides enterprise-level scalability and data center-level availability and the following features:

- ▶ Fibre Channel connectivity for IBM S/390 and IBM @server zSeries, and UNIX-based servers and Intel-based servers running Windows NT
- ▶ Enables connectivity to IBM FICON-attached and Fibre Channel-attached storage and to IBM UltraSCSI-attached storage with gateways and routers
- ▶ Enables connectivity to McDATA ES-1000 Loop Switch for Fibre Channel-attached tape
- ▶ Offers 64-port Fibre Channel switch fabric with full redundancy for all active components
- ▶ Offers rack space saving with Small Form Factor (SFF) LC optical transceivers

Enterprise Fabric Connectivity (EFC) management provides a scalable, modular architecture consisting of the EFC Server PC, the EFC Management Software, and the EFC Product Manager application. EFC Management software centralizes the management of multiple, distributed Directors in an enterprise-wide Fibre Channel fabric.

For more information on products and SAN Fibre Channel implementation, refer to the following Redbooks:

- ▶ *Planning and Implementing an IBM SAN*, SG24-6116
- ▶ *IBM SAN Survival Guide*, SG24-6143

5.4 Building a Storage Area Network

In order to build a Storage Area Network, certain components that are already known from classical communication networks must be considered. The server systems must be equipped with host bus adapters, there must be hubs for port concentration and connection, switches for interconnecting servers, storage devices, and other switches.

5.4.1 SAN connections and services

There are naming conventions for ports in an SAN environment. Usually the following types can be distinguished:

- ▶ Ports that are connected to a node are called N_Ports.
- ▶ Ports for devices that operate in arbitrated loop mode are called L_Ports.
- ▶ Ports that are operating in switched fabric mode are called F_Ports.
- ▶ Ports for the interconnection and extension of switches are called E_Ports.
- ▶ Ports for general use without specifying their operation mode are called G_Ports.

These prefixes can also be combined for specifying the type of connection that is supported on a certain physical port or to describe which device is actually attached to a port, for example, NL_Port describes a node that is attached in an arbitrated loop mode.

Managing and operating a Storage Area Network requires services internal to the network, for example, a login service to enable hosts or devices to join the network. But there are also external services required, for example, for the management and monitoring of the devices in terms of systems management.

In summary, these are the most important services for a Fibre Channel network:

- ▶ Login server for managing the fabric login process
- ▶ Name server for providing directory service of N_Ports and NL_Ports
- ▶ Management server for providing Simple Network Management Protocol (SNMP) access to Management Information Base (MIB) information for each fabric element

The building blocks for a SAN are described in Section 5.3, “SAN products” on page 142. A typical SAN environment for an SAP R/3 environment is shown in Figure 5-3 on page 146, where switches have been used to connect a disk storage subsystem and a tape library to an SAP R/3 system, and two tape libraries to a storage management server.

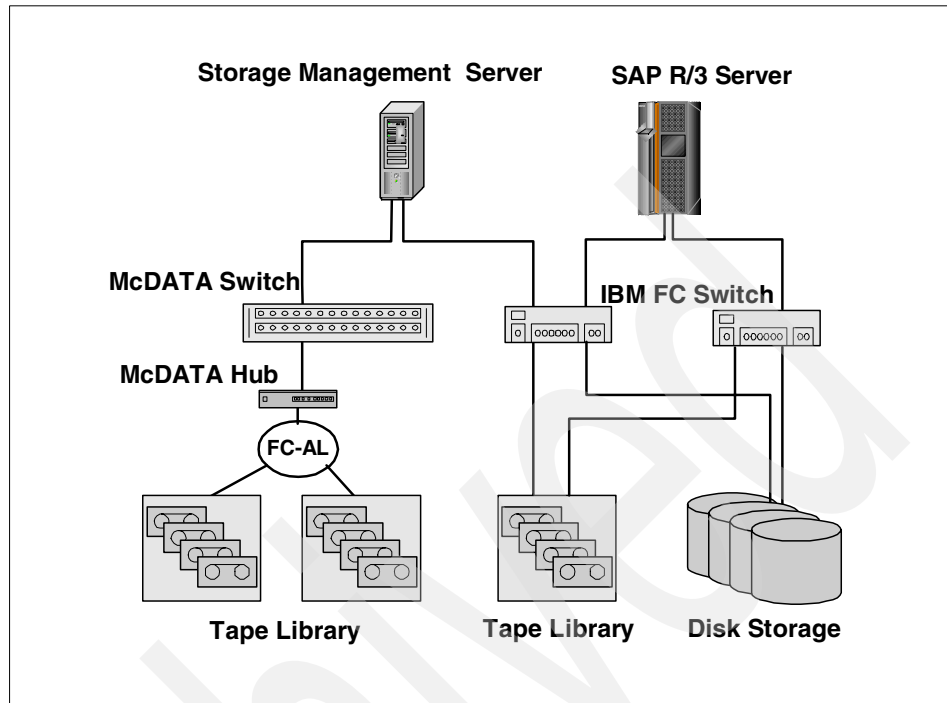


Figure 5-3 SAN scenario in an SAP R/3 environment

5.4.2 Design considerations for a SAN

Designing a Storage Area Network requires that you observe the same basic principles as in designing a traditional communication network. Criteria like multiple paths and the avoidance of single points of failure also apply for SANs. But there are also some differences that are due to physical limitations of the SAN components and due to software and hardware implementations. We give some practical hints and tips for the design considerations with disk and tape storage subsystems.

RS/6000 and IBM @server pSeries connectivity

There are two different Fibre Channel adapters available for RS/6000 and IBM @server pSeries models, a 32-bit version, FC #6227; and a 64-bit version, FC #6228, which requires a 64-bit PCI slot. The 32-bit adapter is a standard 1 Gb/s adapter with SC connectors. The 64-bit adapter supports transfer rates up to 2 Gb/s and is equipped with LC multi-mode fibre connectors. For connecting the 64-bit adapter to an SC cable, there is an LC-SC convertor cable, FC #2456.

Both Fibre Channel adapters only support multi-mode fibers. At a speed of 1 Gb/s, there are the following distance limitations:

- ▶ Multi mode 50/125 µm fiber: 2m - 500m
- ▶ Multi mode 62.5/125 µm fiber: 2m - 175m

The 64-bit adapter running in 2 Gb/s mode has the following limits:

- ▶ Multi mode 50/125 µm fiber: 2m - 300m
- ▶ Multi mode 62.5/125 µm fiber: 2m - 150m

A switch or hub has to be provided for distances in excess of 500 m.

Enterprise Storage Server connectivity

The Fibre Channel Host Bay Adapters of the ESS also provide multi-mode fiber connection only. Both multi-mode fiber types are supported, and the same distance restrictions mentioned in the previous section apply. A switch or hub has to be provided for distances in excess of 500 m.

The Fibre Channel ports of the ESS can be configured as a point-to-point, arbitrated loop or fabric attachment, with the constraint that the arbitrated loop does not support more than one host adapter.

The Enterprise Storage Server is designed with full redundancy for all components. Nevertheless, it may happen that a complete host bay must be placed in an acquiescent state for several seconds for internal recovery procedures. During this time, there is no access to the disks, which are usually recoverable by the operating systems retrying the I/O requests.

There are I/O requests in AIX, such as the writing of JFS-log information which have only a very small retry interval. In order to allow for such I/O operations to complete in time, there should always be a second path available to the disk. This second path should go through a different host bay, because if there is a recovery procedure inside the ESS the whole host bay is usually affected.

Attention: The IBM @server pSeries must be attached to different host bays of the Enterprise Storage Server with multiple paths, and the Subsystem Device Driver must be used.

In order to enable the second path from an operating system point of view, there is a need for software that enables multi-path I/O. In the case of AIX, this is done with the Subsystem Device Driver (SDD). This software implements two features. It enables retry mechanisms on the same path and the selection of a different path (if one has failed). In addition, SDD enables load balancing between multiple paths with different algorithms selectable by the administrator.

Tape library connectivity

The use of tape libraries in a SAN environment provides the freedom to locate the libraries and the media in a different location than the storage server or the disk storage components with the application data. It also offers the possibility to move application data directly from disk storage systems to tape storage systems without using the traditional communication networks. Another feature is the attachment of a tape drive to several servers for fault-tolerant or disaster-tolerant scenarios.

The tape libraries described in “Tape storage products” on page 52 are all available with Fibre Channel attachment. There is a fundamental difference between the supported Fibre Channel topologies of an ESS and the drives of the tape libraries. The tape drives only support the arbitrated loop protocol. This is a restriction if they are used in combination with certain switches. The IBM SAN Switch 2109 supports direct attachment of FC-AL devices and adapts the protocol. The McDATA switches and directors require the interposing of a special device, the McDATA ES-1000 Loop Switch. This device is essentially a hub, which also converts the FC-AL protocol for fabric integration and offers one uplink to a fabric switch.

Attention: If you want to attach Fibre Channel tape drives to a McDATA switch or director, you have to use a McDATA ES-1000 Loop Switch for protocol conversion.

As described in “Tape storage products” on page 52, there is an important difference between the attachment of an LTO Ultrium tape drive and a 3590 Magstar tape drive. The LTO Ultrium drives have only one port for the attachment; the Magstar drives offer two ports. With the additional recoverable path feature of the IBM AIX Enhanced Tape and Medium Changer Device Driver (Atape) for Magstar drives, it is possible to implement a highly available I/O path. More details on this feature can be found in the document *IBM Magstar Tape Drives -- AIX High Availability SAN Failover for 3590*, found at:

http://www.storage.ibm.com/hardsoft/tape/3590/prod_data/magstarwp.pdf

Descriptions on how to implement the features are included in the document *IBM SCSI Tape Drive, Medium Changer, and Library Device Drivers Installation and Planning Guide*, GC35-0154.

Network

In a nutshell:

- ▶ Use separated networks for performance and security.
- ▶ Use Gigabit Ethernet networks where high performance is required.
- ▶ Optimize parameters of network adapters and switch ports.
- ▶ Optimize the network options of AIX.

This chapter describes network layouts for an SAP R/3 environment using IBM @server pSeries.

The SAP R/3 environment is based on a three-tier client/server design. The clients and the servers in the three different tiers communicate over specific networks. The quality of the available networks highly influences the service that is provided to the users.

It is important that the network layout for an SAP R/3 environment fulfills a number of requirements, which are discussed in Chapter 2, “Requirements for a reliable SAP R/3 environment” on page 7. The requirements for the network layout are performance, reliability, availability, scalability, security, and manageability.

We discuss the required network performance for SAP R/3 systems and also present our recommendation for a network layout that fulfills the requirements for a reliable SAP R/3 system. We then present network technologies that can be used with IBM *@server* pSeries servers. At the end of this chapter, we discuss the implementation of networks regarding AIX and the active network components.

This chapter covers the highlighted area in Figure 6-1, which is derived from the SAP R/3 infrastructure model that is used throughout the book.

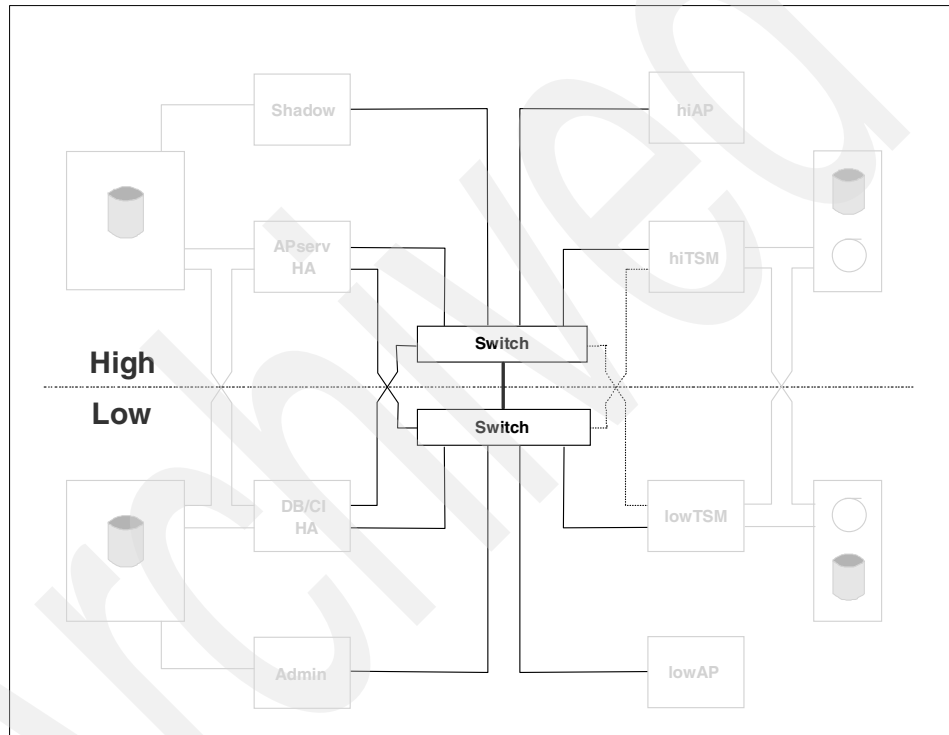


Figure 6-1 The network

6.1 Network requirements

It is important that the performance of the network infrastructure does not limit the operation of the SAP R/3 system. Therefore, the network between the presentation layer and the application layer, the network between the application layer and the database layer, and the network between the database layer and the backup server have to provide appropriate performance.

In this section, we show the requirements for the latency and the bandwidth of SAP R/3 Release 4.6 for these networks. We then review the requirements for the network while performing backup or recovery tasks within the database.

6.1.1 Latency requirements of SAP R/3 Release 4.6

The response time of a dialog step for an end user depends on the time needed for processing the request by the application server and database server, and the round trip time on the network. A small latency in the network technology used is very important for achieving a good response time for the end user. In Section 6.2.2, “Characteristics of switches” on page 157, we will show technologies that fulfill the requirement for small latency.

6.1.2 Bandwidth requirements of SAP R/3 Release 4.6

In SAP R/3 systems, two types of processes are used to work on the data of the database:

- ▶ Interactive work is processed by *dialog processes*.
- ▶ Non-interactive work is processed by *batch processes*.

We now discuss the requirements of both types of processes for the network.

Dialog processes

The network bandwidth that is required by dialog processes for the traffic between the presentation layer and the application layer is described in the SAP document *Network Load for Release 4.6*. We summarize the requirements that will be used in Section 6.3.1, “The front-end network” on page 161 to calculate the required bandwidth for our network layout in an SAP R/3 environment.

The *Enjoy* initiative in SAP R/3 Release 4.6 has introduced an improved interaction design for SAP GUI clients. This interaction design is called *LAN connection*. Since SAP R/3 Release 4.6, many transactions use the control technology that allows local activities on the SAP GUI client without contacting the application server. Examples for control technologies are local scrolling, tree views, text editor, and context menus. The amount of data transferred for each dialog step depends not only on the complexity of the transaction screen, but also on the number of items that have to be displayed.

However, a flag called *Low Speed Connection (LSC)* can be set in the options for an SAP logon connection, so that a modified interaction design is used for a low speed connection to the network. In this mode, only the data actually visible on the SAP GUI screen is transferred from the application server to the SAP GUI. The LSC requires significantly less network bandwidth than the LAN connection.

Information on how to enable and disable the LSC flag in your SAP GUI client can be found in the SAP documentation for SAP GUI clients.

We now take a look at the required bandwidth for the network for the LSC interaction design. It is assumed that the communication between the presentation layer and the application layer creates the only traffic on the network.

For SAP R/3 Release 4.6 and SAP GUI for Windows or SAP GUI for Java clients, which use the Low Speed Connection flag, the following formula has been derived using FI, MM, and SD standard application benchmarks:

$$C_{LSC} = 28.000 * N / [L * (T_{Think} + T_{Res})] / 8 \text{ [bytes/sec]}$$

In this formula, the following definitions are used:

C_{LSC}	Required network bandwidth for LSC measured in bytes/sec.
L	Line utilization ($0 < L < 1$) as a buffer for statistical fluctuation. Data is not transferred in a continuous stream, but rather has peak times within a statistical fluctuation. In order to ensure acceptable network response times, line utilization should not be higher than 0.5.
T_{Think}	The <i>Think</i> time denotes the time between user interactions. This time can vary to a great extent.
T_{Res}	The <i>Response</i> time is the time of system processing from the time a request reaches an application server until data is sent back from the application server to the SAP GUI client.
N	Number of concurrent users with think time T_{Think} and response time T_{Res} .

Using the LAN connection, the amount of data per dialog step transferred from the application layer to the presentation layer increases. SAP has measured that for the modules FI, MM, and SD, this increase is, on average, 106 percent over the LSC. Thus, the following calculation gives us an upper limit for the required network bandwidth for the SAP GUI with LAN connection:

$$C_{LAN} = 2.06 * C_{LSC}$$

In this formula, the following definitions are used:

C_{LAN}	Required network bandwidth for LAN measured in bytes/sec
-----------	--

C_{LSC} Required network bandwidth for LSC measured in bytes/sec

For the LSC, the average number of transferred bytes per dialog step is approximately 2600. At the same time, the average number of transferred bytes between the application layer and the database layer is approximately 20.000. Thus, the required bandwidth C_{AP} for the communication between the application layer and the database layer is 7.7 times the bandwidth C_{LSC} :

$$C_{AP} = 7.7 * C_{LSC} = 7.7/2.06 * C_{LAN}$$

In this formula, the following definitions are used:

C_{AP} Required network bandwidth for communications between the database server and the application server measured in bytes/sec

C_{LSC} Required network bandwidth for LSC measured in bytes/sec

C_{LAN} Required network bandwidth for LAN measured in bytes/sec

In Table 6-1, we show calculations, using the presented formulas, for the required network bandwidth between application and presentation layers, depending on different work demands for SAP R/3 Release 4.6.

Table 6-1 Network bandwidth between the application and presentation layers

N	L	T _{Think}	T _{Res}	C _{LSC} [KB/sec]	C _{LAN} [KB/sec]	C _{AP} [KB/sec]
100	0.5	10	1	62	128	479
30	0.5	10	1	19	38	144
100	0.5	30	1	22	45	170
30	0.5	30	1	7	14	51

Batch processes

The requirements for network bandwidth of traffic caused by batch processes differ due to the business needs and the setup of the SAP R/3 system.

Normally, the batch processes are scheduled to run at times where the additional work load on the SAP R/3 system does not affect the interactive working users. If the SAP R/3 system has a central instance on the database server with batch queues configured, then batch jobs running on this central instance do not produce any network load.

Pitfall ahead!

Scheduling batch processes on application servers results in a demand of network resources. Batch processes are often used to exchange additional data with other systems. These additional bandwidth requirements have to be considered.

Thus, the bandwidth requirements for batch processing cannot be estimated in general.

6.1.3 Bandwidth requirements for backup and recovery tasks

For reliable operation of an SAP R/3 system, the integrity of the data in the database of the SAP R/3 system has to be ensured. If you encounter a problem in the database (for example, a software error, an user error, or a hardware error), you have to be able to recover the database up to the time the error occurred. This requirement can be fulfilled by performing regular backups of the database and archiving the redo log files at specified times throughout the day.

In Chapter 7, “Backup and recovery” on page 185, a strategy for the backup and recovery is presented in detail. The scope of this section is to determine the bandwidth you need to be able to do these tasks in an appropriate time.

The time you need to perform a full backup depends on many aspects, such as the size of the database, the backup tool, the network technology between the database server and the backup server, the sizing of the database server and its storage subsystem, and the sizing of the backup server and its attached storage components.

All components in this backup chain should be sized in such a way that, in case of an error, you can restore the database in a previously agreed period of time, which depends on the business need of the SAP R/3 system (see Section 2.6.2, “General criteria for service level agreements” on page 25).

This agreed period of time defines a requirement for the bandwidth of the network between the database server and the backup server, presuming that all components in the backup chain are well sized.

In Table 6-2, the required sustained bandwidth for some restore scenarios are presented. Compression of the data before transmission is not taken into account.

Table 6-2 Required sustained bandwidth for the backup network

Database size [GB]	Restore time [h]	Required bandwidth [MB/sec]
30	1	8.5
30	0.5	17

Database size [GB]	Restore time [h]	Required bandwidth [MB/sec]
100	2	14
100	1	28
500	3	47
500	2	71

6.2 Network technologies for IBM @server pSeries

In this section, we present the features of the most commonly used network technologies for which network adapters are available for recent IBM @server pSeries machines. We then discuss the characteristics of the switches for these network technologies. We also present the bandwidth you can expect using a switched network of the presented network technologies. At the end of this section, we discuss some aspects of Inter Switch Links, which are important for the network layout of an SAP R/3 environment.

6.2.1 Characteristics of different network technologies

The important characteristics of the network technologies in an SAP R/3 environment are their latency and bandwidth, the cabling between the network adapters and the switches, and the maximum possible distances between two network adapters or switches. In Table 6-3, we have listed these features for the most commonly used network adapters.

Table 6-3 Features of common network adapters for IBM @server pSeries

Adapter	Bandwidth [Mb/sec]	Bandwidth [MB/sec]	Cable type	Max. distance to switch
Ethernet	10 / 100	1.25 / 12.5	CAT 5	100 m
ATM adapter	155 / 622	19.4 / 77.8	Fibre MultiMode	200 m @ 62.5µm
Gigabit Ethernet SX	1000	125	Fibre MultiMode	500 m @ 50µm
SP Switch adapter	1200	150	Copper cable	10 m

Ethernet

Ethernet technologies are frame-based and have varying packet lengths. Network adapters using Ethernet technology are commonly used and vary in their media speed. At present, media speeds of 10, 100, and 1000 Mb/sec are available. The Ethernet technology with a media speed of 100 Mb/sec is called Fast Ethernet, the Ethernet technology with a media speed of 1000 Mb/sec is called Gigabit Ethernet.

Pitfall ahead!

AIX also supports *EtherChannel*. EtherChannel is an aggregation technology that allows you to combine multiple Ethernet adapters together to form a larger pipe. You have to make sure that the switch to which you connect the EtherChannel also supports it.

ATM

Asynchronous Transfer Mode (ATM) offers a highspeed, switch-based networking technology with fixed length cells of 53 bytes. ATM provides a connection oriented transport layer based on the switching of the cells.

There are two different approaches to support the TCP/IP protocol on ATM. The first approach is called Classical IP, where IP packets are directly packed into ATM cells. The second approach is called LAN emulation, where ATM cells are used to simulate standard Ethernet frames. The TCP/IP protocol then uses standard device drivers to run on top of these Ethernet frames.

Detailed information about ATM can be found in the redbook *RS/6000 ATM Cookbook*, SG24-5525.

SP Switch

The SP Switch is designed to connect nodes of an SP to each other. It is possible to attach some of the IBM @server pSeries machines to the SP Switch using SP Switch Attachment adapters.

The basic module of the SP Switch is a *switch board*, which has 16 ports for connecting nodes. With an additional 16 ports, which are used for connecting other switch boards in different frames, it is possible to create bigger networks using several interconnected switch boards. The SP Switch provides a message-passing network that connects all processor nodes with a minimum of four paths between any pair of nodes. It provides low-latency, high-bandwidth communication between nodes.

In contrast to Ethernet and ATM, it is not possible to directly connect two SP Switch adapters to each other. You always have to connect the SP Switch adapter to the switch board. An SP Switch is normally not connected to any other switch; thus, we do not cover the SP Switch in Section 6.2.2, “Characteristics of switches” on page 157.

Detailed information for the SP Switch can be found in the redbook *Understanding and Using the SP Switch*, SG24-5161.

6.2.2 Characteristics of switches

In a modern network environment, switches are mainly used to enhance the performance and manageability of the networks. We discuss some characteristics of switches used for Ethernet and ATM network technology.

As mentioned in the previous section, the SP Switch will not be covered in the following discussion.

Latency of switched networks

Modern switches provide minimal latency and packet loss. These features are essential for mission-critical applications, such as the SAP R/3 systems. The latency of the switches differ due to the technology used. Most of the switches are usually suitable for SAP R/3 systems.

Bright idea!

We assume that all the network adapters of the servers are attached to switches in an SAP R/3 environment.

Bandwidth

Using switches, a packet on the network is only seen by a network adapter if that network adapter is the destination of the packet.

The bandwidth of the switched network can be limited by the bandwidth of the switch backplane. If the switching capacity of a switch backplane is larger than what is needed to switch the total traffic load across all the respective ports, the switch is called *nonblocking*. Otherwise it is called *blocking*.

If the configuration of the network consists of multiple switches connected to each other, the bandwidth between the switches can also limit the bandwidth of the network.

Bright idea!

Some switches prioritize some of their ports and grant them a certain percentage of the available bandwidth of the backplane. These ports can be used for the connection to other switches and for ports on which a high bandwidth demand is expected. These ports could be used, for example, for the connection of an SAP R/3 database server, a TSM server, or for the connection between two switches.

Manageability

Modern switches often support different network technologies concurrently by combining different modules. These modules are then assembled with network ports of a certain network technology. For example, it is possible to set up a switch with one module supporting Fast Ethernet and on a module supporting Gigabit Ethernet using fiber optics. The administrator can then manage different networks with only one switch.

Bright idea!

By using a network of switches from one vendor, it can be possible to manage all switches from one administration console.

We now discuss some features of switches of different network technologies.

Virtual Local Area Networks

Some switches support *Virtual Local Area Networks (VLANs)*. Using VLANs, switches can be split up into multiple logical switches. Each of the ports can be assigned to one VLAN. The VLANs are separated from each other, so no traffic between the VLANs occurs; even the ports are connected physically to the same switch. All VLANs on one switch use the switch backplane. Therefore, for a *blocking* switch, the bandwidth of a VLAN is restricted not only by its own traffic, but also by the bandwidth used by the other VLANs on the same switch.

Pitfall ahead!

If the switch supports different network technologies on its ports, you can mix different network technologies in the same VLAN. Using VLANs exploits the flexibility of the switches and helps to consolidate the switches of different networks.

Inter Switch Links

The connection between two switches is often called *Inter Switch Link (ISL)*. When connecting two switches with VLANs, some vendors support a *trunk protocol* on the ISL. When using such a trunk protocol on the ISL between two switches, packets from all VLANs of the first switch can be exchanged with the corresponding VLANs of the second switch and vice versa.

Pitfall ahead!

If a trunk protocol is not used, you need one ISL for each VLAN.

The Spanning Tree Protocol

Switches which support the *Spanning Tree Protocol (STP)* (IEEE document 802.1D) communicate with each other and dynamically enable exactly one network path between two switches of the same network, thus avoiding bridge loops. All other network paths are blocked. In case that the enabled network path breaks down, the switches automatically negotiate which new network path between the switches has to be set to nonblocking. Each VLAN must have its own spanning tree set up.

6.2.3 Bandwidth of different network technologies

It is a common experience that the theoretical bandwidth of a network technology is not reached. Using the TCP/IP protocol in practice imposes a protocol overhead that degrades the theoretical bandwidth of the network.

In Table 6-4, we present the achievable network bandwidth you can expect using the TCP/IP protocol in practice for the IBM @server pSeries servers. They can vary from environment to environment.

Table 6-4 Bandwidth of common network adapters for IBM @server pSeries

Adapter	Physical bandwidth [MB/sec]	Net TCP/IP bandwidth [MB/sec]
Ethernet	1.25/12.5	1/10
ATM adapter	19.4/77.8	15/60
Gigabit Ethernet SX	125	50
Gigabit Ethernet SX (Jumbo)	125	60
SP Switch adapter	150	34 - 135
SP Switch Attachment	150	74

Ethernet

For 10 Mb/sec Ethernet and for Fast Ethernet connections, the net bandwidths that are given in the table have been seen in real life situations. You can use the Gigabit Ethernet SX adapter in two different configurations. These configurations differ by the *Maximum Transfer Unit (MTU)* and the use of jumbo frames, which are not supported by all switches. In the first configuration, the standard MTU size of 1500 is used, resulting in a maximum net bandwidth of 50 MB/sec. In the second configuration, we use a MTU size of 9000 and jumbo frames, resulting in a maximum net bandwidth of 60 MB/sec.

ATM

The throughput of TCP/IP over ATM depends on the bandwidth of the underlying physical ATM layer (155 Mb/sec, 622 Mb/sec) and the chosen approach of the IP implementation. For more information regarding ATM on IBM @server pSeries, see the redbook *RS/6000 ATM Cookbook*, SG24-5525.

SP Switch

The values for the bandwidth of the SP Switch adapters are taken from the document *RS/6000 SP: SP Switch Performance*. You can see that the net network bandwidth for SP Switch adapters differ between adapters that are installed in SP nodes (up to 135 MB/sec, depending on the used nodes) and adapters that are used in SP attached servers (74 MB/sec). The SP System Attachment adapter attaches on one end to a PCI slot in the server's I/O drawer and on the other to the SP Switch cable. The SP Switch adapter for nodes inside the SP is connected directly to the system bus, so the transferred data does not have to go through the PCI bridges.

6.2.4 Characteristics of Inter Switch Links

Two characteristics of the Inter Switch Links are very important for the network layout of an SAP R/3 environment:

- ▶ The supported distance for an Inter Switch Link
- ▶ The possibility to have redundant Inter Switch Links

Supported distance for an Inter Switch Link

The supported distance for an Inter Switch Link of a given network technology can differ substantially from vendor to vendor. Using monomode fibers, this distances can surpass five km.

The SP Switch is an example of a switch where the connection between switches cannot span long distances. Two SP Switches in two frames of an SP only support a distance of approximately up to 20 m directly.

Redundant Inter Switch Links between switches

Only the network layout of the basic configuration of an SAP R/3 system consists of one switch.

We show in Section 6.4.2, "Network layout for the fault-tolerant configuration" on page 169 that, in a fault-tolerant SAP R/3 system, there are at least two switches that have to be connected to each other. These switches are normally located in one computing center.

In a disaster-tolerant SAP R/3 system, you have the additional requirement of distributing the two switches into two separated computing centers (see Section 6.4.3, "Network layout for the disaster-tolerant configuration" on page 172). To exclude a single point of failure, there has to be two Inter Switch Links between the two switches. If the switches cannot be connected directly over the distance between the computing centers, you obviously get a more complex environment.

6.3 Definition of the different networks

In this section, we describe networks that are necessary to operate an SAP R/3 system landscape on behalf of the agreed service level agreement. We discuss the three different networks that are required by our sample network layout for an SAP R/3 system.

6.3.1 The front-end network

In Section 3.2.4, “The networks (switch)” on page 70, the front-end network and the access network are introduced as a matter of principle. We now define the *front-end network*, the *access network*, and the *server network*, which are closely related to each other.

We then discuss the two main constraints for the front-end network, which are the suitable network technologies and the connection of the front-end network to the access network.

Front-end network and access network

Bright idea!

In an SAP R/3 system, the communication between the SAP GUI clients and the application servers (*access communication*) and the communication between the application servers and the database server (*front-end communication*) take place on the network to which the message server is bound. In Section 3.2.4, “The networks (switch)” on page 70, we defined two different segments of this network. The segment in which the front-end communication takes place is called *front-end network*. The front-end network is located in the computer center. The segments, where communication between the SAP GUI clients and the application servers take place that are outside of the computer center, are called the *access network*. In Figure 6-2 on page 162, we show a schematic network layout with the access network and the front-end network.

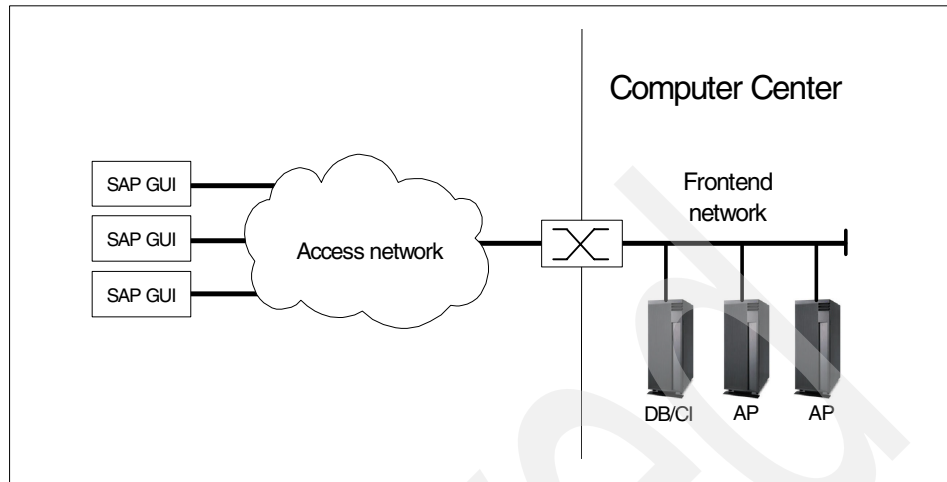


Figure 6-2 Illustration of the front-end network and access network

The access communication uses both the access network outside of the computer center and the front-end network inside the computer center. This is depending on the implementation of the network, that is, how the access network and the front-end network are connected. If both networks are in the same sub-network, switches are used. If the two networks are in different sub-networks, routers are used.

Server network

The communication of the application servers with the database server of an SAP R/3 system is called *server communication*. Normally the access communication and the server communication share the front-end network.

Bright idea!

If circumstances force you to separate the server communication from the access communication, it is possible to redirect the sever communication to a different physical network, which is called *server network*. Figure 6-3 on page 163 shows the additional server network.

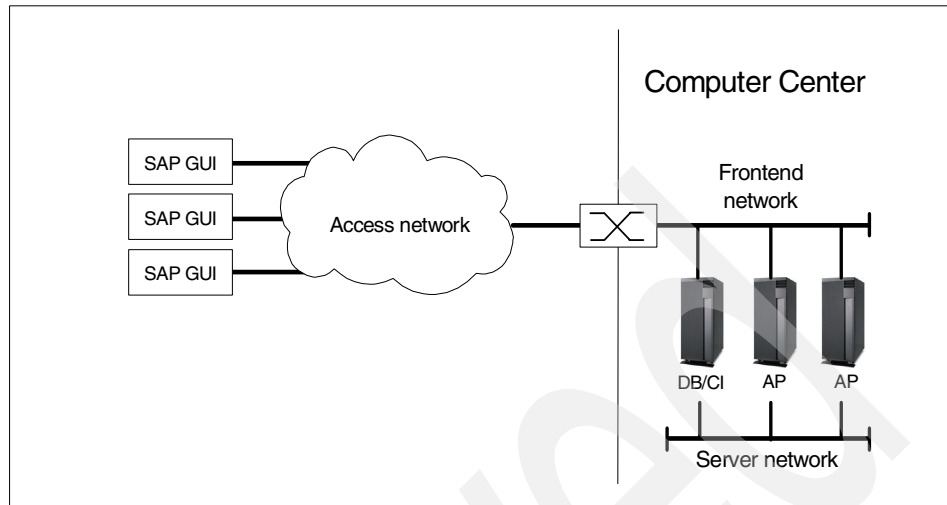


Figure 6-3 Introduction of the server network

Because the message server is bound to the front-end network, all services of an SAP R/3 system are also bound to the front-end network. If the server communication has to use a server network, all the server traffic has to be redirected to the server network. This can be done using host routes on every SAP R/3 server. Details on how to set up these routes can be found in the SAP document *SAP R/3 in Switchover Environments*.

Pitfall ahead!

In most SAP R/3 environments, it is possible to set up a *front-end network* in which the communication of the whole SAP R/3 system can take place. Using only the front-end network will reduce the complexity of the SAP R/3 environment.

If you are running an SAP R/3 system in an SP environment, the use of the SP Switch as a server network is an option.

Required bandwidth for the front-end network

For sizing the required bandwidth of a switched front-end network, you have to calculate the required bandwidth for each server of the SAP R/3 environment. The maximum required bandwidth for the front-end network of all servers in the SAP R/3 environment constrains the choice of the possible network technology. It is a good idea to include additional bandwidth in the sizing that can be used during peak work loads and which intercepts future growth of the SAP R/3 system.

Pitfall ahead!

If you have a central system, only the bandwidth for the access communication is required.

An application server requires bandwidth for the access communication and for the server communication.

In Section 6.1.2, “Bandwidth requirements of SAP R/3 Release 4.6” on page 151, we discussed how to calculate the bandwidth of the communication of SAP R/3 Release 4.6. As an example, we compare the results for three different SAP R/3 systems using these formulas. The comparison is displayed in Table 6-5, together with the available technologies of networks discussed in Section 6.2, “Network technologies for IBM ^ pSeries” on page 155. We use the following values for the parameters of the formulas $L = 0.5$, $T_{\text{Think}} = 10$ and $T_{\text{Res}} = 1$.

Table 6-5 Required bandwidth of the front-end network for interactive work

N	C _{LAN} [MB/sec]	C _{AP} [MB/sec]	C _{LAN} +C _{AP} [MB/sec]	Recommended technology	Net bandwidth [MB/sec]
200	0.3	0.9	1.2	Gigabit Ethernet	50
500	0.6	2.3	3.0	Gigabit Ethernet	50
1000	1.3	4.7	5.9	Gigabit Ethernet	50

As discussed, an SAP R/3 system also needs bandwidth for the batch processing. Even if this required bandwidth for batch processing is eight times higher than the required bandwidth for interactive work, Gigabit Ethernet can still be chosen as the technology for the front-end network for all the SAP R/3 systems shown in the table.

Bright idea!

If one Gigabit Ethernet cannot provide enough bandwidth for the front-end network, a EtherChannel combining two Gigabit Ethernet adapters can be used, as long as the switch supports it.

Bright idea!

The Gigabit Ethernet network technology has sufficient network bandwidth and it has been proven to be reliable in many environments. We suggest you use a switched Gigabit Ethernet network as a front-end network, as long as no constraints of the already existing network environment excludes it.

Connection of the front-end network to the access network

The SAP GUI clients have to communicate with the application servers, and thus the front-end network has to be accessible from the access network, which is outside of the computing centers. This can be accomplished either by using routing of the front-end network to the access network or by using a flat network where the front-end network and the access network are part of the same sub-network. There are pros and cons for both solutions.

- In a flat network, where only switches are used, the network layout is easy to implement. All servers connected to the front-end network inside the computing center are in the same broadcast domain, like all the SAP GUI clients in the access network. Thus, traffic in the front-end network can be hard to analyze.
- If an approach is used where the access network is in a different sub-network as the front-end network, routing between these two sub-networks has to be provided. The front-end network then is an independent sub-network and thus it is its own broadcast domain. This solution has the benefit that the traffic on the front-end network can easily be monitored.

6.3.2 The backup network

Bright idea!

In an SAP R/3 environment, it is necessary that backup and restore tasks do not affect the normal operation of the SAP R/3 systems. Therefore, it is reasonable to set up a separated network for backup and restore operations, which is called the *backup network*. As described in Section 6.1.3, “Bandwidth requirements for backup and recovery tasks” on page 154, the required bandwidth for a database restore is demanding for the network.

In Table 6-6, we show the required bandwidth of three different database sizes for the backup network and give a recommendation for a network technology to use.

Table 6-6 Required bandwidth for the backup network and recommendations

Database size [GB]	Restore time [h]	Required bandwidth [MB/sec]	Recommended network technology	Bandwidth [MB/sec]
30	1	9	Gigabit Ethernet	50
100	1	28	Gigabit Ethernet	50
500	2	71	2* Gigabit Ethernet	2 * 50

Table 6-6 shows that even for a database size of only 30 GB, the required bandwidth for a restore time of one hour reaches nearly the maximum bandwidth of Fast Ethernet of approximately 10 MB/sec. We advise you to use Gigabit Ethernet in order to have spare bandwidth available.

The Gigabit Ethernet technology can also supply the required bandwidth for the second example, consisting of a database size of 100 GB and a restore time of one hour.

In the third example, you can see that a single Gigabit Ethernet adapter connected to the backup network cannot supply the required bandwidth. Combining two Gigabit Ethernet adapters, the required bandwidth can be supplied. This configuration can be used with the backup and recovery solution we present in Chapter 7, “Backup and recovery” on page 185.

Bright idea!

Using Gigabit Ethernet for the backup network gives you the possibility to scale the backup network from small SAP R/3 databases to large SAP R/3 databases without changing the technology.

Thus, we suggest that you use a switched Ethernet Gigabit network as a backup network.

If the SAP R/3 system and the backup server are running in an SP environment with an SP Switch, you can also use the SP Switch as backup network for all examples presented in Table 6-6 on page 165.

6.3.3 The control network

In an SAP R/3 environment, many servers have to be administrated. In our network layout, we introduced an independent physical network called the *control network*, which increases the manageability of an SAP R/3 environment to a big extent.

Bright idea!

It is very useful to be able to connect to all the servers using this control network, which is completely independent of all provided services on the front-end and backup networks and of a possible cluster configuration. The administrator can perform all necessary administration tasks using this independent control network without influencing the operation of the SAP R/3 environment.

If the control network is only routed to the computers of the administrators, the ability to access the servers can be restricted in an efficient way by allowing certain services such as telnet, rsh for the user root, or ftp only for connections through the control network.

In addition to the servers, there are often manageable LAN and SAN components present in modern SAP R/3 environments. These components are predestined to be connected to the control network so that the administrator can control all components of the environment through a single network. A few examples for such manageable components are LAN switches, SAN switches and directors, storage subsystem systems like the ESS, and tape libraries.

In most SAP R/3 environments, a Fast Ethernet network is suitable as a control network. We suggest that you use the Fast Ethernet network adapters that are integrated into most of the common IBM @server pSeries servers for the control network.

6.4 Network layout for different configurations

Usually, in an SAP R/3 system landscape, different SAP R/3 systems are present that have different requirements for their availability (see Chapter 3, “Architecture” on page 35). In this section, we present a scalable network layout that meets the requirements for a basic configuration, a fault-tolerant configuration, and a disaster-tolerant configuration of an SAP R/3 system. By combining the corresponding network layout of the SAP R/3 systems, it is possible to set up a network layout for the entire SAP R/3 system landscape.

For example, there might be an SAP R/3 system landscape that consists of a development SAP R/3 system and a quality assurance SAP R/3 system, which both require a basic configuration. Additionally, the SAP R/3 system landscape has a productive SAP R/3 system that requires a disaster-tolerant configuration. The resulting network layout for the entire SAP R/3 system landscape then consists out of the disaster-tolerant network layout for the productive SAP R/3 system in which the additional components of the network layouts for the development SAP R/3 system and the quality assurance SAP R/3 system are added.

6.4.1 Network layout for the basic configuration

Our basic configuration for the network layout of an SAP R/3 system consists of the three proposed networks: the front-end network, the backup network, and the control network. Figure 6-4 shows this network topology for an SAP R/3 system in a basic configuration.

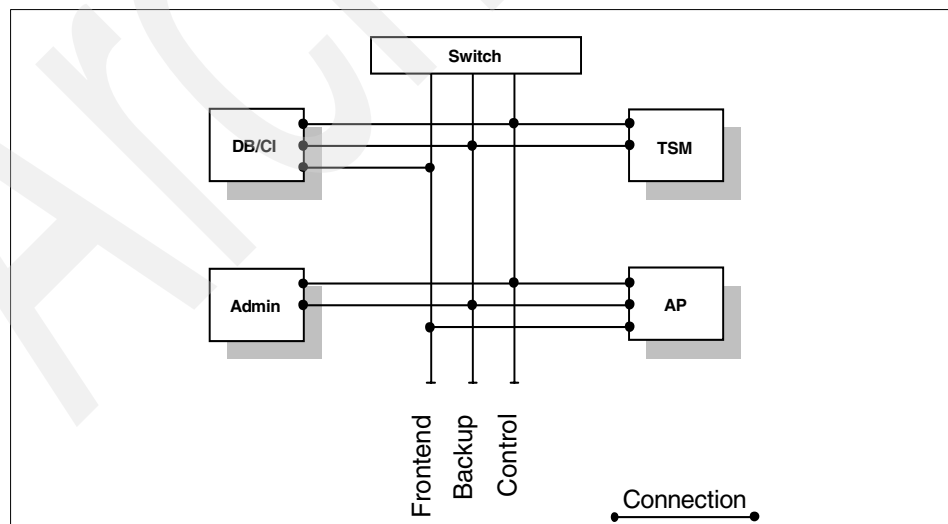


Figure 6-4 Network topology for the basic configuration

Bright idea!

This figure of the network topology is an abstract view of the connections from all servers of the SAP R/3 environment to the switch unit. We distinguish for each server between the front-end network, the backup network, and the control network. These networks are physically separated and can be implemented using different network technologies. The switch shown in Figure 6-4 on page 167 can consist of one or more switches, depending on the technology used. When a server is connected to a switch, a connection between the server and the corresponding network is shown. This connection has a dot at each end.

Each SAP R/3 server is connected to the switch of the control network, the backup network, and the front-end network using one network adapter for each connection.

The server TSM, which hosts the Tivoli Storage Manager (see Chapter 7, “Backup and recovery” on page 185), does not need to be attached to the front-end network, because its service is bound only to the backup network.

As described in Section 6.3.3, “The control network” on page 166, the server Admin from which the administration of the SAP R/3 environment is done should not be connected to the front-end network.

To clarify the physical layout of the switched network topology, we have included Figure 6-5. In this figure, the direct connections of the servers to the switch are plotted. The front-end, backup, and control network are all connected to switches of different sub-networks. If a switch with VLAN capability is used, all networks can be connected to this switch.

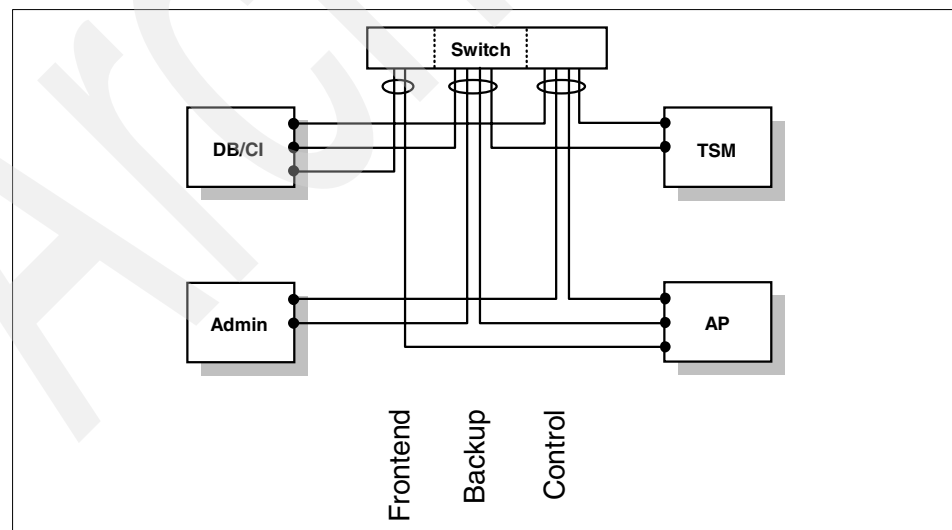


Figure 6-5 Physical layout of network connections from servers to switches

6.4.2 Network layout for the fault-tolerant configuration

If an SAP R/3 system has to be fault-tolerant, all single points of failures have to be eliminated. A fault-tolerant connection of a server to a specific network has to be set up using two network adapters.

For an SAP R/3 system, the front-end network is the most critical network. A loss of the connection to the backup network or the control network of an SAP R/3 server complicates the operation of the system. A loss of the connection to the front-end network of an SAP R/3 server stops its service at once. If the host of the central system loses its connection to the front-end network, the SAP R/3 system cannot be accessed any more. If an additional application server loses the connection to the front-end network, all the processes running on that application server stop and users logged to this server are disconnected.

In Figure 6-6, the connection of a server to the front-end network is schematically shown. The server is connected through one network adapter to a switch. The switch itself connects the front-end network to the access network.

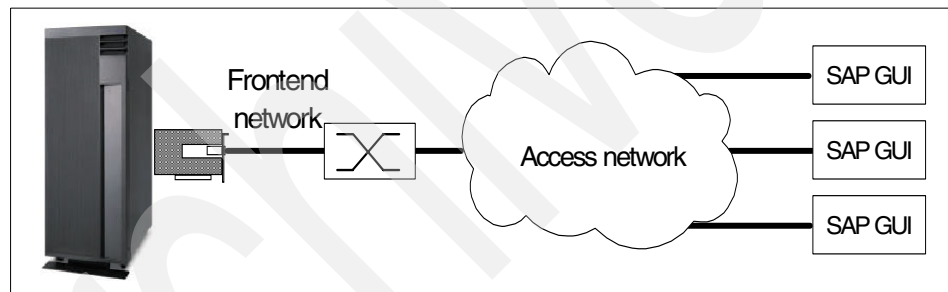


Figure 6-6 Connection with a single network adapter to the front-end network

In this configuration, a defect of either the network adapter, the switch, or a cable between the switch and either the network adapter or the network where the SAP GUI clients are located, cuts the connection of the server to the SAP GUI clients.

Bright idea!

A server can be connected redundantly to the front-end network by using two network adapters in the server, which are connected to two different switches. In Figure 6-7 on page 170, an example of a redundant connection of a server to two switches of the front-end network is illustrated.

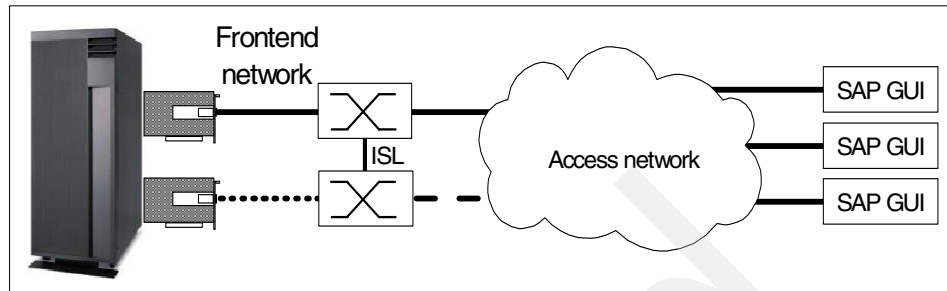


Figure 6-7 Connection of two network adapters to the front-end network

Bright idea!

To achieve fault tolerance, the access network has to be connected to both switches. The switches are connected to each other with an Inter Switch Link (ISL).

In case of a flat network, the switches use the spanning tree protocol for blocking one of the connections during normal operation (see dashed line). When the switch providing the active connection breaks down, the other switch unblocks the standby link.

If traffic to the access network is routed, the switches have to be configured as a high available router unit. Thus, if one switch fails, the other one takes over the router address.

The SAP R/3 service is bound to the upper network adapter in Figure 6-7. The second network adapter operates as a standby adapter, which is indicated by the dotted line.

In this configuration, a defect of either one network adapter, one switch or one cable only cuts one of the connections from the server to the SAP GUI clients. When using cluster software, the loss of the connection between the server and the switch is automatically detected and the standby adapter takes over the service.

In case of a complete breakdown of an SAP R/3 server, the cluster software initiates the transfer of the complete service of that server to its backup server in the cluster.

Pitfall ahead!

We explained that in a fault-tolerant configuration, two network adapters are required for the front-end network in each server. In contrast to most network technologies, only one SP Switch adapter is supported in each node of an SP and in each attached server. Therefore, the SP Switch cannot be used for the front-end network in a fault-tolerant configuration.

Figure 6-8 shows the network layout for a fault-tolerant SAP R/3 system.

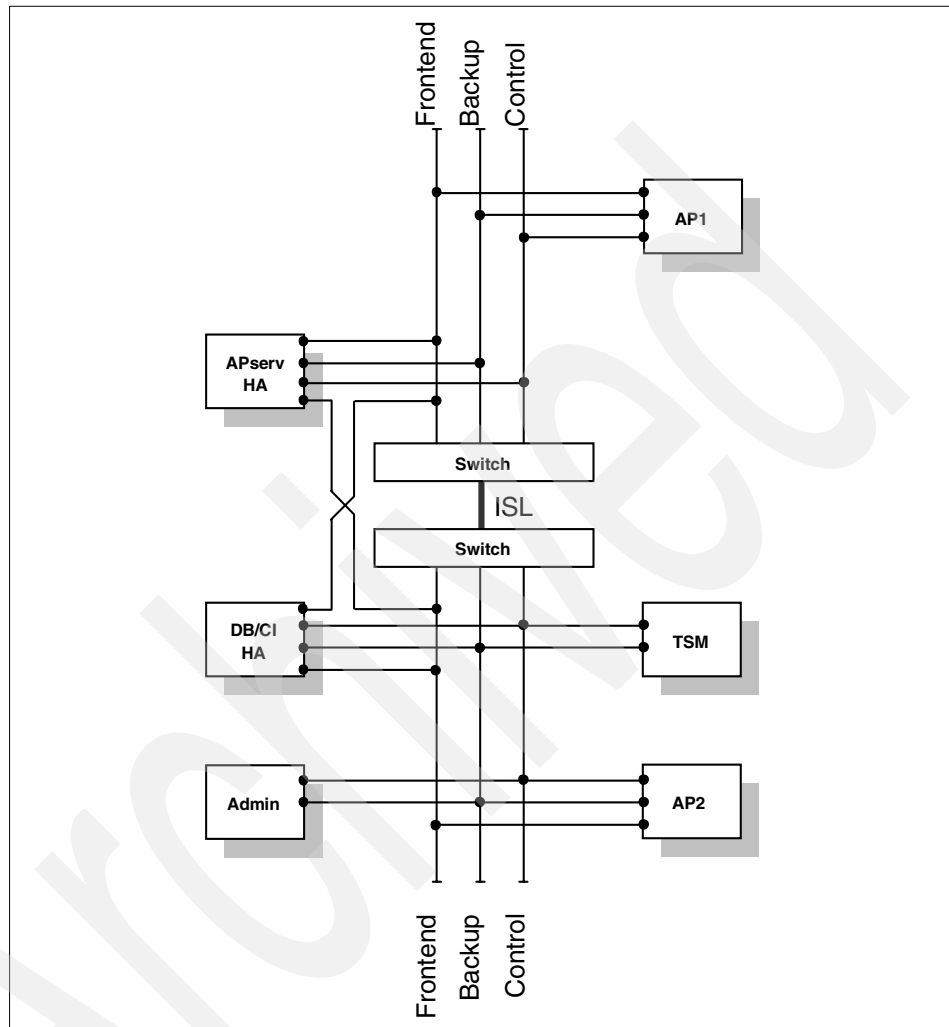


Figure 6-8 Network topology for the fault-tolerant configuration

The network layout for the fault-tolerant configuration differs from the network layout for the basic configuration by the use of two independent switches, which have a fault-tolerant Inter Switch Link.

Bright idea!

The fault-tolerant connection of each of the servers in the cluster to the front-end network is realized through the redundant connection of two network adapters for each server to the two independent switches.

The SAP R/3 servers APserv HA and DB/CI HA are running in a cluster, each of them being the backup server for the other. If one of these servers fails, the backup server takes over the service.

Furthermore, there are two application servers (AP1 and AP2) introduced. By using these two application servers, the workload is distributed to three application servers. If, in this configuration, one application server fails, there are still two application servers available.

6.4.3 Network layout for the disaster-tolerant configuration

Bright idea!

For a disaster-tolerant network layout, all redundant parts of the environment have to be separated in two independent computing centers. If one computing center is lost due to a disaster, the SAP R/3 system runs on the servers that are available in the other computing center.

Figure 6-9 shows the network layout for a disaster-tolerant SAP R/3 system.

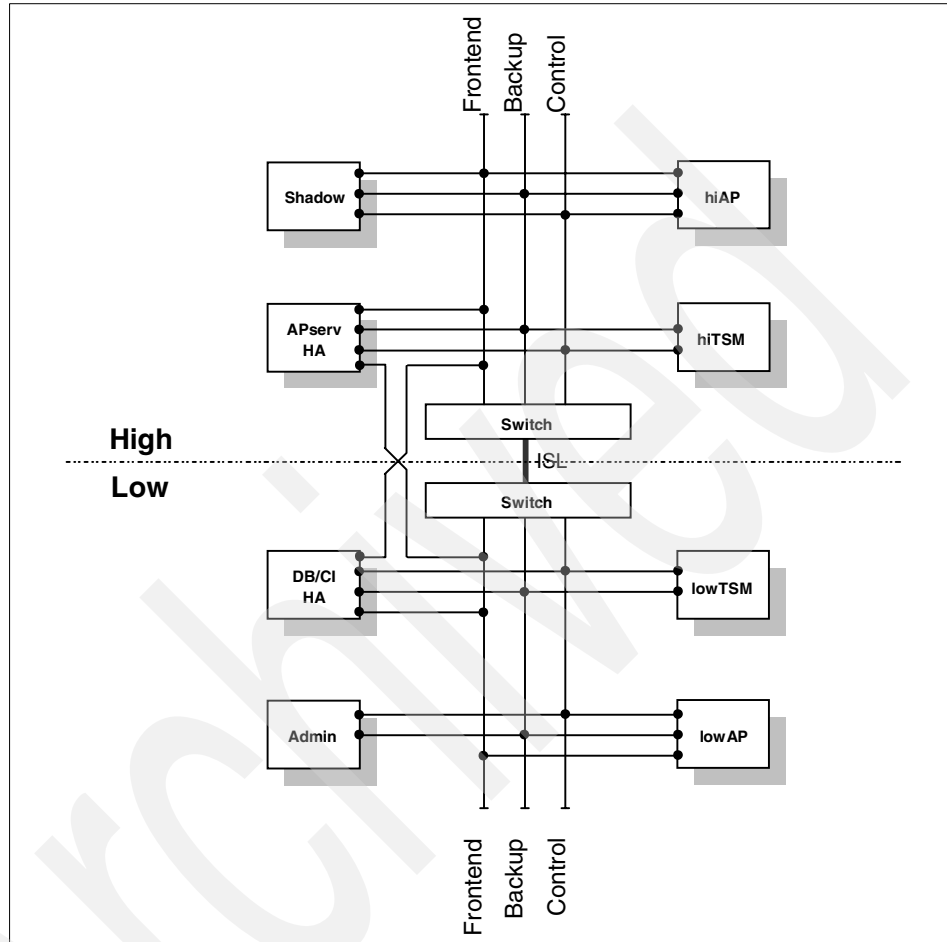


Figure 6-9 Network topology for the disaster-tolerant configuration

The network layout for the disaster-tolerant configuration differs from the network layout for the fault-tolerant configuration by the distribution of the servers to separated computing centers, which are called High and Low. The SAP R/3 servers APserv HA and DB/CI HA are running in a cluster. If one of these servers fails, the backup server takes over its service. The disaster-tolerant connection of the servers in the cluster to the front-end network is realized through the redundant connection of two network adapters to the two switches in the separated computer centers.

In our layout, the connection between the server APserv HA in the computing center High and the switch in the computing center Low has to span the distance between the computing centers. An similar connection is needed for the server DB/CI HA and the switch in the computing center High. Also, the Inter Switch Link has to span the distance between the computing centers.

6.5 Configuration of switches and network adapters

In this section, we point out some important aspects that you have to consider when implementing our sample network layout for the SAP R/3 environment (see Section 6.4, “Network layout for different configurations” on page 167 for more details).

There are two components of the network that have to be configured correctly so that the servers of the SAP R/3 systems can communicate correctly:

- ▶ The switches
- ▶ The network adapters

We discuss these components and their settings in the following sections.

6.5.1 Configuration of switches

This section is restricted to Ethernet switches, which are the most commonly used switches at present.

The configuration of an SP Switch, with all its connections to the SP Switch adapters, is described in detail in the redbook *Understanding and Using the SP Switch*, SG24-5161.

For more information regarding ATM on IBM @server pSeries, see the redbook *RS/6000 ATM Cookbook*, SG24-5525.

Ethernet switches

A connection between two Ethernet network adapters can be characterized by the following parameters:

- ▶ The mediaspeed
- ▶ The parameter duplex
- ▶ The maximum transfer unit (MTU) size
- ▶ The flow control

If you connect two Ethernet adapters with auto-negotiation switched on, all parameters for the connection are dynamically negotiated. The result of this negotiation between the two network adapters should only depend on the capabilities of both adapters and the used cabling.

Using auto-negotiation for a pair of Fast Ethernet adapters that use different chipsets or are produced by different vendors often results in a setting of the parameters that do not correspond to the fastest mode both adapters support.

Pitfall ahead!

Even worse, sometimes the negotiated parameters do not match between the two adapters. Thus, we recommend that you set the parameters for Ethernet ports in the switch to the values given in Table 6-7 for the servers in an SAP R/3 environment.

You have to ensure that the Ethernet adapters connected to these ports are also manually set to the same parameters. Section 6.5.2, “Configuration of network adapters and TCP/IP options” on page 175 explains how to set these parameters in AIX.

Table 6-7 Settings for Ethernet ports

	Fast Ethernet	Gigabit Ethernet
Speed	100	1000
Duplex	Full	Full
MTU size	1500	1500/(9000)
Flow Control	-	Enabled

Some switches support *jumbo frames* for Gigabit Ethernet. Jumbo frames use an MTU size of 9000 instead of the standard MTU size of 1500. AIX supports the jumbo frames for Gigabit Ethernet.

Consider that in a network where Gigabit Ethernet adapters are not only present, a choice of an MTU of 9000 leads to fragmentation of IP-packets, which are transferred to an adapter that is set to the standard MTU size of 1500.

6.5.2 Configuration of network adapters and TCP/IP options

The configuration of a network adapter in AIX has to be performed in two steps. First, the parameters of the device driver for the network interface have to be configured. After enabling the device driver of the network interface, the parameters for the TCP/IP protocol have to be adjusted.

Configuration of network adapter device drivers

In this section, we use examples for the configuration of Ethernet network adapters. The configuration of the SP Switch adapter is explained in detail in the redbook *RS/6000 SP System Performance Tuning Update*, SG24-5340, and the configuration of ATM adapters is discussed in the redbook *RS/6000 ATM Cookbook*, SG24-5525.

The examples we show in this section are based on a configuration that is shown in Example 6-1. To list all available Ethernet adapters in the server, use the **lsdev -Cc adapter | grep ent** command. We use the two adapters (ent0 and ent3) of this configuration for our discussion.

Example 6-1 Output of lsdev -Cc adapter | grep ent

ent0	Available 20-58	Gigabit Ethernet-SX PCI Adapter (14100401)
ent1	Available 20-70	IBM PCI Ethernet Adapter (22100020)
ent2	Available 40-58	Gigabit Ethernet-SX PCI Adapter (14100401)
ent3	Available 60-60	IBM 10/100 Mbps Ethernet PCI Adapter (23100020)
ent4	Available 60-70	IBM 10/100 Mbps Ethernet PCI Adapter (23100020)
ent5	Available A0-60	IBM 10/100 Mbps Ethernet PCI Adapter (23100020)
ent6	Available A0-70	IBM 10/100 Mbps Ethernet PCI Adapter (23100020)
ent7	Available 80-58	Gigabit Ethernet-SX PCI Adapter (14100401)
ent8	Available C0-58	Gigabit Ethernet-SX PCI Adapter (14100401)

The number of device driver attributes for different network adapters differ. To get an impression of the type of attributes, we list all attributes that can be set for a Fast Ethernet adapter and a Gigabit Ethernet adapter.

Displaying the attributes of a device driver

Use the command **lsattr -El ent3** to list the values of all attributes of the device driver for the Fast Ethernet network adapter ent3. The output of this command is shown in Example 6-2.

Example 6-2 Output of lsattr -El ent3 for a Fast Ethernet network adapter

attribute	value	description	user_settable
busio	0xffff000	Bus I/O address	False
busintr	291	Bus interrupt level	False
intr_priority	3	Interrupt priority	False
tx_que_size	8192	TRANSMIT queue size	True
rx_que_size	256	RECEIVE queue size	True
rxbuf_pool_size	384	RECEIVE buffer pool size	True
media_speed	100_Full_Duplex	Media Speed	True
use_alt_addr	no	Enable ALTERNATE ETHERNET address	True
alt_addr	0x000000000000	ALTERNATE ETHERNET address	True
ip_gap	96	Inter-Packet Gap	True

Similarly, we use the `lsattr -El ent0` command to list all values of the attributes of the device driver for the Gigabit Ethernet network adapter `ent0`. The output of this command is shown in Example 6-3.

Example 6-3 Output of `lsattr -El ent0` for a Gigabit Ethernet network adapter

attribute	value	description	user_settable
busmem	0x8f6bc000	Bus memory address	False
busintr	32	Bus interrupt level	False
intr_priority	3	Interrupt priority	False
rx_que_size	512	Receive queue size	False
tx_que_size	8192	Software transmit queue size	True
jumbo_frames	no	Transmit jumbo frames	True
media_speed	Auto_Negotiation	Media Speed (10/100/1000 Base-T Ethernet)	True
use_alt_addr	no	Enable alternate ethernet address	True
alt_addr	0x0004ac002001	Alternate ethernet address	True
trace_flag	0	Adapter firmware debug trace flag	True
copy_bytes	2048	Copy packet if this many or less bytes	True
tx_done_ticks	1000000	Clock ticks before TX done interrupt	True
tx_done_count	64	TX buffers used before TX done interrupt	True
receive_ticks	50	Clock ticks before RX interrupt	True
receive_bds	6	RX packets before RX interrupt	True
receive_proc	16	RX buffers before adapter updated	True
rxdesc_count	1000	RX buffers processed per RX interrupt	True
stat_ticks	1000000	Clock ticks before statistics updated	True
rx_checksum	yes	Enable hardware receive checksum	True
flow_ctrl	yes	Enable Transmit and Receive Flow Control	True
slh_hog	10	Interrupt events processed per interrupt	True

Changing a value for a device driver attribute

If you want to change the value of a device driver attribute of an Ethernet adapter, you have to detach the TCP/IP protocol for that specific interface if it is currently used for TCP/IP. You can check the interface status with the command given in Example 6-4. A sample output of this command is also given in Example 6-4.

Example 6-4 Status of a network interface

```
lsattr -El en3 -H | grep -E "state|value"
```

attribute	value	description	user_settable
state	up	Current Interface Status	True

In case the interface status is up, the following command can be used to set the interface status of the network adapter `ent3` to the value detach:

```
chdev -l en3 -a state=detach
```

Attention: This command will stop all communication of the network adapter ent3.

You can change the values of the device driver attributes for the device ent3 using the command `chdev -l ent3 -a attribute=value` or using the fast path `smitty chgenet`.

Pitfall ahead!

Recommended values for the Ethernet device driver attributes

We recommend that you set the following attribute for a Fast Ethernet network:

media_speed This parameter should be set to `100_Full_Duplex`. At the same time, the corresponding port in the switch must also be set to a speed of 100 Mb/sec with full duplex enabled.

We recommend that you set the following attributes for a switched Gigabit Ethernet network:

rx_checksum This attribute should be set to `yes`. This enables the calculation of the receive checksum of packets on the Gigabit Ethernet SX adapter. Taking advantage of an adapter's TCP checksum capability increases performance because the system CPU no longer has to calculate the checksum for every TCP segment.

media_speed This attribute is ignored for the Gigabit Ethernet SX adapter.

flow_ctrl This attribute should be set to `yes` if the switch supports it.

jumbo_frames This attribute enables the transmission of jumbo frames. It should only be set to `true` if the attached switch supports jumbo frames and is configured correctly. Also, the MTU size has to be adjusted to 9000. Otherwise this attribute should be set to `false`.

After setting all required device driver attributes for the network adapter, you have to enable the TCP/IP protocol again. For example, you can enable the TCP/IP protocol on the network adapter ent3 with the command `chdev -l en3 -a state=up`.

Displaying the current active network device driver values

To display the Ethernet device driver and device statistics you can use the command `entstat`. This command is important in finding out which values of the device driver attributes are used in the case of auto-negotiation. There are similar commands to display this information for token ring, FDDI, and ATM network adapters, which are called, respectively, `tokstat`, `fddistat` and `atmstat`.

Example 6-5 shows a sample output of the command **entstat -d ent0** for a Gigabit Ethernet adapter. Only the part of the output, which is important in this scope, is included.

Example 6-5 Partial output of the entstat -d ent0 command

```
ETHERNET STATISTICS (ent0) :  
Device Type: Gigabit Ethernet-SX PCI Adapter (14100401)  
Hardware Address: 00:04:ac:00:20:01  
Elapsed Time: 59 days 2 hours 20 minutes 53 seconds  
  
... (lines omitted) ...
```

General Statistics:

```
-----  
No mbuf Errors: 0  
Adapter Reset Count: 0  
Adapter Data Rate: 2000  
Driver Flags: Up Broadcast Running  
Simplex AlternateAddress 64BitSupport  
ChecksumTCP ChecksumOffload PrivateSegment  
DataRateSet
```

Adapter Specific Statistics:

```
-----  
Additional Driver Flags: Autonegotiate  
Entries to transmit timeout routine: 0  
Firmware Level: 12.4.15  
Transmit and Receive Flow Control Status: Disabled  
Link Status: Up  
Autonegotiation: Enabled  
Media Speed Running: 1000 Mbps Full Duplex
```

In the General Statistics section of Example 6-5, you can see that for this Gigabit Ethernet adapter, the attributes **ChecksumTCP** and **ChecksumOffload** are set. In the Adapter Specific Statistics section, you can see what attributes the auto-negotiation protocol has set up between the switch and the network adapter.

Tuning of the TCP/IP protocol parameters in AIX

The performance of the transmission of TCP/IP packets can be tuned to some degree. Tuning always influences the entire system. If, for example, some cache sizes are increased, you have less memory available for applications to run. So any tuning should be carefully planned, and the resulting performance of the entire system should be checked after each change of parameters.

The way network parameters are changed can be grouped into two categories:

- ▶ The first category is systemwide network parameters, which affect *all* network adapters. These network parameters can be set or checked with the **no** command. The **no** command changes these parameters only in the currently running environment. Thus, they are only effective until the next reboot.
- ▶ The second category of network parameters, called *Interface Specific Network Options (ISNO)*, can be set for each network adapter separately, if the **no** parameter `use_isno` is set to 1. The number of available ISNO parameters depends on the type of the adapter. The ISNO parameters are set using the **chdev** command and are read out using the **lsattr** command. The **chdev** command writes the ISNO parameters into the Object Data Manager (ODM). Thus, the settings for these parameters are used to set up the network adapter after a reboot. The values of the ISNO parameters set on the individual network adapter overwrite the systemwide values of the corresponding parameters. Some of the network options can also be set using the **ifconfig** command. In contrast to the **chdev** command, the values for the parameter set with **ifconfig** are *not* stored in the ODM, and thus are lost after a reboot.

At this point, the parameters that are important for the network adapters are discussed. For further information, see the *Performance Management Guide*.

use_isno	Enables the use of Interface Specific Network Options. The default value is 1 (enabled). This attribute only applies to AIX Version 4.3.3 and later versions.
MTU	This value of the maximum transfer unit limits the size of packets that are transmitted on the network. The value can be set in a range from 60 to 65536.
tcp_pmtu_discover	Enables or disables path MTU discovery for TCP applications. A value of 0 disables path MTU discovery for TCP applications, while a value of 1 enables it (default for AIX Version 4.3.3 and later versions).
udp_pmtu_discover	Enables or disables path MTU discovery for UDP applications. A value of 0 disables path MTU discovery for UDP applications, while a value of 1 enables it (default for AIX Version 4.3.3 and later versions).
thewall	Specifies the maximum amount of memory, in kilobytes, that is allocated to the memory pool.
sb_max	This value sets the upper bound on the size of TCP and UDP socket buffers. It limits <code>tcp_sendspace</code> , <code>tcp_recvspace</code> , <code>udp_sendspace</code> , and <code>udp_recvspace</code> .

tcp_sendspace	The value of this attribute sets the size of the TCP socket send buffer. The value must be less than or equal to sb_max. If the value exceeds 64 KB, the attribute rfc1323 has to be set to 1.
tcp_recvspace	The value of this attribute sets the size of the TCP socket receive buffer. The value must be less than or equal to sb_max. If the value exceeds 64 KB, the attribute rfc1323 has to be set to 1.
udp_sendspace	The value of this attribute sets the size of the UDP socket send buffer. The value must be less than or equal to sb_max. If the value exceeds 64KB, the attribute rfc1323 has to be set to 1.
udp_recvspace	The value of this attribute sets the size of the UDP socket receive buffer. The value must be less than or equal to sb_max. If the value exceeds 64KB, the attribute rfc1323 has to be set to 1.
rfc1323	A value of 1 indicates that tcp_sendspace and tcp_recvspace can exceed 64KB. The default value is 0.
tcp_mssdflt	The default maximum segment size used in communicating with remote networks.
tcp_nodelay	If the requests or responses are variable-size, use TCP with the TCP_NODELAY option. We recommend that you set this value to 1.

Displaying the values for the network options

You can use the **no -a** command to display the attributes values for the TCP/IP protocol that are set systemwide. A sample output of this command is given in Example 6-6, where only the relevant attributes are shown.

Example 6-6 Sample output of the no -a command (truncated)

```

...(lines omitted)...
sb_max = 1048576
tcp_sendspace = 16384
tcp_recvspace = 16384
rfc1323 = 0
use_isno = 1

```

You can use the **lsattr -El en3** command to list all values of the attributes of the protocol for the Fast Ethernet network adapter ent3. The output of the command is given in Example 6-7 on page 182.

Example 6-7 Output of the lsattr -El en3 command

attribute	value	description	user_settable
mtu	1500	Maximum IP Packet Size for This Device	True
remmtu	576	Maximum IP Packet Size for REMOTE Networks	True
netaddr	14.24.29.231	Internet Address	True
state	up	Current Interface Status	True
arp	on	Address Resolution Protocol (ARP)	True
netmask	255.255.252.0	Subnet Mask	True
security	none	Security Level	True
authority		Authorized Users	True
broadcast		Broadcast Address	True
netaddr6		N/A	True
alias6		N/A	True
prefixlen		N/A	True
alias4		N/A	True
rfc1323		N/A	True
tcp_nodelay		N/A	True
tcp_sendspace		N/A	True
tcp_recvspace		N/A	True
tcp_mssdflt		N/A	True

Similarly, we use the **lsattr -El en0** command to list the values of all attributes of the protocol for the Gigabit Ethernet network adapter ent0. The output of the command is given in Example 6-8.

Example 6-8 Output of the lsattr -El en0 command

attribute	value	description	user_settable
mtu	1500	Maximum IP Packet Size for This Device	True
remmtu	576	Maximum IP Packet Size for REMOTE Networks	True
netaddr	10.1.28.101	Internet Address	True
state	up	Current Interface Status	True
arp	on	Address Resolution Protocol (ARP)	True
netmask	255.255.255.0	Subnet Mask	True
security	none	Security Level	True
authority		Authorized Users	True
broadcast		Broadcast Address	True
netaddr6		N/A	True
alias6		N/A	True
prefixlen		N/A	True
alias4		N/A	True
rfc1323		N/A	True
tcp_nodelay		N/A	True
tcp_sendspace		N/A	True
tcp_recvspace		N/A	True
tcp_mssdflt		N/A	True

Changing a value for a network option

Attributes of the TCP/IP protocol, which can be set systemwide, can be changed using the **no -o attribute=value** command.

Attributes of the TCP/IP protocol, which can be set for the individual network adapter, can be set using the **chdev** command. If you want to change en3, you can use the **chdev -l en3 -a attribute=value** command.

Attention: To enable the use of Interface Specific Network Options, the **no** parameter **use_isno** has to be set to 1.

Recommended values for the network options

Pitfall ahead!

In Table 6-8, the recommended values for the discussed attributes are given. The values are taken from the *Performance Management Guide*. If a value of an attribute does not have to be changed from the standard, a dash (-) is used.

Table 6-8 Recommended values of the attributes for the TCP/IP protocol

	Fast Ethernet	Gigabit Ethernet	Gigabit Ethernet	ATM	SP Switch
Speed	100 Mb	1000 Mb	1000 Mb	155 Mb	1200 Mb
MTU	1500	1500	9000	9180	65520
thewall	-	-	-	-	16384
sb_max	32768	131072	262144	131072	1310720
tcp_sendspace	16384	65536	131072	65536	262144
tcp_recvspace	16384	16384	92160	65536	262144
udp_sendspace	-	-	-	-	65536
udp_recvspace	-	-	-	-	655360
rfc1323	0	0	0	0	1
tcp_mssdflt	-	-	-	-	1448
tcp_nodelay	1	1	1	1	1
tcp_pmtu_discover	0	0	0	0	0
udp_pmtu_discover	0	0	0	0	0

Setting permanent system-wide network options

Pitfall ahead!

As described in the previous section, all attributes changed using the **no** command are *not* stored in the ODM. Thus, a reboot will change the values of these parameters to their kernel default as long as no script sets specific values during system boot.

Bright idea!

For example, the script `/etc/rc.net` is executed during a system boot and changes some of the network options. If you want the changed attributes to be set every time the system boots, one possibility is to include all the required commands in a script that is executed by **init**. It is not advisable to include these parameters in a script that is used by AIX during boot, such as `/etc/rc.net`. If you migrate to a new version of AIX these scripts are likely to be overwritten.

In an SP environment, the tunable values for each node can be set in the `tuning.cust` file. This file is found in the `/fttpboot` directory on each node. It should be created and managed in the `/fttpboot` directory on the Control Workstation (CWS). There are sample files for different environments located in the `/usr/lpp/ssp/install/config` directory if you do not already have a `tuning.cust` file.

Example 6-9 shows the sample entry `rtune` in the `inittab`. The script `/usr/scripts/cust/scripts/rc.tuning.customer` has to contain all systemwide network options you need.

Example 6-9 Sample entry for the inittab

```
rctcpip:a:wait:/etc/rc.tcpip > /dev/console 2>&1 # Start TCP/IP daemons
...(lines omitted)...
rtune:2:once:/usr/scripts/cust/scripts/rc.tuning.customer # Set no parameters
```

You have to make sure that the values for the attributes of the network options set by your script are not overwritten by another script that is called after your script after a system reboot. The easiest way is to check all changed values of network option attributes after a reboot.

Backup and recovery

In a nutshell:

- ▶ Include all systems into your backup and recovery solution.
- ▶ Install all systems with NIM and use `mksysb` for operating system backups.
- ▶ TSM, in conjunction with TDP for R/3, offers a high performance backup and recovery solution for your SAP R/3 environment.
- ▶ Use virtual nodes to tailor your TSM solution.
- ▶ Use tape libraries for an automated, reliable, manageable, and efficient operation.
- ▶ Do regular restore and recovery tests to ensure the operability of your backup and recovery solution.

SAP R/3 systems represent a part of the IT infrastructure that is crucial to the company. Data inconsistencies or loss of data are unacceptable. However, there are plenty of reasons that may lead to a loss of data in an SAP R/3 environment, for example, user errors, software errors, power or hardware failure, or even a disaster. Thus, a proven backup and recovery concept and implementation is an essential part of a reliable SAP R/3 infrastructure.

In this chapter, we discuss the various aspects of an SAP R/3 environment from a backup and recovery point of view. The discussion of the concepts is based on the products Tivoli Storage Manager (TSM) and Tivoli Data Protection (TDP) for SAP R/3.

This chapter covers the highlighted area in Figure 7-1, which is derived from the SAP R/3 infrastructure model that is used throughout the book.



Figure 7-1 Backup and recovery

7.1 Introduction

In this section, we describe a general concept that is usable for all systems of an SAP R/3 landscape. The implementation example is focused to fulfill the production requirements. We will discuss the concepts and the implementation of such a solution. The presented reliable solution meets the key business requirements for protecting your vital data.

Usually, a TSM server implementation is not only used by SAP R/3 systems, but also by other possibly unrelated systems in the same I/T environment. In this section, we will only cover the SAP R/3 systems.

7.1.1 Terminology

To understand the next sections, it is necessary to have a common understanding of terms that are frequently used when discussing database and backup operations. The subsequent descriptions explain the meaning of the following terms:

- ▶ Log files
- ▶ Backup
- ▶ Archive
- ▶ Restore
- ▶ Recovery

Log files

Relational databases use *log files* to record changes made in a database. Figure 7-2 shows a schematic picture where data is written to the database and information about the change is written to log files.

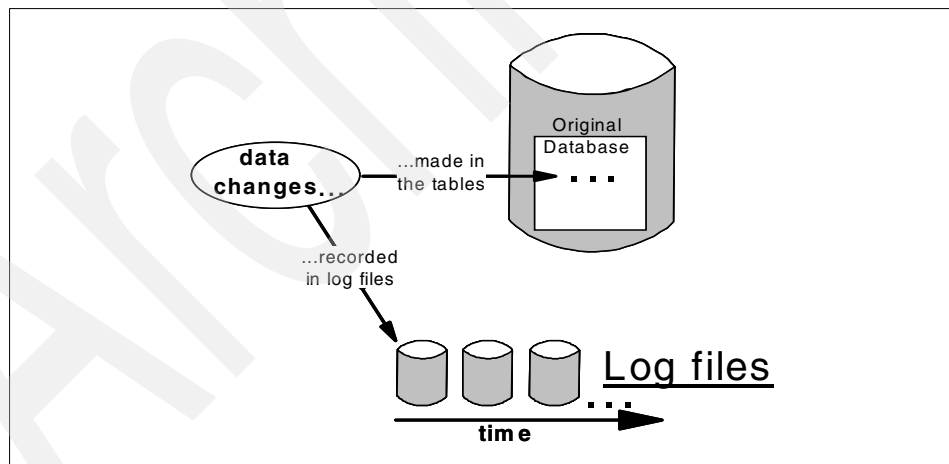


Figure 7-2 Normal database operation

Different database vendors use different terms for log files, for example, *database logs* for DB2 databases and *redo log files* for Oracle databases.

Backup and archive

For data integrity reasons, you have to create *backups* of the database and its log files. This allows you to recreate the database after a failure without any loss of data. The backup of the log files is usually called an *archive*. This is shown in Figure 7-3.

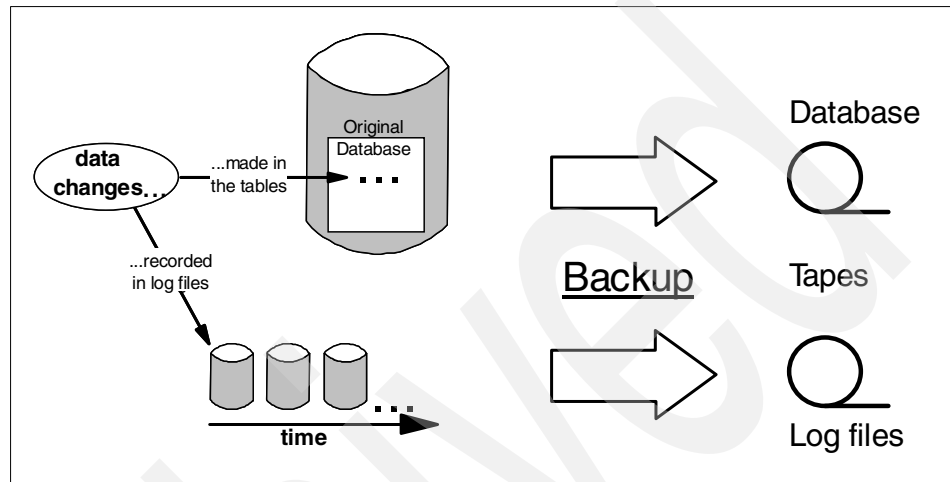


Figure 7-3 Backup of a database

Restore and recovery

In case of a failure, it may be necessary to *restore* the database and its log files from a backup copy. The process of reconstructing the data to a certain point in time by applying log files is called *recovery*. This is illustrated in Figure 7-4 on page 189.

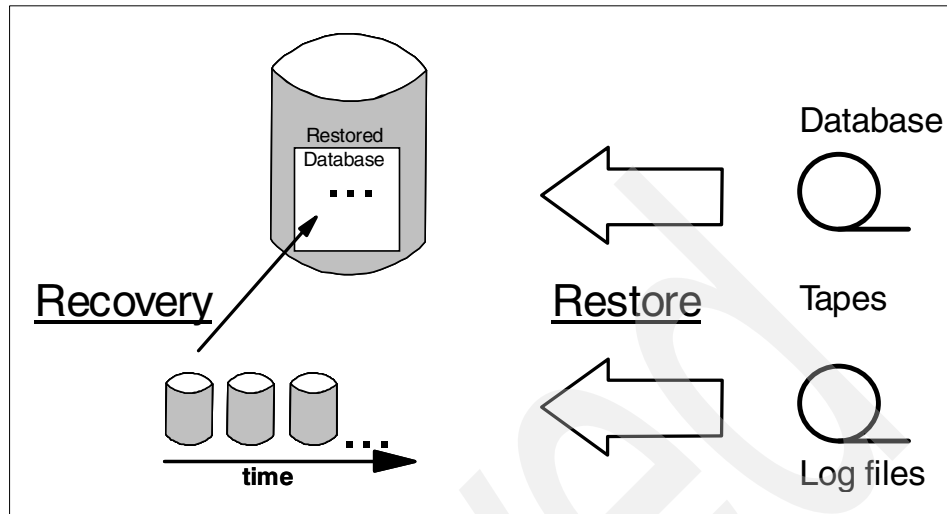


Figure 7-4 Restore and recovery of a database

TSM uses some of these terms in its own context. The expression *TSM backup* is used for a backup of objects, such as files and directories. Expiration of these backup objects is controlled by a configured number of versions and a retention period. The expression *TSM archive* is used for a backup of data for long-term storage. The server retains archives to a certain point in time as defined in the corresponding archive copy group.

To distinguish between the different terms, we explicitly use TSM backup or TSM archive for TSM specific actions in this chapter.

7.1.2 General aspects

The general requirements for SAP R/3 environments are discussed in Section 2.2, “Definition of requirements” on page 12.

This leads to the following key requirements for a backup and recovery solution in an SAP R/3 environment:

- ▶ No loss of data in SAP R/3 systems, inside and outside of the database
- ▶ Minimum downtime for SAP R/3 systems during backup or recovery
- ▶ Automatic operation sequences to reduce system management effort
- ▶ Enable on demand backups to hold data over a long period
- ▶ Usage of similar procedures for all SAP R/3 systems
- ▶ Consideration of disaster recovery management

- ▶ Include all systems necessary for a consistent overall environment

Pitfall ahead!

The last point in the list is very important for the possibility of a point in time recovery of the complete environment. The SAP R/3 system has usually many inbound and outbound interfaces to different supplying systems. If a point in time recovery of the database has to be performed and some of your connected systems are not able to resend interface data or to handle data sent to them again, inconsistencies in your overall environment will occur. In this case, you have to define synchronization points or synchronization procedures.

More information on that subject can be found in the SAP white paper *Backup & Restore Concepts for mySAP.com System Landscapes*, found at:

<http://service.sap.com/atg>

Each installation has its own requirements for a backup and restore solution. The main issues that have to be taken into account are:

- ▶ The available time for backup
- ▶ The available time for restore and recovery
- ▶ The type, characteristics, and amount of the data
- ▶ The frequency of data modifications
- ▶ The hardware and software available for backup and recovery

All different SAP R/3 systems in an SAP R/3 environment basically have the same requirement for the backup and recovery solution. The requirements for the individual issues listed above can differ. The frequency of backups or the number of versions retained will differ between a development system and a production system.

The backup and recovery solution that we describe in this chapter can be used for all SAP R/3 systems of your SAP R/3 landscape.

The implementation example in this chapter is based on TSM and TDP for R/3. As a prerequisite for reading this chapter, we presume a good understanding of these products and of the basic principles they use.

Detailed information is available in the product manuals and the following Redbooks:

- ▶ *Tivoli Storage Management Concepts*, SG24-4877
- ▶ *Getting Started with Tivoli Storage Manager: Implementation Guide*, SG24-5416
- ▶ *Using ADSM to Back Up Databases*, SG24-4335
- ▶ *Backing Up DB2 Using Tivoli Storage Manager*, SG24-6247

- ▶ *R/3 Data Management Techniques Using Tivoli Storage Manager*, SG24-5743
- ▶ *ADSM Server-to-Server Implementation and Operation*, SG24-5244

7.2 Backup objects and methods

Designing a backup and recovery concept for an SAP R/3 environment requires that you consider of different types of data. We distinguish between the following objects:

- ▶ AIX operating system, including all Licensed Program Products (LPP)
- ▶ SAP R/3 and database executables, including configuration files
- ▶ SAP R/3 database, including log files
- ▶ Archived log files of the SAP R/3 database
- ▶ Files of the SAP R/3 Transport Management System (TMS)
- ▶ Files of the interfaces of the SAP R/3 systems
- ▶ TSM database and recovery log and TSM server configuration files
- ▶ Logical Volume Manager (LVM) configuration information

All these objects have different requirements regarding backup and recovery and will be processed using different methods.

The following sections provide a more detailed description of each classification.

7.2.1 AIX operating system

In this section, we will explain how TSM incremental backups can be used for a point in time recovery of files of the AIX operating system.

The AIX operating system is the basis for the installation and operation of an SAP R/3 system on IBM @server pSeries. The backup of this component comprises all elements of the customized AIX operating system, especially:

- ▶ Characteristics of the operating system
- ▶ Created groups and users
- ▶ Network configurations
- ▶ Printers and printer queues
- ▶ Installed Licensed Program Products

- Information documenting the logical and physical structure of the whole system (refer to Section 7.2.8, “LVM configuration” on page 196)

Generally all the relevant information is stored in the volume group rootvg and will be backed up with the AIX command **mksysb**. With the **mksysb** command, a bootable system backup can be created if directed to a tape. Using a bootable system backup on tape, you can easily perform a restore of AIX. AIX Version 4.3.3 also supports creating a bootable system backup to a CD-R device.

Bright idea!

Normally in an SAP R/3 environment, there are a number of SAP R/3 servers installed and additional servers are used to control and maintain the environment. In such an environment, a possible alternative is to write your system backups into files and implement a AIX NIM server for required restores. This solution enables you to implement an automated procedure with more than one backup version and without manual tape management. Additionally, you save the expenses for attaching tape drives directly to all servers. The NIM server can furthermore act as a central software repository for all servers running AIX (see Section 3.2.6, “Administration server (Admin)” on page 73).

In addition to the image created by the **mksysb** command, the files on all servers can be backed up using the TSM incremental backup. Using incremental backups, you only back up the files that have been changed since the last backup. Thus, you have much less data to backup. This means you are able to perform an incremental backup faster than a **mksysb** system backup. In the case of a restore, these incremental backups enable a point-in-time restore of single files. A prerequisite for the restore of data using the TSM client is the restore of the AIX system, including a TSM client, which can be done using a **mksysb** system backup.

7.2.2 SAP R/3 database including log files

The database is the core of every SAP R/3 system. It contains not only the application data of the implemented business processes, but also the customizations and all ABAP programs, screens and so on, which, in total, determine the functionality of an SAP R/3 system. The availability of the database and the ability to recover it in the event of a failure is therefore of crucial importance for the SAP R/3 system. The following objects belong to a database from the backup point of view:

- Database control files (Oracle only)
- Database data files (Oracle data files and DB2 containers)
- Database log files (Oracle redo log files and DB2 database logs)
- Administration database (DB2 only)

- Information logs of the backup runs (Oracle SAPDBA logs and DB2 recovery history files)

SAP provides a so called *backint* interface to enable the backup of SAP R/3 databases into storage management systems, such as TSM, using the SAP R/3 administration utilities. The product TDP for R/3 is based on this interface. TDP uses the TSM archive command to back up the database. TDP for R/3 is available for Oracle or DB2 databases.

For a DB2 database, you can also use the integrated DB2 command **backup** for backing up the database into TSM. In this case, the TSM backup is executed by DB2.

The restore of an Oracle database can be performed with the utility **backfm** or the appropriate SAP R/3 utilities. The SAP R/3 utilities require that the backup log files are already available. If necessary, the backup log files have to be restored in advance. Restores of DB2 databases are always based on the DB2 command **restore**. If you restore a database backup that was made in online mode, the database has to be recovered or rolled forward after the restore with all necessary log files.

7.2.3 Archived log files of the SAP R/3 database

The archived log files contain information about changes to your database data files. A database can be recovered to a particular point in time using these archived log files. Depending on the number and size of the necessary log files, a recovery can be very time-consuming.

If all available log files are filled up, the database hangs. Section 13.2.2, “Checking file systems” on page 410 and Section 14.3.1, “Archive file system full” on page 448) deal with this subject. Thus, it is very important to be able to archive the log files continuously.

The log files are archived and restored with TDP for R/3 using the SAP R/3 *backint* interface. As with the database backups, the TSM archive mechanism is used. Because the log files are essential for a recovery of the database, always make at least two copies of each log file and have them stored on different tapes. Using TSM, you have to define the same number of management classes as the number of copies of log files you want to archive. In this way, you ensure that each copy of a log file is placed on a different media. Using DB2 as the database, there is also the possibility to archive the log files into TSM with the **brarchive** utility.

7.2.4 SAP R/3 and database executables

The executables of SAP R/3, the executables of the database, and all associated profiles, message logs, trace files and so on are normally stored in special file systems. The physical disks on which these file systems reside are pooled into volume groups in which no other file systems should be placed (see also Section 3.2.3, “The storage subsystem (disk storage)” on page 69). Accordingly, the backup of these files cannot be performed together with the backup of the operating system.

The following objects are part of this component:

- ▶ Database executables and configuration files
- ▶ SAP R/3 executables and profiles
- ▶ Contents of home directory of user <sid>adm
- ▶ Information about tape management of SAP R/3 backup utilities
- ▶ Log files of SAP R/3 batch jobs
- ▶ Output of ABAP reports on file
- ▶ Spool requests and protocols of SAP R/3 batch input sessions (depending on setting of the SAP R/3 parameter `rspo/store_location`)

Additionally, all files that were customized so that the specific SAP R/3 service on the dedicated server can be provided are members of this backup object. Examples of such customized files are the files `/etc/services` or `/etc/inittab`.

In this way, it is possible to restore the SAP R/3 system on any other server in case of a disaster or if a system copy is needed (see Chapter 12, “SAP R/3 system copy” on page 365).

These objects are backed up using TSM incremental backup to hold a defined number of versions over a defined retention period.

Because the requirements of this component and the AIX component differ, two different TSM management classes have to be used.

7.2.5 SAP R/3 Transport Management System files

The files in the directory of the SAP R/3 TMS include all the dictionary, repository, and customization changes, which have to be distributed to the systems in your SAP R/3 landscape. Although most of these files can be regenerated in case of a disaster, it can be very helpful to back up these files on a regular basis. This is especially important during a time critical rollout of new implementation projects.

These files should be backed up using the TSM incremental backup. Depending on your requirements or company specific regulations, you may want to keep these files for long-term storage. In this case, use TSM archive with an appropriate retention period (on-demand backups).

7.2.6 Interfaces of the SAP R/3 system

The SAP R/3 system is usually tightly coupled to other systems in your IT infrastructure. Normally, several interfaces are customized to receive or distribute data. In addition to online interfaces, files are often used to exchange data. To back up the files of these inbound or outbound interfaces, TSM incremental backup can be used. As a consistent restore depends on the functionality of your interfaces and the status of the corresponding systems, coordination between both is mandatory.

7.2.7 TSM database and recovery log

The TSM database is the core of a TSM implementation. It contains the information needed for server operations and the complete information about all client data that has been backed up or archived. This information includes versions, ownerships, permissions, and location. If this database is destroyed or in an inconsistent state, you have no way to restore your objects without restoring the TSM database first. Beside the database, TSM uses a recovery log for logging transactions.

TSM provides the **backup db** command to perform a consistent backup of the TSM database. Using the **backup db** command with the option `type=full` backs up the database, including the recovery log. When using the **backup db** command with the option `type=incremental`, only the recovery log is backed up.

For a restore of the TSM database, the following server configuration files are necessary:

- ▶ Volume history file (as configured in `dsmserv.opt`)
- ▶ Device configuration file (as configured in `dsmserv.opt`)
- ▶ Server option file (`dsmserv.opt`)
- ▶ Server disk file (`dsmserv.dsk`)

7.2.8 LVM configuration

TSM can only restore into available file systems. To provide the required file systems of your environment with the same characteristics as initially installed, you need the following information:

- ▶ Definitions of volume groups, logical volumes, and file systems of your system
- ▶ Distribution of the data over the available disks
- ▶ Permissions and ownerships of mount points

The first step is to have an up-to-date system documentation, as described in Section 3.2.6, “Administration server (Admin)” on page 73. To enable a point in time recovery, you should hold several versions of this system documentation. However, in case of a restore, you have to recreate the environment manually with the risk of human error.

Bright idea!

Therefore we recommend an automatic procedure. With the **savevg -mef <vg_map_file> <vg_name>** command, you can create a map file of a volume group without creating a backup. You need an exclude file named `/etc/exclude.<vgname>`, in which all files are excluded from the backup. The resulting map file enables a rebuild of the environment with the **restvg -qf <vg_map_file>** command. If the map will be stored in the rootvg, it is included as part of the system backup method described in Section 7.2.1, “AIX operating system” on page 191. Additionally, it can be copied to other servers, such as the administration server, to be able to access it in case of a disaster.

7.2.9 Summary

Table 7-1 shows a collection of previously described backup objects and their respective backup and restore methods. The numbers in the first column will be used to reference these objects in the following sections.

Table 7-1 Backup objects and methods

No.	Backup object	Backup method	Restore method
1	AIX operating system (image level)	mksysb and TSM (selective backup)	NIM
2	AIX operating system (file level)	TSM (incremental backup)	TSM (restore)
3	LVM configuration information	savevg	restvg
4	SAP R/3 database (DB2) including log files	TDP for R/3 (db2 backup db)	TDP for R/3 (db2 restore db)

No.	Backup object	Backup method	Restore method
4	SAP R/3 database (Oracle) including log files	TDP for R/3 (SAP R/3 brbackup)	TDP for R/3 (backfm / SAP R/3 brrestore)
5	Archived log files of the SAP R/3 database	TDP for R/3 (SAP R/3 brarchive)	TDP for R/3 (backfm / SAP R/3 brrestore)
6	Executables of SAP R/3 and database	TSM (incremental backup)	TSM (restore)
7	SAP R/3 TMS files	TSM (incremental backup)	TSM (restore)
8	Interface files	TSM (incremental backup)	TSM (restore)
9	TSM database	TSM (backup db) and remote copy	dsmserv restore db
10	TSM configuration information files	remote copy	remote copy

7.3 TSM concept

There are a plenty of ways to implement a solution that complies with the requirements defined in Section 7.1.2, “General aspects” on page 189. In this section, we will make a suggestion for implementing a TSM based solution for an SAP R/3 environment. This is one possibility that has been proven to work in several installations.

7.3.1 Virtual nodes

Pitfall ahead!

We recommend that you use TSM virtual node names, for the following reasons:

- ▶ It allows you easier control of the space used per node at the TSM server.
- ▶ You have better possibilities to control the restore of all objects of a virtual node. For example, you can simply restore the SAP R/3 relevant files to another node for duplicating a system.
- ▶ It enables you to create different schedules for incremental backups (see Section 7.3.5, “Scheduling” on page 205).
- ▶ You can easily move the objects and schedules of one virtual node from one host to another, for example, associated with HACMP resources.

The different characteristics of the virtual node are defined through specific TSM client configuration files:

dsm.sys One per system
dsm.opt One per virtual node
inclexcl One per virtual node

A naming convention for the files of a virtual node, such as dsm.<v_node>.opt, makes handling easier. The inclexcl file controls which files are part of the specific virtual node. Additionally, you are able to define the management class to which the data will be backed up (refer to Section 7.3.3, “Domain, management classes and copy groups” on page 201). The access to these configuration files is controlled by the following environment variables:

DSM_DIR Points to directory of dsm.sys
DSM_CONFIG Points to dsm.opt

The location of the inclexcl file is specified in dsm.sys per virtual node stanza.

The drawbacks of this solution are:

- ▶ You have to maintain complex inclexcl files and you have to ensure that all files will be backed up at this level.
- ▶ If collocation is enabled for your storage pools, you have an increased consumption of tapes.

We group the backup objects into the virtual nodes shown in Table 7-2.

Table 7-2 Virtual Nodes for an SAP R/3 environment

Backup object no.	Backup object	Virtual Node
1	AIX operating system (Image level)	AIX
2	AIX operating system (file level)	
3	LVM configuration information (indirectly)	
4	SAP R/3 database including log files	DB
5	Archived log files of the SAP R/3 database	
6	Executables of SAP R/3 and database	SAP
7	SAP R/3 TMS files	
8	Interface files	

Backup object no.	Backup object	Virtual Node
9	TSM database	
10	TSM configuration information files	

Because of the requirements in your specific environment, additional groups can be created if necessary.

7.3.2 Storage pools

In common installations, you have disks and tapes as storage media. TSM enables you to pool storage media of the same type in so called *storage pools*. By dividing the available media into different storage pools used by different nodes, you can configure different characteristics. The following parameters are relevant regarding performance and reliability of your backup and recovery solution:

Collocation

With collocation enabled, the TSM server tries to join the information from one node in a small set of volumes. This parameter can improve the performance of the restore by reducing the necessary tape mounts.

Collocation should be activated for storage pools used for incremental backups, as the data stored in these pools will be distributed over the available media. Collocation is not useful for SAP R/3 database backups. These are big files and usually stored together during the backup.

Reclamation

The reclamation process copies the valid data of a tape to a new tape, if the amount of valid data has fallen short of the defined threshold.

This recycling of fragmented tapes is useful for all storage pools with irregular expirations.

Reuse delay

Reuse delay defines the period for which Wtapes are kept after all data on the tape has expired.

We suggest that this period corresponds to the retention period of your TSM database backups. When restoring an old TSM database, you ensure that no tapes have been reused in the meantime.

Bright idea!

For the design of the TSM server storage for an SAP R/3 environment, we suggest the following storage pool configuration:

- ▶ Use *disk pools* for all incremental backups, such as AIX files and SAP R/3 executables. This enables you to perform several schedules in parallel without being restricted by the number of available tape drives.

The data is migrated to tapes during periods of lower backup activity. This can be customized by adapting thresholds for administrative schedules (refer to Section 7.3.5, “Scheduling” on page 205). This procedure is necessary to get an empty disk pool for the next backup run and to copy all of the data from the tape pool to the associated copy pool.

- ▶ Use *tape pools* for all SAP R/3 database backups. For large files, the performance is better if using tape pools with fast tape drives rather than using disk pools. Normally, tape pools used by TSM archive will be defined to not use reclamation, because most of the data on one tape is backed up with the same retention period.
- ▶ *Copy pools* are recommended for at least all data that has been saved by incremental backups. In this case, you always have a copy available for all backed up data, and thus reduce the potential for data loss due to media failure.

If you want to use TSM Disaster Recovery Manager for your disaster recovery management, you have to use copy pools for all pools that have to be checked out. The possibilities to be prepared for a disaster will be discussed in Section 7.4, “Preparing for disaster” on page 210.

- ▶ The creation of copy pools for tape pools used for SAP R/3 database backups depends on your environment and requirements. As a rule of thumb, you need twice the number of tape drives as used during the backup or twice the time. Other factors you have to take into account are the frequency of your backups or your disaster recovery plan.

We suggest that you define two storage pools for the database backups and use them in an alternating fashion to enable vaulting or the distribution to two libraries (refer to Section 7.4.2, “Dual library and TSM server/library sharing” on page 213). You have to take into account that, depending on the number and size of requested log files, a recovery procedure can be very time-consuming.

- ▶ TDP requires one storage pool per copy of an archive log file. SAP recommends that you have two copies on different media. Therefore, two storage pools are defined for this backup object.
- ▶ Define an extra tape pool for the ability to store data over a long period, for example, several years. Such special requests outside the regular scheduling are also called on demand backups.

Table 7-3 contains examples for storage pools to back up an SAP R/3 environment based on the suggestions made above.

Table 7-3 Storage pools

Backup object no.	Storage Pool	Next Pool	Copy Pool	Collocate	Reclaim
1,2,3,6,7,8	STG.DISK	STG.TAPE			
1,2,3,6,7,8	STG.TAPE		STG.TAPE.CP	yes	yes
1,2,3,6,7,8	STG.TAPE.CP			no	yes
4	STG.BACK.1			no	no
4	STG.BACK.2 ^a			no	no
5	STG.ARCH.1			no	no
5	STG.ARCH.2			no	no
all	STG.EVER		STG.EVER.CP ^b	no	no
all	STG.EVER.CP			no	no

a. Only necessary for disaster considerations

b. Only necessary for disaster considerations

7.3.3 Domain, management classes and copy groups

The virtual nodes for all SAP R/3 systems will be grouped together in one policy domain. In our example, the policy domain is called SAP. The default management class and all additional management classes are defined in the corresponding policy set.

TSM management classes and the associated copy groups are used to map the different requirements of the backup objects, defined in Section 7.2, “Backup objects and methods” on page 191, to the TSM server storage configuration. Which objects can be grouped together in which management classes depends on the requirements of the specific installation. You need different management classes access different storage pools in order to define different characteristics for expiration.

In our example, which is shown in Table 7-4 on page 202, either a backup copy group or an archive copy group is defined for a management class. To represent the association between the management class and its copy group and the corresponding storage pool, you may use a similar name with a unique prefix in each case for all of them.

Bright idea!

A special case is the management class MC.EVER, which handles on demand backups. You have two possibilities to perform such backups for storing data over a long period.

- ▶ One alternative is to use TSM archive with an appropriate retention period. In this case you have to take into account that the TSM archive will follow links, which are often used in SAP R/3 file systems.
- ▶ Another alternative is to create a backup set, which copies the last active version of backed up files from the TSM server storage. You need at least one tape per backup set that can be restored independently from the TSM server.

Table 7-4 Management classes and copy groups

Backup object no.	Management class and type	Backup versions				Archive	Storage pool
		vere	verd	rete	reto	retver	
1,3	MC.IMG backup	1	1	31	31		STG.DISK
2,3	MC.AIX backup	3	2	31	92		STG.DISK
6,7,8	MC.SAP backup	5	2	31	92		STG.DISK
4	MC.BACK.1 archive					30	STG.BACK.1
4	MC.BACK.2 ^a archive					30	STG.BACK.2
5	MC.ARCH.1 archive					30	STG.ARCH.1
5	MC.ARCH.2 archive					30	STG.ARCH.2
all	MC.EVER ^b archive (backup)					nolimit	STG.EVER

a. Only necessary for disaster considerations

b. Necessary copy groups, depending on your requirements

An explanation of the expiration characteristics follows:

- | | |
|-------------------------------|---|
| vere (version exist) | Maximum number of stored versions while a file exists |
| verd (version deleted) | Maximum number of stored versions after a file has been deleted |

rete (retain extra)	Retention period (days) of extra versions of an inactive file
reto (retain only)	Retention period (days) of last version of an inactive file
retver (retain version)	Retention period (days) of an archived file

For example, to back up all the SAP R/3 relevant files to TSM, the management class MC.SAP is defined with the backup copy group CP.SAP. A maximum of five versions of these files are stored in the TSM server, reduced to two versions if the file is deleted. This is relevant, for example, for SAP R/3 profiles or to have several versions of SAP R/3 kernel executables. Many other files, such as SAP R/3 job logs, have unique names and only one version exists normally. If the file is inactive, all extra versions will be stored for 31 days (1 month). The last version will be deleted after 92 days (3 months).

Pitfall ahead!

Before implementing your TSM solution, you have to analyze the different requirements of the SAP R/3 systems of your landscape, such as the development, the quality assurance, and the production system. After this analysis, create a table like the one shown in Table 7-4 on page 202 for all the different SAP R/3 systems. It can be necessary to define more management classes than those mentioned above to fulfil the different requirements.

As an alternative to defining a retain version for the TDP related management classes, you can set it to no limit and use the version control mechanism of TDP. We do not recommend you use this mechanism, because it makes the control of your storage management much easier if you have a clear idea when your data will expire. Additionally, it is simpler to perform on demand backups if not using TDP version control. On the other hand, the definition of retain version requires a proper check of the state of your SAP R/3 backup schedules to detect failed backup operations early enough.

The relationship between backup objects, management classes, and through copy groups associated storage pools of our implementation example is shown in Figure 7-5 on page 204.

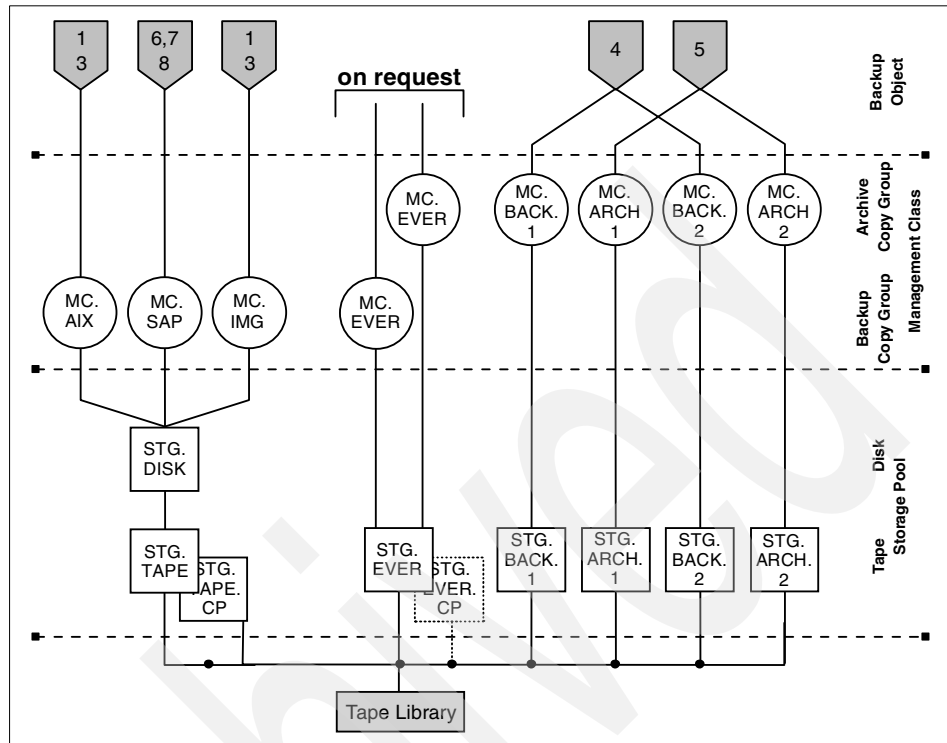


Figure 7-5 Backup objects, management classes, and storage pools

7.3.4 Devices

Pitfall ahead!

To meet the requirements regarding reliability, manageability, and efficiency, we strictly recommend that you use tape libraries as backup devices. Other solutions, such as single tape drives and tape stackers, impose a higher administrative effort and exposes your data to human faults.

The available technology for tape drives and libraries applicable in SAP R/3 environments is described in Section 3.1.3, “Storage subsystems” on page 46. If you use more than one library, for example, with library sharing, you have to define at your TSM server one device class per library. Library sharing is part of the preparation for disaster discussed in Section 7.4, “Preparing for disaster” on page 210.

7.3.5 Scheduling

The TSM scheduler facility provides a central point of control for scheduling and checking all your backup operations in your environment. On the other hand, it is possible to use the SAP R/3 scheduler (using SAP R/3 transaction DB13 and SM37) or the AIX crontab. These alternatives can only take care of your backups at a server level. Therefore, we suggest that you use the TSM scheduler to fulfill the requirement for automated client backup operations.

TSM scheduling is divided into two different categories: administrative and client scheduling. An administrative schedule performs an action on the TSM server, while the client schedule is executed on a TSM client.

Client schedules

In our SAP R/3 environment, a client schedule is used to trigger a backup operation on a group of TSM clients. On the client machine, a scheduler daemon has to be active to be ready to perform the actions defined by the scheduler.

With the definition of virtual nodes, one scheduler daemon per virtual node is necessary. The daemons can be started by scripts inserted in /etc/inittab or crontab. Special considerations are necessary if you implement a fault-tolerant model from the SAP R/3 point of view, as described in Section 3.3, “Building a fault-tolerant model” on page 77. Some of the virtual nodes are associated with resources taken over by a standby node in case of a takeover. Therefore, the HACMP application server scripts have to make sure that the daemons will be stopped on the node releasing the resources, and started on the node acquiring the resources.

Client schedules can fail due to differences in time settings on the TSM server and TSM client. By implementing, for example, NTP time synchronization, you prevent client schedules from failing by exceeding the TSM defined time frame for execution (see also Section 10.2.2, “Time synchronization” on page 284).

For more information on client schedules and their components, see Table 7-5.

Table 7-5 Client schedules

Backup object no.	Schedule name	Frequency	Association
1	<hostname>_IMG	Weekly (Sunday)	All AIX server
2	ALL_AIX_INC	Daily	All AIX server
3	ALL_LVM_INFO	Weekly (Sunday)	All AIX server
6,7,8	ALL_SAP_INC	Daily	All SAP server

Backup object no.	Schedule name	Frequency	Association
4	BKP_DB_<SID>_1	Daily (alternating between the storage pools)	Per SAP DB server
4	BKP_DB_<SID>_2 ^a		Per SAP DB server
5	ARC_DB_<SID> ^b	Two times the day	Per SAP DB server

a. Alternative to divide online/offline backups or to use two libraries or different media

b. Depends on the frequency of changes in your database

Administrative schedules

Any action that is used on a regular basis to manage the TSM server storage should be defined as an administrative schedule. Table 7-6 contains suggestions for important schedules according to the defined environment. The sequence is crucial when implementing disaster proven concepts discussed in Section 7.4, “Preparing for disaster” on page 210.

Table 7-6 Administrative schedules

Schedule name	Description	Frequency sequence
MIGR_STG.DISK_B	Sets the migration threshold of storage pool STG.DISK to 0 percent in order to force migration to tape.	Daily, after all backups have finished
MIGR_STG.DISK_E	Sets the migration threshold of pool STG.DISK back to 90 percent.	Daily, after migration has finished
COPY_STG.TAPE	Copies of pool STG.TAPE to pool STG.TAPE.CP.	Daily, after migration has finished
BKP_DBTSM_TAPE	Backs up the TSM database to tape.	Daily, after copy has finished
DEL_OLD_DB_BKPS	Deletes old TSM database backups from the history file.	Daily, if TSM database backup was successful
EXPIRE_INVENTORY	Executes expired TSM inventory.	Daily, after TSM database backup has finished

Schedule name	Description	Frequency sequence
RECL_STG.TAPE_B	Sets the reclaim value of pool STG.TAPE to 51 percent to force the tape reclamation	Daily, after expire inventory has started
RECL_STG.TAPE_E	Set the reclaim value of pool STG.TAPE back to 100%	Daily before backup operations start
RECL_TAPE.CP_B	Set the reclaim value of pool STG.TAPE.CP to 51% to force the tape reclamation	Daily after expire inventory has started
RECL_TAPE.CP_E	Set the reclaim value of pool STG.TAPE.CP back to 100%	Daily before backup operations start

On demand backups and restores

During normal operation in an SAP R/3 environment, you can perform all your backups and archives using client schedules. But in certain situations, you want to be able to start a backup or restore of certain data individually. For this situation, it is possible to start individual backups or restores that do not have to be triggered by schedules.

In general, it is a good idea to back up each component of your environment before you change some important features of it. Thus, if you upgrade AIX, you should also make a **mksysb** system backup before you start the upgrade.

Some examples of times when you want to make individual backups of the database are:

- ▶ Before and after an upgrade of the database
- ▶ Before and after an upgrade of SAP R/3
- ▶ Before and after an Euro Conversion
- ▶ Before and after changing the hardware

7.3.6 TSM database and recovery log

As already stated in Section 7.2.7, “TSM database and recovery log” on page 195, the TSM database is crucial for your backup and restore environment. For the implementation of a TSM database with high reliability, we make the following suggestions:

- ▶ Operate the TSM database in the *roll-forward* mode.

- ▶ Perform regular backups of the TSM database.
- ▶ Use different disks for the following TSM server parts:
 - TSM database
 - TSM recovery log
 - Disk storage pools
- ▶ Mirror all these parts at the AIX LVM or TSM level.

AIX mirroring may fit in your overall definition of your environment; mirroring on the TSM level is easier to handle from a TSM point of view.
- ▶ Install the parts on their own volume group, including all configuration files.

If you operate the TSM database in roll-forward mode, you are able to perform a roll-forward recovery or a point-in-time recovery. In contrast to the operation in normal mode, you require more space in the recovery log to record all the activity.

In roll-forward mode either a full or an incremental TSM database backup resets the recovery log back to empty. A TSM database used only for SAP R/3 systems usually has a size of 3 to 4 GB. Hence, it is always possible to perform a full TSM database backup instead of incremental. We recommend that you perform full backups, as the handling of incremental backups is more complicated in case of a restore and recovery.

TSM database backups should be performed regularly after essential changes in the database. These backups are a part of the administrative schedules (see Section 7.3.5, “Scheduling” on page 205). By setting a database backup trigger, you ensure that the recovery log does not run out of space and that a backup is performed after a certain amount of transactions within the TSM database. If the database backup trigger automatically runs backups more often than you want, you have to increase the value of the trigger or the recovery log size.

Bright idea!

We recommend that you back up the TSM database to tape (to enable vaulting) and to disk. When backing up to disk, it should be directed to an NFS-mounted file system from another server. The TSM database backup to disk can be coupled with storing the following information at the same location:

- ▶ Volume history file generated with the **backup volhistory** command
- ▶ Device configuration file generated with the **backup devconfig** command.
- ▶ TSM server options as a copy of the file `dsmserv.opt`
- ▶ TSM database and recovery log setup information as a copy of file `dsmserv.dsk`

For example, this automated procedure can be implemented with a script started as a client schedule for the TSM server. In the basic model, which is described in Section 3.2, “Building a basic model” on page 65, the NFS-mounted file system can be a part of the administration server *Admin*.

In the disaster-tolerant model, which is described in Section 3.4, “Building a disaster-tolerant model” on page 86, server lowTSM can backup its information to a file system located on the server hiTSM and vice versa. This is described in more detail in Section 7.4.2, “Dual library and TSM server/library sharing” on page 213.

7.3.7 Using NIM as part of the TSM concept

NIM (Network Installation Manager) is a very powerful tool covering all possible installation, backup and recovery, and maintenance tasks of the AIX operating system. In the scope of this section, we consider the capability of network boot and installation of an operating system backup (**mksysb**). TSM can only restore files to a system that is installed and has the TSM client installed. Thus, the capability of NIM to supply each server in the SAP R/3 environment a bootable image is essential.

A general description of NIM functionality is provided in Section 3.2.6, “Administration server (Admin)” on page 73.

Figure 7-6 illustrates our suggestion of how such a solution can be implemented.

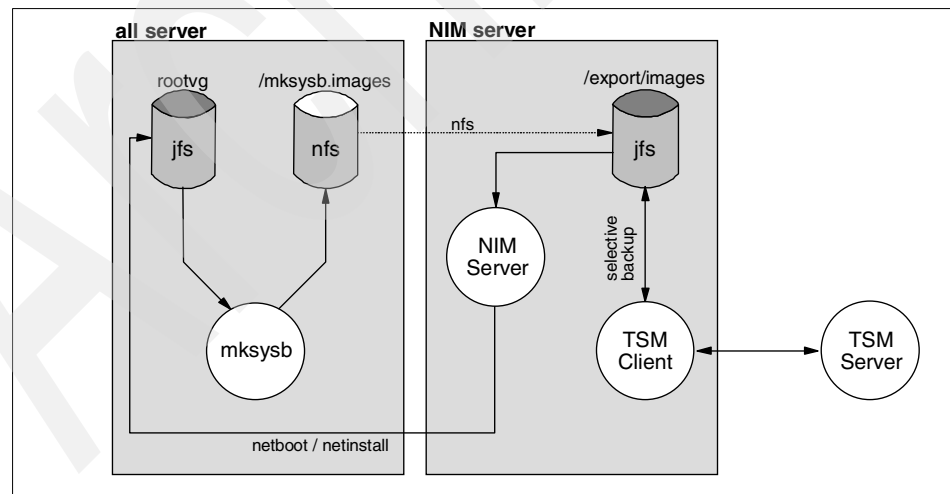


Figure 7-6 NIM solution for restore of system backups (mksysb)

Create an AIX system image with the **mksysb** command in a file system named `/mksysb.images`. This file system is located on the NIM server and mounted with NFS. Sufficient disk space to store the system images of all servers is a prerequisite. On the NIM server the system images will be backed up with TSM selective backup to be able to restore them in the case of a disaster.

Pitfall ahead!

We recommend that you hold at least two versions of the system images of all clients in the file system on the NIM server. This ensures that after a failure during the creation of system backup, a valid file is still available. The current version can be linked to a file name to which the NIM resource definition refers.

Configure a NIM server according to the *AIX V4.3 Network Installation Management Guide and Reference*, SC23-4113. When defining the installation network, refer to the backup network, as explained in Section 3.2.4, “The networks (switch)” on page 70 and Section 6.3.2, “The backup network” on page 165.

For the restore, in addition to the created master, you also need at least the definition of the client machine and the association to the resources of the type `lpp_source`, `spot`, and `mksysb`. The procedure to initiate the network boot of the server depends on the installed IBM *@server* pSeries model. Afterwards, install the image as normal.

To cover the restore of your NIM server, you have to regularly create bootable system backups to tape. This applies also to the Control Workstation (CWS) of a RS/6000 SP, which has the same NIM functionality. Another alternative is to implement a secondary NIM server, as described in Section 7.4.2, “Dual library and TSM server/library sharing” on page 213.

7.4 Preparing for disaster

Section 3.2.5, “The backup subsystem” on page 71 and Section 3.4.3, “Dual library and TSM server” on page 89 give suggestions on how to achieve disaster safety or disaster tolerance for your backup and recovery infrastructure. There are many possibilities for how to set up disaster safe or disaster-tolerant solutions. In this section, we discuss, in more detail, recommendations on how to implement and operate TSM in two different solutions.

7.4.1 Vaulting

The implementation of many SAP R/3 environments is based on the basic model introduced in Section 3.2, “Building a basic model” on page 65. One solution to achieve disaster safety for such an environment is to vault parts of your backup data and store it outside the computing center. For this procedure, you have to define dedicated storage pools in order to separate tapes for checkout from tapes that remain in the library. We use the terms of checkout tapes and resident tapes to differentiate between these two kinds of tape.

Bright idea!

There are different possibilities to create checkout tapes:

1. Create copy storage pools for all desired primary storage pools. Decide which of the copy storage pools are candidates for your checkout pools.
2. Write two copies into different primary storage pools during the backup operation. This can be done using TDP for R/3 with the configuration of two or more log copies. One of the storage pools is your candidate for checkout.
3. Back up your data, alternating to two different primary storage pools. This is only useful if the object is a database and can be recovered forward after restoring this backup. One of the storage pools is your candidate for checkout. This solution is only sustainable if the necessary database log files can be applied in a reasonable processing time, which should be defined in the service level agreement.

Another alternative is to operate a shadow database at a remote location. This topic is discussed in Chapter 9, “Shadow database” on page 259. In our example, which is described in Table 7-3 on page 201, you can use the following storage pools to check out tapes for vaulting:

1. Copy storage pool STG.TAPE.CP for all your miscellaneous files, such as AIX and SAP R/3 executable, AIX images, interface files, and TMS files
2. Storage pool STG.BACK.2 for the SAP R/3 database and log files
3. Storage pool STG.ARCH.2 for one copy of the archived log files of the database

In the worst case, you have to apply database log files of one day based on the scheduling defined in “Client schedules” on page 205. Running a recovery procedure may be very time-consuming, depending on the number and size of requested log files. These considerations have to be checked against the requirements of the service level agreement.

Additionally, you need the appropriate TSM database backups and configuration information and a tape with a bootable system backup of your TSM server and your NIM server. Thus, you have all the media available at the off-site location needed to rebuild your TSM server environment in case of a disaster.

Pitfall ahead!

The sequence of your administrative schedules is a very important part of keeping a set of media checked out that represents a consistent state of your TSM server. Table 7-6 on page 206 contains a suggestion for a practical sequence.

Tape management

You need a properly coordinated concept to enable the tracking of the tapes stored outside the computing center. These tapes, belonging to TSM storage pools, are still controlled by the TSM server. To handle a tape management procedure, such as needed for check in and check out, and for searching the required tapes, you can create your own scripts or you can evaluate the Tivoli Disaster Recovery Manager (DRM). The DRM supports you in automating the handling of tape movement and in prescribing actions for disaster recovery. A prerequisite for the usage of the DRM are copy storage pools you want to be checked out. Thus, our suggestions are not applicable for using DRM, as we use the primary storage pools STG.BACK.2 and STG.ARCH.2 for checkout.

Further information about vaulting and the Tivoli Disaster Recovery Manager is available in the publication *Tivoli Storage Manager for AIX Administrator's Guide, Version 4 Release 1*, GC35-0403 and the redbook *Tivoli Storage Management Concepts*, SG24-4877.

Recover from a disaster

The following steps are necessary to recover the TSM server from a disaster:

1. Select the hardware for restore.
2. Detect the latest backup tapes.
3. Restore the TSM server, including TSM LPPs and the device drivers for the library with the help of your NIM server or from a bootable tape backup.
4. Restore all configuration information files of the TSM server backed up together with the TSM database.
5. Recreate and initialize the files for the TSM database volumes and recovery log volumes with the **dsformat** command.
6. Restore the TSM database from the detected backup tape or from a backup copy directed to disk.
7. If applicable, redefine the missing storage pools.
8. If necessary, restore other nodes with NIM and TSM.

7.4.2 Dual library and TSM server/library sharing

In SAP R/3 environments with higher requirements regarding availability in case of a disaster, we recommend that you implement two shared tape libraries and two TSM servers distributed to different locations (see Figure 7-7 and Section 5.4.2, “Design considerations for a SAN” on page 146).

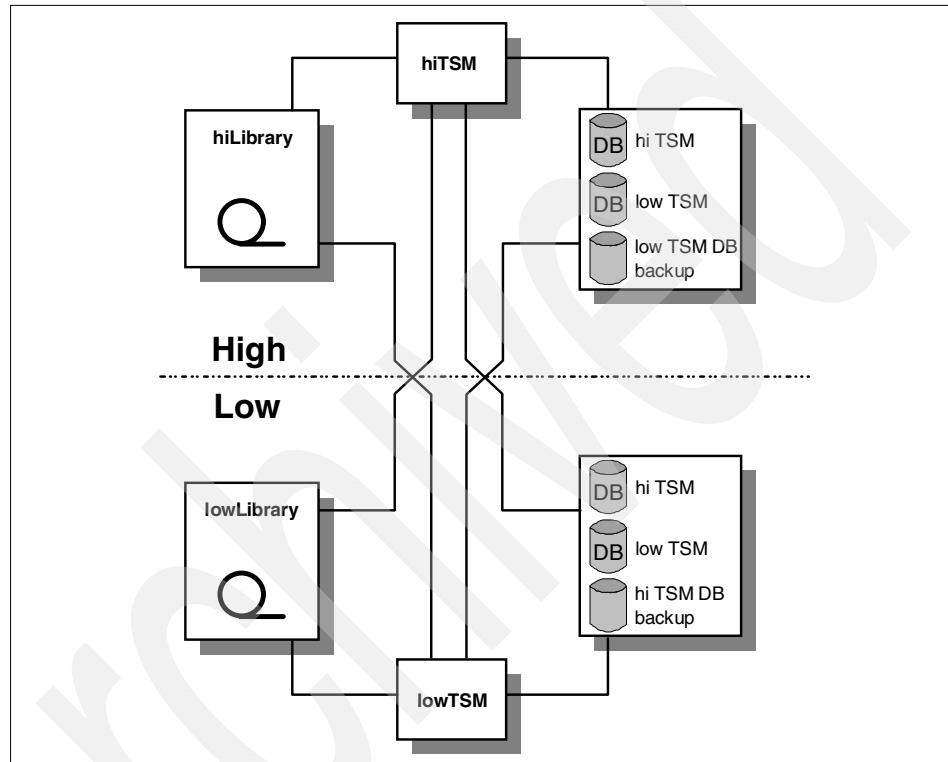


Figure 7-7 Dual library and TSM server

A library is defined as shared if it is to be shared with another Tivoli Storage Manager server over a storage area network (SAN) or a dual SCSI or FC connection to library drives. Depending on the type of library device, there are differences in how to define and use the library with TSM. If you implement the library sharing based on the library manager and library client components, lowTSM should act as the library manager for the low library and hiTSM should act as the library manager for the high library. With this component distribution, you enable the library manager operations to the remaining library in case of a disaster. Setting up SAN shared libraries is discussed in detail in the redpaper *Tivoli Storage Manager SAN Tape Library Sharing*, REDP0024.

The configuration components of the two TSM servers are described in Section 3.4.3, “Dual library and TSM server” on page 89. In addition to the shared libraries, you have two external storage subsystems attached to both servers. Each TSM server is keeping its TSM database and recovery log, the disk storage pools, and all configuration information in a separate volume group on the external storage. All this data is mirrored to the storage subsystem located in the other computing center. Considering the restrictions discussed in Section 4.10.2, “A disaster-tolerant ESS configuration” on page 134, this configuration enables a manual takeover of the TSM server instance to the other TSM server.

In the next section, we discuss our suggestions on how to implement and operate the suggested TSM solution.

Normal operation

Bright idea!

The TSM server lowTSM is used to back up the production SAP R/3 database. Daily alternating client schedules back up the database one day on the low library, the other day on the high library. Log files are always archived simultaneously to both libraries. Referring to our example, that means the storage pool STG.BACK.1 is associated with the device class of the low library and STG.BACK.2 is associated with the device class of the high library. The same applies to the storage groups STG.ARCH.1 and STG.ARCH.2.

Although a restore of the SAP R/3 database is not necessary in case of a disaster, because the production database is mirrored on both storage subsystems, as shown in Figure 7-8 on page 215, the solution of alternate backups to both libraries and the available log files guarantees the possibility to restore and recover the database from the remaining library, even in the case of a double fault.

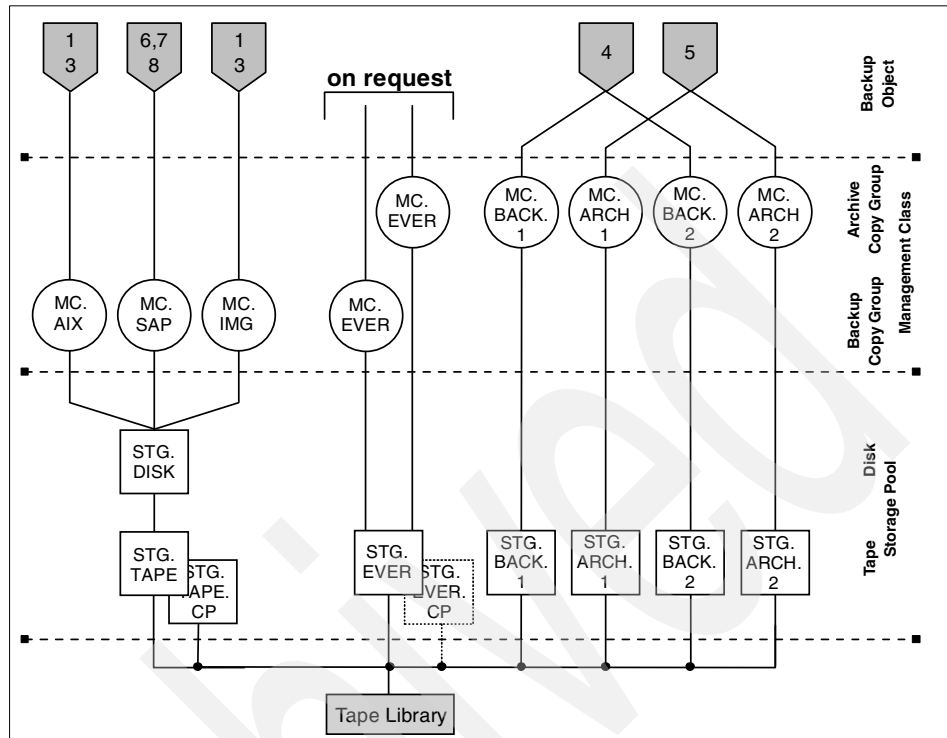


Figure 7-8 Distribution of storage pools to shared libraries

During normal operation, the other objects will be backed up incrementally to the disk storage pool STG.DISK as defined. After migration to the tape pool STG.TAPE in the low library the backup data will be copied to the copy pool STG.TAPE.CP located in high library.

This normal operation is presented in Figure 7-9 on page 216.

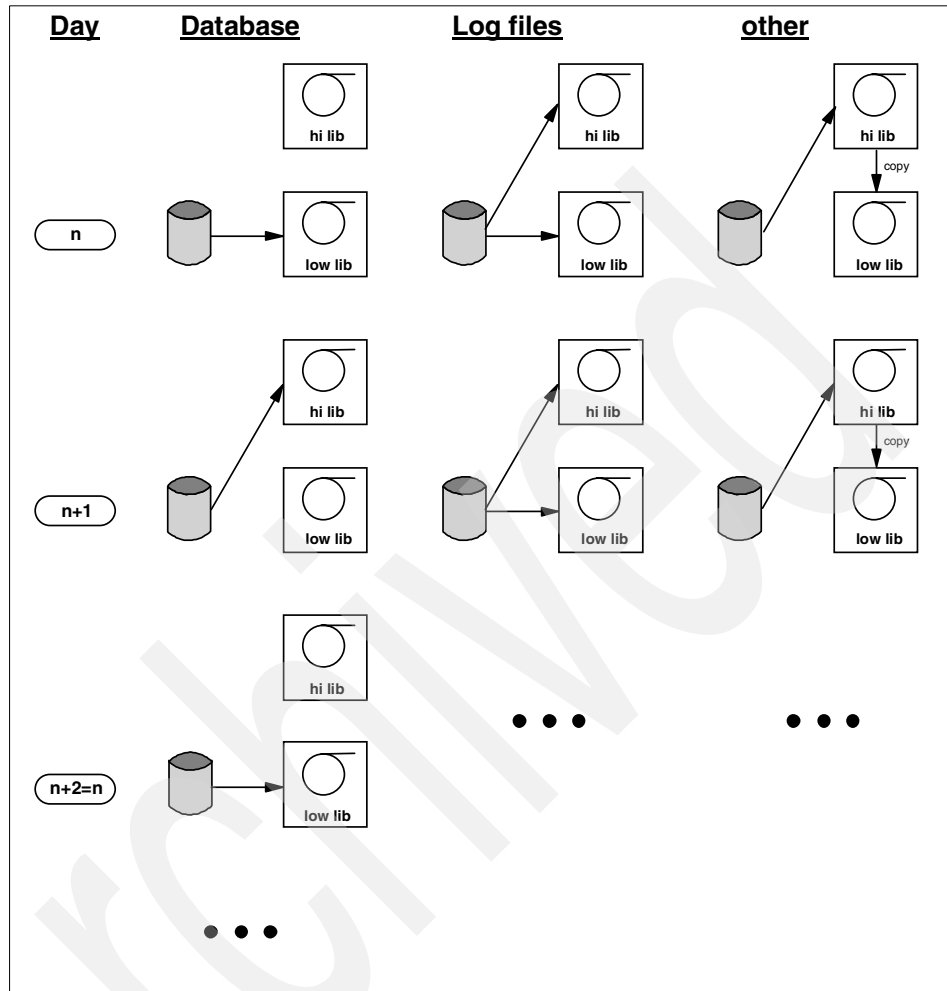


Figure 7-9 Schematic picture of backup paths to the libraries lowlib and hilib

The backups of systems other than production will be treated in the same way. Normally, there would be a number of non-production systems that are implemented according to the basic model (refer to Section 3.2, “Building a basic model” on page 65) distributed to both computer centers. You can distribute the necessary storage pools to both libraries to balance the load between both libraries. As a minimum solution to be disaster proven, you have to back up all objects to the library of the remote side.

For example, the backups of a server in computer center Low would be stored by the TSM server lowTSM on tapes residing in the library in computer center High and vice versa. Thus, in case of a disaster, you would lose the servers of the affected center, but not their backed up data.

Although the databases of both TSM server instances are mirrored on both storage subsystems, they should be backed up on tapes of both libraries and to disks in the remote storage subsystem in each case. Having some backup generations of a TSM database available even protects you from corruption of the database brought about by an application failure.

This solution normally removes the necessity to check out tapes of production backups from any library. It is still possible to introduce additional vaulting for one or both libraries, but only for the highest security demands or in environments where a coincidence could lead to a loss of both libraries at the same time.

Recovering from a disaster

Assume that the computer center Low is lost. The HACMP cluster for SAP R/3 production takes over the SAP R/3 database and central instance to the node APserv. The TSM server instance from node lowTSM is transferred manually to the node hiTSM according to the manual procedure described in Section 8.2, “Manual and automatic takeover” on page 224. Now you can start the TSM server instance of lowTSM. If the TSM database is corrupted, you can use the available backups on tape or disk for the restore. After some minor configuration adaptations within the TSM server instance lowTSM (reflecting the loss of the tape drives and tapes in the lower library), restores are possible for all backup objects from the storage pools of lowTSM associated with the high library.

To enable backup operations again, some adjustments are necessary. Starting from the condition that the disk storage pool STG.DISK was as well mirrored to computing center high, you will still miss the storage pools associated with the low library. You have to redefine and restore the storage pools STG.TAPE and STG.EVER from their copy storage pools. This is a long running process! The pools STG.BACK.1 and STG.ARCH.1 can be recreated or you have to adjust the appropriate TSM server configuration. The approach depends on the available library resources in the computing center. The same applies to the storage pools of the TSM server instance of hiTSM associated with the low library.

Handling of operating system images

The NIM server that normally is implemented on the Admin server should be replaced by two NIM servers, one on each TSM node. One of these NIM servers has to be defined as a *primary NIM server*, from which the *secondary NIM server* is synchronized. All nodes are defined as clients and normally their system images are backed up to the primary NIM server in the same way, as described in Section 7.3.7, “Using NIM as part of the TSM concept” on page 209.

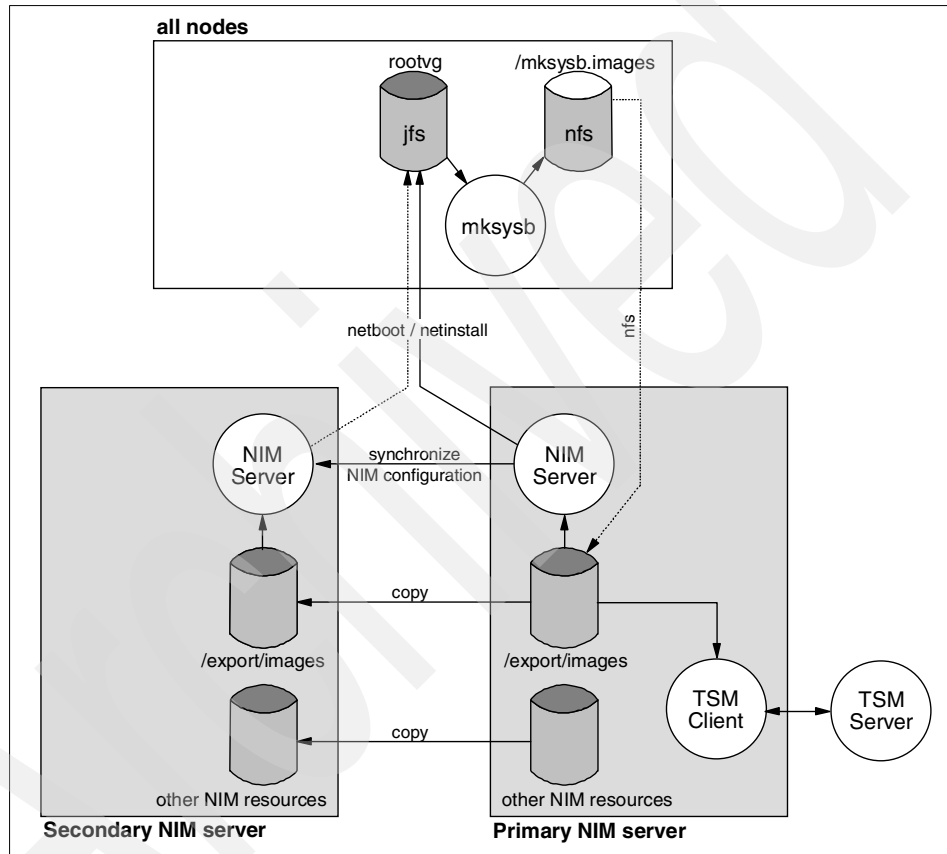


Figure 7-10 Primary and secondary NIM server

To keep the secondary NIM server in sync, you have to copy all the defined NIM resources, such as LPP sources and system backup (**mkysyb**) images, to this server and you have to synchronize the NIM configuration information. We suggest that you update the NIM configuration on the secondary server by backing up the NIM database on the primary NIM server and regularly restoring it

on the secondary NIM server. Afterwards, you have to adjust and activate the NIM master configuration on the secondary NIM server to represent the different server. The primary NIM server is defined as a client machine on the secondary server and vice versa.

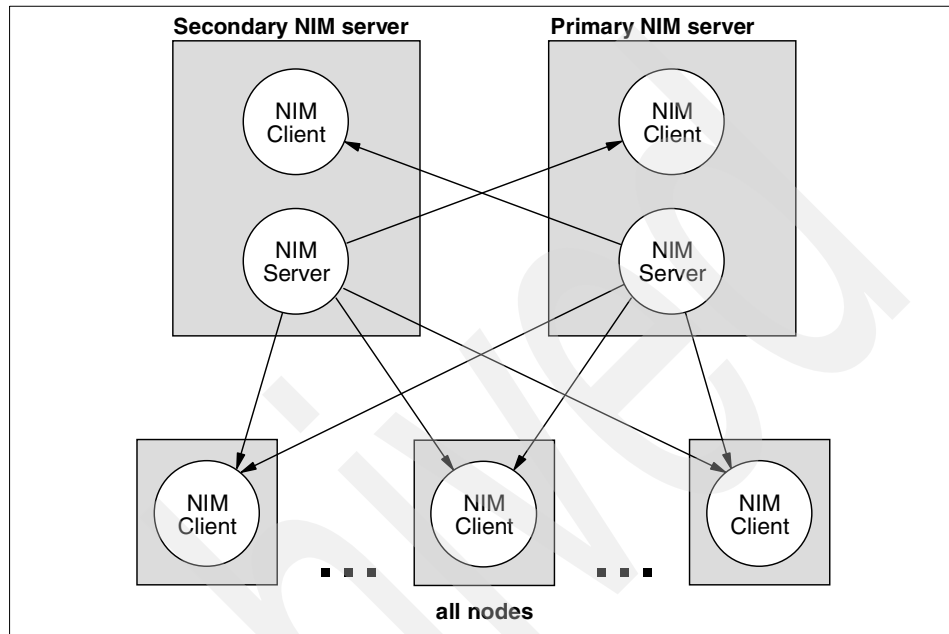


Figure 7-11 NIM client server relationships

Attention: A backup of a NIM database should only be restored to a system with a NIM master fileset that is at the same level as the level from which the backup was created. Therefore, maintain the same level of AIX software on these systems.

With such a synchronization, both NIM server have access to all resources and have all other machines defined as NIM clients. This solution enables the reconstruction of AIX on lost servers through the remaining NIM server also after a disaster.

7.5 Test and operation of the implemented solution

One of the challenges that you must consider when designing a backup and recovery solution is to make sure that it is tested. You have to test the whole implemented solution, not only parts of it. Only complete restore and recovery tests enable you to build up, document, and verify the necessary procedures usable in case of an error requiring a recovery.

Pitfall ahead!

We recommend regular restore and recovery tests to ensure the operability and the training of your staff. A good approach is to create all documents, backup media, special hardware requirements, and installation scripts and send them to an off-site location as a "Disaster Recovery Starter Kit." Then, once a year, perform a complete recovery test to make sure the documents are accurate for recovery. You should then incorporate any changes that were uncovered during your test.

Tests are essential after changes of the software components involved in your backup and recovery solution, such as AIX, TSM, TDP for R/3, and the SAP R/3 tools **brbackup**, **brarchive**, and so on. With the test, you have to verify:

- ▶ The completeness of your backups
- ▶ The procedures for restore and recovery
- ▶ The duration of backup and recovery

During the operation of the TSM Solution, you have to make sure that all schedules have been successfully completed. See Section 13.5, "Backup" on page 429 for a detailed description of tasks that have to be performed regularly.

High availability

In a nutshell:

- ▶ Use HACMP to automatically respond to hardware failures.
- ▶ Customize HACMP to fulfill your requirements.
- ▶ Combine SAP R/3 central instance and database on one cluster node.
- ▶ Avoid a dedicated server network in an HACMP cluster.
- ▶ Test each HACMP cluster regularly and especially after changes in the cluster configuration.

This chapter provides information on design, implementation, and configuration of highly available SAP R/3 environments.

In the first part, we introduce techniques of implementing fault tolerance for applications through redundancy in hardware components. Using these techniques, a manual takeover scenario can be implemented. In environments where an automatic takeover of an application is required, the software package High Availability Cluster Multi Processing (HACMP) for AIX can be used.

We discuss the implementation of HACMP for SAP R/3 and consider the implementation of a special treatment for SAP R/3 NFS mounts. Furthermore, we describe important tasks of SAP R/3 application start and stop scripts through a step list.

We then provide state-transition diagrams that are an efficient way to document and test a cluster. These tests have to be performed after the implementation of the cluster and after every change to the environment.

In the last part of this chapter, we discuss the extension of the presented solution in the case where a disaster-tolerant model has to be used.

This chapter covers the highlighted area in Figure 8-1, which is derived from the SAP R/3 infrastructure model that is used throughout the book.

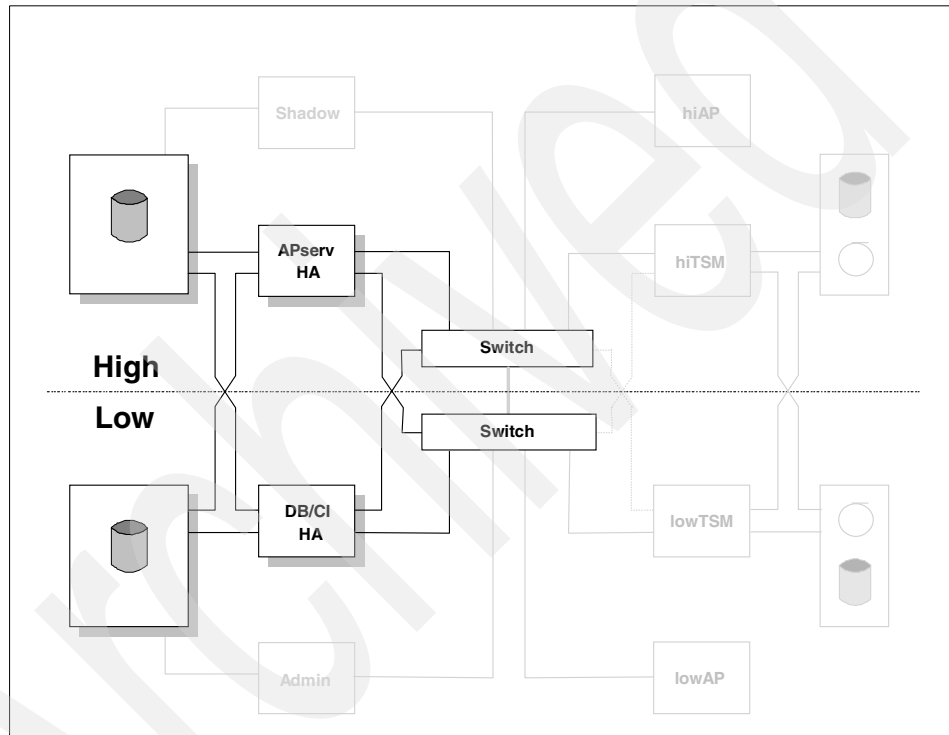


Figure 8-1 High availability

8.1 Introduction

Fault tolerance is a characteristic of a system that allows an application, such as SAP R/3, to run without unplanned downtime. This characteristic is discussed in depth in Section 2.2.3, “Definition of high availability and downtime” on page 13.

It is not possible to avoid unplanned downtime completely, but it is possible to remove all single points of failure in an environment to achieve a high availability of the server and the application.

In the previous chapters, we have discussed the architecture and the requirements for storage and networks, but we have not explained the details about the implementation of fault tolerance.

For example, if a network adapter fails, a second adapter in the server must be configured in the same way as the failing one in order to offer the same service. If one server fails, a second server must be similarly configured and the application must be able to start on the second server. This is called a *takeover* of a service.

In Figure 8-2, you find a comprehensive overview of a hypothetical cluster.

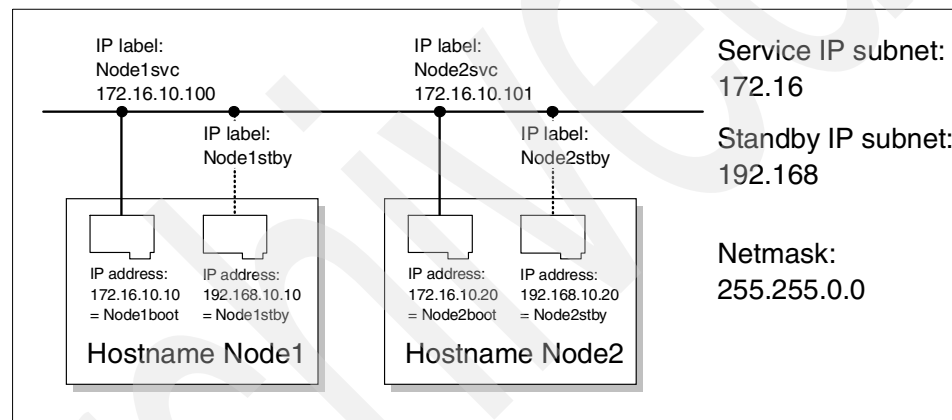


Figure 8-2 IP labels and subnets

Every client communicates over the network with both servers. The server Node1 has a dedicated Internet Protocol (IP) address over which he provides its service to the clients. This dedicated IP address, the so called *service IP address* Node1svc, is bound to one of the adapters on the network. The service IP label is the DNS (Domain Name Services) name of the service IP address, which is usually equal to the host name.

The server Node1 has two IP addresses for the two network adapters at startup time. The IP address 172.16.10.10 belongs to the IP label Node1boot and the IP address 192.168.10.10 belongs to the IP label Node1stby. For server Node2 similar labels for boot and standby are applied. The subnet mask of the service and the standby subnet must be the same (255.255.0.0 in our example).

The two cluster nodes with the host names Node1 and Node2 are both connected to the service network, over which these hosts provide their service. The standby network adapters are connected to the same physical network as the service network adapters in order to be able to take over the service address. However, the standby adapters have IP addresses that belong to an own IP subnet. The IP labels (the already mentioned DNS registered names) are visible for every IP client in the network domain. Each IP label has to be assigned to an adapter.

8.2 Manual and automatic takeover

The basic requirement for a failover is the possibility of a manual takeover of the application between the two servers Node1 and Node2 in our cluster. In order to prepare the environment for an application start or stop, the steps have to be planned in the right sequential order:

For the start of an application on Node1, the following steps have to be executed:

1. Assign the service IP label Node1svc.
2. Assign the disk resources.
3. Start the application.

For the stop of an application on Node1 the following steps have to be executed:

1. Stop the application.
2. Release the disk resources.
3. Release the service IP label Node1svc.

For a manual takeover, the step-by-step list of both sequences leads to the necessary actions. In this case, the application is taken over from the primary server Node1 to the backup server Node2:

For Node1:

1. Stop the application.
2. Release the disk resources.
3. Release the service IP label Node1svc.

For Node2:

1. Assign the service IP label Node1svc.
2. Assign the disk resources.
3. Start the application.

For a takeover in the opposite direction, the steps have to be applied in a vice versa fashion.

Bright idea!

These steps have to be executed manually in order to perform a manual takeover between the two servers in the cluster. It is advisable to put all these tasks in one or more Korn shell scripts to reduce the possibility of making mistakes during execution.

When these scripts are completed and well tested, an operator or administrator is able to initiate a takeover with one command. This is acceptable in a non-critical environment, if the operator is skilled enough to recognize an error. After the recognition and the identification of the error, the operator can react according to the steps listed above.

With an insufficient number of system operators or in a 24x7 environment, it is not possible to wait one or two hours before a failure gets recognized and the according manual procedures are performed. Therefore, a cluster software has to be introduced. For an automatic takeover, a cluster software must be used with two main features:

- ▶ Monitoring of the hardware in order to recognize errors
- ▶ Reacting depending on the failing components

A cluster software is able to manage resources, but has to be customized to handle an application takeover.

The scripts for the start, stop, and takeover of the application are under the control of a cluster software. Figure 8-3 on page 226 gives a rough overview of the two possibilities for a takeover: a manual takeover or an automatic takeover.

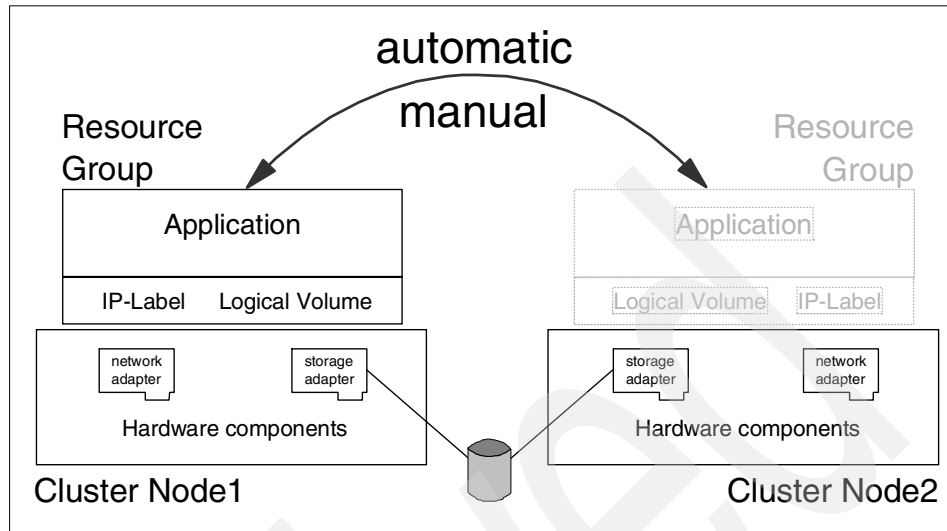


Figure 8-3 A resource group can be taken over manually or automatically

The term *resource group* was already introduced in Section 3.3.2, “Cluster solution for SAP R/3” on page 80. A resource group contains all resources needed for an application to run. These resources have to be handled by the cluster software, which uses individually developed scripts to start and stop the application.

IBM offers the cluster software HACMP (High Availability Cluster Multi-Processing) for the AIX operating system to fulfil the necessary tasks for an automated takeover.

8.3 HACMP basics

In this section, we outline some of the features of HACMP that are needed for an SAP R/3 environment to show, what is possible with HACMP and what are the requirements.

8.3.1 Standard features

HACMP monitors hardware through heartbeats and through scanning and observation of errorlogs, such as the AIX error report `errpt`.

HACMP detects:

- Network adapter failures

- ▶ Node failures
- ▶ Messages in the errorlog

HACMP is able to:

- ▶ Take over disks (with `scsi_disk_reset`)
- ▶ Take over volume groups, logical volumes, and file systems with the necessary clean up procedures (`syncvg` or `fsck`)
- ▶ Change IP adapter information and swap them (IP label or network adapter hardware address takeover)
- ▶ Start or stop applications through scripts

For every change of the cluster state, for example, if an adapter failure is detected, HACMP triggers an event. Every event initializes the execution of an event script. There are predefined event scripts for managing resources, such as service IP addresses or file systems. You only have to provide event scripts for starting and stopping your applications.

It is possible in HACMP to initiate a graceful takeover of a resource group for maintenance reasons, for example.

8.3.2 Required customization

There is no out-of-the-box standard configuration that just requires one setup script and everything is up and running.

Pitfall ahead!

HACMP has to be customized to the respective requirements.

HACMP needs appropriate, well tested scripts that are especially created for the applications, which can be executed according to the occurrence of the following events:

- ▶ For starting and stopping the application during cluster start, stop, and takeover
- ▶ For reactions to a network down event

8.3.3 Application monitoring with HACMP/ES

HACMP classic provides only hardware monitoring. HACMP/ES also provides application monitoring. That is, HACMP/ES is able to trigger events on the system and to perform a certain reaction to these events in a previously defined way. When configuring HACMP/ES, scripts for monitoring an application and for performing the appropriate reaction (in case of an application failure) have to be created.

Looking at the difference of hardware and application monitoring, it is relatively easy to react on hardware events, because hardware failures can be reliably detected. If, for example, a network adapter fails, it has to be replaced or the configuration information has to be switched to a standby adapter. On the other hand, if an application fails, the reasons are not so obvious. A thorough analysis is necessary in order to determine the reason for a failure and to decide on the appropriate reaction.

HACMP/ES only allows pre-defined reactions to certain events. However, the cause of an application failure could be more complex than what can be triggered by an automated event. An automated reaction to this failure may not solve the situation and even impose a certain risk, such as the loss of data.

If an application is unstable, monitoring the application, manually or automatically, will not solve the problem in the long run. A redesign of the application should be considered in this case.

Bright idea!

Application monitoring and notification in case of an error is a useful and necessary operational task. We do not, however, recommend automatic recovery after failures of the database or the SAP R/3 application for the previously mentioned reasons. Thus, we do not recommend that you use the HACMP/ES features for current releases of SAP R/3

The introduction of new highly availability techniques for the enqueue server may, however, require the usage of HACMP/ES.

Monitoring tasks may, however, be facilitated with already customized tools. Refer to Chapter 13, “Daily tasks to prevent error situations” on page 407 for more information to this topic.

In this redbook, we only discuss the use of HACMP classic.

Attention: HACMP/ES is different in its handling than HACMP classic. Our considerations in this redbook are based on several years of experience on HACMP classic and may not apply to HACMP/ES. Further information on HACMP/ES and SAP R/3 can be found in the redbook *HACMP/ES Customization Examples*, SG24-4498

8.4 Implementation of HACMP for SAP R/3

For the implementation of HACMP for SAP R/3, all the considerations for the infrastructure from Chapter 3, “Architecture” on page 35, Chapter 4, “Disk storage” on page 95, and Chapter 6, “Network” on page 149 build the basics for this sections.

8.4.1 Topology

The cluster topology of an HACMP cluster comprises the following components:

- ▶ The cluster definition
- ▶ The cluster nodes
- ▶ The network adapters

The following sections describe these components in more detail.

Cluster topology

In this section, we discuss the fault-tolerant configuration of Section 3.3, “Building a fault-tolerant model” on page 77 in detail. In Figure 8-4, the database server with the central instance (DB/CI HA) is the primary node. The backup node for the DB/CI resource group in the fault-tolerant environment is the server APserv HA with an application server instance on it. These two nodes build up the high availability cluster. Both servers have two network adapters to the front-end network and two paths to the storage subsystem.

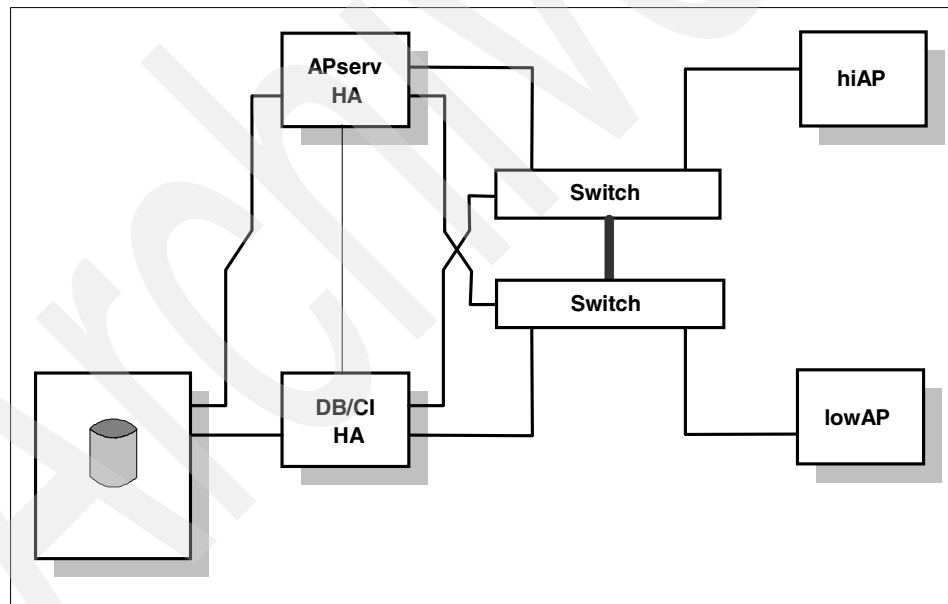


Figure 8-4 Scope of the high available environment

The background and cabling information for doubled network adapters is introduced in Section 4.9, “Fault-tolerance requirements” on page 131 and Section 6.4.2, “Network layout for the fault-tolerant configuration” on page 169.

Server network

According to the definitions in “Server network” on page 162, the server network is a high throughput network for the traffic between the application and database servers. The front-end network connects the clients to the application servers.

In standard installations, there is only one physical network if this network offers enough bandwidth for the front end and the server traffic. Gigabit Ethernet provides this bandwidth in most situations. This network, which combines both traffics, is then referred to as the front-end network.

Bright idea!

Since Gigabit Ethernet implementations are used in many actual environments, we recommend the use of only one physical network, if the bandwidth requirements allow this.

The bandwidth requirement for the server traffic is about ten times as high as the bandwidth requirements for the front-end traffic. If the front-end traffic uses more than ten percent of the bandwidth that a Gigabit Ethernet network offers (that is, 100 Mbps), both traffics do not fit together on one physical Gigabit Ethernet network. This is because the server traffic would be ten times as much, which means more than 1000 Mbps.

If the bandwidth requirement for a separated server network is more than 1000 Mbps, a single Gigabit Ethernet network cannot satisfy the requirements anymore. In this case, a different technology has to be used for the server network, such as the high performance SP Switch.

Pitfall ahead!

If the introduction of a separate server network is necessary, you should consider the severe impact on complexity. When the server traffic is moved to a separate network (rerouted), the heartbeat for monitoring the front-end network adapters is also rerouted (in an HACMP scenario). In this case, the front-end network is not monitored by HACMP anymore and it is not possible to react to network adapter or network failures on the front-end network.

The IBM SAP International Competence Center (ISICC) introduced the usage of IP aliases on the network adapters. This causes the heartbeat to remain on the front-end network, and therefore still enables the monitoring of the adapters, even if the server traffic is rerouted to a separate server network. HACMP is not yet able to handle IP aliases, so the handling of these aliases has to be managed in scripts that have to be developed or adapted according to the individual requirements.

SAP describes the separation (rerouting) of the server traffic in *SAP R/3 in Switchover Environments*, Document 50020596, found at: <http://service.sap.com/systemmanagement>. Refer to *SAP R/3 and HACMP Setup and Implementation*, by ISICC, found at (on the IBM intranet) <ftp://siccserv.isicc.de.ibm.com/perm/hacmp> for more information on using IP aliases and for alias handling scripts.

Adapters

Each of the cluster nodes must have two network adapters attached to the front-end network. The first adapter carries the service address of the resource group, the second one works as a standby adapter during normal operation. The standby adapters of both machines are configured to be part of an IP subnet that is distinct from the subnet used for the front-end network. The standby adapters also exchange heartbeat packets. In case of an adapter failure within a cluster node, the service and standby addresses are swapped by HACMP. This swap is transparent for the applications. In case of a node takeover, if the entire node fails, the service address of the failing node is bound to the standby adapter of the remaining node.

To be able to distinguish a crash of a cluster node from a breakdown of the front-end network, it is highly recommended to connect both cluster nodes, for example, via a point to point serial link. This link provides an additional way for exchanging heartbeats between the cluster nodes and is independent from any IP network. Refer to “Heartbeats with other than a direct cable” on page 255 for more information.

8.4.2 Resource groups

It is very easy to determine the single points of failure in the physical layer of the hardware. To determine the single points of failure in an application, you need an in-depth knowledge of the application. In SAP R/3, the central instance, comprising the message and the enqueue server and the database, are the single points of failure for which a backup server is necessary.

Pitfall ahead!

It is possible to configure a resource group for the database (DB) and another one for the central instance (CI). However, this separation introduces an additional single point of failure, because another hardware component has to be added to the SAP R/3 system. Figure 8-5 on page 232 illustrates the Single Points of Failure (SPoFs) for both configurations.

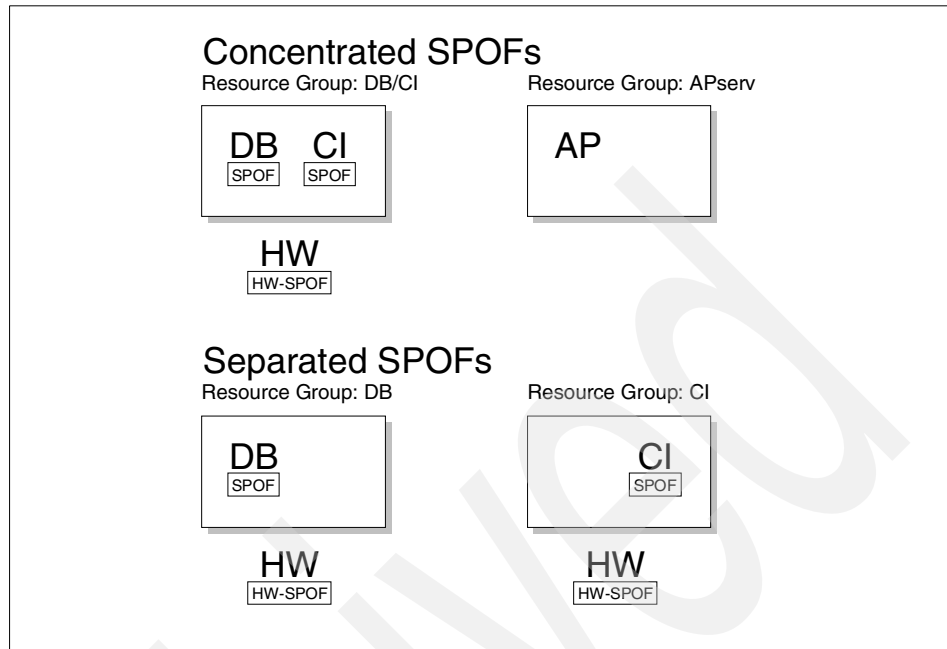


Figure 8-5 Concentrated versus separated SPoFs

If an application has two single points of failure, such as the database and the central instance in SAP R/3, and if these two components run on the same server, we have three SPoFs in total. If we separate the two components (DB and CI) of the application (of which each represents an SPoF) onto two servers, we introduce another SPoF by adding another hardware component to the configuration, because the whole system fails if one server fails. So the probability of a system failure caused by hardware increases twofold.

The administration of the database is easier if there is a central instance on the same server. In this case, the database and central instance build a unit and are started and stopped together. If both are separated, additional logic is necessary to start the database first and then the central instance.

Many of the older documents recommend a separation of database and central instance for performance reasons. With today's powerful servers, the impact of a small central instance on the database system performance is minimal. Thus, the historic and more complicated separation should be avoided.

Bright idea!

In our example (Figure 8-6), two resource groups are configured. One contains the database server and the central instance (DB/CI HA), and the other contains an application server instance (APserv HA). The server DB/CI HA has the host name SIDdb and the server APServ HA has the label SIDap. For the server DB/CI HA, the IP label SIDdb is defined, and for the APServ HA, the label SIDap is defined. It is a requirement of SAP that SAP R/3 runs on a server on which the host name is the same as the service IP label.

Figure 8-6 gives an overview of the service, boot, and standby IP labels. The boot IP labels are valid after the boot of a node. The service IP label is assigned to the adapter, to which the boot IP label is assigned.

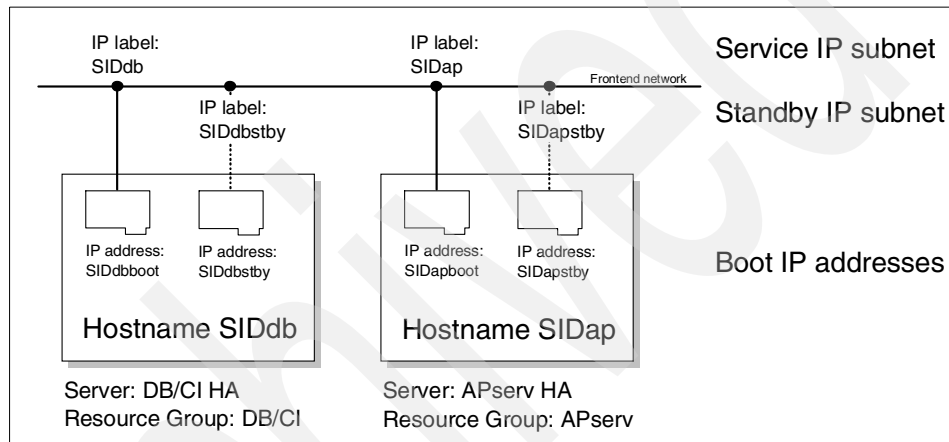


Figure 8-6 Service, boot, and standby IP label in an SAP R/3 cluster

The primary node for the resource group DB/CI is the node DB/CI HA, and the backup node is APServ HA. The primary node for the resource group APServ is the node APServ HA, and the backup node is DB/CI HA.

For both resource groups, an IP label is defined, over which the service is delivered. This label is assigned to an adapter and can be monitored through heartbeats. HACMP is able to make this IP label high available with network adapter hardware address takeover and IP address takeover. For more information on network components, refer to Section 6.5.1, "Configuration of switches" on page 174.

The resource groups have to be defined as cascading resource groups, which is explained in the next section.

Cascading versus rotating resource groups

Cascading and rotating resource groups differ in the way they are configured and the way they operate.

In case of a rotating cluster, there is only one resource group allowed, that is, one service IP label per network. Only the cascading configuration allows two (or more) resource groups in one network, which are needed in a DB/CI and an APserv resource group environment.

In cascading resource groups, there are priorities for the assignment of resources to the cluster nodes, which are defined at installation time. Cascading resource groups have the attribute that when a new node joins the cluster, first its resource group priority is checked. If the new node's resource group priority is higher than the priority of the already running system, the resource group is moved from the running node to the joining node.

With rotating resource groups, the priorities are ignored and the resource group stays where it is. Later, if the node with the assigned resource group fails, the resource group will move to the node with the highest priority for the resource group.

When using NFS mounts across the cluster nodes (NFS cross mounts), only cascading resource groups support automatic NFS mounting across servers during failover. Rotating resource groups do not provide this support. Instead, you must use additional post events or perform NFS mounting using normal AIX routines to support NFS cross mounts for rotating resource groups.

Resources

Both resource groups DB/CI and APserv contain file systems, a service address for the front-end network, and application start and stop scripts. All files that have to be taken over must reside in one or more volume groups within the storage subsystem. In case of a takeover, the remaining node accesses the volume groups of the crashed node and mounts their file systems.

Pitfall ahead!

The service addresses are IP addresses that are bound to a front-end network adapter during the start of a cluster node. The service addresses are used by SAP R/3 for communication between the database, central instance, and other application servers and for the access of the SAP GUIs to the message service. In case of a takeover, the service address of the failing node is transferred to a standby adapter of the remaining node. For a detailed description of service addresses, refer to *SAP R/3 in Switchover Environments*, Document 50020596, found at: <http://service.sap.com/systemmanagement>.

Application start and stop scripts are used to start or stop the database, the central instance of the resource group DB/CI, and the dialog instances of the resource group APserv when starting or stopping the cluster nodes. HACMP uses the same scripts whether a resource group is transferred to the backup node on request or in case of a takeover after a crash. These (cluster) application scripts are not identical with the standard SAP R/3 start and stop scripts. The cluster application scripts integrate standard SAP R/3 start and stop scripts, but also include additional check and cleanup routines to be able to handle all failover scenarios.

Table 8-1 summarizes the contents of the resource groups.

Table 8-1 Contents of the HACMP resource groups

Content	Resource group DB/CI	Resource group APserv
Service address for front-end network	IP address: a.b.c.x IP label: SIDdb	IP address: a.b.c.y IP label: SIDap
Volume groups and file systems	DB volume group SIDdbvg: Database executables Database tablespaces CI volume group SIDcivg: SAP R/3 executables Instance file systems Interface file systems	AP volume group SIDapvg: Instance file systems Database client file systems
Application start and stop scripts	Start and stop scripts for the database and central instance	Start and stop scripts for dialog instance

Service addresses have to be part of the IP subnet used for the front-end network (here: a.b.c). The IP label is the DNS name of the service IP address. We suggest you use SIDdb and SIDap where SID is a placeholder for the system ID of your SAP R/3 system.

Implementation

This section lists the required main entries of the SMIT configuration panels for an HACMP installation with SAP R/3. The SMIT fastpaths are mentioned in brackets ()

The configuration entries are based on our example in Figure 8-6 on page 233 and Table 8-2 on page 236, and cover the several IP labels and definitions of the file systems belonging to the resource groups.

Configure Adapters (smit sm_config_adapters)

An example for the network configuration is shown in Table 8-2.

Table 8-2 HACMP public network configuration

IP label	Network	Function	Attribute	Node
SIDdb	Front end	service	public	DB/CI HA
SIDap	Front end	service	public	APserv HA
SIDdbboot	Front end	boot	public	DB/CI HA
SIDapboot	Front end	boot	public	APserv HA
SIDdbstby	Front end	standby	public	DB/CI HA
SIDapstby	Front end	standby	public	APserv HA
SIDdbtty1	Serial	service	private	DB/CI HA
SIDaptty1	Serial	service	private	APserv HA

Bright idea!

HACMP needs its own IP subnet with dedicated addresses for the monitoring of the standby interfaces. The traffic between the standby interfaces must not be routed. Thus, it is recommended that you use 10.x.x.x, 172.16.x.x, or 192.168.x.x subnets. Use a subnet for the standby addresses that cannot be mistaken accidentally with productive subnets of the company.

Define resource groups (smit cm_add_res)

Resource group definition for SIDdb:

Cascading Primary: SIDdb, Backup: SIDap
IP label SIDdb
File systems All from volume groups SIDdbvg and SIDcivg
Application Server SIDdb

Resource group definition for SIDap:

Cascading Primary: SIDap, Backup: SIDdb
IP label SIDap
File systems All from volume group SIDapvg
Application Server SIDap

Define Application Servers (smit cm_cfg_app)

Application Server definitions for SIDdb:

Start script /usr/scripts/cluster/start_DBCI

Stop script /usr/scripts/cluster/stop_DBCI

For SIDap:

Start script /usr/scripts/cluster/start_AP

Stop script /usr/scripts/cluster/stop_AP

The cluster application scripts for starting and stopping must be used to start and stop SAP R/3 either on one or the other node. The reason is that no identity takeover is performed and, therefore, the host name on the backup node is another than the one on the primary node. The standard SAP R/3 environment provides scripts for starting and stopping the application, which have the host name coded in the name of the script.

The cluster application scripts have to take the host name on the backup node into account in case of a takeover. This can be done, for example, by adding symbolic links for the files starting with startsap_<hostname>, stopsap_<hostname>, .dbenv_<hostname>, and .sapenv_<hostname> that refer to the host name of the backup node.

SAP R/3 profiles

As both cluster nodes have different host names, the environment configuration files have to be adjusted. In the previous section, this is done for the shell environment. The application parameter for SAP R/3 are stored in profiles (/sapmnt/SID/profiles) and have to be adjusted as well. In Chapter 12, “SAP R/3 system copy” on page 365, all the files that need this treatment are listed.

Bright idea!

For SAP R/3, it is necessary that the host name is the same as the service IP label. In case of an takeover, this is not true anymore, and SAP R/3 will not start. For this reason, the parameters SAPLOCALHOST and SAPLOCALHOSTFULL have to be set according to the service label of the primary node for both resource groups. These parameters reference the host name and the service IP label.

In our example, SAPLOCALHOST is set to SIDdb in the DB/CI resource group and to SIDap in the APserv resource group. The SAP R/3 parameter SAPLOCALHOSTFULL corresponds to parameter SAPLOCALHOST extended by the network domain name. The parameter SAPDBHOST references the database instance and is set to SIDdb too. Table 8-3 summarizes the changes.

Table 8-3 Modified SAP R/3 profiles

Profile	Parameter	DB/CI HA	APserv HA
Default profile ^a	SAPDBHOST	SIDdb	SIDdb

Profile	Parameter	DB/CI HA	APserv HA
Instance profile	SAPLOCALHOST	SIDdb	SIDap
	SAPLOCALHOSTFULL	SIDdb.domain	SIDap.domain

a. There is only one default profile in the whole SAP R/3 system, so the SAPDBHOST parameter is the same everywhere.

NFS considerations

Across the nodes, an NFS cross mount is often used to allow users and processes to have access to the same data and executables.

If application servers come into play, NFS is needed to provide the same executables, interface data, log files, and printing data to more than one server. HACMP can handle NFS exports and NFS cross mounts within cascading resource groups, but we recommend that you not use it, because the experience shows that HACMP handling of NFS file systems is not highly sophisticated and leads, in certain cases, to HACMP event errors.

All application servers mount the file systems with the SAP R/3 executables and other data from the central instance node. NFS exports and mounts can therefore be handled within the application start and stop scripts of the cluster nodes.

Problems occur in case of the primary node failure, when the surviving node might have to take over and mount the file system it previously had mounted through NFS. The surviving node cannot mount the file system locally until it has performed an NFS unmount.

Bright idea!

In order to avoid conflicts with hanging NFS mounts, the mount points of the standard SAP R/3 file systems are changed. Global file systems, which contain all other file systems below a /global mount point, and symbolic links, which point from the SAP R/3 standard locations, to these file systems, are introduced.

Figure 8-7 on page 239 shows the handling of NFS mounts. For example, the file system /sapmnt/SID is shown. This file system contains, by default, the subdirectories /sapmnt/SID/exe, /sapmnt/SID/profile, and /sapmnt/SID/global.

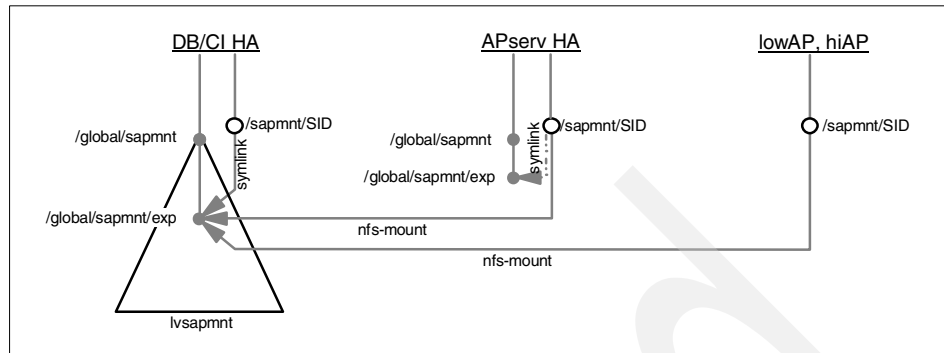


Figure 8-7 Handling of NFS mounts in the highly available state

The NFS export has to be in a subdirectory of the file systems mount point. Only in this case, the file system itself can be unmounted, if an NFS export exists. The primary node DB/CI HA has the file system `/global/sapmnt` mounted and exports the directory `/global/sapmnt/exp` via NFS. The symbolic link points from `/sapmnt/SID` to the directory `/global/sapmnt/exp`.

On the backup node APserv HA, the configuration is similar. Thus, the directory `/global/sapmnt` is the mount point for the file system and the symbolic link points to the directory `/global/sapmnt/exp` too. The backup node APserv HA mounts the file system `/global/sapmnt` from the primary node through NFS to the mount point `/sapmnt/SID`. The underlying symbolic link is resolved and the NFS file system is mounted to `/global/sapmnt/exp`.

All other application servers (lowAP and hiAP, in our example) mount the NFS exported file system on the default mount point `/sapmnt/SID`.

Figure 8-8 on page 240 shows two scenarios, in which either the DB/CI HA node or the APserv HA node fails. The backup server APserv HA takes over the volume group, including file systems and mounts the file system `/global/sapmnt` to its mount point. It is not preliminary for a takeover to take care for the NFS mount in `/global/sapmnt/exp`. The new file system overmounts the directory tree and the symbolic link from `/sapmnt/SID` points to the directory `/global/sapmnt/exp` in the local file system.

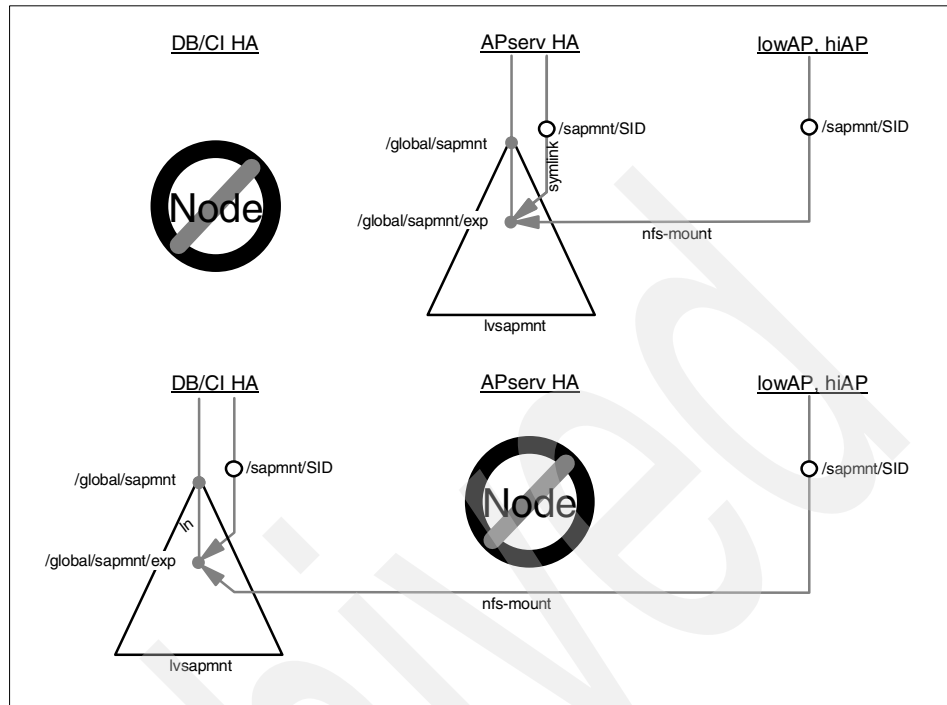


Figure 8-8 Handling of NFS mounts in the takeover scenarios

As this example demonstrates, several small file systems can be consolidated in one large file system. This also applies to file systems for interface data, database executables, and database log files.

With this construction, NFS mounts can be handled error free via HACMP, but only for the cluster nodes. Further application servers must be handled with a dedicated script. For these application servers, a script can be used to handle all of these NFS tasks without HACMP.

Every device has a major and a minor device number as an access address to this device. The major device numbers of the takeover volume groups have to be identical and NFS is the reason. The NFS clients get a file handle from the NFS server and the major device number is part of this handle. After a takeover in the cluster, the NFS clients still have a valid file handle.

Pitfall ahead!

The requirement to make any shared volume group have the same major device number throughout the cluster is only critical when NFS is used. It is, however, a very good practice to follow in general, because it is sometimes hard to predict (in the beginning of an implementation) that shared data will be used.

8.4.3 Application start and stop handling

Considerations for script implementations of the start and stop process comprise a complete sequence of actions.

Starting SAP R/3 on DB/CI HA

The start procedure contains, at least, the following steps to start the database management system and the central instance on a cluster node. In case of a takeover of the resource group DB/CI HA to the backup node, the outlined cleanup actions can be omitted.

Actions on local server

1. Conditional cleanup actions
 - a. Clean up processes belonging to the database or the central instance according to SID (*not at takeover*).
 - b. Clean up shared memory and semaphores belonging to the database or the central instance according to SID (*not at takeover*).
 - c. Remove orphaned **brbackup** and **brarchive** lock files.
 - d. Remove an orphaned Oracle SGA lock file
\$ORACLE_HOME/dbs/sgadefSID.dbf (*Oracle only*).
 - e. Remove the SAP R/3 stop program /usr/sap/SID/D*/work/kill.sap.
 - f. Remove the SAP R/3 systemlog collector lock file
/usr/sap/SID/D*/data/rslgcpid.
 - g. Remove the SAP R/3 systemlog sender lock file
/usr/sap/SID/D*/data/rslgspid.
2. Startup actions
 - a. Varyon the SAP R/3 volume groups (*manual takeover mode only*).
 - b. Mount the SAP R/3 file systems (*manual takeover mode only*).
 - c. Check to see if file systems are mounted correctly.
 - d. Export the NFS file systems.
 - e. Start the Oracle listener (*Oracle only*).
 - f. Start the database management system.
 - g. Start the DB2 administration instance (*DB2 only*).
 - h. Start the SAP R/3 central instance.
 - i. Start the SAP R/3 performance collector saposcol.
 - j. Start the SAP R/3 saprouter.
 - k. Start the TSM client scheduler.

Actions on remote application server instances (except APserv HA)

1. Cleanup actions
 - a. Clear the ARP cache to remove old IP label entries.
 - b. Clean up processes belonging to the application server instance according to SID.
 - c. Clean up shared memory and semaphores belonging to the application server instance according to SID.
 - d. Remove the SAP R/3 stop program `/usr/sap/SID/D*/work/kill.sap`.
 - e. Remove the SAP R/3 systemlog collector lock file `/usr/sap/SID/D*/data/rslgcpid`.
 - f. Remove the SAP R/3 systemlog sender lock file `/usr/sap/SID/D*/data/rslgspid`.
2. Startup actions
 - a. Mount the NFS file systems.
 - b. Check to see if file systems are mounted correctly.
 - c. Start the SAP R/3 application server instance.
 - d. Start the SAP R/3 performance collector `saposcol`.
 - e. Start the SAP R/3 `saprouter`.
 - f. Start the TSM client scheduler.

Starting SAP R/3 on APserv HA

The start procedure contains, at least, the following steps to start the application server instance on a cluster node. If the resource group APserv HA is taken over to the backup node, the marked cleanup actions can be omitted.

Actions on local server

1. Conditional cleanup actions
 - a. Clean up processes belonging to the application server instance according to SID (*not at takeover*).
 - b. Clean up shared memory and semaphores belonging to the application server instance according to SID (*not at takeover*).
 - c. Clear the ARP cache to remove old IP label entries.
 - d. Remove the SAP R/3 stop program `/usr/sap/SID/D*/work/kill.sap`.
 - e. Remove the SAP R/3 systemlog collector lock file `/usr/sap/SID/D*/data/rslgcpid`.

- f. Remove the SAP R/3 systemlog sender lock file
/usr/sap/SID/D*/data/rslgspid.
2. Startup actions
- a. Mount the NFS file systems.
 - b. Check to see if the file systems are mounted correctly.
 - c. Start the SAP R/3 application server instance.
 - d. Start the SAP R/3 performance collector saposcol.
 - e. Start the SAP R/3 saprouter.
 - f. Start the TSM client scheduler.

Stopping SAP R/3

The stop procedures contain the same steps as the start procedures, in reverse order, with stop actions instead of start actions, unmounts instead of mounts, and varyoffs instead of varyons.

The cleanup actions are done in the stop procedures as well as in the start procedures. As the reason for start and stop may result from a takeover, the condition of the resources is not clearly defined. For reliably working scripts, the cleanup is urgently recommended.

Database and enqueue reconnect

Until SAP R/3 Release 4.0A, there is a need for the application server instances to be restarted, in case of a central instance takeover, to avoid locking problems in the SAP R/3 system. If the application server instance connects within 300 seconds (5 minutes), it receives a transaction reset message from the central instance and causes open transactions to be aborted and rolled back. This will also release enqueue locks. If the reattaching does not succeed within this time limit, the application servers must be restarted as the cleanup procedure is not triggered through the central instance after 300 seconds. The restart action is necessary to guarantee database consistency.

SAP R/3 offers, since Release 4.0B patch level 85 and Release 4.5A, the full support of DB reconnect and enqueue reconnect. Under normal circumstances, it should not be necessary to restart the application server instances. This will keep the buffers filled and offers a better performance after a takeover.

Bright idea!

To avoid any possible error situations with locks, we still recommend the restart of all application server instances. The only constraint is reduced performance in the first minutes of an application server instance restart (until the SAP R/3 buffer caches are filled again). Users on the application server instances are logged out

and are reassigned to another instance in the SAP R/3 production system with the next logon. In case the application server instances are not restarted, the working users may get an hourglass on first access and have to deal with a hanging SAP GUI.

The reliability of the whole SAP R/3 system is more important than the performance gained in the first minutes after an instance restart.

Refer to *SAP R/3 in Switchover Environments*, Document 50020596, found at <http://service.sap.com/systemmanagement>, for more information on this issue.

8.4.4 Odds and ends

In this section, several hints and tips are given that may help you during the installation of an HACMP cluster.

Node synchronization

After the installation of SAP R/3 on the primary node DB/CI HA, the backup node has to be synchronized with the primary node. In “Master for distribution of global configuration files” on page 75, we discuss the necessity for a synchronized operating system environment in the whole landscape concerning users, user limits, printers, and other settings. In Section 10.1.2, “Importance of synchronizing” on page 279, these tasks are covered in depth. In the case of an overall synchronized environment, many tasks are already done for the synchronized cluster nodes. Chapter 12, “SAP R/3 system copy” on page 365 also gives hints on the different tasks to complete when moving an SAP R/3 permanently to another host.

Both nodes need at least:

- ▶ The same external volume groups (with same major device numbers) with same configuration (auto varyon is set to off)
- ▶ The same file system and same file system configuration (file systems are not checked and not automatically mounted at server reboot)
- ▶ The same NFS mounts and NFS exports configuration
- ▶ The same users and user limits
- ▶ The same configuration for:
 - Maximum number of processes
 - Asynchronous I/O state at reboot and minservers/maxservers
 - I/O pacing with high and low watermark
 - Autoreboot should be off

- ▶ DB2 only:
 - DB2 Software is installed in the rootvg and therefore has to be explicitly installed on the backup node with the `installp` command.
 - DB2 patches need to be applied on both nodes.
 - The DB2 administration instance has to be installed on the backup node.
- ▶ Oracle only:
 - Oracle executables are stored in file systems in the external volume group and are switched over.
 - The load extension and other related files have to be copied from the primary node to the backup node, and the owner and permissions have to be set on the backup node according to the values on the primary node. The following files are relevant: `/etc/loadext`, `/etc/pw-syscall`, `/etc/oratab`, `/etc/ora_kstat`, and `/etc/orainst.loc`.
 - The entries from the `/etc/inittab` also have to be set on the backup node:


```

/usr/sbin/mkitab strload:2:once:"/usr/sbin/strload"
/usr/sbin/mkitab orapw:2:wait:"/etc/loadext /etc/pw-syscall"
/usr/sbin/mkitab orakstat:2:wait:"/etc/loadext /etc/ora_kstat"
          
```
- ▶ Adapted paging space configuration
- ▶ Adapted DNS search logic (if applicable):
 - For `/etc/environment`:


```
NSORDER=local4,local6,bind4,bind6
```
 - For `/etc/netsvc.conf`:


```
hosts=local4,local6,bind4,bind6
```

I/O pacing and syncd

Some AIX parameters have to be adjusted when HACMP is implemented.

By default, the I/O pacing in AIX is turned off, which means the values for low water mark and high water mark are both zero (0/0). This is the default as well if HACMP is installed.

Bright idea!

If there is a heavy write workload on the disks, it is recommended that you turn I/O pacing on to protect your system from crashes caused by the dead man switch. A crash indicates that the system was overloaded and the cluster was not able to fulfil its system alive check tasks in a reasonable amount of time.

I/O pacing may have a serious impact on I/O performance. The configuration of reasonable I/O pacing values is discussed in depth in Section 11.4.4, “Disk I/O pacing” on page 342.

Changing the time interval for **syncd**, for example, from 60 to 10 seconds to avoid problems with the dead man switch usually has no effect in an SAP R/3 installation that is set up according to the recommendations in this book. **syncd** writes dirty pages in this time interval from the AIX Virtual Memory Manager to disk. If the file system cache is reduced to a minimum, as recommended in Section 4.6.1, “Journaled file systems versus raw logical volumes” on page 120, a value of 10 or 60 seconds makes no difference and cause no harm, such as flooding the I/O subsystem. The dirty pages from the buffers of the database management system are flushed to the disk by the database management system itself and are not affected by the **syncd**.

Heartbeat detection rate

The heartbeat detection rate can be set *Low*. In case of a short network interference, the cluster reacts less swiftly, and causes no actions if the problems are resolved within a few seconds.

To set the detection rate, enter the following command:

```
smitty sm_config_networks.chg.select
```

Then select the following menus:

- ▶ ether
- ▶ Failure Detection Rate

Spanning Tree Protocol (STP)

In “The Spanning Tree Protocol” on page 158, we discuss the need for STP in our network environment. This protocol enables the exploration of new network paths, if a switch has a failure, and it avoids loops in the network paths. If a server is connected to a switch port, the STP must be disabled for this port.

Pitfall ahead!

In case of a switch failure, the remaining switches start a negotiation with their network neighbors to determine the new topology. During the time of negotiation, the port, where a server is connected to, may be unavailable if the STP is enabled on this port. This leads to unnecessary adapter swaps initiated by the cluster software and to an unstable state of the cluster.

SAP R/3 license

The SAP R/3 license key includes information about several fundamental aspects of your SAP R/3 system, such as setup and the underlying hardware of the message server host. Any change to a key parameter or the location and specification of the message service host will invalidate the installed license. A takeover of the DB/CI resource group will affect the licensing mechanism.

For a takeover of the DB/CI resource group, two licenses need to be available. Both can be in place at the same time, as they are stored in the database with these steps:

1. Get the hardware keys of both servers with the **saplicense -get** command as user <sid>adm.
2. The keys are sent to SAP with a request for the licenses.
3. Run the installation of the licenses as user <sid>adm for every license key you obtain from SAP with the command **saplicense -install** on the node where the database is active.
4. Check the installed license as user <sid>adm with the command:
saplicense -show

SAP GUI client requirements

When it comes to fault- or disaster-tolerant environments, the clients have to be considered as well. Depending on the network structure and the installed components, such as switches or routers, the requirements differ significantly.

In a flat network (without routers between clients and servers), a major problem is the cache on the clients, where the network adapter hardware addresses are stored in the so-called Address Resolution Protocol (ARP) cache. In case of an IP takeover without adapter hardware address takeover in the cluster, the client has stored the invalid adapter hardware address. The default cache resident time is 20 minutes, and within this time, no connect to the cluster node is possible. This problem can be resolved by lower cache resident times on every client, for example, five minutes, or by adapter hardware address takeover, if this is supported by the network components.

Bright idea!

To avoid problems of this kind, an adapter hardware address takeover is highly recommended.

The same problems may apply if source routing is activated on the clients. Even if the cluster executes a takeover, a client tries to reach the server on the previous network path and cannot connect. Therefore, source routing should be switched off or ARP caches should be cleared in a script during the start of the SAP GUI.

Configuration takes too long

If, for some reason, the cluster needs more than 360 seconds for an action, the cluster gives a warning and is in the state config-too-long. To raise this time interval to 3600 seconds (one hour), the following command is needed:

```
/usr/bin/chssys -s clstmgr -a -u 3600
```

Cluster manager not visible if inactive

By default, the cluster subsystems are not visible in the listing of the `lssrc` command when they are inactive.

To make the cluster manager and the other cluster subsystems visible, the execution of the following commands is necessary at installation time:

```
/usr/bin/chssys -s clstrmgr -d
/usr/bin/chssys -s clinfo -d
/usr/bin/chssys -s clsmuxpd -d
```

The effect of these commands is shown in Example 8-1.

Example 8-1 Output from `lssrc -g cluster`

# lssrc -g cluster				
Subsystem	Group	PID	Status	
clinfo	cluster		inoperative	
clstrmgr	cluster		inoperative	
clsmuxpd	cluster		inoperative	

8.5 Testing of the cluster

The testing of the cluster environment is an essential task. Scenarios have to be worked out and every step has to be tested. To describe all possible states and test scenarios of a cluster, a state-transition diagram is very useful.

8.5.1 Cluster diagram (state-transition diagram)

Bright idea!

One part of the cluster documentation could contain a state-transition diagram. The state-transition diagram is neither a completely new method nor the ultimate solution for system documentation. It is very often used to describe the behavior of complex systems, which an HACMP cluster undoubtedly is. The behavior within an HACMP environment can be completely described in a state-transition diagram. Compared with a description in text form, a state-transition diagram gives an overview and detail at a glance. It is a very compact form of visualization.

Two element types are used for the state-transition diagram: circles and arrows. A circle is the description of a state with particular attributes. A full grey circle (or state) means that the cluster or, better, the service delivered by the cluster is in a highly available state. A partially grey shaded state represents limited high availability. States drawn in white are not highly available. State-transition diagrams contain only stable states. The state-transition diagram does not contain temporary or intermediate states.

The second type of element is the arrow, which represents a single transition from one state to another. All bold arrows are automatically initiated transitions. Non-bold drawn arrows are manually initiated transitions. The diagram should contain only relevant transitions, as well as the states.

A state-transition diagram is used to document a cluster. The diagram itself has the following rules to describe the behavior of a cluster:

- ▶ An arrow must start and end at a circle. There should be at least one arrow from and to every state (no orphaned states).
- ▶ All states in the state-transition diagram must be valid according to the user's expectations for the level of service. All required states must be included.

A minimal set of test cases are defined in the state-transition diagram. This means that every arrow describes a transition that should be tested independently. This applies whether the transition is initiated automatically or manually.

More possible states and transitions exist in a cluster than the diagram contains. Most of the ones not shown in the diagram contain multiple errors at the same time or make no sense. The diagram must cover all possible single faults and can cover selected multiple error situations. If all reasonable states and the ways to reach them are documented in the diagram, the tests can be started.

Diagram for automatic operation

Figure 8-9 on page 250 describes a sample cluster diagram for our scenario. The state-transition diagram describes all allowed states of the cluster for documentation and testing purposes. A detailed legend for Figure 8-9 on page 250 can be found in Figure 8-10 on page 251.

The start and stop scripts mentioned in the diagram are based on the procedures, which are described in "Application start and stop handling" on page 241.

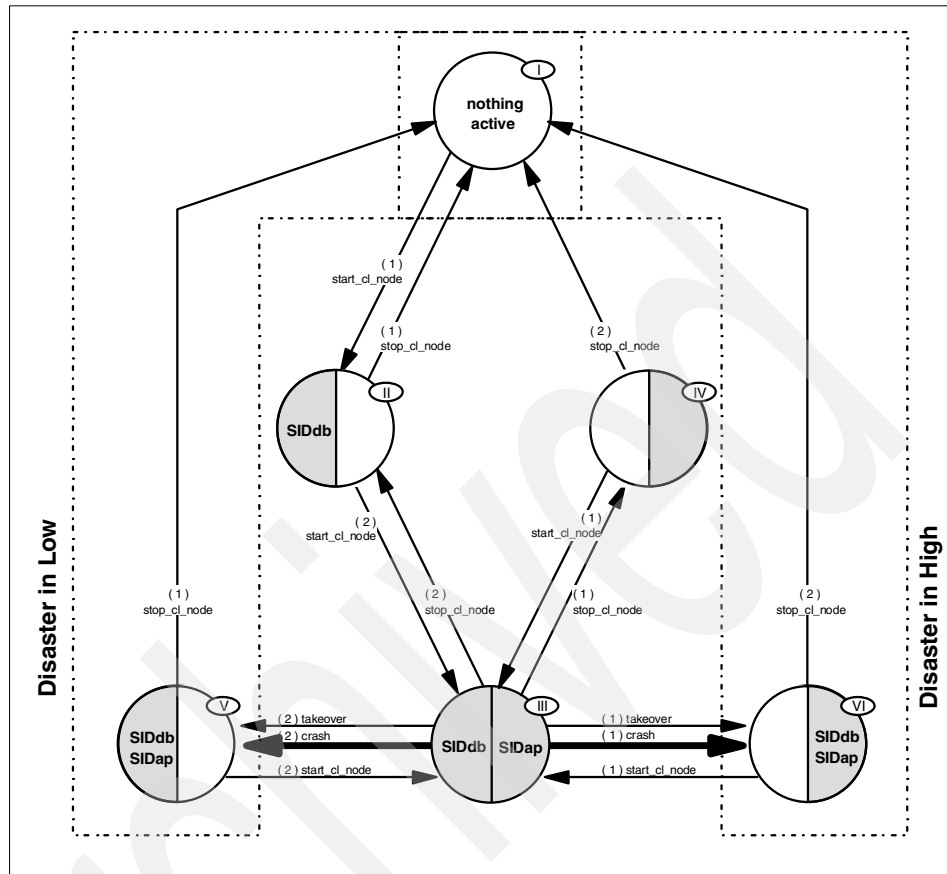


Figure 8-9 State-transition diagram for automatic operation

To start the cluster from the “nothing active” state (I) to a fully HACMP monitored state (III), the following steps have to be executed:

1. The cluster software has to be started on node DB/CI HA (start_cl_node).
2. The cluster must be started on the backup node APserv HA (start_cl_node). The high available state (III) is plotted at the center bottom of Figure 8-9 and describes the active resource group SIDdb on the node DB/CI HA and the resource group SIDap on the node APserv HA.

In case of a crash of the node APserv HA, an automatic takeover is initiated by HACMP. This leads to a state (V), where the resource groups SIDdb and SIDap are active on node DB/CI HA.

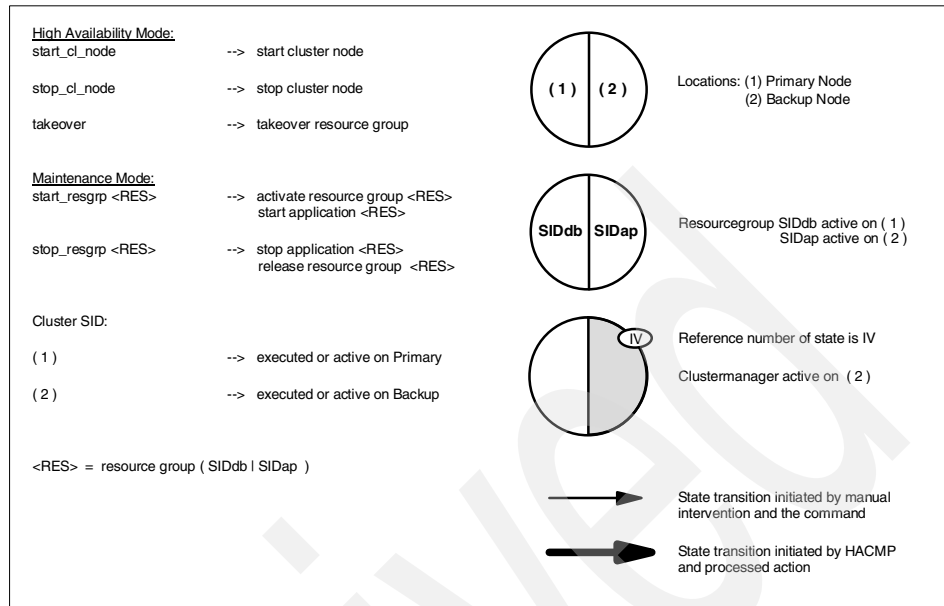


Figure 8-10 Legend for the state-transition diagram

Diagram for manual operation

If the cluster cannot be started or is not started for some reason, the SAP R/3 application cannot run, because the resource group is not assigned to a cluster node. For startup, manual intervention is required, and the IP label has to be assigned, the volume group must be activated, and the file systems must be mounted. For this and other cases, it is necessary to create scripts, which take care of the resource handling. These are the same scripts that are used for the manual takeover.

If the primary server DB/CI HA is not up and running at all, for example, due to a power outage or a hardware defect, it is not possible to start the application SAP R/3 on the backup server APserv HA via HACMP. In this case, the scripts can be used for the manual mode of the cluster and are also needed in case of upgrades of the cluster software itself.

Thus, a second transition diagram (Figure 8-11 on page 252) is necessary to represent the cluster in the several states of manual operation.

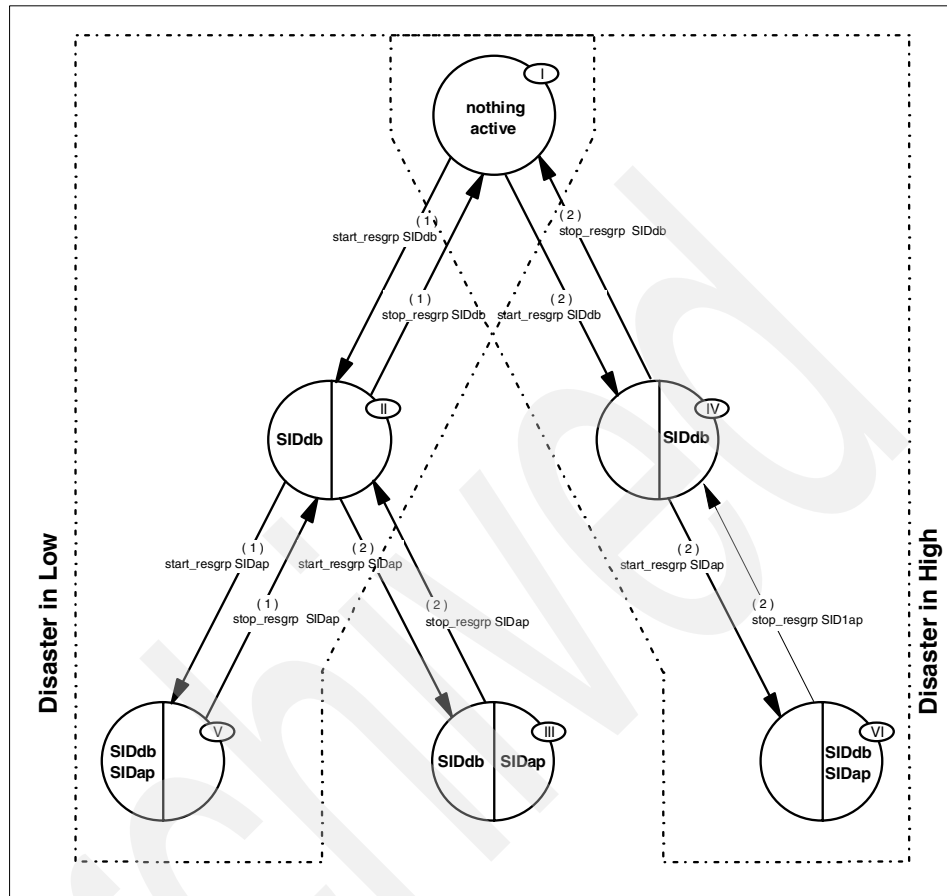


Figure 8-11 State-transition diagram for manual operation mode

With the start and stop scripts, as in the automatic operation mode, it is possible to run the application SAP R/3 on the cluster nodes. In case of hardware failures, it is also practical to circumvent this error situation by assigning the resource group to another node.

8.6 Extensions for the disaster-tolerant model

In the previous sections, we have discussed the fault-tolerant model, which means a configuration in which single points of failures are avoided. In this section, we discuss the extension to a disaster-tolerant configuration.

Topology

The hardware topology has to change to fulfill the requirements of business continuance in case of a disaster in one computing center. Figure 8-12 gives an overview of the extended configuration.

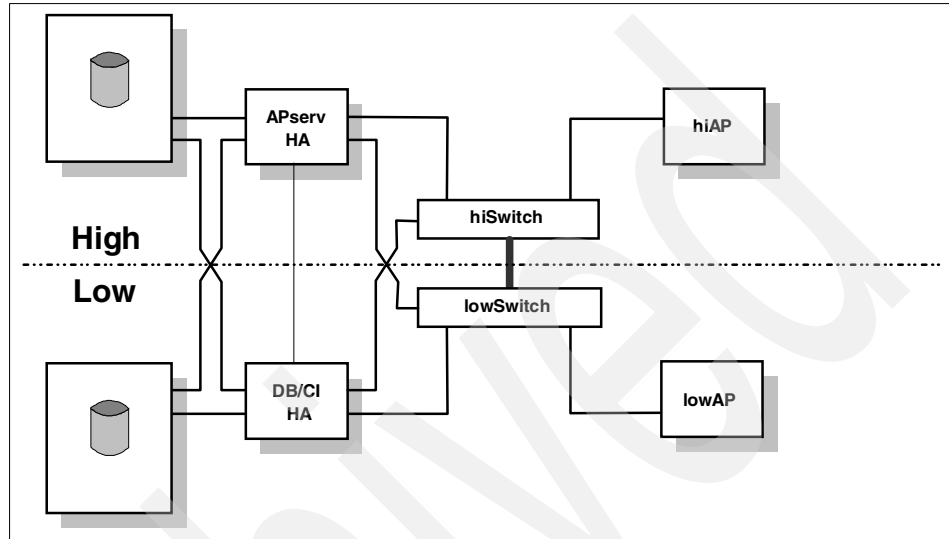


Figure 8-12 Disaster-tolerant model

One design change affects the disk storage subsystems. For a disaster-tolerant solution, a second disk subsystem has to be introduced. The nodes must be connected to the second subsystem and, for redundancy, both connections have to be doubled. So four adapters, for the connection from the node to the storage subsystems, are necessary per node. This is described in more detail in Section 4.10.2, “A disaster-tolerant ESS configuration” on page 134.

Another design change has to be made for the network connections. In Section 6.4.3, “Network layout for the disaster-tolerant configuration” on page 172, we discuss the requirements for the disaster-tolerant model. Due to the physical separation of the two computing centers, the network cabling must be extended in one point. The connection from a cluster node to the switch in the same computing center has to be doubled.

In case of an outage in one data center, the backup node has to take over the resource group from the failing primary node. The network IP label, which belongs to this group, is assigned to a remaining standby adapter. If a cluster node has only two adapters (one is connected to the network switch hiSwitch and the other to the lowSwitch), only one adapter is active in case of a computing center outage. This active adapter already has a service IP label assigned. This is the reason why a further network adapter in both nodes is necessary.

Figure 8-13 gives an overview of the configuration.

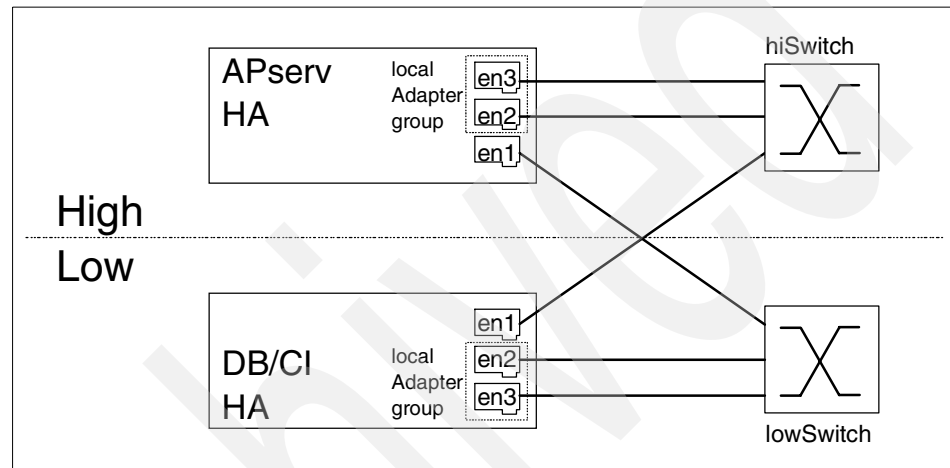


Figure 8-13 Adapter groups

Bright idea!

It is assumed in Figure 8-13 that the adapter en3 in the cluster node DB/CI HA already has the IP label SIDdb assigned. Due to an computing center outage in computing center High, HACMP tries to assign the IP label SIDap to a standby adapter on node DB/CI HA. If standby adapter en2 is chosen accidentally, everything is fine. If HACMP chooses adapter en1 for the assignment, this adapter has no active connection to a switch and the failover fails.

For this scenario, a post event script for HACMP has to be created. In this post event script, the status of the adapters is checked and it is determined that the newly configured adapter en1 has no connection to a switch. This is always the case if an adapter is chosen that does not belong to the local adapter group and the remote switch is not reachable. The script then initiates the configuration swap of the adapter.

Bright idea!

Configuration with no quorum buster

If you do not have a quorum buster, you have to enable the force varyon in the cluster event script `/usr/sbin/cluster/utilities/clvaryonvg`. This will force the varyon during the next takeover, even if half of the disks are missing or not reachable. After the switch, the missing disks are deactivated in the volume group. If the disks or the disk subsystem is again reachable, the following cleanup actions have to be done for the reintegration in order to enable the fully mirrored state of the volumes again:

- ▶ Remove the mirrors of the logical volumes from the lost hdisk (`<hdisk_lost>`) in the application volume group (`<appl_vg>`):

```
/usr/sbin/unmirrorvg <appl_vg> <hdisk_lost>
```
- ▶ Remove the lost disks (`<hdisk_lost>`) of the volume group (`<appl_vg>`):

```
/usr/sbin/reducevg <appl_vg> <hdisk_lost>
```
- ▶ Varyon the volume group (`<appl_vg>`) to clean up zombie disks:

```
/usr/sbin/varyonvg <appl_vg>
```
- ▶ Configure all disks to make the new disk available:

```
/usr/sbin/cfgmgr
```
- ▶ Extend the application volume group (`<appl_vg>`) with the new disks (`<hdisk_new>`):

```
/usr/sbin/extendvg <appl_vg> <hdisk_new>
```
- ▶ Add the mirrors to the logical volumes on the new disks (`<hdisk_new>`) of the application volume group (`<appl_vg>`):

```
/usr/sbin/mirrorvg <appl_vg> <hdisk_new>
```
- ▶ Synchronize the stale mirrors of the logical volumes in the application volume group (`<appl_vg>`) again:

```
/usr/sbin/syncvg -v <appl_vg>
```

Heartbeats with other than a direct cable

Sometimes it is necessary to send the heartbeats over distances larger than the distance allowed for serial RS-232 cabling.

Other methods for the transfer of heartbeats are:

Target mode SSA

In this case, the heartbeat is transferred over the SSA cabling, which can reach distances up to 10 km with fiber optic extenders.

Fiber optic converters

There are products of hardware manufacturers that convert the serial RS-232 signals from copper to fiber

optic cabling. The fiber optics allow distances much longer than the copper cabling.

Classic modem

It is possible to use two modems to elongate the serial RS-232 cable. A leased line or a twisted copper cable is necessary and offers distances up to 200 m.

If you wish to use RS-232 for serial heartbeats across something other than a direct RS-232 cable, you should be aware that HACMP classic and HACMP/ES treat RS-232 serial networks differently.

For HACMP classic, including Version 4.4, the default behavior when HACMP starts is to set the speed of a RS-232 serial network tty to 9600 bps. When a second node is started, it also sets the speed of its end of the serial line to 9600 bps and the nodes communicate at this speed. HACMP/ES up to and including Version 4.3.0 used to behave in the same manner as HACMP classic.

Pitfall ahead!

As of HACMP/ES Version 4.3.1, the default behavior is to start the ttys at 9600 and then, when the contact is established, the speed of each tty is raised to 38400. Therefore, for the RS-232 serial network to operate successfully, the cable network between the two ttys must be able to operate at both 9600 bps and 38400 bps. This is always true for standard cables (FC 3124 & FC 3125). However, it is not true for other serial cabling, for example, line drivers or asynchronous leased lines, which are normally configured for one specific line speed.

8.7 Maintenance of clusters

You must exercise extreme caution in making changes in an HACMP cluster. Both cluster nodes have to be kept absolutely synchronized in their configurations; otherwise, a takeover fails. For example, if a disk is added to a volume group and used within a file system on the primary node DB/CI HA, and the disk is not known on the backup node APserv HA, a takeover cannot take place.

8.7.1 Install programs and patches

The installed operating system level, the licensed program packages (LPP), and the patches have to be the same. A discrepancy in the installed program levels may lead to serious damages or inconsistencies.

This high level list gives an overview of the steps and applies to all kind of patches:

Test system Install and apply the patches without committing them using the following procedure:

1. **smitty update_all**
2. Input device/directory: /patch_directory
3. Commit software update?: no
4. Save replaced files?: yes
5. Reboot, if necessary

Test system Check the errorlogs from AIX and HACMP.

Test system Check the results for a longer time period.

If the results are satisfying, proceed with the next steps or remove the uncommitted software from the test system:

Node APserv HA Stop the cluster software (stop cluster node). Check to see if everything is stopped with **lssrc -g cluster**.

Node DB/CI HA Stop the cluster software (stop cluster node). Check to see if everything is stopped with **lssrc -g cluster**.

Node DB/CI HA Install and apply the patches without commit. Reboot, if necessary.

Node APserv HA Install and apply the patches without commit. Reboot, if necessary.

Node DB/CI HA Start the cluster software (start cluster node). Check the errorlogs from AIX and HACMP.

Node APserv HA Start the cluster software (start cluster node). Check the errorlogs from AIX and HACMP.

Pitfall ahead! **Cluster** Test the most important takeover scenarios to guarantee the functionality of the cluster.

Cluster Check the results for a longer time period.

If the results are satisfying, proceed with the next steps or remove the uncommitted software from the cluster nodes and the test system:

Node DB/CI HA Commit the installed software updates.

Node APserv HA Commit the installed software updates.

Test system Commit the installed software updates.

Additionally, all files and definitions have to be the same, such as /etc/hosts, /etc/environment, or the configured users and printers. Refer to Section 10.1.2, “Importance of synchronizing” on page 279 for synchronizing the environment.

8.7.2 File systems

If you make any changes to the file system definition, the automatic takeover lasts longer, because HACMP recognizes the changes and makes an adjustment through a lazy update of the volume group. With large volume groups, the takeover of the file systems to the backup node may last very long. This can be avoided through a proactive update of the volume group definition on both nodes.

In the following example, the file system /global/sapmnt is extended by 500 MB. The following steps extend this file system and make the appropriate changes on both cluster nodes for a working takeover:

Node DB/CI HA	Extend the file system /global/sapmnt by 500 MB: <code>/usr/sbin/chfs -a size=+1000000 /global/sapmnt</code>
Node DB/CI HA	Activate the volume group <appl_vg> for access by another node: <code>/usr/sbin/varyonvg -b -u <appl_vg></code>
Node APserv HA	Import the volume group definition in learning mode from a disk <hdiskx>: <code>/usr/sbin/importvg -L <appl_vg> <hdiskx></code>
Node DB/CI HA	Update the cluster volume group timestamp: <code>/usr/sbin/cluster/utilities/clupdatevgts <appl_vg></code>
Node APserv HA	Update the cluster volume group timestamp: <code>/usr/sbin/cluster/utilities/clupdatevgts <appl_vg></code>
Node DB/CI HA	Activate the volume group appl_vg for exclusive access: <code>/usr/sbin/varyonvg <appl_vg></code>
Node DB/CI HA	Verify the cluster topology and resources: <code>/usr/sbin/cluster/diag/clconfig -v -tr</code>
Node DB/CI HA	Synchronize the cluster if necessary: <code>/usr/sbin/cluster/utilities/clhare -t</code> <code>/usr/sbin/cluster/utilities/clhare -r</code>

The same steps apply when a new file system is added to the volume group. Additionally, the file system has to be registered to the resource group. If this step is omitted, the takeover fails.

Shadow database

In a nutshell:

- ▶ Shadow databases help in recovering from logical errors.
- ▶ Consider the implications of setting a production database back in time.
- ▶ Shadow databases offer more flexibility for the production backup.
- ▶ Snapshot mechanisms offer fast creation of shadow databases.

This chapter provides information on shadow databases. It describes the business benefits you can gain from deploying a shadow database. It shows the basic technical principles for an implementation and discusses commercial products that help to realize a shadow database.

This chapter shows the implications and the resource consumption of a shadow database system and how it can be used to reach synergies in a data center environment.

In this chapter, there are several practical hints for the integration of a shadow database with a fault- or disaster-tolerant scenario.

We describe other techniques than shadow databases that are suitable to meeting the business requirements for availability. We compare these techniques with the shadow database mechanism.

This chapter covers the highlighted area in Figure 9-1, which is derived from the SAP R/3 infrastructure model that is used throughout the book.

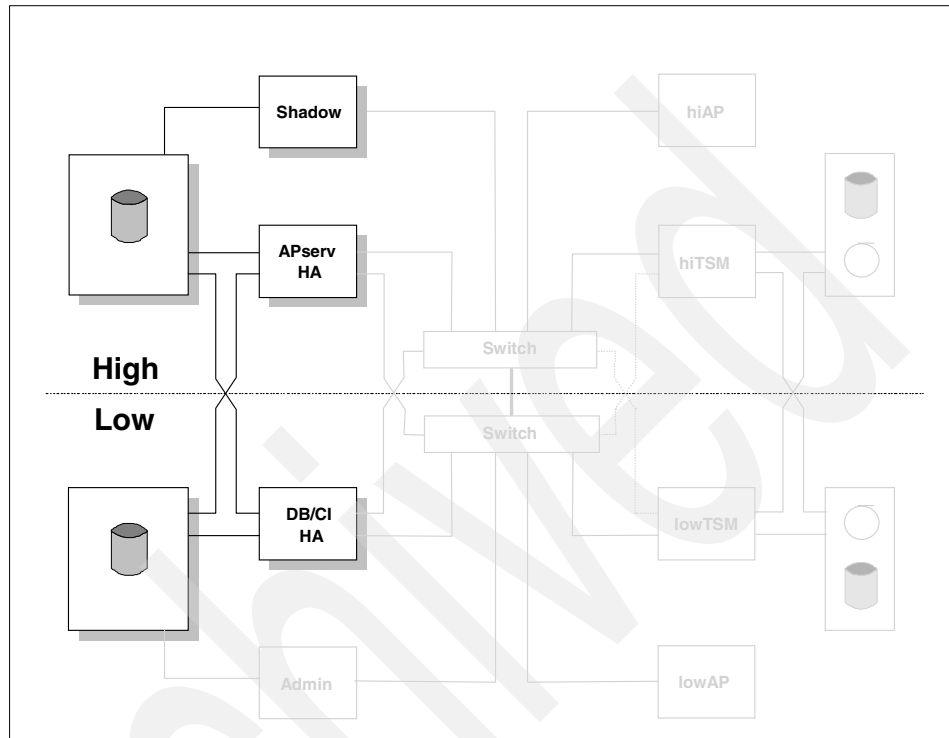


Figure 9-1 The shadow database

9.1 Business needs for a shadow database

In Section 3.4, “Building a disaster-tolerant model” on page 86, we describe a disaster-tolerant infrastructure configuration. The disaster-tolerant solution does not require the implementation of a shadow database. Nevertheless, there are good reasons for the deployment of a shadow database, even in cases where there is only a requirement for fault-tolerance. In Figure 9-2 on page 261, we show a chart which describes the main reasons for system outages, taken from the *Oracle8 Backup and Recovery Handbook*.

The pie chart shows that only 25 percent of outages are attributable to physical errors, and one percent to environmental problems that could be summed up as hardware failures. The scenarios in Section 3.3, “Building a fault-tolerant model” on page 77 and Section 3.4, “Building a disaster-tolerant model” on page 86 cover only these cases.

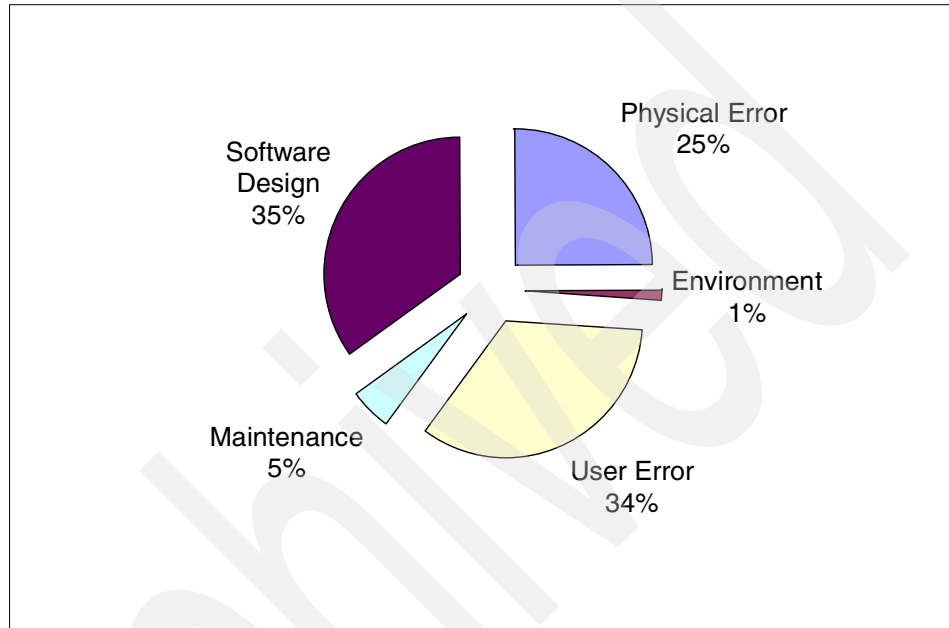


Figure 9-2 Main reasons for system outages in database environments

System outages due to maintenance work account for five percent of outages. This is usually counted as planned outage and does not need to have a negative impact on business critical processes.

Thirty-five percent of the outages are due to software designs that incorporate software defects. Thirty-four percent of user errors lead to system outages. Users in this context describe persons that are installing, administrating, managing, or using software or hardware systems.

These two sources of errors add up to approximately 70 percent of the outages. They usually occur inside the applications and have an impact on the content of the business data. They are not automatically detectable by a monitoring system, it is not easy to react to these errors with a predefined action.

Pitfall ahead!

If business data has been deleted in the application by a user or software error, no mirroring of the data or RAID 5 implementation can recover from that action, because this is not a hardware issue.

In this case, only a backup of the data from a previous day can restore the data. This is one of the reasons why a reliable backup and recovery concept has to be implemented. Only if the backups are working correctly and the recovery has been tested successfully is there a possibility to restore the business data.

In order to restore the data quickly for large databases with a size of hundreds of GB, it is necessary to implement a high performance backup and recovery system with exceptional restore characteristics. Even then, it might take hours to restore the data and apply redo logs to the point in time where the incident happened.

The expected recovery time that is specified in the service level agreements might be so demanding that it is either extremely costly to provide the needed restore capacity or it might not be technically feasible. This is where a shadow database can be very valuable.

9.2 Definition of a shadow database

In this section, we want to give a definition of a shadow database and want to explain how it is working.

Shadow database An instance of a database that has been created as a copy of the original database on a different server system. It runs in constant recovery mode, which means it applies redo logs that are permanently copied from the original database system to the shadow database server machine.

Figure 9-3 on page 263 shows the basic principles of a shadow database and depict the mechanisms to build and run it.

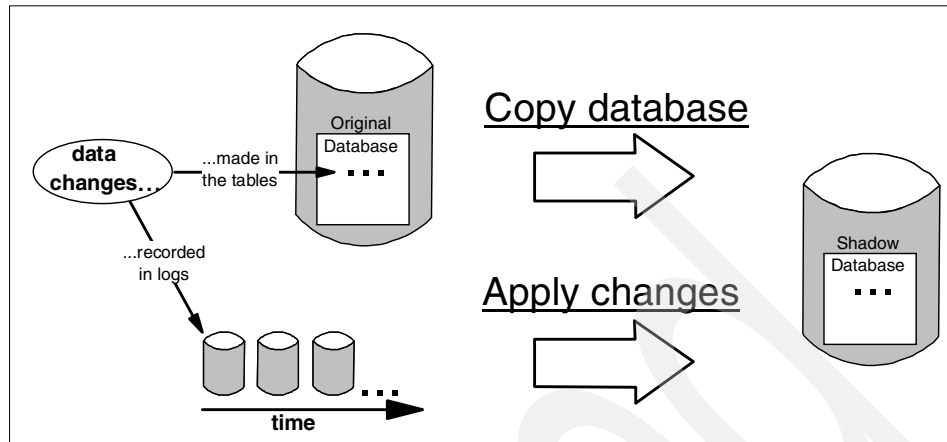


Figure 9-3 Basic principle of a shadow database

In the first step, the database software is installed on a separate machine, the shadow database server. Then the original database is copied to disks that are used by this machine. This can be done by means of an online or offline backup of the original database from a tape library or by just copying the database files from the disks of the original system.

The changes made to the tables in the RDBMS are recorded in redo log files. These files are copied to the shadow database server and are then applied to the shadow database. There is usually a time delay between the copying of the files and the applying of the changes to the shadow database.

9.3 Benefits of a shadow database

There are plenty of possible uses for a shadow database, and in this section, we describe some of the most important features that are provided by it. A shadow database offers a lot of flexibility in the case of database malfunctioning.

9.3.1 Logical database errors

The most obvious implementation is to recover from a *logical database error*. This term is used for software or user initiated data loss situations inside a database caused by accidentally deleting or corrupting data in the database.

To facilitate the recovery from such a situation, it is important to apply the redo logs on the shadow database with a delay of several hours. A long enough time offset should be chosen to ensure that the error can be discovered on the production system before it is applied to the shadow database.

In such a logical error situation, it is extremely important to first stop the recovery mode of the shadow database. The next step is to find out what time the error occurred on the original database. With that knowledge, it is possible to run a point-in-time recovery on the shadow database, just before the error occurred.

For major incidents, the original database has to be shut down and the shadow database must be activated and started. In this case, the original database can be used to find the cause of the error or to extract the last transactions that have been processed after the incident. These transactions can then be rerun manually on the shadow database system.

For minor incidents, such as the accidental deletion of only a few master data records, you can open the shadow database in a read-only mode and extract the data from the shadow database without switching the production to the shadow database server.

9.3.2 Physical database errors

As mentioned before, the shadow database is also suitable for fault-tolerant scenarios. In these scenarios, the disk subsystem is usually protecting the data by RAID 5 or mirroring. Single points of failure are eliminated, but the shadow database adds availability in case of double failures, because it could happen that the whole RAID 5 storage subsystem or two disks of a mirror fail at the same time.

A recovery in this case would be possible from tape backups, with the additional application of offline and online redo logs, but the time for the restore would be considerable. If there is a shadow database, the recovery actions would be very simple and fast. It is only necessary to run a full recovery on the shadow database system to obtain an up-to-date status of the production database.

9.3.3 Database consistency checks

There are two different ways of backing up a database: the online and the offline backup. In case of an online backup, it is essential to have the archived redo log files available that have been written during the time of the backup in order to bring the database into a consistent state after a restore.

In the case of an offline backup, it is important to have all archived redo logs available in between two successive backups. They are needed in order to have the option for a point-in-time recovery, where the database should be restored to a state that is in between the time stamps of two offline backups.

Bright idea!

It is not only important to archive these redo logs regularly and reliably, it is also extremely important that these archived redo logs are correct and complete. A shadow database offers the opportunity to check the completeness and the correctness of the redo log files by applying them on the shadow database before archiving them on tape. The constant recovery mode of the shadow database is also checking the consistency of the original database.

9.3.4 Independent backup

The shadow database offers two more advantages in terms of backups. In many SAP R/3 environments, the time window for backups of the production systems is constantly decreasing, while the amount of data is increasing. Through the use of online backups, it is possible to run backups without any downtime of the system, but the I/O subsystem is heavily used during backups and may have a negative impact on the application performance.

A possible solution is to use the shadow database for backups. Due to the fact that the shadow database is not opened and not used for normal operation, it is possible to stop the recovery at any time. The database can then be backed up in offline mode to tape storage at any time of the day. In addition, it does not stress the storage subsystem of the production system, because the shadow database is implemented on its own storage system. This mechanism is supported by the SAP R/3 tools **brbackup** and **sapdba**, which include the recording of the backup activity and return codes in the original database.

9.3.5 Geographical disaster-tolerance

The mechanism of the shadow database allows you to copy the redo log files to hosts that are located at a geographical distance. In this case, the shadow database host has to be connected via a Wide Area Network (WAN) leased line. The availability of high speed networks and the relatively small size of the redo log data volume makes it possible to easily transfer the log information via the WAN connection and apply it at the remote shadow database site.

9.3.6 Upgrade improvement

In an SAP R/3 environment, there are upgrades of various software components from time to time. This includes the operating system, the database software, and SAP R/3 software. A shadow database can be used as a fallback system in case of a failure of the upgrade. The use of a shadow database helps to expand the time window for the upgrade, because the time to go back to the original database content is much shorter with a shadow database than with the restore of the original database with conventional tape backups.

Another scenario for the use of a shadow database is the acceleration of hardware upgrades. If new hardware systems are necessary to provide more performance for the SAP R/3 database, the switch to the new system can be made with reduced downtime, when a shadow database is created on the new server. At the time of the switch, the old database is shutdown, the last redo logs are copied over to the new system, and the database on the new hardware is just recovered to the actual time.

9.4 Implications of a shadow database

There are several implications when a shadow database is used in an SAP R/3 environment. One area covers the technical implications, the other covers the organizational area.

9.4.1 Technical implications

In order to integrate a shadow database into your SAP R/3 environment, you have to provide hardware resources for running the shadow database.

Disk storage

It is necessary to provide the same amount of disk space as the original database plus space for the copied redo log information.

Pitfall ahead!

The amount of space for the redo logs can be quite substantial if you decide to run the shadow system with a long time delay. Very active SAP R/3 systems have been observed to produce as much as 50 GB of redo log information per night. It is recommended to monitor the redo log volume of the original database (if possible) and size the disk storage of the shadow database system accordingly.

If you want to be completely independent of any failure of the disk subsystem where the original database runs, you have to use a separate storage subsystem for the shadow system. This also guarantees that there is no negative performance impact on the original database if there is a backup running from the shadow database, because the reads are coming from a different storage subsystem.

Server hardware

An additional server system is needed for the shadow database. The resource consumption for applying the redo logs is not as high as for the standard database processing on the original database. This is mainly due to the fact that there is no read activity on the shadow database; thus, a lot of the internal database activity is avoided.

Due to the interdependencies in the data, there are frequent locks at the table and row level during the processing of the recovery transactions. Thus, the recovery mechanism of the RDBMS runs, to a large extent, in a serialized way.

Bright idea!

Therefore, the shadow database systems usually require only a fraction of the number of CPUs of the production database system, typically a factor of four to eight less, depending on the ratio of write to read activity.

9.4.2 Organizational implications

Implementing and operating a shadow database is relatively easy and straightforward. It is much more difficult to define a process for the management of error situations. The decision to stop an SAP R/3 production system has very severe consequences. The most obvious consequence is that the SAP R/3 system is no longer available for interactive users and external systems.

The person in charge of the SAP R/3 operation has to evaluate the error situation and has to find out whether the error situation is severe enough to justify a shutdown of the application. In case of hardware failures, it is much easier to make a decision, because there are no alternatives than to initiate a failover. In case of user or software errors, there is sometimes the possibility to repair the damage in the system without stopping the operation.

The decision to stop the production database implies a recovery scenario on the shadow database system to a point-in-time. This is the next difficulty, because a point-in-time must be specified up to which the shadow database should be recovered. On the one hand, this time should be as close to the time where the error occurred; on the other hand, it must not be later, because then the same error would be introduced into the shadow database and this means that the shadow database is also defective.

Pitfall ahead!

Another important aspect is the consistency of external systems that have interfaces to the SAP R/3 system and exchange data. A point-in-time recovery of the production database ensures consistency of the data and might recover the data, except for maybe the work of a short period of time, which can be manually reentered. However, the connected external systems are usually also running databases, so they should also be reset to the same point-in-time. This may be possible for systems that are also running a shadow database; other systems would have to be restored from tape storage and recovered. This is a lot of tedious work that would involve many systems, and the likelihood of failure exists.

Even more serious is the fact that some systems just cannot be reset to a certain point-in-time, such as material inventory systems, where real goods are taken out of the warehouse or brought into the warehouse. These kinds of transactions cannot easily be repeated. In such cases, it does not make sense to run a shadow database system.

Important: If external systems do not allow a restore back in time, it does not make sense to run a shadow database for the SAP R/3 system. Organizational precautions have to be established to cover for such cases.

In conclusion, it must be pointed out that the decision to use a shadow database is mainly a question of business recovery possibilities. The decision to go “back in time” has many consequences and it is very difficult to gather enough information for a correct decision, especially in a stress situation when a severe user or software error has been detected.

9.5 Implementation of a shadow database

In this section, we describe ways to implement a shadow database with different techniques and show the advantages of the various solutions.

9.5.1 General considerations

We have not yet discussed the implementation details for shadow databases. The basic principles are identical for both Oracle and DB2.

The built-in tools of Oracle allow the automatic transfer of the redo log files to the shadow database system, but do not provide mechanisms for retries in case of transfer failures. In order to improve this situation, the transfer may also be managed by self-written tools, which allow a more sophisticated exception handling.

DB2 uses so-called *User Exits* for attaching self-written programs to the database. This is also the standard mechanism for the management of redo log files, which is used by all backup and archiving tools. This program can be extended to implement the data transfer to the remote shadow database system.

The most difficult part of running a shadow database is to maintain it. The information changes of the original database are logged in the redo log files and applied to the shadow, but the structural changes of the database are not covered in this process. Examples for structural changes are the renaming of

data files or the creation of new data files. Also, the change of configuration parameters in the initialization files, the change of software versions, or the application of fixes, creates differences between the two database instances and a successful redo log application or activation of the shadow database may fail.

Pitfall ahead!

The usability and manageability of a shadow database is strongly connected with the possibility to monitor the status of recovery on the shadow database. A method should be implemented to see whether there are any problems in the data transfer from the original to the shadow database system and whether there are any problems while applying the redo logs.

There are commercial tools on the market that solve all the above mentioned problems. They implement the missing functions and offer additional features for setting up and monitoring shadow databases without the need of detailed database recovery functions. A reliable product for Oracle and DB2 is called *DBshadow* and is offered by the company Libelle. Oracle and IBM also plan to offer the missing features and monitoring tools in the next releases of their database management systems.

In the next sections, we describe two different possibilities when selecting the size of hardware for the shadow database server.

9.5.2 Large shadow database server

“Server hardware” on page 266 explains the required performance for the shadow database. It may nevertheless be reasonable to select a shadow database server that has the same performance as the production system. This is necessary if the shadow database server is supposed to take over the production workload in case of a logical error in the production database.

In this case, the network configuration also has to be changed in a way so that the SAP R/3 application server instances and other clients can connect to the shadow database, which acts as the new database server.

9.5.3 Small shadow database server

Bright idea!

Another option can be used when running the shadow database on server hardware that is less powerful than the production database server. Instead of swapping the network identity of the system to the identity of the former production database, it is also possible to swap the disk storage space, so that the contents of the database are moving from the shadow database system to the production database system. This is facilitated by the use of Fibre Channel based SAN environments, where all the disk space can be connected to any machine on the SAN. This technique is shown in Figure 9-4 on page 270.

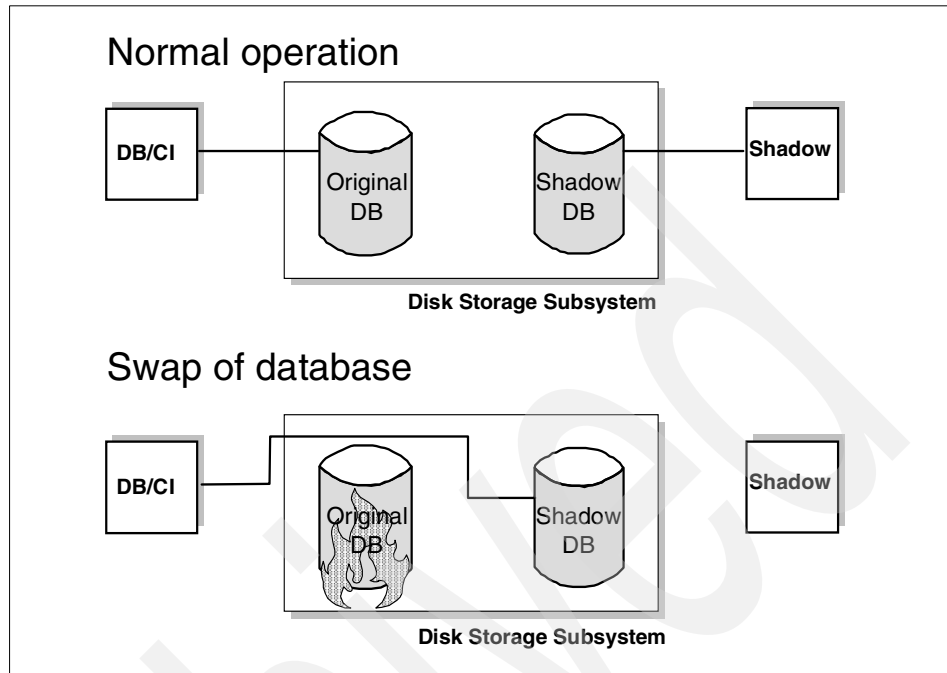


Figure 9-4 Database swap in case of an error in the production database

During normal operation, each of the database servers is working on its own disk area, where the respective databases are located. In case of a failure of the original database, the shadow database is activated and the disk storage area of the shadow database is physically and logically attached to the DB/CI database server.

For a complete recovery from the error, it is necessary to attach the storage area of the former original database to the shadow database server system and re-initialize the shadow database. This corresponds to a complete swap of the two disk storage areas between the two servers, as shown in Figure 9-5 on page 271.

All the steps necessary to perform the swap of the disk storage areas can be done with standard AIX Logical Volume Manager commands.

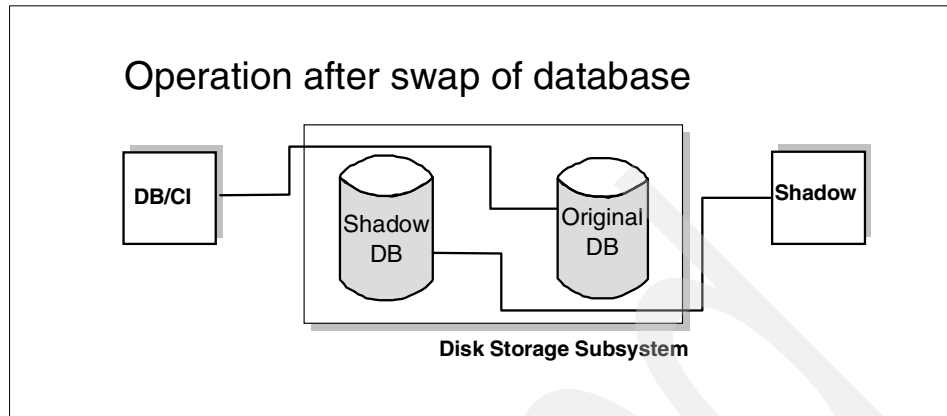


Figure 9-5 Disk assignment for operation after a swap

9.5.4 Considerations for a disaster-tolerant scenario

In a disaster-tolerant environment, there is an additional requirement for a shadow database solution. In the previous section, we have replaced the storage area of the production system with the shadow database storage area. In order to still have a disaster-tolerant configuration, it is also necessary to have the disk storage of the shadow database mirrored between the two data centers. This is a very costly configuration, as this would add the size of the production database to the storage volume requirements.

The best way to run a shadow database in a disaster-tolerant environment is to leave it unmirrored, but implement a procedure for mirroring the shadow database storage area after it has become the new production database. The storage area for the mirroring can be taken from the unmirroring of the old original database.

9.5.5 Performance enhancements

Considering the fact that today's SAP R/3 database sizes are of the order of several hundred GBs, it is important to speed up some of the time critical operations in a shadow database environment. The step that consumes most of the time is the creation of an initial copy of the production database.

This can be managed by either one of the following steps:

- ▶ Restoring files from a backup
- ▶ Copying the data files from the production database while it is in a consistent state

- Creating an instant copy of the disk content by using a feature of the storage subsystem

Bright idea!

The last item is the fastest way to generate a copy. The speed of this mechanism depends on the performance of the subsystem. With the IBM Enterprise Storage Server, it is possible to create a so called *FlashCopy* in a couple of seconds.

The FlashCopy mechanism requires that the source volumes and the target volumes are located in the same logical subsystem of the ESS. This restricts the use of the shadow database to the same ESS as the production database.

9.6 Comparison with FlashCopy

Many companies consider the implementation of other techniques than shadow databases to ensure a tolerance in case of logical errors. As already pointed out, the mirroring of the data or a RAID 5 implementation does not protect against logical errors.

Snapshot mechanisms similar to FlashCopy in the ESS are implemented in many advanced storage subsystems. They offer the possibility to make frequent and instant copies of databases that can be used as sources for a fast recovery. We discuss the differences between FlashCopy and a shadow database implementation in the following sections.

9.6.1 Time delay

Shadow databases work in a continuous way. The time delay between the state of the original database and the shadow database state is always fixed. With FlashCopy or similar tools, there is always a discrete interval in which the copy is made and, thus, the time difference between the actual state and the copy varies. An example can illustrate this:

- We assume a configuration where we create a FlashCopy at 0:00, 8:00, and 16:00, and further assume that at 17:00, we discover a logical error that happened at 15:00.
- If we want to restore the database status with the timestamp of 14:59, we need the FlashCopy of 8:00 to recover forward in time, because the 16:00 copy is too late in time. This means that at least two versions of the FlashCopies had to be kept. If we have only the latest version from 16:00, we have to restore the database from tape storage and have to recover it. In this case, we did not benefit from the FlashCopy feature.

- ▶ A shadow database follows the original database, for example, in a fixed eight hours time interval, and therefore offers always the possibility of recovering to any point-in-time in these eight hours.

9.6.2 Hardware separation

FlashCopy offers no protection against a hardware failure in the same storage subsystem. If one rank in the ESS fails, both the original and the copy are not available. A shadow database can be physically separated from the original database and therefore offers more safety.

9.6.3 Consistency check

If a database is copied with a tool like FlashCopy, the corruption of a database remains undetected. If the shadow database recovery mechanism is used, a defect in a redo log file can be detected.

9.6.4 Backup from copy

The FlashCopy on an ESS can either be initiated with the *NoCopy* option, where the data is not physically moved inside the storage subsystem, or the *Copy* option, where the data is moved. The movement stresses the storage subsystems' internal bandwidth and might lead to a negative performance impact. In both cases, the backup from the same storage subsystem might influence the performance of the production system.

The storage area of the shadow database may be located on a different storage subsystem. During the initialization of the shadow database, there might be a performance impact on the production system, while files are copied from there. However, the backup from the distinct shadow database storage subsystem does not interfere with the performance of the production database.

Thus, a backup from a shadow database with a separated disk storage subsystem may be faster than from a FlashCopy.

9.6.5 Creation of test systems

The use of a shadow database for creating a quality assurance or test system is possible, but the effort for creating the database for a one time action is relatively high. In this case, FlashCopy offers the same functionality in a less complex way.

9.6.6 Summary

We summarize the strengths of the two methods in Table 9-1.

Table 9-1 Protection against failures

Error condition	Shadow DB	FlashCopy
ESS rank failure	Yes	No
Logical error	Yes	Yes
Disaster safe	Yes	No
Redo log error	Yes	No
Structural changes	Manual	Yes

“Shadow DB” in this table describes an implementation with a separate disk subsystem, which explains the disaster tolerance of the solution. The manual intervention needed for structural changes to the database in case of a shadow database can be circumvented if a third party tool is used.

Depending on the business needs and the involved costs, the most suitable implementation should be chosen.

Hints and tips

In a nutshell:

- ▶ Use a central point of management for your SAP R/3 environment.
- ▶ Keep the configuration of the environment synchronized.
- ▶ Create scripts to automate repetitive tasks and to prevent errors.

In this chapter, we present hints and tips for implementing and operating a reliable SAP R/3 infrastructure.

In the first section, we discuss the principles of administration in the SAP R/3 environment. We present a concept for naming, central management of configuration files, and documentation.

Based on these concepts, we go into details for the implementation of AIX and SAP R/3.

In the last section, we present hints and tips for operating the SAP R/3 infrastructure.

This chapter covers the highlighted area in Figure 10-1 on page 276, which is derived from the SAP R/3 infrastructure model that is used throughout the book.

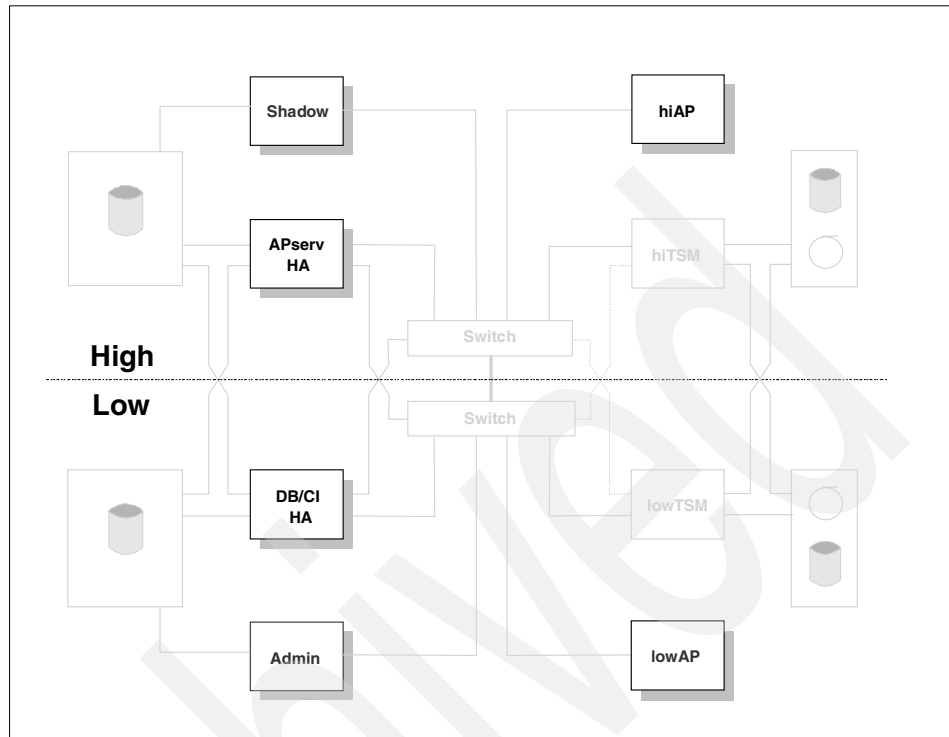


Figure 10-1 Hints and tips

10.1 Principles of administration

In Section 2.4, “Guidelines for the implementation” on page 17, the basic principles for the implementation and administration of a complex environment are outlined. In this chapter, we take a closer look on the details of most of these principles:

- ▶ Obey the SAP standards.
- ▶ Use reasonable concepts for naming (refer to “Naming concept” on page 277).
- ▶ Use unification (refer to “Importance of synchronizing” on page 279).
- ▶ Implement a high degree of automation (refer to “Easy interface for administration” on page 279).
- ▶ Manage global configuration information in a single place and introduce administration domains (refer to “Principles of the SP system” on page 280).

- Observe obviousness and avoid side effects.
- Introduce housekeeping (refer to Section 13.4.1, “Housekeeping batch jobs” on page 422 and Section 13.6, “Cleaning the log files” on page 434).
- Create documentation and keep it up-to-date (refer to “System documentation” on page 282).

10.1.1 Naming concept

A self-explanatory convention of names that are easily remembered may cause less daily effort. In the following, a few recommendations for naming conventions are presented.

IP labels

A rule for building IP labels can be `<hostname><suffix>` or `<SID><suffix>`, where:

- `<hostname>` is the host name of the machine, while `<SID>` is the SAP R/3 system ID.
- `<suffix>` is the label suffix for:

bkp	Backup network
ctl	Control network
db	Database and central instance
ap	Application server instance

For example: `sap100bkp`, `sap50ctl`, or `PRDdb`

Volume groups

A rule for building volume group names can be `<application><type>vg<number>`, where:

- `<application>` is an application, such as:

sapr3	SAP R/3
catia	CATIA server
nfs	NFS server
tsm	Tivoli Storage Manager
nim	Network Installation Manager

- `<type>` means:

p	Productive
q	Quality assurance
d	Development

t	Test
w	Data warehouse
e	Education

- ▶ **<number>** is an optional continuous number
For example: `sapr3pvg01`, `sapr3evg01`, or `catiavg`

Logical Volume

A rule for building logical volume names can be `lv<content><SID>.<number>`, where:

- ▶ **<content>** is the volume's content, such as:

data	Database file systems, such as <code>/db2/SID/sapdata4</code>
oracle	Oracle directory, such as <code>/oracle/SID</code>
sapmnt	Global SAP R/3 directory, such as <code>/global/sapmntPRD</code>
spool	Print spooling directory, such as <code>/global/SID/sapspool</code>
trans	Transport directory, such as <code>/usr/sap/trans</code>
jfslog	Journalized File System log
- ▶ **<SID>** is the SAP R/3 system identifier, such as:

PRD	Productive
QAS	Quality assurance
DEV	Development
- ▶ **<number>** is an optional continuous number
For example: `lvdataDEV.01`, `lvoracleQAS`, `lvtransPRD`, and `jfslogPRD.01`

Script names

We advise you to think about a naming convention for scripts that you create on your own. A possible convention can be `<action>_<object>`, where:

- ▶ **<action>** is what the script does, such as:
start, stop, backup, sync, show, document, mount, or manage
- ▶ **<object>** is the target object of the action, such as:
sap, database, hosts, errorlog, system, scripts, file system, or resgroup

For example: `start_backup`, `check_mounts`, `show_logs`, or `sync_printers`

10.1.2 Importance of synchronizing

It is essential for a complex environment to keep all parts as uniform as possible. This section describes some problem areas that can be avoided through synchronization.

In a uniform environment, it is important to have the same installed program levels as well as the same installation of printer or device drivers.

A printer driver for a new printer model can be installed on all of your systems in parallel instead of only on the development system. The additional effort can be neglected in comparison to the effort, which is needed to find out why a printer is working correctly on the development system but is producing senseless output in the production system. Different printer definitions, such as a changed paper size from letter to A4 on only one application server, require a lot of time for error investigation.

If NFS is used in your environment, you must synchronize the user IDs of all users among all servers. Otherwise, the different permissions of different users with their user IDs can cause for example hanging transports in SAP R/3's Transport Management System (TMS).

Different times or time zones on different servers can cause trouble in normal operation and will cause trouble in case of restore and recovery action.

A takeover in a highly available environment fails or leads to serious damage if the two cluster nodes are not configured in the same way or with the same values.

The time for investigating problems, which arise from an unsynchronized environment, can easily be saved. The search for errors in the configuration can be reduced dramatically, and this time can be used for a deeper check of the system, such as performance observations, or for checking the error reports for predicted hardware failures.

10.1.3 Easy interface for administration

In AIX, there are lot of commands, and all of these commands can be executed with a wealth of parameters. In daily operation, only a few are used for administration purposes. It is very easy, and also advisable, to put the daily tasks in tiny scripts to fulfill the different administrative tasks more reliably and without mistakes caused by typing errors.

If these scripts are written in a flexible way, they can be used for different systems and can save a lot of time. An example for a script can be **manage_sap -start PRD** to start the productive SAP R/3 system with the system identifier PRD. The script can be used to control the development system too, for example, **manage_sap -stop DEV**.

10.1.4 Principles of the SP system

The SP concepts were created by IBM based on many years of experience with mainframe operation and their maintenance.

In an open world, several servers will go out of synchronization within days or hours and will cause harmless and also harmful problems, which have to be solved manually by patching the differences. To support the daily work of administrating of more than one server, the following features are offered in an SP environment.

Control Workstation

The Control Workstation in an SP environment offers:

- ▶ A single point of control.
- ▶ Central user ID management.
- ▶ A central source for Licensed Program Products, patches, and backups to maintain different machines with different operating system images.
- ▶ A central access point to the hardware (Perspectives) to switch on/off the machines remotely.
- ▶ File collections, to synchronize several files among all servers to prevent error situations.
- ▶ A dedicated internal LAN connection to the servers (nodes).
- ▶ Serial connections to every server for initial administration tasks, even if an IP is not set up yet.
- ▶ Special parallel commands and working collections to work with in a distributed server environment, such as **pcp**, **prm**, **pmv**, **p1s**, **pdf**, **pfind**, or **pps**.
- ▶ Time master functionality.

There are servers that are connected to an SP frame or are inside an SP frame. These servers:

- ▶ Are independent and work in a shared-nothing architecture.
- ▶ Are integrated in a space-saving way in a rack (up to 16 autonomous servers).

- ▶ Can use a high speed server interconnection.
- ▶ Work on their own operating system image.
- ▶ Allow the execution of several software products, like Lotus Domino, WebSphere Application Servers, and SAP R/3 (in one SP-Frame on different machines). Nevertheless, all servers can be managed and maintained from one single point of control.
- ▶ Allow easy administration, as only a few different operating system images can be used for dozens of servers.

If there is no SP in your environment, you can make use of these features, but in a less comfortable way. Individual tools have to be built for some of the tasks; some others are standard features in AIX.

Administrator server

In Figure 10-1 on page 276, the administration server *Admin* fulfills the function of the Control Workstation in a non-SP environment. It is the central point of management for synchronizing purposes, the NIM and Web server for the environment, and the primary workstation for the administrator. The concepts of an administration console are outlined in Section 3.2.6, “Administration server (Admin)” on page 73.

The server *Admin* can be configured as a central mailbox for mails from all servers. This mail rerouting has the advantage that it is not necessary to log on to several servers to check the local mail there.

The central mailbox leads to less effort for the system administrator to check out mails from the failure prediction diagnosis, which come with AIX and IBM @server pSeries, to avoid unplanned system outages.

As with for the SP Control Workstation, it is possible to install and configure an 8-port or an 128-port adapter (IBM FC 2943 or FC 2944) in order to increase the number of serial ports for connecting the server Admin with the serial port of other servers in the environment. Then you are able to access the service processor of each server using the AIX commands **ate** or **cu**. The access to the service processor offers the ability to start up the server remotely, even if this server is switched off.

Due to cable length limitations of serial cables, the server has to be in the same area as the RS/6000 or IBM @server pSeries servers. There are several possibilities to extend serial lines, for example, serial-to-optical or serial-to-ethernet converters, if necessary.

If the administrator prefers to manage the cluster remotely, he can login through a TCP/IP connection to the Admin server and issue the `cu` command in his telnet session.

An alternative for accessing the service processors of the different servers is to use a terminal server. In our scenario, a terminal server can be connected to the control network via serial connections to all service processors. A terminal server provides a TCP/IP connection to serial or graphical devices, so that remote access is possible.

For more information concerning Service Processor and a detailed description of the configuration, refer to the white paper *IBM @server Clustered Computing*, found at:

<http://www-1.ibm.com/servers/de/eserver/pseries/library/specsheets>.

10.1.5 System documentation

The environment and the settings of a server must be well documented in order to install a new server with the same settings in case of a hardware failure or a disaster.

In a synchronized environment, many settings, such as users and printers, are equal on all servers, so the documentation of these items is not necessary. However, the configurations of the network or storage adapters are individualized on each server. The configuration of the Workload Manager, for example, may only be used on one server throughout the environment. A failure of a server requires a lot of manual effort to achieve the same state it was at before.

It is a good idea to automate the task of documentation in a script, which dumps the configuration information into a flat file. This file can be copied to the administration console. With a few lines of added HTML code (Hyper Text Markup Language), it is possible to scan through the files with the Netscape Navigator Web browser, which is delivered with the AIX Bonus Pack.

In a standard AIX environment, it is recommended that you set up a Web server for the AIX documentation (manual pages). This Web server can be customized to include the system documentation files. Then it is possible to access the documentation from your personal computer.

The next step to an efficient documentation server is the inclusion of all other product documentation in electronic form. This can be the SAP R/3 Online Documentation, the DB2 and Oracle online help, the TSM server documentation, or the HACMP documentation.

With the installation of a new software version in your environment, the information on the Web server must be updated too. This way, it will always reflect the actual documentation of your environment.

10.2 AIX implementation

The AIX operating system provides a reliable basis for the implementation of an SAP R/3 system. The following sections contain some hints and tips to make the administrator's life easier.

10.2.1 Network Installation Manager (NIM)

The Network Installation Manager (NIM) enables the remote installation and update of servers without any additional media. A NIM server provides the boot and installation service and the NIM client is booted over the network and is installed with the sources provided by the server.

This centralized approach for the handling of program sources and backups offers the opportunity to get the same installed program level on all servers.

A NIM server installation consists of:

- ▶ The NIM server software
- ▶ The different scripts for configuration
- ▶ The sources or filesets of the Licensed Program Products (LPP sources)
- ▶ The small operating system image called Shared Product Object Tree (SPOT)
- ▶ The `mksysb` system backup images of the NIM clients

As a golden rule, for an optimal operation in a NIM environment, the operating system level of the LPP sources, from the SPOT and from the clients, must match.

There are several possibilities for handling the LPP sources in a NIM environment. The following tips help to reduce the needed space and offers a more flexible way for administrating the different levels of filesets.

Instead of using only one lppsource directory on the NIM server for all sources, a single directory tree for every maintenance level should be created. So base level and update filesets should first be separated and then linked to a common directory. This provides the opportunity to use the base level filesets multiple times to build individual LPP sources for different update levels, and it protects existing LPP sources and allows different OS levels without wasting space. Figure 10-2 gives an overview of the directory structure.

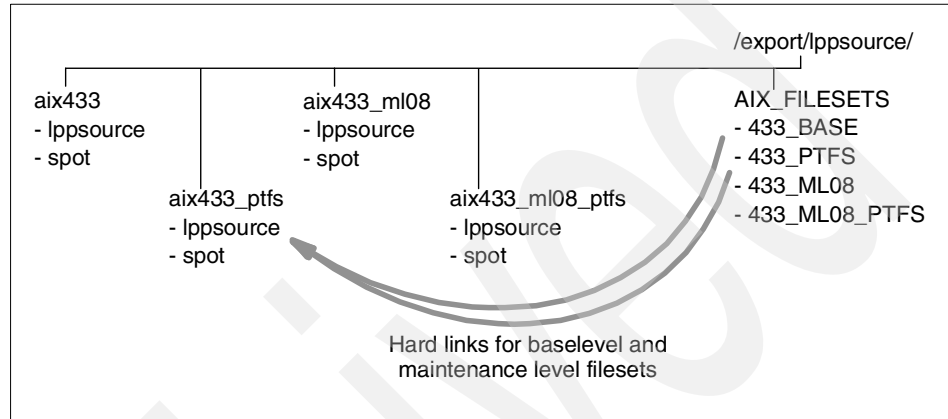


Figure 10-2 LPP source handling

The filesets are placed in the directories written in capital letters. For the different maintenance levels, the sources are linked via hard links into the target lppsource directories written in lower case small letters. Then the SPOT is created in the particular spot subdirectory for every level.

The advantages of this setup are low space requirements, easy administration for installation, and easy housekeeping. If a specific level is outdated and not needed anymore, the directories with the outdated filesets can be simply deleted.

10.2.2 Time synchronization

A synchronized and correct time throughout the entire environment is very important for proper operation. The communication between servers, the timestamps of files in local or NFS file systems, or the backup filesets need the same trusted time base. The Network Time Protocol (NTP) is used for clock synchronization across servers.

An NTP server in your local environment or in the Internet can be used as a time master. The IP address of the time server is defined in the file `/etc/ntp.conf`.

A sample configuration file consists of the following line:

```
# the server timeserver.ibm.com is an hypothetical example
server timeserver.ibm.com
```

If there is no connection to the Internet and no NTP server connected to an CUT (Coordinated Universal Time) server by radio is installed in your location, you can choose the admin console with its installed clock as a server in order to have the identical time in your environment. In the directory `/usr/samples/xntp`, you will find an example of how to treat the local clock as a reference in the file `localclock.conf`:

```
# treat the local clock as a reference
server 127.127.1.0 prefer
```

This configuration is set on the server Admin. All other servers in your environment then reference the server Admin as a time server.

Refer to Section 10.4.3, “Time differences and daylight saving time” on page 302 for the SAP R/3 view on this topic.

10.2.3 File system for logs and documentation

It is recommended that you create a separate file system for all kinds of log files, script return codes, and server documentation. This avoids the overflow of standard file systems, such as `/` and `/var`, which may have serious impacts to normal operation.

The use of the file system `/tmp` should be avoided too, because of the resulting problems from filling up the `/tmp` file system, for example, aborting the system `mksysb` backup command. Also, the execution of cleanup jobs can unintentionally destroy necessary documentation information.

A file system in the `rootvg` is created on every server with the mount point `/var/domain`. If a server belongs to an SAP R/3 system, it uses the subdirectory `/var/domain/SAP`. A Tivoli server or an NFS server could use similar directories, such as `/var/domain/TIV` or `/var/domain/NFS`.

10.2.4 File systems

It is possible to define mount groups for several file systems. These mount groups allow you to categorize file systems and mount them in a single step.

For example, the mount group `SID` on server `DB/CI` containing a DB2 database can consist of the following file systems:

- `/usr/sap/SID`

- ▶ /sapmnt/SID
- ▶ /usr/sap/trans
- ▶ /home/sidadm
- ▶ /db2/SID
- ▶ /db2/SID/log
- ▶ /db2/SID/sapdata1 to sapdata<n>

The mount group SID_AP on the application servers may consist of these file systems:

- ▶ /usr/sap/SID
- ▶ /sapmnt/SID as NFS mount from DB/CI
- ▶ /usr/sap/trans as NFS mount from DB/CI
- ▶ /home/sidadm as NFS mount from DB/CI

For NFS, we recommend the usage of soft mounts. If hard mounts are used and the exported file system is not available, the access to a file system may hang endlessly if no time out value is set. This can cause a server to hang in the reboot process, if the automount option is set to true for the NFS file system.

For all servers, the file system attribute *automount* should be set to *false*. Otherwise, server reboots may lead to lock situation and the sequence of the reboots may have an severe impact on other servers, which should be avoided.

10.2.5 Redirect the console

In an AIX default installation, the console messages are displayed on the device that has been defined as the AIX console. It is not possible to analyze errors once they have been overwritten on the console by subsequent messages. Therefore, it is strongly recommended that you redirect the console output to a file, which enables further investigation in case an error arises.

The standard AIX command **lscons** outputs the current console device:

```
/usr/sbin/lscons
/dev/lft0
```

With the command **swcons**, the new device or a log file can be set:

```
/usr/sbin/swcons /var/domain/SAP/logs/console.log
swcons: console output redirected to: /var/domain/SAP/logs/console.log
```


To redirect the console log at every system reboot add a line to the `/etc/inittab` by using the command:

```
/usr/sbin/mkinitrd 'swcons:2:once:/usr/sbin/swcons  
/var/domain/SAP/logs/console.log'
```

10.2.6 Clever tools

A lot of tools, for different purposes, can be found on the Internet. Check out these links for more information:

- ▶ The AIX toolbox for Linux applications:
<http://www.ibm.com/servers/aix>
- ▶ The University of California in Los Angeles:
<http://aixpdslib.seas.ucla.edu/aixpdslib.html>
- ▶ The Web site of the IBM partner Bull:
<http://www.bull.de/pub/>

The following tools can be recommended for daily tasks and can be downloaded from the above mentioned sites:

- ▶ An easy to use system monitoring tool that is called **monitor**.
- ▶ A program, which lists the open files in a file system, that is called **lsof**.
- ▶ For interconnection to the world of OS/2 and Windows, for example, for data interchange, the software **samba** can be used.

10.2.7 Remote commands

In an environment with a central administration server, it is very practical to set up permissions for the remote execution of commands. If all servers are connected over a dedicated control network that cannot be directly accessed from outside the computing center, this is an easy task. Furthermore, this separated control network offers the necessary security to prevent unauthorized access.

For the permission to execute remote commands on a server, the `/.rhosts` file is necessary. The `/.rhosts` file has to contain all IP labels of all adapters, which are connected to the control network. The file itself is then distributed to all servers and then the user `root` is able to execute remote commands like **rsh**, **rlogin**, **rexec**, or **rdist** from the administration server in order to fulfill administration tasks more effectively without entering passwords all the time.

The `.rhosts` file itself should have the permission 600 or 644 and must not have the permission 664, 666, or 777, because these permissions do not work.

Often, it is necessary to perform checks through commands on several nodes at the same time. Therefore, it is useful to create a command script, which executes a command on more than one server at a time. Several parallel commands have been introduced into the IBM @server pSeries SP environment. The command **dsh** is such a powerful command, which can be used in a distributed environment. A working collection of servers has to be defined for this command, on which the command is executed. The working collection is defined through the environment variable **WCOLL**.

For example, the command **export WCOLL="SIDdbct1 SIDapct1"** sets the working collection to the servers **SIDdb** and **SIDap**. It makes sense to define this variable in your default environment. The server list can also be put in a file called **/wcoll_all**. Then the **WCOLL** variable is set with the command:

```
export WCOLL=$(/usr/bin/cat /wcoll_all)
```

A very simple version of a **dsh** script without any error handling is the script in Example 10-1.

Example 10-1 Simple example for a dsh script

```
#!/usr/bin/ksh
for host in $WCOLL
do
    /usr/bin/echo "Output from:" $host
    /usr/bin/rsh $host $1
    /usr/bin/echo
done
```

With the script or **dsh** command it is now possible to execute one command on two servers, for example, **dsh "errpt"** lists the output of the AIX error report from both servers.

If one server is down for maintenance, the execution of **dsh** hangs until the **rsh** time out is exceeded. To avoid hangs during execution, an enhanced script version of the **rsh** command can be created. A script called **rsh_rc**, for example, first checks, via a single ping, if the host is reachable and then calls the **rsh** command.

It is recommended that you put all self created scripts (like **dsh** and **rsh_rc**) and programs in a single directory tree, for example, **/usr/scripts**. This tree can then easily be replicated to the other servers in an environment.

10.2.8 Configuration files

Generally, in AIX, many configurations are defined in flat files. Some of these files are mentioned in detail in the following sections. A useful collection of files, which should be synchronized over several servers in an environment, is listed too.

IP names and resolution

It is important that all local network interface addresses, IP labels, and networks must be resolvable locally. Commands such as **netstat**, **ping**, **df**, **exportfs**, or NFS mounts may hang if IP names are used that cannot be resolved due to a Domain Name Server outage or misconfiguration.

The relevant entries have to be made in the files `/etc/hosts` and `/etc/networks`.

The order of the name resolution mechanism is determined by the environment variable `NSORDER` (refer to “`/etc/environment`” on page 290) or the entries of the files `/etc/netsvc.conf` and `/etc/irs.conf`.

To set the order to local resolution and then to name server resolution, the `/etc/irs.conf` file has to contain the following line:

```
hosts local4,local6,bind4,bind6
```

To set the order to local resolution and then to name server resolution, the `/etc/netsvc.conf` file has to contain the following line:

```
hosts=local4,local6,bind4,bind6
```

The values of `/etc/irs.conf` are overwritten by the values of `/etc/netsvc.conf`.

Environment variables setting

This section presents useful settings for the environment, which are placed in the following files:

- ▶ `/etc/environment`
- ▶ `/etc/kshrc`
- ▶ `/etc/security/login.cfg`
- ▶ `/etc/motd`

/etc/environment

Many environment variables concerning the whole operating system are defined in the file `/etc/environment`.

- ▶ To set the order for IP name resolution to local and then to name server resolution, set the variable `NSORDER`. This variable overrides the order set in `/etc/netsvc.conf`. For example:

```
NSORDER=local4,local6,bind4,bind6
```

- ▶ Change the language variable of the system to `en_US`. For the different users in the system, this variable may be overwritten in their personal environment. For example:

```
LANG=en_US
```

- ▶ The following is an example for the setting of Daylight Saving Time for Central Europe. In the time zone of Middle Europe (MEZ), the Daylight Saving Time has a 1 hour offset. It starts in March (M3) and ends in October (M10) on the last Sunday (.5) of the month at 2 o'clock (02:00:00). For example:

```
TZ=MEZ-1MES,M3.5.0/02:00:00,M10.5.0/03:00:00
```

- ▶ Change the command line prompt to reflect the host name and the actual path. For example:

```
PS1="[ $(hostname) ] $PWD #"
```

This will produce the following output on host `zombie` in the directory `/sapmnt/ZMB`:

```
[zombie] /sapmnt/ZMB #
```

/etc/kshrc

The session specific environment is set in `/etc/kshrc`.

- ▶ A useful function for a recursive grep in the files of the current subdirectory tree is called `rgrep`:

```
function rgrep
{ /usr/bin/grep "$*" $( /usr/bin/find . -type f ) }
typeset -xf rgrep
```

- ▶ The history of commands entered on the command line is very useful and may save a lot of typing. To set the command line editing to your preferred editing tool the command shell built-in command `set` is used:

```
set -o vi
```

or

```
set -o emacs
```

- ▶ If two different users work with the same user ID, which should be avoided in normal cases, or if a user works with the same user ID in two different

```
/usr/bin/rm -f $HOME/.sh_history
```

Change the login prompt at telnet time to reflect the name of the system in the file `/etc/security/login.cfg`. The lines in Example 10-2 have to be added to the default and the `/dev/console` stanzas.

```
default:  
herald = "\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\nAIX  
Version 4.3.3\n\rWelcome to the IBM eServer pSeries 680\n\rlogin [DB/CI HA]:"
```

Customize the `/etc/motd` (message of the day) for a welcome message to point out the actual server and the assignment of the server, such as *SAP R/3 Test system* or *TSM Server*.

The table does not claim to be complete. It is our recommendation that you synchronize these files across the servers in your environment in order to decrease administration overhead. If the file synchronization is introduced after the servers are already in productive use, you must take care not to overwrite important information in the destination files.

Files or directories	Content
/etc/environment	Definition of the system environment
/etc/profile and /.profile /.dtpprofile	Definition of the (user) specific environment

Files or directories	Content
/etc/kshrc and /.kshrc /etc/csh.login and /.login /etc/csh.cshrc and /.cshrc	Definition of the (user) specific session environment
/etc/ntp.conf	NTP configuration
/etc/hosts /etc/networks	All interfaces, IP labels, and networks of all servers in the environment
/etc/resolv.conf /etc/netsvc.conf	Configuration for IP name resolution and order of resolution
/etc/services	All used service ports for all applications in the environment
/etc/passwd /etc/security/passwd /etc/group /etc/security/group /etc/security/envIRON /etc/security/limits /etc/security/users	User definitions with passwords and limits
/.rhosts /etc/hosts.equiv /etc/hosts.lpd	Authorizations for remote commands, remote logon, and remote printing
/wcoll_all	Working collection
/etc/qconfig /var/spool/lpd/pio/@local/ddi /var/spool/lpd/pio/@local/dev /var/spool/lpd/pio/@local/custom	Printer configuration ^a
/usr/local/bin /usr/local/lib /usr/local/man	Directories for tools and programs like monitor, lsof, or samba
/usr/tivoli/tsm/client/ba/bin/dsm.sys	TSM client definitions
/usr/tivoli/tsm/client/ba/bin/inclexcl.list	Include/Exclude lists for the TSM client
/usr/scripts /usr/scripts/define_alias_common	All self written scripts and command alias definition
/usr/scripts/db /usr/scripts/db/define_alias_db	Self written scripts and command alias definition for database handling
/usr/scripts/sap /usr/scripts/sap/define_alias_sap	Self written scripts and command alias definition for SAP handling

Files or directories	Content
/usr/scripts/cluster /usr/scripts/cluster/define_alias_cluster	Self written scripts and command alias definition for cluster handling
/usr/scripts/doc /usr/scripts/doc/define_alias_doc	Self written scripts and command alias definition for documentation handling
/usr/scripts/lib	Self written script libraries

a. After synchronizing the printer definition files, the ODM has to be updated.

ODM update after printer synchronization

After a successful distribution of the configuration files belonging to the printing environment (refer to Table 10-1 on page 291), the ODM has to be updated and a new configuration binary for the **qdaemon** has to be created.

The ODM update is done with the executable **piodigest**. For the creation of a new configuration binary **/etc/qconfig.bin**, the **qdaemon** has to be restarted. Stopping and starting the **qdaemon** during normal daytime operation may cause hanging print jobs or hanging print queues. Therefore, the execution of the commands should be done at a less critical time, for example, every night at one o'clock or only on demand, after a change or an addition of a printer.

The command sequence shown in Example 10-3 schedules the execution of the **piodigest** command and the restart of the **qdaemon** at 1:00 AM.

Example 10-3 Refresh printer definitions

```
/usr/bin/echo '/usr/bin/stopsrc -cs qdaemon ;  
cd /var/spool/lpd/pio/@local/custom ;  
/usr/bin/ls | /usr/bin/xargs -i /usr/lib/lpd/pio/etc/piodigest {} ;  
/usr/bin/startsrc -s qdaemon' \  
| /usr/bin/at 1:00" ;
```

Synchronization of configuration files

In order to synchronize the environment between the servers, the configuration files have to be copied from the master server (Admin) to all other servers. This can be done manually after a change, but can be managed faster and more reliably with a script. For remote copies of some files, the **rcp** command can be used. If a list of files or directory trees has to be kept in sync, the **rdist** command offers easier handling. Example 10-4 on page 294 contains the content of the distribution file **DISTHOST**. The synchronization is started with the command:

```
/usr/bin/rdist -f DISTHOST
```

Example 10-4 Distribution file DISTHOST for the rdist command

```
# ----- Host list
ALL = ( SIDdbctl SIDapctl hiAPctl lowAPctl hiTSMctl lowTSMctl )

# ----- File list
FILES      = ( /etc/hosts /etc/services /usr/scripts )
FILES_EXCL = ( /usr/scripts/cluster )

# ----- Distribution
${FILES} -> ${ALL}
    install -wRy;
    except ${FILES_EXCL} ;
```

The execution of the command copies the files `/etc/hosts` and `/etc/services` on all servers listed in the variable `ALL` and also distributes the entire directory tree under `/usr/scripts` to all servers. The subdirectory `/usr/scripts/cluster` is excluded from the distribution. Extraneous files in the destination directory are deleted and copies of files that are more recent on the destination than on the source are not overwritten.

10.2.9 Printing time out

The default time interval that a print job can last in AIX may be exceeded when printing large outputs on slower printers. The default time out can be changed to another value, such as 60 minutes, by editing the file `/usr/lib/lpd/pio/etc/piorlfb`. The time out flag of the printer backend can be changed like this:

```
typeset piorlfb_rbflags="-T 60" # rembak flags
```

10.2.10 Address Resolution Protocol cache (ARP cache)

In “Starting SAP R/3 on DB/CI HA” on page 241, the whole ARP cache is deleted in case of an automatic or manual takeover scenario to remove obsolete ARP entries from the cache. This can be done by executing the following command:

```
/usr/sbin/arp -a | /usr/bin/grep -v "?" | /usr/bin/awk ' { print $1 } '\
| /usr/bin/xargs -I {} /usr/sbin/arp -d {}
```

10.3 SAP R/3 implementation

The installation of SAP R/3 has some requirements, which are explained in the following sections.

10.3.1 Multiple SAP R/3 systems on one server

Today, IBM offers servers that are powerful enough to serve multiple SAP R/3 instances on one physical server. There are no restrictions from SAP to put two or more non-productive SAP R/3 systems, such as two development systems or a development and a quality assurance system, on the same server. In the past, there were some restrictions for running two or more productive SAP R/3 systems on the same server. These restrictions do not apply anymore.

This is an excerpt of the official statement from SAP about Server Consolidation:

“SAP supports several SAP Systems and DB Instances on the same server as long as there are no restrictions on the database side. Furthermore, this kind of configuration may diverge from SAP standards. Due to this, SAP has written dedicated SAP Notes (see SAP Note 21960. Apart from describing technical requirements, this SAP Note also contains a list of further relevant related notes) where possible restrictions and eventual administration complexity related to server consolidation are documented. In our SAP Notes we may describe "recommended" or "not recommended" configurations, depending on the customer landscape. Nevertheless, SAP leaves the latest decision to customers: SAP will offer support in any case.”

For running more than one SAP R/3 system on the same server, there are some constraints. These are mentioned in SAP Note 21960.

Note: All SAP Notes can be found at <http://service.sap.com/notes>.

In case of an Oracle database server, the standard listener configuration does not work and has to be enhanced. This is described in SAP Note 153835.

If the server does not have enough resources to provide enough memory and CPU power for all installed SAP R/3 systems, the installation of more than one system on a single server is not recommended.

If you plan to install two or more SAP R/3 systems on the same server, you should consider the use of AIX Workload Manager (WLM). The AIX WLM provides the possibility to distribute the hardware resources of a server equally to several applications, and it allows you to prioritize one application over the other too.

Refer to the Redbooks *Workload Management: SP and Other RS/6000 Servers*, SG24-5522 and *AIX 5L Workload Manager (WLM)*, SG24-5977 for more information concerning this topic.

10.3.2 Paging space

The following rules apply for the paging space in AIX:

- ▶ Avoid paging if possible, because this brings the performance of the system down. Paging in an SAP R/3 system can be usually avoided by more physical memory, an adopted memory management in SAP R/3, or less configured services (processes).
- ▶ Paging space should be mirrored to avoid a system crash caused by a failure of one of the paging disks.
- ▶ Paging space should be separated on an extraneous disk with no other workload on it.
- ▶ If the paging space consists of more than one volume:
 - Place only one paging space on one physical disk.
 - The paging space volumes should be equal in size.

The paging space volumes are written in a round-robin fashion. Different sized volumes cause a higher workload on the larger paging volume or on a disk with multiple paging spaces.

- ▶ The general rule for paging space sizing is:
 - SAP R/3 kernel for 32-bit: three times the number physical memory plus an additional 500 MB. For instance, 2 GB physical memory requires 6.5 GB paging space.
 - SAP R/3 kernel for 64-bit: A minimum of 20 GB is necessary.

10.3.3 User limits

For all users associated with SAP R/3, such as <sid>adm, db2<sid>, ora<sid>, the limits should be set to the following for a 32-bit SAP R/3 kernel:

- ▶ Soft CPU = -1
- ▶ Soft CORE file size = 2097151 (this is the default)
- ▶ Soft FILE size = 4194302
- ▶ Soft DATA segment size = -1
- ▶ Soft STACK size = -1

If you are using the 64-bit SAP R/3 kernel, then all limits should be set to -1.

For example, the limit for fsize can be changed with the command:

```
/usr/bin/chuser fsize=<newsize> <username>
```

The parameter <newsiz> is the new limit in 512 byte blocks. For this parameter to take effect, the user must logout and login again. If the limits of the root user are changed, the values does not take fully effect for already running processes. If this is necessary, the server has to be rebooted

These limits can also be changed by using the system administration tool SMIT.

Refer to SAP Note 323816 for more information concerning this topic.

10.3.4 Easy access to installation media of SAP R/3

When an installation or an upgrade of an SAP R/3 system is performed, the installation program (**R3setup**) allows and recommends the copy of the SAP R/3 CD-ROMs to a disk area. Due to internal installation and program logic, the copied files must all be in uppercase letters. The conversion from lower case to capital letters of files in a directory tree can be done via the script shown in Example 10-5.

Example 10-5 Script for converting file names to capital letters

```
#!/bin/sh
# Shell script to uppercase directory letters.
# Give full directory pathname as input.
STARTDIR=$1
if [[ -d $STARTDIR ]]; then
    set -x
    cd $STARTDIR
    /usr/bin/chmod -R 777 *
    for DIRNAME in `usr/bin/find $STARTDIR -type d -depth -print`
    do
        cd $DIRNAME
        for FILEORDIRNAME in `usr/bin/ls`
        do
            /usr/bin/mv $FILEORDIRNAME `usr/bin/echo $FILEORDIRNAME | \
            /usr/bin/tr "[a-z]" "[A-Z]"`
        done
    done
    set +x
else
    /usr/bin/echo "Directory does not exist!"
fi
```

Since AIX 4.3.3, it is possible to mount a CD-ROM in uppercase mode. So after the copy to disk, the destination files are uppercase too. The command is:

```
/usr/sbin/mount -v cdfs -p -r -o upcase /dev/cd0 /cdrom
```

10.3.5 Oracle recommendations

Some recommendations concerning Oracle are discussed in the following sections.

Directories

The following directories must exist on a server with the write and execution permission set to user ora<sid>. In case of a system copy or a system restore these directories may be missing and cause an abort of the Oracle database management system with the error message ORA-600:

- ▶ /oracle/<SID>/saptrace/background
- ▶ /oracle/<SID>/saptrace/usertrace
- ▶ /oracle/<SID>/saparch

TCP nodelay (Nagle algorithm)

To optimize the data transfer between the different instances of an SAP R/3 system, the parameter tcp.nodelay=true should be set in the file protocol.ora, which must be placed in the directory /oracle/<SID>/8xx_xx/network/admin. The file must have the name .protocol.ora (with a leading dot) in some releases (caused by a minor bug). This issue can be resolved by a symbolic link from one file to the other in order to avoid problems. Refer also to SAP Note 198752 for more information.

Asynchronous I/O

The usage of asynchronous I/O should be switched on for performance reasons. As the default and the name can change to an Oracle upgrade, this parameter should be checked on the database to avoid performance problems. This check can be done as in Example 10-6.

Example 10-6 Check of asynchronous I/O

```
/usr/bin/su - ora<sid>
svrmgrl
> connect internal
> select * from v$parameter where name like '%_io';
```

The values of use_async_io or disk_async_io have to be set to *true* in the file \$ORACLE_HOME/dbs/init<SID>.ora.

Control file

For former releases of Oracle databases, the control file of the database was a very critical file. A lost control file resulted in an unrecoverable and useless database. Therefore, the recommendation was to mirror the control file three times on different disk areas.

Since a few years ago, a new control file may be created anytime. So the importance of the control file is not as high as it was. For performance reasons, the number of control files should be two (maximum). If you are familiar with the new creation of the control file, you can run your database with only one control file and perform a weekly job to back up the control file to a trace file. With the content of the trace, a new control file can be created.

For more information concerning the creation of new control files, refer to Chapter 12, “SAP R/3 system copy” on page 365.

10.3.6 Number range buffering

For performance reasons, a large number of number range objects are buffered in SAP R/3. When buffering a number range object, numbers are not updated individually in the database; rather, the first time a number is requested, a preset group of numbers are reserved in the database (depending on the number range object) numbers and these are made available to the application server in question. The following numbers can then be taken directly from the application server buffer. If the application server buffer is exhausted, new numbers are generated in the database.

Pitfall ahead!

The following effects may be seen due to number buffering:

- ▶ If an application server is shut down, the numbers that are left in the buffer (that is, that are not yet assigned) are lost. As a result, there are gaps in the number assignment.
- ▶ The status of the number range interval reflects the next free number that was not yet transferred to an application server for intermediate buffering. Therefore, the current number level does not display the number of the *next* object.
- ▶ If several application servers are used, the chronological insert sequence is not determined by the numerical sequence on the individual hosts, due to the separate buffering on the different servers.

Buffering the number range objects has a positive effect on the performance, as no database access (on the number range table NRIV) is required for every posting. Furthermore, a serialization of this table (database blocking) is prevented to a large extent so that posting procedures can be carried out in parallel.

If you still require a continuous allocation because of any business needs, you can deactivate the number range buffering deliberately for individual objects. Number range objects that must be continuous due to legal specifications or a corresponding application logic, must not be buffered with the buffering type *Main memory buffering*.

Not buffering the number ranges may lead to blocked work processes on an instance. (They are waiting for the DB table NRIV to be released.) In this case, report SAPLSNR3 appears in the process overview and the performance gets very bad.

For more information, refer to SAP Notes 37844 and 62077.

10.3.7 Clean up jobs

For a reliable system, it is necessary to clean up temporary data and obsolete information, such as background job protocols. The clean up batch jobs are called in SAP R/3 reorganization jobs and are discussed in Section 13.4.1, “Housekeeping batch jobs” on page 422, because they have to be monitored regularly. These jobs must be set up in every SAP R/3 system to prevent table overflows or longer processing times due to an innumerable amount of table entries.

The default jobs can be scheduled in SAP R/3 Release 4.6C with the System Administration Assistant (SAA) using transaction SSAA. The variants of these default jobs do not clean up logs of event driven jobs. Thus, these variants have to be adapted or new variants have to be created to match these jobs too.

10.3.8 Update V3

In SAP R/3, there are function modules of the type *collective run (V3)*, in addition to V1 and V2 update function modules. These are not updated automatically in contrary to the function modules mentioned first, but only when a special report triggers the update. Then all calls of the function module are collected, summarized, and updated at once. They are treated like V2 update modules.

To trigger the execution of these V3 update requests, the RSM13005 report has to be scheduled regularly in the SAP R/3 system. The report is client dependent, so a job with this report as a job step has to be created in every productive client.

Refer to SAP Note 140357 for more information.

10.4 Operation of the SAP R/3 infrastructure

The operation of an SAP R/3 infrastructure requires a skilled team of administrators and operators. The following sections discuss some hints concerning the operation of an environment. These hints are meant to ease the daily operation.

10.4.1 Server journal

For all servers in the environment, it is a recommended task that you keep an account of changes made to the environment. This concerns the installation of new or exchanged hardware, the installation of fixes, or the change of a script.

Changes to a script can be documented in the header of the script itself with the appropriate user ID and the date of change. Other changes to the environment can be recorded in a *server journal*. It is possible to have a dedicated book for these tasks, but it is even easier to have a file, which is kept on a centrally managed server.

An administrator may not remember changes in detail one month after the changes. To avoid double work, to simplify troubleshooting in case of an error, and to not forget changes in other systems, this journal is preliminary.

It is also much easier to follow a well determined change strategy for the installation of fixes. A new fix can be installed, for example, in a test system and can be tested there for a period of three days. After the successful completion of the test, the fix can be applied to the production and development system. Without a server journal, the tracking of these tasks is more complicated and maybe impossible.

10.4.2 Starting point for profiles

Keep in mind that the SAP R/3 profiles stored in /sapmnt/SID/profile after the installation are basic profiles for the first start of SAP R/3. Normally, they do not offer performance optimized tuning or further consideration for an optimized memory usage. They are delivered in a very basic version and serve as a starting point.

The adoption to the individual hardware environment has to be done from the administrator of the SAP R/3 system. After the installation, the GoingLive check and EarlyWatch service provided by SAP can give further hints for a more optimized configuration.

In order to enable these checks, the basic customization has to be performed. The report RTCCTOOL has to be executed, which guides you to further actions and links to further SAP Notes.

For SAP R/3 Release 4.6C, the following SAP Notes apply: 197886, 91488, 116095, and 187939.

After a successful installation of all necessary tools, the Early Watch Alert has to be configured within the transaction SDCC. This step enables some periodic jobs that collect system statistics as input for the EarlyWatch sessions.

For an optimized utilization of the hardware resources, a deep understanding of the memory management and the processes is necessary. This information is discussed in detail in Chapter 11, "Performance" on page 305.

10.4.3 Time differences and daylight saving time

SAP recommends that you stop the SAP R/3 system during the change to daylight saving time and back.

During the time change, the local time changes at once, whereas the UTC (universal time coordinator or GMT) does not change at once, but continuously. The local time is stored internally in SAP R/3. Therefore, problems may arise if the SAP R/3 is up and running during the change.

For example, the run-time error "ZDATE_LARGE_TIME_DIFF" occurs at the end of the daylight saving time period (within a period of an hour, for example, from 2.00-3.00). This possibly only appears in the system log or as a message in the status line. A too big time difference between the servers may lead to the message "TIME_DIFF_TOO_LARGE" and result in an unusable system. You can execute the report RSDBTIME to check the times within an SAP R/3 system.

The time cannot be changed while the SAP R/3 System is running. The system must be restarted.

For more information, refer to SAP Notes 7417, 101726, 102088, and 398374.

10.4.4 Background processing

In an SAP R/3 standard installation, the system default printer LP01 is defined. This printer must not be deleted. If a user schedules a batch job, the user specific default printer is the spool output printer for the batch job. If the user has no specific default printer customized, the printer LP01 is the hard coded system default spool output printer for batch jobs.

Do not schedule a batch job with a spool output printer, which is defined with the *Host spool access method* “F”. The method “F” means printing on a front-end computer (SAP GUI) and can cause problems with the batch job when the front-end computer is switched off at batch run time, for example.

For background processing, additional tools may be necessary. These tools monitor jobs, check customized conditions, and are able to restart jobs. They can schedule and control the workload of background processing.

If HACMP is used in an automatic takeover environment, a tool may also become mandatory. Monitoring and restarting of background processing is a necessary task in case of an takeover, which cannot be managed with HACMP. If a large amount of background processing is done at night without operational assistance, the implementation of some additional tools should be considered. In case of a takeover, these tools can be customized to reschedule aborted tasks in order to continue background processing at night.

10.4.5 Avoid client copies

In former releases of the SAP R/3 system or database, copies could not be performed easily due to some restrictions concerning the renaming of databases on some database management platform. Therefore, client copies have to be used often to create new clients within the same or other SAP R/3 systems.

When SAP R/3 is first introduced in your environment, it is very useful to use the client copy functionality for creating new (empty) customizing clients in your systems. When the clients later are filled with data, the clients will probably become too large for a client copy. SAP recommends, for larger clients, the use of system copy or database copy as a standard client copy may take several days.

The following excerpt is from SAP Note 67205:

“Copying a productive client is only practical if you want to build a new test system from a productive client. But also in this case, a database copy is preferable to a client copy.”

Refer to Chapter 12, “SAP R/3 system copy” on page 365 for more information concerning this topic.

Performance

In a nutshell:

- ▶ Get an understanding of the AIX shared memory management.
- ▶ Configure SAP R/3 to use the alternative AIX ES Extended Memory.
- ▶ Use optimized settings for the AIX virtual memory management.
- ▶ Enable and configure asynchronous disk I/O in AIX.
- ▶ Learn to use basic performance monitoring tools provided by AIX.

In this chapter, we describe basic concepts of AIX and of both the 32-bit and the 64-bit implementation of SAP R/3 Release 4.6. Understanding these concepts helps set up a high performing SAP R/3 system in terms of memory management and I/O throughput. This chapter is not intended to be a troubleshooting guide if your systems show an unexpected low performance.

The understanding of shared memory usage within SAP R/3 is one of the most important concepts in tuning the performance of an SAP R/3 system. In the first section, we therefore discuss, in detail, the shared memory management facilities of AIX. After explaining the concept of SAP R/3 instance buffers, we discuss the different SAP R/3 Extended Memory models for 32-bit and 64-bit implementations that are available for AIX.

In the second part of this chapter, we discuss the important concepts of the AIX operating system that are relevant for performance. We explain the different modes of a CPU, the concept of virtual memory management, and the concept of disk I/O operations. We then give recommendations for parameter values that allow smooth operation of an SAP R/3 system.

In the last part of this chapter, we describe some AIX performance monitoring tools.

This chapter covers the highlighted area in Figure 11-1, which is derived from the SAP R/3 infrastructure model that is used throughout the book.

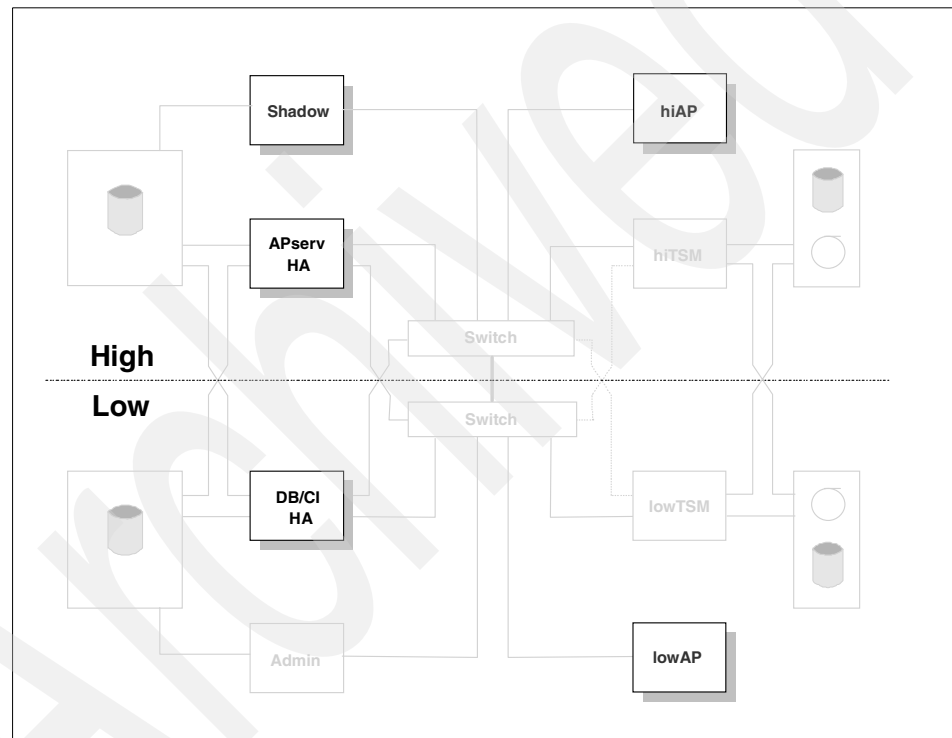


Figure 11-1 Performance

11.1 AIX shared memory management

In AIX, the *virtual memory (VM)* consists of real physical memory and the available paging space. The virtual address space is the range of addresses that a process and the kernel is allowed to reference.

11.1.1 Virtual memory

The virtual address space is partitioned into segments. A segment is a contiguous portion of the virtual-memory address space and has a size of 256 MB. Data objects can be mapped into these segments and can be shared between processes or maintained as private. A process has shared segments and private segments.

All segments are partitioned into fixed-size units called *pages*, which have a size of 4096 bytes. Because the virtual memory consists of real memory and paging space, a page can be located in real memory or on disk. The *AIX Virtual Memory Manager (VMM)* keeps track of the location of the individual pages in the virtual memory and resolves references of processes to pages. If a process accesses a page that is located on disk, the VMM reads the page from paging space or from the accessed file into real memory. The process then can access the page.

Since AIX 4.3, processes using 32-bit address space and processes using 64-bit address space can run simultaneously on 64-bit hardware. A 32-bit process can access 16 segments and a 64-bit process can access up to 16^{23} segments. A process using 32-bit addresses can address 4 GB, a process using 64-bit addresses can address 16 exabytes.

11.1.2 Segment layout of processes

Most of the first 16 segments are used in the same way when using 32-bit address space and using 64-bit address space. A process using 64-bit address space is not limited to the number of segments, but to the virtual memory available in a machine.

In this section we describe the segment layout of 32-bit processes. The differences of the layout for 64-bit processes are explained in the redbook *AIX Version 4.3 Differences Guide*, SG24-2014.

When a 32-bit process is initialized, it can access up to 16 segments. Some of the segments are used for accessing the AIX kernel and memory mapped I/O. In Figure 11-2 on page 308, the layout of the segments for a process is shown.

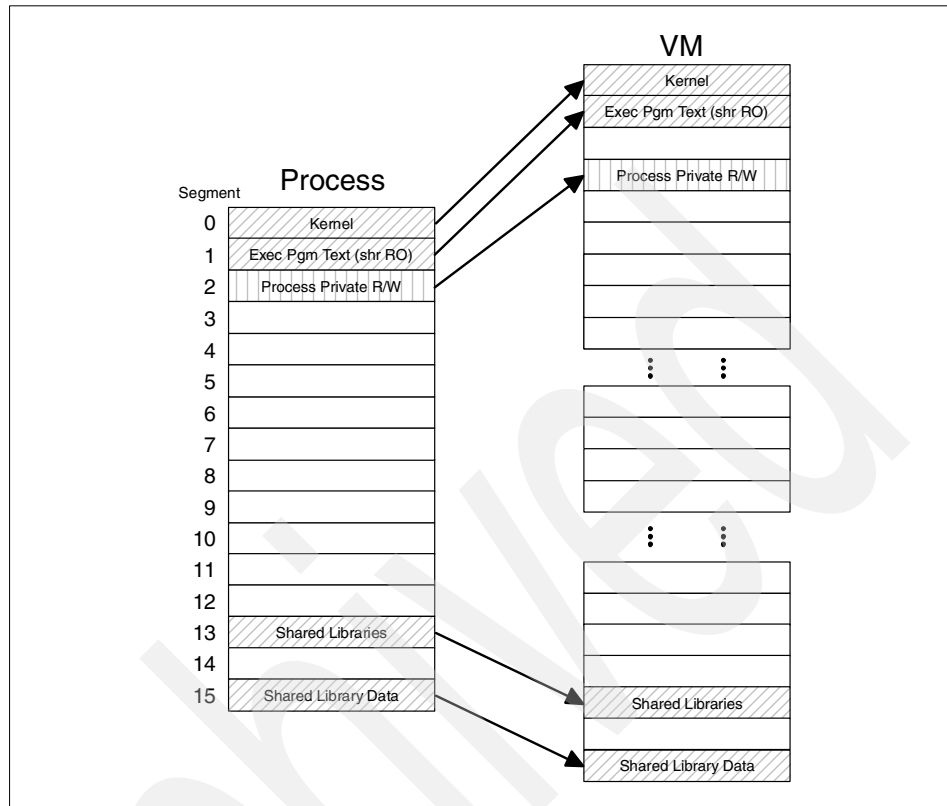


Figure 11-2 Segments of a process

Segments of the virtual memory can be mapped to the free segments of the process. The following segments can be used by the process when the process is running in user mode:

Segment 0

Segment number 0 is used for the first kernel segment in user mode. It is read-protected in user mode, except for the system call (SVC) tables, svc_instruction code, and system configuration structure.

Segment 1

This segment contains the executable program object code text. This segment is read only and can be shared for equal processes. In this way, the segment only has to be loaded once into real memory if multiple instances of the same program are started.

Segment 2

Segment number 2 is used as the process private segment. This segment contains most of the per-process

information, including user data, user stack, user heap, kernel stack, and user block.

Segment 3-12

The process can use these segments to attach shared memory segments and memory mapped files.

Segment 13

This segment is used to access shared libraries.

Segment 14

Since AIX Version 4.2.1, the process can use this segment to attach shared memory segments and memory mapped files. Before AIX Version 4.2.1, this segment was reserved.

Segment 15

This segment is used to access data of shared libraries.

Thus, during run time, a process can use the segments 3 to 12 and 14 with a total amount of 2.75 GB memory.

11.1.3 Sharing segments between processes

It is possible to share segments between processes. In Figure 11-3 on page 310, we show a configuration where two processes share two segments. These shared segments are shaded grey.

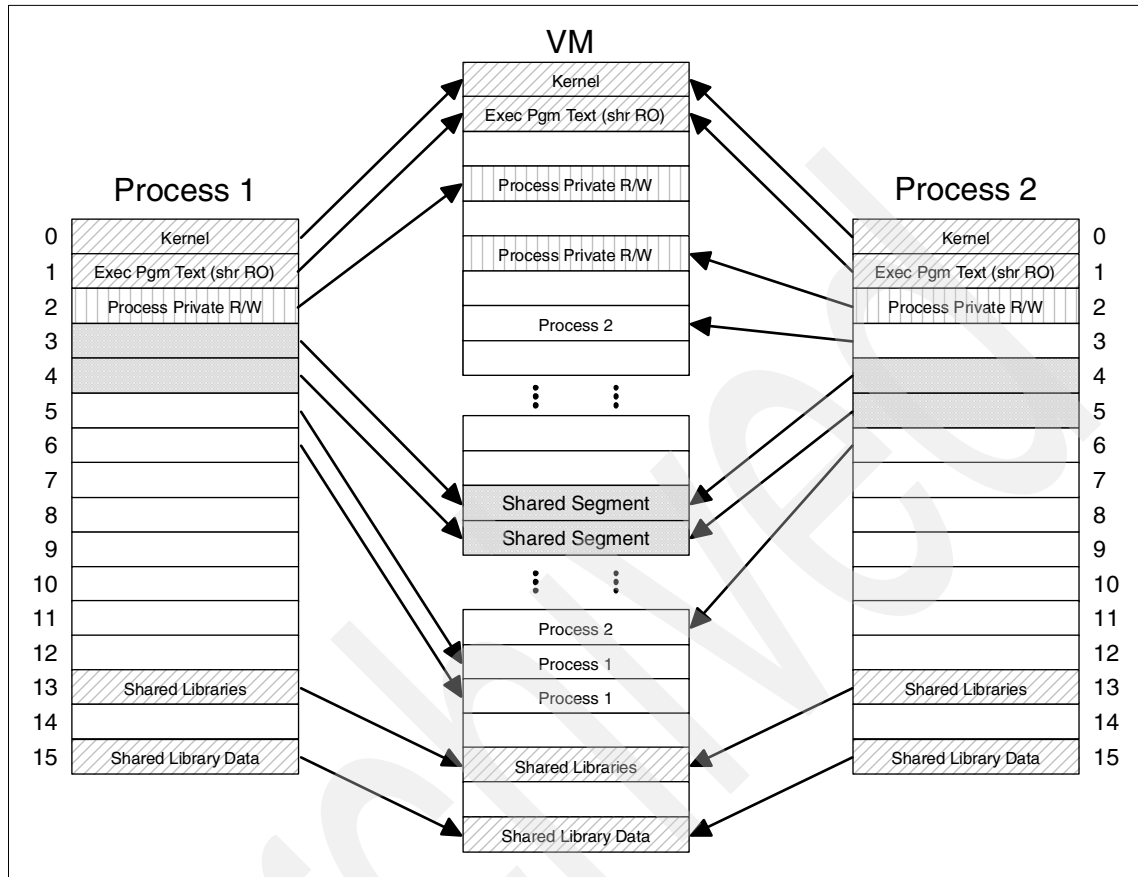


Figure 11-3 Two processes sharing segments

In this example, the segments 3 and 4 of process 1 point to two shared segments, which are also accessed by process 2 using segment 4 and 5. However, the synchronization of the access to the shared segments between the processes needs to be done at the application level.

11.1.4 Shared memory regions

When a process wants to access shared memory, it has to create a *shared memory region* using the `shmget()` routine. This shared memory region is then attached to the process's address space using the `shmat()` routine.

In AIX Version 4.2, the maximum size of a shared memory region is limited to 256 MB, which corresponds to one memory segment. Since AIX Version 4.3, shared memory regions can be up to 2 GB. In this case, more than one memory segment is used.

Normally, each shared memory region consumes 256 MB of address space, independent of the actual size it needs. Since AIX Version 4.2.1, you can use *Extended Shared Memory (ESM)* to define shared memory regions in a more granular way. The size of the memory regions can be set, corresponding to pages, in steps of 4096 bytes. Also, multiple shared memory regions can share one segment. There is the restriction that Extended Shared Memory cannot be used for shared memory regions bigger than 256 MB.

The Extended Shared Memory can be turned on by setting the environment variable EXTSHM=ON. Using Extended Shared Memory will have a very small negative performance impact.

Extended Shared Memory essentially removes the limitation of only 11 shared memory regions, and provides a better exploitation of the available memory segments for a process.

A summary of the characteristics of the shared memory management for AIX is given in Table 11-1.

Table 11-1 Summary of shared memory management for 32 bit processes

AIX version	4.2.1	4.2.1	4.3.1	4.3.1	Since 4.3.2
EXTSHM	OFF	ON	OFF	ON	ON
Segments	10	10	11	11	11
Maximum usable process memory	2.50 GB	2.50 GB	2.75 GB	2.75 GB	2.75 GB
Maximum number of memory regions	10	4096	11	32768	131072
Minimum size of memory regions	256 MB	4 KB	256 MB	4 KB	4 KB
Maximum size of memory regions	256 MB	256 MB	2 GB	2 GB	2 GB

11.2 SAP R/3 instance buffers

SAP R/3 instance buffers are memory areas of an application server instance. They contain data and program code that is frequently used by all work processes. The SAP R/3 instance buffers are important for a high interactive performance. The access to buffered data that is stored in the main memory of the application server is much faster than the retrieval of data from the database. Even if the requested data is in the database block buffer, so that no disk access is necessary on the database server, the data has to be transferred over the network for every request. Thus, SAP R/3 instance buffers also significantly reduce network load between application servers and databases.

Restarting an instance should be avoided, as all buffer contents are destroyed. They are initialized, and it will take a considerable time until they contain the same data again.

Because SAP R/3 buffers are not shared between different instances, modifications of some of the buffer areas are regularly synchronized. Normally, the buffer contents are rather static.

11.2.1 SAP R/3 buffer types

These are the important buffer types:

Repository buffer This buffer contains table and field definitions of the ABAP dictionary. The database access agents of a work process consult the repository buffer before accessing a specific database table to detect whether the contents of the table are buffered. The repository buffer is also known as *nametab* and consists of the four shared buffers *TTAB*, *FTAB*, *IREC* and *short nametab*.

Table buffers These buffers store the content of database tables. There are two types of table buffers: *partial* and *generic*. Partial table buffers store single table records or keys. Generic table buffers hold sequences of records or even all records of single tables.

Program buffer This buffer stores the compiled and, thus, executable versions of ABAP programs. As the whole business logic and application processing of the SAP R/3 modules is based on ABAP programs, this buffer is very important. There is a *least recently used* (LRU) algorithm implemented for reusing the program buffer.

GUI buffers There are two buffers that store GUI objects. The *presentation buffer*, also known as *Screen* or *Dynpro*

buffer, contains dynamically created SAP GUI screens. The *menu buffer* contains GUI elements, such as menu and button definitions.

There are some more buffers for specific services, such as dispatcher, message, spool, and enqueue, and for the administration of the different SAP R/3 memory areas, such as *roll*, *paging*, and Extended Memory.

11.2.2 Checking SAP R/3 buffer configuration

The size of the individual buffers can be tuned through the modification of instance profile parameters. The **sappfpar** tool can be used to check the buffer configuration for an instance. As user <sid>adm, you can issue the following command, specifying the full path to the instance profile:

```
sappfpar pf=/sapmnt/<SID>/profile/<Instance_Profile> check
```

The buffer related output of the command is shown in Example 11-1.

Example 11-1 Output of sappfpar - buffer related

```
tequila:slzadm 3> sappfpar pf=SLZ_DVEBMGS10_tequila check

=====
=
==  Checking profile:      SLZ_DVEBMGS10_tequila
=====
=
Shared memory disposition overview
=====

Shared memory pools
Key:   10  Pool
      Size configured.....: 140000000 ( 133.5 MB)
      Size min. estimated.:  136694014 ( 130.4 MB)
      Advised Size.....:    140000000 ( 133.5 MB)

Key:   40  Pool for database buffers
      Size configured.....:  324000000 ( 309.0 MB)
      Size min. estimated.:  321000288 ( 306.1 MB)
      Advised Size.....:    324000000 ( 309.0 MB)

Shared memories inside of pool 10
Key:    1  Size:      2000 (   0.0 MB) System administration
Key:    3  Size:   13714400 (  13.1 MB) Disp. communication areas
Key:    4  Size:    514648 (   0.5 MB) statistic area
Key:    7  Size:    14838 (   0.0 MB) Update task administration
Key:   11  Size:    500000 (   0.5 MB) Factory calender buffer
Key:   12  Size:   3000000 (   2.9 MB) TemSe Char-Code convert Buf.
Key:   13  Size:  10500000 (  10.0 MB) Alert Area
```

Key:	14	Size:	20000000 (19.1 MB)	Presentation buffer
Key:	16	Size:	22400 (0.0 MB)	Semaphore activity monitoring
Key:	17	Size:	376932 (0.4 MB)	Roll administration
Key:	18	Size:	57444 (0.1 MB)	Paging administration
Key:	19	Size:	50000000 (47.7 MB)	Table-buffer
Key:	31	Size:	4806000 (4.6 MB)	Dispatcher request queue
Key:	33	Size:	20480000 (19.5 MB)	Table buffer, part.buffering
Key:	34	Size:	4194304 (4.0 MB)	Enqueue table
Key:	51	Size:	3200000 (3.1 MB)	Extended memory admin.
Key:	52	Size:	40000 (0.0 MB)	Message Server buffer
Key:	54	Size:	4202496 (4.0 MB)	Export/Import buffer
Key:	55	Size:	8192 (0.0 MB)	Spool local printer+joblist
Key:	57	Size:	1048576 (1.0 MB)	????????????
Key:	58	Size:	4096 (0.0 MB)	????????????

Shared memories inside of pool 40

Key:	2	Size:	1000208 (1.0 MB)	Disp. administration tables
Key:	6	Size:	255590400 (243.7 MB)	ABAP program buffer
Key:	41	Size:	9010000 (8.6 MB)	DB statistics buffer
Key:	42	Size:	3620592 (3.5 MB)	DB TTAB buffer
Key:	43	Size:	32534392 (31.0 MB)	DB FTAB buffer
Key:	44	Size:	7958392 (7.6 MB)	DB IREC buffer
Key:	45	Size:	4886392 (4.7 MB)	DB short nametab buffer
Key:	46	Size:	20480 (0.0 MB)	DB sync table
Key:	47	Size:	3073024 (2.9 MB)	DB CUA buffer
Key:	48	Size:	300000 (0.3 MB)	Number range buffer
Key:	49	Size:	3000000 (2.9 MB)	Spool admin (SpoolWP+DiaWP)

Shared memories outside of pools

Key:	1002	Size:	400000 (0.4 MB)	Performance monitoring V01.0
Key:	58900110	Size:	4096 (0.0 MB)	SCSA area

Nr of operating system shared memory segments: 4

[...]

11.2.3 Arrangements of SAP R/3 buffers in pools

As described in Section 11.1, “AIX shared memory management” on page 306, when not using Extended Shared Memory, there are only eleven shared segments that can be attached to a 32-bit AIX process. Therefore, not every buffer can occupy a single shared segment. Rather, they are normally grouped in so called *pools*, which are mapped to shared segments.

A predefined shared memory key is assigned to every buffer. The keys 10 and 40 are reserved for pools. The instance profile determines the association between buffers and pools. The following statements declares that the table buffer (key 19) is part of pool 10 (refer to Example 11-1 on page 313):

```
ipc/shm_psize_19 = -10
```

The size of pool 10 and pool 40 has to be configured so that all buffers fit in. For example, a size of 140 MB for pool 10 is specified by the line:

```
ipc/shm_psize_10 = 140000000
```

After changing the size of an SAP R/3 buffer you also have to adjust the pool size. You can use the command given in Section 11.2.2, “Checking SAP R/3 buffer configuration” on page 313 to check the advised and actually configured size.

11.2.4 Implications of the AIX version for SAP R/3 buffers

Depending on the AIX version, there are some implications for the setup of SAP R/3 buffers.

AIX 4.2.1 and EXTSHM=ON

Since AIX Version 4.2.1, eleven instead of ten shared memory segments are available for the work processes of an SAP R/3 instance.

During instance startup, SAP R/3 occupies two very small shared memory regions for the SAP R/3 operating system performance collector process *saposcol* and the *Shared Common System Area (SCSA)*. As they cannot be configured to be part of a pool, there are normally two whole 256 MB segments used (see keys 1002 and 58900110 in Example 11-1 on page 313). The remaining space of these segments cannot be used by the instance, thus wasting almost 512 MB of address space.

AIX 4.2.1 introduced the possibility for small shared memory regions to share a single 256 MB segment. To activate this feature for SAP R/3, the environment variable **EXTSHM** has to be set to 'ON' for user <sid>adm. Now the shared memory regions for *saposcol* and *SCSA* can be shared with pool 10, which saves two segments for SAP R/3 Extended Memory. For details, see SAP Note 95260.

Unfortunately, the **sappfpar check** command only analyzes the instance profile on file but does not output the actual state of the shared memory distribution of the running instance. Therefore, the returned number of operating system shared memory segments may be wrong.

Attention: EXTSHM=ON must not be set for a DB2 or Oracle database run-time environment. Therefore, it must not be set within the environment of user db2<sid>, ora<sid>, and <sid>adm on the database host. If you want to use EXTSHM=ON for a central instance that runs on a database server, you have to set the variable in the instance start profile. See SAP Notes 174882 and 95260.

AIX 4.3.1 and large shared memory segments

Before AIX Version 4.3.1, the SAP R/3 pools were not allowed to be larger than a single shared memory segment, that is 256 MB. This often leads to problems with the program buffer because, as of SAP R/3 release 4.0x, even a whole 256 MB segment is sometimes too small for the program buffer of a production system. If the program buffer is filled up completely, ABAP program code in the buffer is replaced on a least recently used basis. This leads to a fragmentation in the buffer, which makes the space problem even worse. As of SAP R/3 kernel 3.11, there was a work around. Pool 40 was implemented to consist of two adjacent 256 MB segments which allowed larger program buffers inside pool 40.

Pitfall ahead!

AIX Version 4.3.1 abolished the 256 MB limit for shared memory regions. Thus, both pool 10 and pool 40 can exceed 256 MB if there are enough adjacent 256 MB segments available. But still only eleven segments can be used for an instance based on the 32-bit SAP R/3 kernel with the standard installation. See SAP Note 117267 for details.

As the environment variable EXTSHM is ignored for shared memory segments larger than 256 MB, it is generally recommended that you configure pool sizes as multiples of 256 MB to avoid the waste of shared memory address space.

A large program buffer can be configured outside of a pool as a dedicated shared memory segment. Example 11-2 shows the instance profile parameters to create a 500 MB program buffer (key 06).

Example 11-2 Configuration of a large program buffer

```
# configure program buffer (key 06) as a dedicated shared memory segment
ipc/shm_psize_06 = 0

# size of program buffer: 500 MB
abap/buffersize = 512000000
```

11.2.5 Buffer tuning

SAP R/3 transaction ST02 calls the buffer monitor, which comprises all buffer related information for an instance. There you find the following important attributes for several buffers:

Hit ratio	This value is the ratio between the number of application requests that were successfully served out of the buffer and the number of all requests.
Buffer quality	This value describes the buffer hit rate in terms of database requests. An application request can demand several single database requests. The buffer quality is the ratio between the number of avoided database requests and the number of all database requests. The buffer quality is more important than the hit rate because the benefit from the buffer is calculated with weighted requests.
Allocated size	This value specifies the amount of shared memory in KB that is allocated by the buffer. The allocated size is independent of the current usage of the buffer because the buffer allocates the whole amount of memory according to its configured size during the start of the instance.
Free space	This value provides the current free buffer space in KB and percentage of the whole buffer size. It does not consider fragmentation.
Number of directories	A directory entry is required for every buffered object that points to the memory address where the object is stored. This value specifies the overall number and number of available directory entries for the buffer.
Swapping	This value represents the number of objects that have been swapped out of the buffer in order to provide space for new objects since the start of the instance.

Before a buffer can be tuned, it has to be monitored for a sufficiently long period of time during normal operation. The buffer monitor does not provide meaningful values until after this period of time. A good value for the buffer quality depends on the buffer type.

Pitfall ahead!

As a rule of thumb, the buffer quality of all buffers should be better than 95 percent after times of excessive system usage. A well tuned buffer provides a high buffer quality, a value for swapping that should be almost zero, and a relatively small amount of free space in order not to waste shared memory space. For more details of buffer tuning, refer to the SAP Library.

11.3 Models for SAP R/3 Extended Memory

The memory management of SAP R/3 has evolved considerably since its first releases. SAP R/3 Release 3 introduced the concept of Extended Memory, which provides a common memory area for all work processes.

There are different ways in which *SAP R/3 Extended Memory* can be configured on AIX. SAP implemented a special model for AIX to allow 32-bit application server instances to address more than 2.7 GB of shared memory, which corresponds to eleven segments with a size of 256 MB each. In this section, we explain what the SAP R/3 Extended Memory is and describe the three different models it can use.

11.3.1 A definition of SAP R/3 Extended Memory

While the buffers described in Section 11.2, “SAP R/3 instance buffers” on page 312 store user independent data, we now look at how SAP R/3 handles user dependant data. The user sessions within SAP R/3 are not bound to a dedicated work process of an instance. Instead, a user session only attaches to a work process while processing a single dialog step. The next dialog step of the same user session is often connected to a different work process of the same instance. This mechanism allows the system to handle a large number of users with a relatively small number of work processes.

In general, a simple SAP R/3 transaction consists of several dialog steps. During the processing of a transaction the user works with a set of data, the so called *user context*, that has to be accessible by the different work processes. SAP R/3 implements user context switching using shared memory. User contexts can vary considerably in size. To be able to serve a lot of users with a single instance, it is necessary to configure as much shared memory address space as possible. The amount of addressable shared memory that is available for storing user contexts is called SAP R/3 Extended Memory.

Sequence of allocating memory for user contexts

Extended Memory is not the only space where user contexts are stored within a work process. Figure 11-4 on page 319 shows the different memory types. The following discussion refers to dialog work processes.

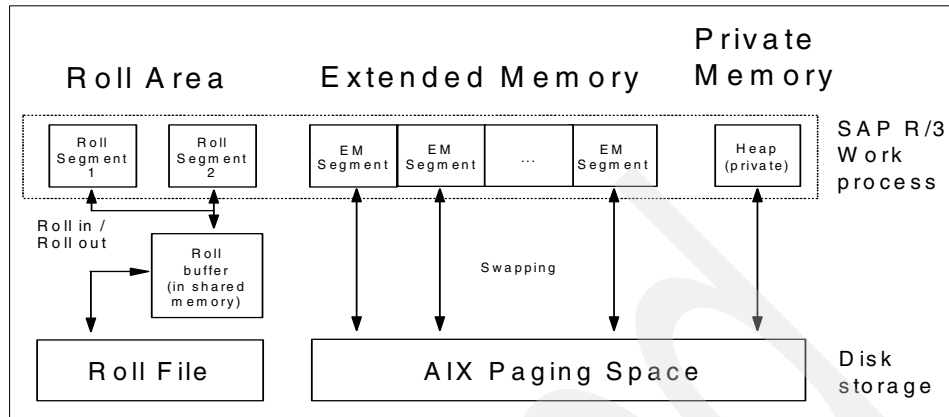


Figure 11-4 Memory types of SAP R/3 work processes

Due to the former memory model of SAP R/3 Release 2, there is still a *Roll Area*. During a context switch, the initial memory (*SAP R/3 roll area*) of a user context is taken from the Roll Area. This SAP R/3 roll area is copied from the SAP R/3 Roll buffer into the so called *Roll Segment 1* that is part of the process private segment of the work process. This mechanism is called *Roll in*. If there is not enough space in the SAP R/3 Roll buffer in shared memory for all roll areas, some are swapped to the instance's *Roll File* on disk. Roll area swaps slow down user context switches considerably and should be avoided by tuning the SAP R/3 Roll buffer.

Since SAP R/3 Release 3, the Roll Area is no longer used for data of a user context. It is only deployed for storing the pointers into the Extended Memory, where the new context data resides. If the Extended Memory is used up and the user context demands further space, it uses the old SAP R/3 Release 2 mechanism and allocates from the roll area in *Roll Segment 2*. If the size limit of the roll area is reached, the last possibility to enlarge the user context is to occupy *Private Memory*, which is located on the *Heap*. Because the heap is part of the process private segment, which is not sharable with other processes, the work process has to be switched to *private mode (PRIV)*. In private mode, the work process is reserved for this user context until the end of the transaction and is not available for processing dialog steps of other users.

After finishing the dialog step, the work process detaches the user context by copying the roll area back to the Roll buffer (*Roll out*). In the event that the processing of a dialog step ends with allocating private memory, the process is killed and restarted automatically by the SAP R/3 system.

The sequence of allocating memory for the user contexts is different for batch processes. There are no user context switches, because a background task is processed by a single batch process from the beginning until the end. Again, the roll area is used first. If the batch process expands further, private memory is allocated in the second place. In the last step, Extended Memory is occupied.

There are plenty of instance profile parameters to configure the amount of memory that can be allocated by a single user context for each memory type. There are also size limits for each memory type that can be configured on an instance level. Refer to “Memory Management” in the SAP Library for a detailed description of the parameters.

Tip: You can display the memory allocation sequence and the configured size limits within SAP R/3 by executing the ABAP report RSMEMORY with Transaction SE37.

Tuning SAP R/3 memory management

These are the main goals when tuning SAP R/3 memory management:

Bright idea!

- ▶ Avoid swapping of roll areas to the Roll file in any case. These swaps make the frequent user context switches very expensive in terms of performance. Either enlarge the size of the SAP R/3 Roll buffer to fit more roll areas or decrease the size for a single roll area.
- ▶ Avoid the use of private memory for user contexts, if possible. As dialog work processes are occupied by a single user when operating in private mode, the overall performance of the instance is dramatically reduced. A further drawback is the time it needs to restart a work process that was in private mode after finishing a transaction to free the allocated heap space for other processes.
- ▶ As Extended Memory is the most efficient memory type for storing user contexts, it is important to provide as much of it as possible for an instance.
- ▶ There should be as little swapping of Extended Memory to the AIX paging space as possible during the main times of interactive usage of the instance. The advantage of the direct access to a user context in a shared Extended Memory segment is lost, if it has to be paged in from disk.

The next sections describe three models for configuring Extended Memory on AIX, the 32-bit standard configuration, the 32-bit alternative configuration AIX ES Shared Memory, and the 64-bit alternative configuration AIX ES Shared Memory.

11.3.2 Standard Extended Memory configuration (32-bit)

The standard 32-bit Extended Memory configuration is depicted in Figure 11-5 on page 322. There are a maximum of eleven AIX shared memory segments available to a work process of an SAP R/3 instance. The figure shows the maximum amount of Extended Memory that can be configured with the standard 32-bit model.

All work processes attach to the same eleven segments that are accessed via a shared mapped file. Using EXTSHM=ON, segment 3 contains the SCSA area and the Roll buffer that is configured to build up a stand-alone, 255 MB buffer. Segments 4, 5, and 6 are occupied by the buffer pools 10 and 40.

The remaining seven segments 7 to 12 and 14 are configured for Extended Memory. Thus, the instance has a maximum of 1.75 GB shared memory for storing user contexts. This is also, roughly, the limit for the size of a single user context that is created by a large batch process for instance.

As all user contexts stored in Extended Memory can be accessed by all work processes at a time, they have to be protected from each other. If one work process starts the processing of a certain user context, it removes the protection. At that time, the context is locked for all other work processes. After finishing the dialog step, the work process protects the context again. The protection is implemented through a system call on page level. The overhead during context switches due to protection can be considerable for large Extended Memory configurations, because it increases with the number of managed pages.

Bright idea!

An application server that runs a fully configured instance should have at least 3 GB of main memory and, according to SAP recommendations, about 10 GB of paging space.

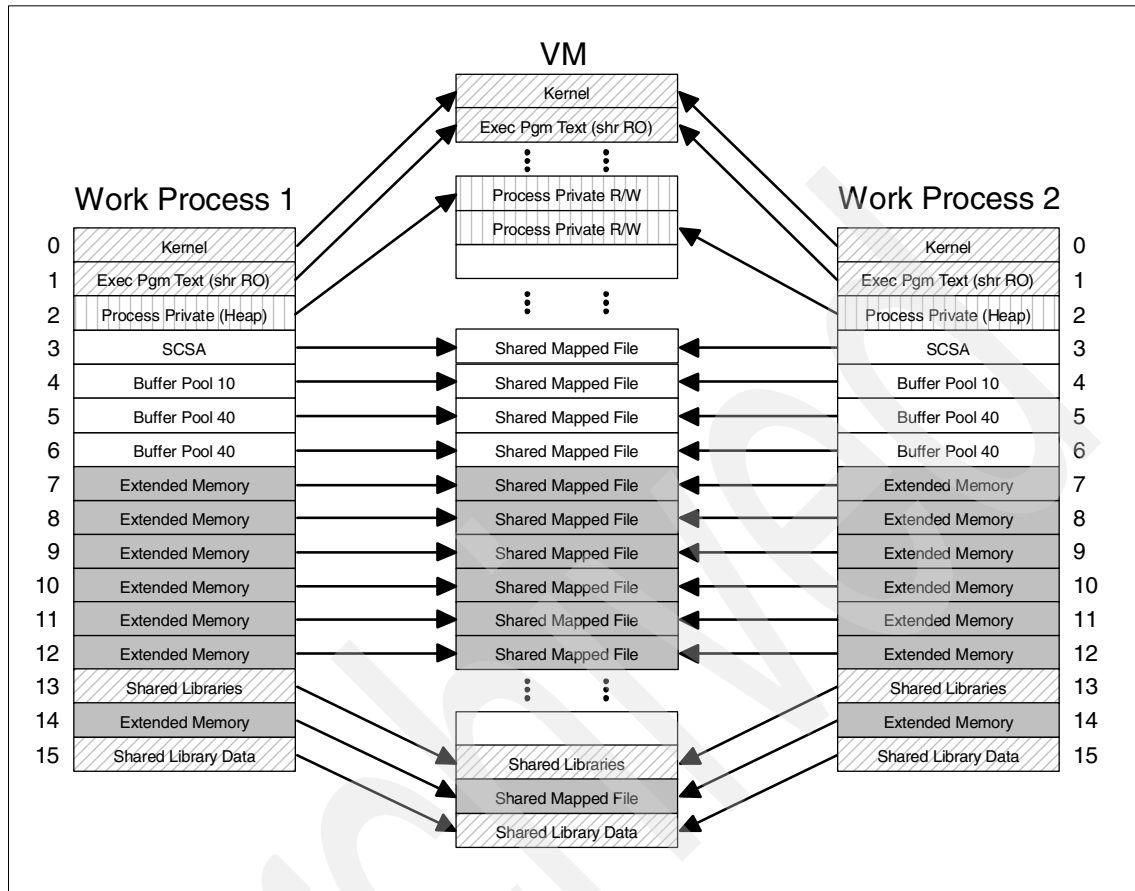


Figure 11-5 Standard Extended memory configuration (32-bit)

The configuration of Extended Memory can also be checked with the **sappfpar** command. The output is shown in Example 11-3. The tool shows the number of shared segments used for Extended Memory. There is also a rough estimation of the paging space requirements.

Within SAP R/3, the actual used amount and the overall available amount of Extended Memory can be checked with transaction ST02.

Example 11-3 Output of sappfpar - Extended Memory related

```
tequila:slzadm 3> sappfpar pf=SLZ_DVEBMGS10_tequila check
```

```
=====
=
==  Checking profile:      SLZ_DVEBMGS10_tequila
=====
```

[...]

```
Shared memory resource requirements estimated
=====
Nr of shared memory descriptors required for
Extended Memory Management (unnamed mapped file): 7

Total Nr of shared segments required.....:      11
System-imposed number of shared memories.:      11
Shared memory segment size required min...: 140000000 ( 133.5 MB)
System-imposed maximum segment size.....: 2147483648 (2048.0 MB)
R/3-imposed maximum segment size.....: 2147483647 (2048.0 MB)

Swap space requirements estimated
=====
Shared memory.....: 444.8 MB
..in pool 10  130.4 MB,   97% used
..in pool 40  306.1 MB,   99% used
..not in pool   0.4 MB
Processes.....: 247.9 MB
Extended Memory .....: 1792.0 MB
-----
Total, minimum requirement.....: 2484.7 MB
Process local heaps, worst case..: 762.9 MB
Total, worst case requirement....: 3247.6 MB

Errors detected.....: 0
Warnings detected.....: 0
tequila:slzadm 4>
```

11.3.3 AIX ES shared memory implementation (32-bit)

In the 32-bit alternative Extended Memory implementation for AIX, there is still a maximum size of approximately 1.5 GB for a *single* user context because of the architectural limits in addressable shared memory of a single work process imposed by the 32-bit address space. The main difference to the standard model is that the work processes do not share *all* of the user contexts at the same time. Figure 11-6 on page 324 shows the alternative model.

Again, all work processes share the segments 3 to 6 containing the SAP R/3 instance buffers. But during the processing of a user dialog step, the work process attaches dynamically and only to the segments that contain the related user context. In order to find the dedicated segments, the work processes have

to share only one segment simultaneously (segment 7 in Figure 11-6 on page 324). This segment contains the meta information for the user contexts, which provides the number of allocated 256 MB segments and their location in virtual memory.

Our example shows a single point in time where work process 1 works with user context 4. This context contains two 256 MB segments located somewhere in virtual memory that are attached to segments 8 and 9 in the address space of the work process. At the same time, work process 2 holds user context 1, which is spread over three segments attached to segment 8, 9, and 10 in the address space of the work process. There are further contexts (for example, context 3 and context 5) stored in Extended Memory, which are not attached to any work processes at all. They will be attached dynamically to work processes (as soon as a dialog step for one of these contexts is executed).

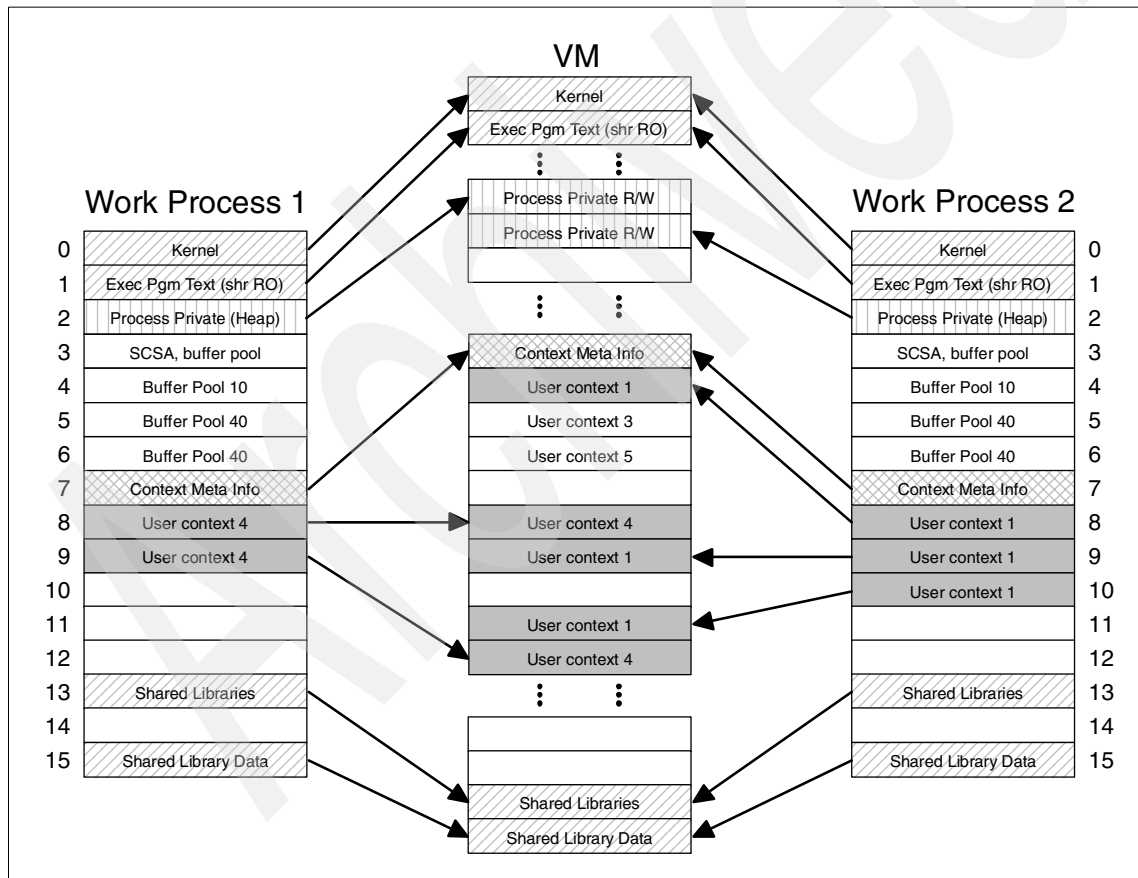


Figure 11-6 Alternative AIX ES Extended Memory configuration (32-bit)

These are the advantages of the AIX ES Extended Memory configuration:

- ▶ There is no architectural limit for the total quantity of Extended Memory that can be configured. The overall size depends only on the available main memory and paging space of the application server.
- ▶ The overhead introduced by the protection of segments during context switching remains nearly constant, because only the shared segment containing the meta information has to be protected. This is independent of the overall size of the Extended Memory and is comparatively inexpensive.
- ▶ There is no need to configure process private memory on the heap for batch processes to avoid the usage of Extended Memory. Batch user contexts no longer compete for memory with interactive user contexts.
- ▶ The alternative model is easy to configure but also flexible at the same time. For a sample configuration, refer to Section 11.3.5, “AIX ES Extended memory configuration recommendations” on page 327.

Important: Pay attention to the following configuration details if using the alternative implementation:

- ▶ You must *not* set the environment variable PSALLOC=EARLY. (See SAP Note 95454)
- ▶ Use EXTSHM=ON (Refer to Section 11.2.4, “Implications of the AIX version for SAP R/3 buffers” on page 315 and SAP Note 95260)

11.3.4 AIX ES shared memory implementation (64-bit)

In a 64-bit environment, it is also highly recommended that you use the AIX ES shared memory implementation. In comparison to the standard model, context switches are much more faster with large contexts.

The main difference from the AIX ES shared memory implementation for 32-bit is the fact that AIX allows a virtually unlimited number of shared memory segments for Extended Memory to be addressed by a single 64-bit work process. The 64-bit SAP R/3 Release 4.6D kernel technically limits the maximum size of a single user context to 256 segments, that is, 64 GB, at the moment. There are quota parameters that restrict the size of a user context, such as `ztta` or `roll_extension`, which practically limit the user context to about 8GB.

Figure 11-7 on page 326 shows the alternative AIX ES shared memory model for 64-bit. Again, there is one segment containing the meta information of all user contexts which is shared by all work processes. Work process 1 in the figure works with a very large user context built out of nine or more 256 MB segments. The other work process handles a small user context with only three segments.

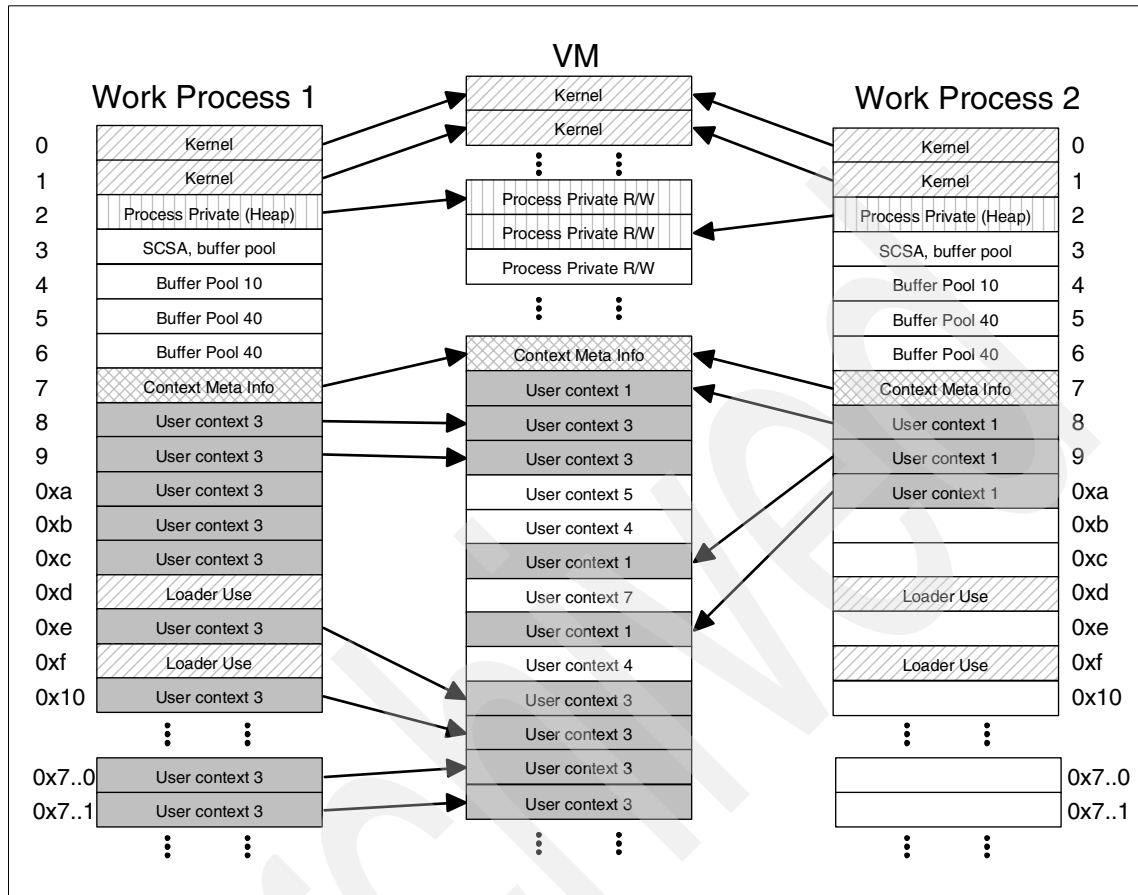


Figure 11-7 Alternative AIX ES Extended Memory configuration (64-bit)

Very large batch jobs in particular benefit from the removed size limit for user contexts. But be aware that badly or incorrectly coded ABAP reports can create large user contexts that can lead to serious performance problems caused by paging.

Important: Pay special attention to the available paging space when using the alternative AIX ES Extended Memory model for 64-bit. Running out of paging space may lead to a crash of the SAP R/3 instance because AIX may start to kill processes. Limit the total size of the Extended Memory with the instance profile parameter EM/MAX_SIZE_MB. Monitor the allocation of paging space regularly using the transaction ST06 or AIX monitoring tools, such as **topas**.

Using the 64-bit SAP R/3 kernel and the alternative Extended Memory model for AIX, it is possible to configure very large instances on an application server.

Pitfall ahead!

Although there are nearly no limits any more concerning Extended Memory, there are two SAP R/3 architectural limits that you should be aware of when planning to configure a very large instance:

- ▶ The maximum number of work processes per SAP R/3 instance
- ▶ The dispatcher of an SAP R/3 instance does not support more than 2048 user connections at a time.

These limits may lead to the setup of several instances per physical server system.

11.3.5 AIX ES Extended memory configuration recommendations

In this section, we provide instance profile parameter recommendations for the alternative AIX ES model, which are shown in Table 11-2. They apply to both 32-bit and 64-bit models (if not otherwise stated).

The recommendations are based on practical experience of the authors and on the following references:

- ▶ IBM SAP Marketing White Paper: *64-bit SAP R/3 on the RS/6000*, Version 1.3, April 2001
- ▶ SAP Notes 95260, 95454, 117267, 128935, 146289, 146528, 174882, 191801, 396983

There is no guarantee that these parameter values work for every specific environment.

Important: Some of the profile parameters are *case sensitive*.

Table 11-2 Parameter recommendations for AIX ES shared memory

Parameter (Instance profile)	Recommended value (64-bit / 32-bit)	Explanation
ES/TABLE	SHM_SEGS	Enables the alternative AIX ES Extended memory configuration.
ES/SHM_SEG_SIZE	256	Segment size (256 MB). This value must <i>not</i> be changed.
ES/SHM_USER_COUNT	2000	Maximum number of all open user sessions. Default value is 1024. Maximum number is 4000.

Parameter (Instance profile)	Recommended value (64-bit / 32-bit)		Explanation
ES/SHM_SEG_COUNT	Depends on the available main memory and paging space.		Specifies the maximum number of 256 MB segments that are used for Extended Memory for the whole instance. Default is 1024. See also EM/MAX_SIZE_MB.
ES/SHM_PROC_SEG_COUNT	64 (16 GB)	7 (max.) (1.75 GB)	Specifies the maximum number of 256 MB segments that are used for Extended Memory within a work process. Default is 3.
ES/SHM_MAX_PRIV_SEGS	63	6	Specifies the maximum number of 256 MB segments that can be dedicated to a user context. The value is usually ES/SHM_PROC_SEG_COUNT minus one.
em/blocksize_KB	4096		Specifies the block size for storage allocation if a user context demands further space. The recommended value reduces the number of expensive allocation system calls in comparison to the default value of 1024.
em/initial_size_MB	256 * ES/SHM_SEG_COUNT		Although this parameter is ignored with the ES model, it should be specified through the given formula to enable the sapfpar tool to correctly calculate the size for the Extended Memory.
EM/MAX_SIZE_MB	Depends on the available main memory and paging space.		This parameter limits the amount of actually used Extended Memory for the whole instance to the available space. It adapts the concept of virtual memory consumption to the real configured paging space. See also EM/TOTAL_SIZE_MB.
EM/TOTAL_SIZE_MB (as of Kernel 4.6D, patch level 570)	Depends on the available main memory and paging space.		This parameter replaces the parameter EM/MAX_SIZE_MB as of Kernel 4.6D, patch level 570, because it is not able to maintain the parameter using transaction RZ10/RZ11. The old parameter can still be used for compatibility reasons. See SAP Note 396983.
ztta/roll_first	1		Specifies the size of the initial roll segment. The value of 1 reduces the size of the roll memory to a minimum and speeds up the roll-in during a context switch.
ztta/roll_extension	4294967295 (4 GB) - for 64-bit 1610612736 (1.75 GB) - for 32-bit		Specifies the maximum size in bytes of a user context. Transactions or batch processes creating larger contexts are terminated by the instance. This is a protection from so called runaways. Runaways are badly implemented ABAP reports that consume too much memory. This parameter can be dynamically changed within SAP R/3 with the report RSMEMORY. Theoretically, this parameter could be calculated by the formula $ES/SHM_PRIV_SEGS * 256 * 1024 * 1024$.

Parameter (Instance profile)	Recommended value (64-bit / 32-bit)	Explanation
ztta/roll_area	3000000 (approx. 3 MB)	Limits the consumption of roll memory for background processes.
abap/ heap_area_dia	10000000 (approx. 10 MB)	Limits the amount of private memory (heap) that a dialog work process can consume. For the AIX ES model, the recommendation is to use a very low value to avoid heap bottlenecks.
abap/ heap_area_nondia	10000000 (approx. 10 MB)	Limits the amount of private memory (heap) that a background process can consume. For the AIX ES model, the recommendation is to use a very low value to avoid heap bottlenecks.
abap/ heap_area_total	2000000000 (approx. 2 GB)	Limits the total amount of private memory (heap) for all work processes.

Pay attention to the following instructions when configuring the alternative AIX ES Extended Memory model for an SAP R/3 instance.

Important:

- ▶ Implement the latest SAP R/3 kernel patches that are available at the time of configuration.
- ▶ Observe the SAP paging space recommendations. For the 64-bit kernel, at least 20 GB of paging space is required.
- ▶ Use EXTSHM=ON for all SAP R/3 instance processes.
- ▶ Do *not* use PSALLOC=EARLY for SAP R/3. (See SAP Note 95454)

11.4 AIX concepts

In this section, we discuss some features of the AIX operating system that are important in an SAP R/3 environment and that have not been discussed in the previous chapters. We briefly introduce some concepts of AIX and show what should be adapted for servers in the SAP R/3 environment.

You can find more information on how to understand the performance of your AIX system in the redbook *AIX 5L Performance Tools Handbook*, SG24-6039 and in the *Performance Management Guide* in the AIX 5L Version 5.1 online documentation. Even though they comprise the Version 5.1 of AIX, most of the information is also applicable for Version 4.3.3 of AIX.

11.4.1 CPU

The IBM @server pSeries and IBM RS/6000 servers are symmetric multiprocessor (SMP) machines based on 64-bit processor and memory architectures. An overview of servers that are normally used in an SAP R/3 environment can be found in Section 3.1.2, “IBM ^ pSeries and RS/6000” on page 42.

On an SMP server, only one copy of the operating system runs across all of its processors. The processors are coupled with a high-speed bus and share the same global memory, disks, and I/O devices. All of the processors are essentially identical and perform identical functions.

AIX has a multi threaded operating system design to exploit the architecture of the IBM @server pSeries and IBM RS/6000 servers.

To be able to determine possible bottlenecks in a server, we first introduce the fundamental definitions that are used to describe the work of an SMP system.

Processor modes

The fundamental dispatchable entity of AIX Version 4 is the *thread*. A thread can be thought of as a low-overhead process. It requires fewer resources to create than a process. Each new process is created with a single thread that has its parent process priority and contends for the CPU with the threads of other processes. The process owns the resources used in execution; the thread owns only its current state.

In AIX, each user process's thread that is dispatched to a CPU will be executing in either user mode or system mode. The kernel can also create its own processes, which can be single or multi-threaded, but these threads will always stay in system mode. Although user processes' threads run in user mode, these threads will run in system mode when they need services from the kernel. For example, when a user process needs to access a file, the kernel provides a system call for this purpose. During run time, the access to the file will be performed by a thread in system mode.

Starting with AIX 4.3.3, each CPU has its own *run queue*. Each thread of a process is placed according to its priority into the run queue. The processors then fetch the thread with the highest priority from the run queue and execute it for a time slice, which is usually 10 ms. After its completion, the next thread is fetched. If a thread has to wait for I/O, it is put on the *wait queue*. If the I/O request finished, the thread is put again on the run queue.

Using asynchronous disk I/O, the process does not have to wait for the completion of its disk I/O request and can continue running in its user or system mode. Thus, the process is not put on the wait queue during the asynchronous disk I/O.

When many processes run on an SMP system, it is no longer easy to track the properties of each process. It is therefore common to look at the characteristics of the processes running on all processors.

One processor can execute one thread at a time. The processor can execute a thread in system mode or in user mode.

If the processes have to access data from disk, the processor has to wait for the data. It is also possible that the processor has nothing to do; in other words, there are no runnable threads.

We now define these four modes of a processor more precisely:

User	The processor is in user mode, if the running thread executes within its application code and does not require kernel resources to perform computations, manage memory, or set variables.
System	There are two types of processes that can be in the system mode. The kernel processes (kprocs) itself, and the threads that have performed a system call to access kernel resources. In both cases, the processor is said to be in system mode.
Wait	If all the threads that are scheduled on a processor are in the wait queue, the mode of the processor is wait. With AIX 4.3.3 and later, NFS goes through the buffer cache. If waits in the buffer cache routines occur, the thread is put on the wait queue. Before AIX Version 4.3.3, the thread has been accounted to be idle.
Idle	The processor is idle, or waiting, without pending I/O requests (disk or NFS). If the run queue is empty, the system dispatches a process called wait.

These modes are used to describe the load of a processor. Commonly, the percentage of the time in which a processor is in each mode is calculated. In an SMP server, these values are often summed over all available processors.

Since AIX Version 4.3.3 and later, the method used to compute the percentage of CPU time spent waiting on disk I/O was modified. Before AIX Version 4.3.3, with one pending disk I/O request, all processors have been accounted to be in wait mode. The change in AIX 4.3.3 is to only account an idle CPU as in wait mode, if an outstanding I/O was started on that particular CPU.

11.4.2 The Virtual Memory Manager

Simply put, the function of the Virtual Memory Manager (VMM) is to manage the allocation of real memory page frames and to resolve references from a program to virtual memory pages. Typically, this happens in an instance where pages are not currently in memory or do not exist in the case where a process makes the first reference to a page of its data segment.

The amount of virtual memory used can exceed the size of the real memory of a system. The function of the VMM from a performance point of view is to:

- ▶ Minimize the processor use and disk bandwidth resulting from paging
- ▶ Minimize the response degradation for a process resulting from paging

Virtual memory segments can be of three types, defined as:

- ▶ *Persistent segments*

Persistent segments are used to hold file data from the local file systems. Because pages of a persistent segment have a permanent disk storage location, the VMM writes the page back to that location when the page has been changed, if it can no longer be kept in memory. When a persistent page is opened for deferred update, changes to the file are not reflected on permanent storage until an fsync subroutine operation is performed. If no fsync subroutine operation is performed, the changes are discarded when the file is closed. No I/O occurs when a page of a persistent segment is selected for placement on the free list if that page has not been modified. If the page is referenced again later, then it is read back in.

- ▶ *Working segment*

These segments are transitory and only exist during use by a process. Working segments have no permanent storage location and are hence stored in paging space when real memory pages need to be freed.

- ▶ *Client segments*

These segments are saved and restored over the network to their permanent locations on a remote file system rather than being paged out to the local system. CD-ROM page-ins and compressed pages are also classified as client segments.

The free list

The VMM maintains a list of free memory pages available to satisfy a page request. This list is known as the *free list*. The VMM uses a page replacement algorithm to determine which pages in virtual memory will have their page frames reassigned to the free list.

Page replacement

When the number of pages in the free list becomes low, the page stealer is invoked. The page stealer is a mechanism that moves through the Page Frame Table (PFT) looking for pages to steal. The PFT contains flags that indicate which pages have been referenced and which have been modified.

If the page stealer finds a page in the PFT that has been referenced, then it will not steal the page, but, rather, it will reset the reference flag. The next time that the page stealer passes this page in the PFT, if it has not been referenced, it will be stolen. Pages that are not referenced when the page stealer passes them the first time are stolen.

When the modify flag is set on a page that has not been referenced, it indicates to the page stealer that the page has been modified since it was placed in memory. In this instance, a page out is called before the page is stolen. Pages that are part of a working segment are written to paging space, while pages of persistent segments are written to their permanent locations on disk.

There are two types of page fault, a *new page fault*, where the page is referenced for the first time and a *repage fault*, where pages have already been paged out before. The stealer keeps track of the pages paged out, by using a history buffer that contains the IDs of the most recently paged out pages. The history buffer also serves the purpose of maintaining a balance between pages of persistent segments and pages of working segments that get paged out to disk. The size of the *history buffer* is dependent on the amount of memory in the system, for example, a memory size of 512 MB requires a 128 KB history buffer.

When a process terminates, its working storage is released, and pages of memory are freed up and put back on the free list. Files that have been opened by the process can, however, remain in memory.

On a symmetrical multiprocessor (SMP) system, the *lrud* kernel process is responsible for page replacement. This process is dispatched to a CPU when the minfree parameter threshold is reached. This process continues to steal pages until the free list has at least the number of pages specified by the maxfree parameter.

Starting with AIX Version 4.3.3, the lru process is multi-threaded. In this case, there is one lru thread per memory pool, which is then dispatched to the CPU. The number of memory pools is, by default, based on the amount of physical memory in the system and the number of CPUs. Each memory pool has its own minfree and maxfree, so when an individual memory pool's freelist reaches minfree, that pool's lru thread gets dispatched to do page replacement.

In the following list, we show possible values and restrictions of the parameters:

minfree	Specifies the minimum number of frames on the free list. This number can range from 8 to 204800. The default for minfree is 120.
maxfree	Specifies the number of frames on the free list at which page stealing is to stop. This number can range from 16 to 204800, but must be greater than the number specified by the minfree parameter by at least the value of maxpagehead (default 8). The default for maxfree is 128.

The page replacement algorithm is most effective when the number of repages is low. The perfect replacement algorithm would eliminate repage faults completely and would steal pages that are not going to be referenced again.

Paging space

The operating system supports three paging space allocation policies:

- ▶ Late Paging Space Allocation (LPSA)
- ▶ Early Paging Space Allocation (EPSA)
- ▶ Deferred Paging Space Allocation (DPSA)

The late paging space allocation policy (LPSA)

With the LPSA, a paging slot is only allocated to a page of virtual memory when that page is first touched. The risk involved with this policy is that when the process touches the file, there may not be sufficient pages left in paging space.

The early paging space allocation policy (EPSA)

This policy allocates the appropriate number of pages of paging space at the time that the virtual memory address range is allocated. This policy ensures that processes do not get killed when the paging space of the system gets low. To enable EPSA, set the environment variable PSALLOC=early. Setting this policy ensures that when the process needs to page out, pages will be available. The recommended paging space size when adopting the EPSA policy is at least four times the size of real memory.

The deferred paging space allocation policy (DPSA)

This is the default policy in AIX 4.3.3. The allocation of paging space is delayed until it is necessary to page out, hence no paging space is wasted with this policy. Only when a page of memory is required to be paged out will the paging space be allocated. This paging space is reserved for that page until the process releases it or the process terminates.

Attention: The early paging space allocation policy must not be used for SAP R/3. Our recommendation is to use the deferred paging space allocation policy. Because this policy is used by default, you do not have to change it at all.

File system caching

Computational pages can be defined as working storage segments and program text segments. *File pages* are defined as all other page types, usually persistent and client pages.

The AIX operating system will leave pages in memory that have been read or written to. If these file pages are requested again, then this saves an I/O operation.

The minperm and maxperm values control the level of this file system caching. The value of minperm is not a strict limit; it is only considered when the VMM needs to perform page replacement. The unit of minperm and maxperm is the percentage of total real memory.

The thresholds set by maxperm and minperm can be considered as the following:

- ▶ If the percentage of file pages in memory exceeds maxperm, only file pages are stolen by the page replacement algorithm.
- ▶ If the percentage of file pages in memory is in the range between minperm and maxperm, the page replacement algorithm steals only file pages, unless the number of file repages is higher than the number of computational repages.
- ▶ If the percentage of file pages in memory is less than minperm, both file pages and computational pages are stolen by the page replacement algorithm.

Pitfall ahead!

In some cases, applications such as databases cache pages by themselves. Therefore, there is no need for the file system to cache pages as well. In this case, the values of minperm and maxperm can be set to low values.

The meaning of `maxperm` can be changed to define a hard limit. When the value of the parameter `strict_maxperm` is changed from its default 0 (zero) to 1 (one), the value of the parameter `maxperm` becomes a hard limit for the size of memory that is used for a persistent file cache. When the upper limit is reached, the least recently used (LRU) algorithm is performed on persistent pages.

Pitfall ahead!

When enabling `strict_maxperm`, it should be done right after the machine has been booted or before too much of the memory is used by applications; otherwise, there may be a huge amount of page replacement activity to force the `maxperm` value to be adhered to if the VMM is told that `maxperm` must be a strict limit.

The `numperm` value that is displayed by the `vm tune` command represents the number of non-text persistent or file pages. This value is not tunable. It is the actual percentage of pages in memory that are classified as file pages.

Attention: On the database server of an SAP R/3 system, duplicate buffering of the database files should be avoided by correctly setting the `vm tune` parameters.

Bright idea!

Recommendations for SAP R/3

The file system cache of the VMM should not duplicate the caching of data, which is already done by the database.

If your SAP R/3 database server has more than 2 GB of real memory and is used only for SAP R/3 services, then you should set the values of `minperm` and `maxperm` so that:

- ▶ The value of `minperm` is equivalent to 150 MB.
- ▶ The value of `maxperm` is equivalent to 250 MB.

Because the unit of these values is a percentage of real memory, this means, for example, that for a system with 2 GB real memory, the value of `minperm` has to be set to 7, and the value of `maxperm` has to be set to 12.

We recommend that you not use the strict limit for `maxperm`; thus, you have to ensure that `strict_maxperm` is set to 0 (zero):

```
strict_maxperm=0
```

If you have a database server with less than 2 GB of real memory, or if you have non SAP R/3 services running on the database server, we recommend that you start with the default values of AIX and only start tuning later on.

For information on how to set the parameters, refer to Section 11.5.2, “The `vm tune` command” on page 347.

Sequential-Access Read Ahead

The VMM tries to anticipate the future need for pages of a sequential file by observing the pattern a program uses to access the file. When the program accesses two successive pages of the file, the VMM assumes that the program will continue to access the file sequentially, and the VMM schedules additional sequential reads of the file. This is called *Sequential-Access Read Ahead*. These reads are overlapped with the program processing, and will make the data available to the program sooner than if the VMM had waited for the program to access the next page before initiating the I/O. The number of pages to be read ahead is determined by two VMM thresholds:

minpgahead	A number of pages read ahead when the VMM first detects the sequential access pattern. If the program continues to access the file sequentially, the next read ahead will be for 2 x minpgahead, the next for 4 x minpgahead, and so on until the number of pages reaches maxpgahead. This value can range from 0 through 4096. It should be a power of 2.
maxpgahead	A maximum number of pages the VMM will read ahead in a sequential file. This value can range from 0 through 4096. It should be a power of 2 and should be greater than or equal to minpgahead.

If the program deviates from the sequential-access pattern and accesses a page of the file out of order, sequential read ahead is terminated. It will be resumed with minpgahead pages if the VMM detects a resumption of sequential access by the program.

Attention: Due to limitations in the kernel, the maxpgahead value should not exceed 512. The difference between minfree and maxfree should always be equal to or greater than the value of maxpgahead.

Bright idea!

Recommendations for SAP R/3

For a database server in a SAP R/3 system running Oracle, the following values should be set:

minpgahead	2
maxpgahead	16

For information on how to set the parameters, refer to Section 11.5.2, “The vmtune command” on page 347.

For a detailed discussion of the settings, refer to the redbook *Database Performance on AIX in DB2 UDB and Oracle Environments*, SG24-5511.

11.4.3 Asynchronous disk I/O

The disk I/O subsystem of AIX is built up using different software layers. Figure 4-11 on page 122 gives an overview of the different layers involved when an I/O request of an application is processed. In this section, we first show how the asynchronous disk I/O system works in AIX and how it can be configured. We then give recommendations for the attributes in an SAP R/3 environment.

How does asynchronous disk I/O work?

If the asynchronous disk I/O has been enabled in the run-time environment, all applications using the application interface for asynchronous disk I/O will profit from an enhanced performance for disk I/O operations.

Normally, an application has to wait for the I/O subsystem of the operating system to complete its disk I/O request before it can continue. If an application is programmed to use asynchronous disk I/O, the I/O request runs in the background and does not block the application. Thus, the disk I/O requests and the application run simultaneously, which can improve performance to a great extent.

The application has to know in which status each of its I/O requests are. Therefore, every I/O request has its control block in the application user space. Using these control blocks, an application can determine whether and how its I/O operations are completed. There are three ways of acquiring the information:

- ▶ The application can poll the status of the I/O operation.
- ▶ The operating system asynchronously notifies the application when the I/O operation is complete.
- ▶ The application can wait until the I/O operation is complete.

Using asynchronous disk I/O on JFS, the I/O requests of an application are placed in a queue. The requests in that queue are processed by kernel processes (KPROC) called *aio*servers. An aio server takes one request from the queue and is in charge of this request until it is completed. The same aio server cannot process another request during this time. Thus, the number of aio servers limit the number of JFS disk I/O operations that can be in progress simultaneously in the system.

There are some attributes which can be changed to influence the asynchronous disk I/O:

minservers	The minimum number of aio servers that are started for asynchronous disk I/O. The default value is 1.
maxservers	The maximum number of aio servers that are started for asynchronous disk I/O. The default value is 10. Since

each aioserver uses memory, this number should not be much larger than the expected amount of simultaneous asynchronous disk I/O requests.

maxreqs	Maximum number of asynchronous disk I/O requests that can be stored in the queue. The default value is 4096.
kprocprio	The process priority with which the aioservers are started at initialization. The default value for kprocprio is 39.
autoconfig	Indicates if the aio0 device will be in the status defined or available after a reboot. The default value is defined. If you want to use aio0 directly after each reboot, you should set the value to available.

When the asynchronous I/O system is enabled, the kernel starts the number of aioservers, specified by the attribute minservers. If a process queues an I/O request while no free aioservers are available, one additional aioserver will be forked, as long as the maximum number of aioservers specified in the attribute maxservers is not exceeded. These new aioservers will be started in the context and the priority of the initiating process. After the maximum number of aioservers are started, an additional asynchronous I/O request has to wait for an aioserver to complete an I/O operation.

It is obvious that you have to adapt the value of maxreqs, depending of the settings of maxservers.

The aioservers that are started at the initialization of the aio0 device run with the user ID of root and with a process priority of 39. The process priority can be changed using the attribute kprocprio.

Attention: It is not recommended to decrease the value for kprocprio, because system hangs or crashes can occur if the aioservers are favored too much. We recommend that you not change the value of kprocprio at all.

If a process submits an asynchronous disk I/O request while no aioservers are idle, an additional aioserver is started. This aioserver runs with the user ID and the process priority of the process that submitted the asynchronous disk I/O request. The default for the process priority of normal users is 40, with a nice of 20, which leads to a total process priority of 60.

Asynchronous disk I/O before AIX 4.3.3

The asynchronous disk I/O has been changed between AIX 4.3.2 and AIX 4.3.3. Before AIX Version 4.3.3, the algorithm did not work asynchronously if two I/O operations accessing the same file were queued. When an application queued a second I/O operation on a file, where another I/O request was not completed yet, the application had to wait for the completion of the first I/O request. This is also called *blocking disk I/O*.

In AIX Version 4.3.3 and all subsequent releases, the algorithm is changed to also allow asynchronous I/O operations to be performed on the same file, which is called *non-blocking disk I/O*.

For some applications, this change can increase performance up to 50 percent. Because databases have to make small changes in large files, many asynchronous disk I/O requests per file are often scheduled.

Bright idea!

Recommendations for an SAP R/3 system

In an SAP R/3 environment, it is very important that you set up the asynchronous disk I/O on the database servers correctly. The amount of needed aioservers strongly depends on the size and work load of the database and the used machine, but normally scales with the number of used data files.

Attention: If you change values of the attributes of aio0, they will be used after the next restart of the system.

We recommend that you start the tuning procedure with the following settings:

1. The value for maxservers should be set to two times the number of files, which are accessed asynchronously.
2. Let C be the number of CPUs on the server. Then the value for minservers should be set to C-1 or to 2, whichever is larger.
3. Make sure that the maximum number of processes allowed per user maxuproc is set appropriately. For a large SAP R/3 system, we recommend that you set the value of maxuproc to 2000 using the following command:

```
chdev -l sys0 -a maxuproc=2000.
```
4. Set the value of the parameter maxreqs to 12288.

You then can monitor the amount of used aioservers for a period when normal operation is going on.

Configuration of the asynchronous disk I/O

In this section, we describe how to configure the asynchronous disk I/O using the command line. You can also configure the asynchronous disk I/O using `smit` with the fastpath `smitty aio`.

The asynchronous disk I/O is defined as a device of the class `aio`. To check the state of the asynchronous disk I/O, you can use the following command:

```
lsdev -Cc aio
```

If the asynchronous disk I/O device `aio0` is not enabled, you will get the following output:

```
aio0 Defined Asynchronous I/O
```

To add the device `aio0` to the system, you have to use the command:

```
mkdev -l aio0
```

After having added the device `aio0`, its state is changed to `Available`, which is shown in the following:

```
lsdev -Cc aio
aio0 Available Asynchronous I/O
```

Once the device is added, you cannot disable it from the system without a reboot. Also, all changes of attributes of the device `aio0` can only be stored in the ODM and will be used when initializing the device `aio0` again after a restart.

You can have a look at the attributes of the device `aio0` which are stored in the ODM:

```
lsattr -El aio0 -H
```

The output of this command is shown in Example 11-4.

Example 11-4 Output of the `lsattr -El aio0 -H` command

attribute	value	description	user_settable
minservers	1	MINIMUM number of servers	True
maxservers	10	MAXIMUM number of servers	True
maxreqs	4096	Maximum number of REQUESTS	True
kprocprio	39	Server PRIORITY	True
autoconfig	defined	STATE to be configured at system restart	True
fastpath	enable	State of fast path	True

Attention: The `lsattr` command displays the values that are stored in the ODM. If the values of the device `aio0` have been changed since the last initialization of `aio0`, the values displayed with the `lsattr` command are not identical to the values actually used by `aio0`!

To change an attribute of the device aio0 in the ODM, you can use the command

```
chdev -l aio0 -P -a attribute=value
```

To change, for example, the maximum number of aioservers in the ODM to a value of 200, you can use the following command:

```
chdev -l aio0 -P -a maxservers=200
```

Important: You have to change the attribute autoconfig to the value Available to have asynchronous I/O enabled after a restart. Use the command:

```
chdev -l aio0 -P -a autoconfig=available
```

You can check the value of maxuproc (maximum number of processes allowed per user) in your system using the command:

```
lsattr -El sys0 -H
```

The output of this command is shown in Example 11-5.

Example 11-5 Output of the lsattr -El sys0 -H command

attribute	value	description	user_settable
maxuproc	500	Maximum number of PROCESSES allowed per user	True
[...]			

Depending on the version of AIX, you can check the number of aioserver which are running using the following commands:

AIX 4.3.3 `pstat -a | grep aioserver | wc -l`

AIX 5.1 `ps -k | grep aioserver | wc -l`

11.4.4 Disk I/O pacing

In AIX, a disk I/O pacing algorithm is not used by default. It can be used to prevent processes from saturating the I/O subsystem.

The disk I/O pacing algorithm

Some processes generate a very large amount of disk I/O requests, which can deteriorate the response times of other programs.

There are two parameters used in the disk I/O pacing algorithm that can be adjusted:

- maxpout** HIGH water mark for pending I/O write requests per file
- minpout** LOW water mark for pending I/O write requests per file

The algorithm works on a per file basis. If the number of pending I/O write requests for a file reaches the value of maxpout, all processes submitting a write I/O request for that file are put to sleep. If enough write I/O operations for this specific file have been completed, so that fewer than minpout pending write I/O requests remain, all processes which had submitted write I/O requests for that file will be put on the run queue again.

Thus, using disk I/O pacing will slow down processes that are writing heavily on a few files. In the time these processes are put to sleep, other processes can run. Thus, using disk I/O pacing can improve the response times of these processes comparing to the AIX default.

The default for the values of maxpout and minpout is 0. This means that the algorithm is not used at all.

Using disk I/O pacing, you tune the system. Disk I/O Pacing sacrifices some throughput on I/O intensive programs to improve the response time of the other programs. The challenge for a system administrator is to choose settings that result in a throughput/response time trade-off that is consistent with the organization's priorities. However, choosing the wrong values for the algorithm can deteriorate both the throughput and the response time drastically.

Bright idea!

Recommendations for an SAP R/3 system

In an SAP R/3 system, only the database server will experience high disk I/O activity. Therefore disk I/O pacing should stay turned off on all application servers by default.

The following recommendations apply to SAP R/3 systems without an HACMP cluster:

- ▶ If you use a central system and you experience slow response times of the application server during high loads on the database, see whether using disk I/O pacing improves the response times.
- ▶ If you use a distributed SAP R/3 system, where the database server and the application servers are separated, there normally is no need to activate disk I/O pacing on the database server.

If HACMP is used on the database server, you have to make sure that even during high load periods the HACMP Cluster Manager Process is scheduled to run regularly. Otherwise, the dead man switch is triggered, which halts the system instantly, resulting in a false takeover situation.

Attention: Disk I/O pacing is not used per default in an HACMP cluster. However, if you have a high I/O load on a cluster node, or when the Dead Man Switch is triggered on a cluster node, it is recommended to use disk I/O pacing.

The following recommendation and a further discussion is given in the *Performance Management Guide*:

“The high- and low-water marks were chosen *by trial and error*, based on the knowledge of the I/O path. Choosing them is not straightforward because of the combination of write-behind and asynchronous writes. High-water marks maxpout of $(4 * n) + 1$, where n is a positive integer, work particularly well.”

As a starting point for the value of n, you can use the number of active disks the application is writing to. You should then adjust the value of n so that it is a multiple of 4.

For the value of minpout, a good starting point is $3*n$. In Table 11-3, we give an example for different numbers of active disks.

Table 11-3 Example for a starting point for disk I/O pacing

Number of disks	maxpout	minpout
8	33	24
32	129	96
40	161	120
80	321	240

Attention: The settings maxpout=33 and minpout=24, which can be found in many references, were determined many years ago, by trial and error, on far less powerful systems than are typically used today. Using these settings will reduce the I/O throughput severely. We have experienced reductions of the I/O throughput up to 80 percent on some systems!

Setting the attributes for disk I/O pacing

You can check the actual values of maxpout and minpout using the following command:

```
lsattr -El sys0 -H
```

The output of this command is shown in Example 11-6.

Example 11-6 Output of the lsattr -El sys0 -H command

attribute	value	description	user_settable
maxpout	0	HIGH water mark for pending write I/Os per file	True
minpout	0	LOW water mark for pending write I/Os per file	True
[...]			

In the following example, we set maxpout to 161 and minpout to 120:

```
chdev -l sys0 -a maxpout=161 -a minpout=120
```

These values for the attributes maxpout and minpout are active immediately. If the -P flag is specified for the **chdev** command, only the database is updated to reflect the changes, and the device itself is left unchanged. The -T flag of the **chdev** command is used to make a temporary change in the device without the change being reflected in the database.

11.5 Tools to monitor AIX

In the previous sections, we gave you some advice on how to set up a well tuned and reliable SAP R/3 system. Nevertheless, every workload on a system is different, so that the previous suggestions are only a starting point.

In this section, we briefly present some performance monitoring tools. It is neither intended to give a full review of the available tools or to guide you while analyzing the performance of your system, but to show you which values are reasonable to be checked first in an SAP R/3 system.

You can find more information on performance monitoring and tuning tools in *Performance Management Guide, AIX 5L Version 5.1* (found in the online documentation) and in the redbook *AIX 5L Performance Tools Handbook*, SG24-6039.

11.5.1 The vmstat command

The **vmstat** command reports statistics about kernel threads, virtual memory, disks, traps, and CPU activity. Reports generated by the **vmstat** command can be used to balance system load activity. These system-wide statistics (among all processors) are calculated as averages for values expressed as percentages, and as sums otherwise.

If the **vmstat** command is invoked without flags, the report contains a summary of the virtual memory activity since system startup (see Example 11-7).

Example 11-7 Output of the vmstat command

kthr		memory			page				faults				cpu			
r	b	avm	fre	re	pi	po	fr	sr	cy	in	sy	cs	us	sy	id	wa
1	5	3788039	245	0	70	93	1638	34272	0	2131	9205	3921	2	11	47	40
0	4	3788039	262	0	57	119	434	8105	0	1930	5238	1766	0	4	75	21
0	6	3788039	244	0	58	101	964	25034	0	2167	13396	4728	2	2	60	37
0	10	3788039	24	0	83	59	1683	28003	0	2174	8696	4454	3	2	51	44
2	10	3788044	247	0	52	180	1486	35867	0	2139	6941	3770	5	10	35	49
2	6	3788044	243	0	38	190	994	20806	0	2152	6140	2753	10	11	43	36
1	6	3788044	152	0	26	224	985	9582	0	2083	7601	2099	9	6	50	35
1	5	3788044	217	0	17	242	743	14459	0	2258	6824	2990	5	2	65	27

The columns of the output show the following attributes:

kthr	Kernel thread state changes per second over the sampling interval. r Number of kernel threads placed in run queue. b Number of kernel threads placed in wait queue (awaiting resource, awaiting input/output).
Page	Information about page faults and paging activity. These are averaged over the interval and given in units per second. re Pager input/output list. pi Pages paged in from paging space. po Pages paged out to paging space. fr Pages freed (page replacement). sr Pages scanned by page-replacement algorithm. cy Clock cycles by page-replacement algorithm.
Faults	Trap and interrupt rate averages per second over the sampling interval. in Device interrupts. sy System calls. cs Kernel thread context switches.
CPU	Breakdown of percentage usage of CPU time. us User time.

- sy** System time.
- id** CPU idle time.
- wa** CPU idle time during which the system had outstanding disk/NFS I/O request(s). See detailed description in “Processor modes” on page 330.

11.5.2 The vmtune command

The **vmtune** command changes operational parameters of the Virtual Memory Manager and other AIX components.

Attention: Misuse of this command can cause performance degradation or operating system failure. Before experimenting with **vmtune**, you should be thoroughly familiar with both Performance Overview of the Virtual Memory Manager (VMM) and Tuning VMM Page Replacement.

The **vmtune** command is included in the fileset bos.adt.samples. You have to check to see if it is installed on your server. You always should use the fileset bos.adt.samples from the same version of AIX as your installed kernel. To check if the fileset bos.adt.samples is installed, you can use the command:

```
# lsllpp -l bos.adt.samples
```

The output of this command is shown in Example 11-8.

Example 11-8 Output of the lsllpp -l bos.adt.samples command

# lsllpp -l bos.adt.samples			
Fileset	Level	State	Description

Path: /usr/lib/objrepos			
bos.adt.samples	4.3.3.0	COMMITTED	Base Operating System Samples

Displaying attributes

If you call the **vmtune** command without any parameters, the values of all parameters that can be changed are displayed:

```
/usr/samples/kernel/vmtune
```

In Example 11-9 on page 348, the output of the **vmtune** command is displayed. The most important attributes for an SAP R/3 system are already explained in Section 11.4, “AIX concepts” on page 329. The other attributes can be checked using the manual pages.

Example 11-9 Output of the vmtune command

vmtune: current values:									
-p	-P	-r	-R	-f	-F	-N	-W		
minperm	maxperm	minpgahead	maxpgahead	minfree	maxfree	pd_npages	maxrandwrt		
52219	208876	2	8	120	128	524288	0		
-M	-w	-k	-c	-b	-B	-u	-l	-d	
maxpin	npswarn	npskil	numclust	numfsbufs	hd_pbuf_cnt	lvm_bufcnt	lrubucket	defps	
209696	16384	4096	1	93	64	9	131072	1	
-s	-n	-S	-h						
sync_release_ilock	nokillroot	v_pinshm	strict_maxperm						
0	0	0	0						
number of valid memory pages = 262119				maxperm=79.7% of real memory					
maximum pinable=80.0% of real memory				minperm=19.9% of real memory					
number of file memory pages = 14672				numperm=5.6% of real memory					

Changing attributes

To change a value of a parameter using **vmtune**, you have to use the appropriate flags, which are given in Example 11-9. The following command will, for example, set the value of the parameter **minfree** to the value of 120:

```
/usr/samples/kernel/vmtune -f 120
```

Attention: The changes which are made using **vmtune** are not stored. To set up the system after a reboot, you can invoke a script with appropriate **vmtune** commands from the **/etc/inittab** file.

11.5.3 The iostat command

The **iostat** command is used for monitoring the load on the system input/output device by observing the time the physical disks are active in relation to their average transfer rates. The **iostat** command generates reports that can be used to change the system configuration in order to better balance the input/output load between physical disks and adapters. The syntax of the command is:

```
iostat
```

An output of this command is shown in Example 11-10.

Example 11-10 Output of the iostat command

tty:	tin	tout	avg-cpu:	% user	% sys	% idle	% iowait
	0.0	1.3		0.0	0.1	99.9	0.0

Disks:	% tm_act	Kbps	tps	Kb_read	Kb_wrtn
hdisk0	0.2	1.0	0.2	175446	377645
hdisk1	0.0	0.1	0.0	32779	57
hdisk2	0.0	0.0	0.0	2167	4
hdisk3	0.0	0.0	0.0	2175	17
cd0	0.0	0.0	0.0	0	0

To improve performance, the collection of disk I/O statistics can be disabled. You can use the **lsattr** command to check if the collection of disk I/O statistics is enabled:

```
lsattr -E -l sys0 -a iostat -H
```

The output is:

```
attribute value description user_settable
iostat true Continuously maintain DISK I/O history True
```

If the value of the attribute **iostat** is set to **false**, you can enable the collection of disk I/O data using the following command:

```
chdev -l sys0 -a iostat=true
```

11.5.4 The filemon command

The **filemon** command records a trace of file system and I/O system events, and reports on the file and I/O access performance during that period.

In its normal mode, the **filemon** command runs in the background while one or more application programs or system commands are being executed and monitored. The **filemon** command automatically starts and monitors a trace of the program's file system and I/O events in real time. By default, the trace is started immediately; optionally, tracing may be deferred until the user issues a **trcon** command. The user can issue **trcoff** and **trcon** commands while the **filemon** command is running in order to turn off and on monitoring, as desired. When tracing is stopped by a **trcstop** command, the **filemon** command generates an I/O activity report and exits.

The **filemon** command can also process a trace file that has been previously recorded by the trace facility. The file and I/O activity report will be based on the events recorded in that file.

In Example 11-11, we give an output of the **filemon** command.

Example 11-11 Output of the filemon command

```
# Cpu utilization: 1.3%
[filemon: Reporting started]
111071 events were lost. Reported data may have inconsistencies or errors
```

Most Active Segments

#MBs	#rpgs	#wpgs	segid	segtype	volume:inode
125.2	0	32052	25a2	page table	

Most Active Logical Volumes

util	#rblk	#wblk	KB/s	volume	description
1.00	0	256416	12061.6	/dev/test	/test

Most Active Physical Volumes

util	#rblk	#wblk	KB/s	volume	description
0.99	0	256448	12063.1	/dev/hdisk1	N/A

Detailed VM Segment Stats (4096 byte pages)

SEGMENT: 25a2 segtype: page table
segment flags: pgtbl
writes: 32052 (0 errs)
write times (msec): avg 15.645 min 4.792 max 499.696 sdev 20.421
write sequences: 2
write seq. lengths: avg 16026.0 min 448 max 31604 sdev 15578.0

Detailed Logical Volume Stats (512 byte blocks)

VOLUME: /dev/test description: /test
writes: 8421 (0 errs)
write sizes (blks): avg 30.4 min 8 max 32 sdev 5.2
write times (msec): avg 15.942 min 4.781 max 812.042 sdev 26.013
write sequences: 44
write seq. lengths: avg 5827.6 min 192 max 8192 sdev 3161.2
seeks: 44 (0.5%)
seek dist (blks): init 105096,
avg 153991.3min 8 max 426520 sdev 156919.3
time to next req(msec): avg 1.262 min 0.001 max 799.493 sdev 10.158
throughput: 12061.6 KB/sec
utilization: 1.00

Detailed Physical Volume Stats (512 byte blocks)


```

-----
VOLUME: /dev/hdisk1  description:
                        sg245050 pain is temporary, pride is forever
writes:                2511      (0 errs)
  write sizes (blks):  avg  102.1 min    32 max    128 sdev   35.3
  write times (msec):  avg  4.217 min   0.039 max 800.684 sdev  18.831
  write sequences:     43
  write seq. lengths:  avg 5963.9 min    544 max   8192 sdev 3075.7
seeks:                 43      (1.7%)
  seek dist (blks):    init 109352,
                      avg 157459.6 min    8 max  426520 sdev 157134.9
  seek dist (%tot blks):init 0.61523,
                      avg 0.88589 min 0.00005 max 2.39966 sdev 0.88406
time to next req(msec): avg  1.326 min   0.007 max 799.492 sdev  10.409
throughput:            12063.1 KB/sec
utilization:           0.99
[filemon: Reporting completed]

[filemon: 10.629 secs in measured interval]

```

The **filemon** command monitors logical I/O operations on logical files. The monitored operations include all read, write, open, and lseek system calls, which may or may not result in actual physical I/O, depending on whether or not the files are already buffered in memory.

Also, the physical I/O operations between segments and their images on disk are monitored, which corresponds to paging. The **filemon** command monitors I/O operations on logical volumes as well as on physical volumes. At this level, physical resource utilizations are obtained.

The report gives numbers on the average (avg), maximum (max), minimum (min) and standard deviation (sdev) of each value, which is helpful to estimate the distribution of the measured quantity. The sequences and the sequence length give an indication whether the I/O is random or sequential, which is helpful to know if slowdowns in the read or write performance are observed. If, for example, the write sequence is one and the amount of seeks is equal to the total amount of write requests, then the access is definitely random.

For a detailed discussion of the **filemon** command, refer to the manual pages.

11.5.5 The netstat -m command

The **netstat** command symbolically displays the contents of various network-related data structures for active connections.

The network subsystem uses a memory management facility that revolves around a data structure called an *mbuf*. Mbufs are mostly used to store data for incoming and outbound network traffic. Having mbuf pools of the right size can have a very positive effect on network performance.

The **netstat -m** command displays the statistics for the communications memory buffer (mbuf) usage. Each processor has its own mbuf pool. If the network option `extendednetstats` is set to 1, then a summary over all processors is collected and displayed. For performance reasons, the network option `extendednetstats` is set to 0 (zero) in `/etc/rc.net`.

In Example 11-12, a sample output of the **netstat -m** command is shown for the first CPU0.

Example 11-12 Output of the netstat -m command

Kernel malloc statistics:

***** CPU 0 *****

By size	inuse	calls	failed	delayed	free	hiwat	freed
32	98	2952	0	0	30	1440	0
64	45	247	0	0	19	720	0
128	27	11174	0	0	293	360	0
256	34	1290968	0	0	510	864	0
512	30	2963	0	0	10	90	0
1024	6	206	0	0	6	225	0
2048	0	9710	0	0	224	225	80
4096	22	101516	0	0	39	270	0
8192	2	17	0	0	1	22	0
16384	0	0	0	0	40	54	0
65536	1	1	0	0	0	2047	0

***** CPU 1 *****

[...]

In the example, you can see that for sizes of 32 bytes up to 65536 bytes, mbufs have been created. The sizes of the mbufs are given in the column **By size**. In the column **inuse**, the number of used mbufs of each size are given. The column **calls** shows the number of time a call accesses a mbuf of the given size. The column **free** shows the number of free already allocated memory areas for each mbuf size.

If there are entries in the column **failed** there should be a closer monitoring of the mbufs or the network parameters `sb_max` or `thewall` should be changed.

11.5.6 The topas command

The **topas** command reports selected statistics about the activity on the local system. This tool and the **monitor** tool, which is discussed in the next section, can be used to get a first impression of the system performance.

The command uses the curses library to display its output in a format suitable for viewing on an 80x25 character-based display or in a window of at least the same size on a graphical display. The **topas** command requires the `perfragent.tools` fileset to be installed on the system.

An example for the output of **topas** can be found in Example 11-13.

Example 11-13 Output of the topas command

Topas Monitor for host: saphost						EVENTS/QUEUES		FILE/TTY	
Thu Aug 9 16:36:25 2001 Interval: 2						Cswitch	46900	Readch	15073447
						Syscall	172997	Writech	15072257
Kernel	28.3	#####				Reads	86295	Rawin	0
User	15.8	#####				Writes	86290	Ttyout	210
Wait	0.0					Forks	0	Igets	0
Idle	55.7	#####				Execs	0	Namei	0
						Runqueue	1.6	Dirblk	0
Interf	KBPS	I-Pack	O-Pack	KB-In	KB-Out	Waitqueue	1.2		
tr0	1.0	7.5	0.5	0.8	0.2				
lo0	0.0	0.0	0.0	0.0	0.0				
						PAGING		MEMORY	
						Faults	3594	Real,MB	1023
Disk	Busy%	KBPS	TPS	KB-Read	KB-Writ	Steals	0	% Comp	12.7
hdisk1	100.0	14377	263.0	0	14377	PgspIn	0	% Noncomp	27.7
hdisk0	0.0	0.0	0.0	0.0	0.0	PgspOut	0	% Client	0.5
hdisk3	0.0	0.0	0.0	0.0	0.0	PageIn	0		
hdisk2	0.0	0.0	0.0	0.0	0.0	PageOut	3591	PAGING SPACE	
hdisk4	0.0	0.0	0.0	0.0	0.0	Sios	950	Size,MB	2048
								% Used	0.5
								% Free	99.4
dd	(5274)	100.0%	PgSp:	0.1mb	root				
dd	(11728)	74.0%	PgSp:	0.1mb	root				
topas	(15286)	0.5%	PgSp:	0.4mb	root				
syncd	(3206)	0.0%	PgSp:	0.1mb	root				
dtgreet	(5726)	0.0%	PgSp:	1.1mb	root				
						Press "h" for help screen.			
						Press "q" to quit program.			

We will now briefly explain the important areas of the screen of **topas**. For elements which are not discussed here, please refer to the manual pages.

- Processors

In this area, the percentage of Kernel (System), User, Wait and Idle time of the processors are displayed.
- Interf

In this area of the screen network interface, statistics are displayed. The interface that transferred most bytes over the interval is listed first. KBPS denotes the actual transfer

rate. I-Pack and O-Pack denote the number of incoming and outgoing packets. KB-In and KB-Out are the amounts of incoming and outgoing KBs.

Disk

In this area of the screen, disks statistics are shown. The disks are ordered according to their activity during the monitoring interval. Busy% indicates the percentage of time the physical disk was active (bandwidth utilization for the drive). KBPS denotes the actual transfer rate. TPS shows the number of transfers per second to a physical disk. KB-Read and KB-Writ denote the amount of KBs read and written to an hdisk.

Processes

In the bottom left corner, processes are displayed as an ordered list according to their CPU usage during the monitoring interval. For each process, the name, the process ID, the percentage of CPU utilization, the paging space used, and the user are listed.

Events/Queues

The runqueue denotes the average number of threads that were ready to run but were waiting for a processor to become available. The waitqueue similarly denotes the average number of threads that were waiting for I/O to complete.

Memory

Real,MB denotes the size of real memory in MBs. % Comp, % Noncomp and % Client show the percentage of real memory currently allocated to computational, non-computational, and client page frames.

Paging Space

Displays the size and utilization of paging space. Size,MB denotes the sum of all paging spaces on the system. %Used and %Free give the percentage of total paging space currently in use or free.

PAGING

Displays the per-second frequency of paging statistics. PgspIn and PgspOut denote the number of 4 KB pages read and written from paging space per second. PageIn and PageOut denote the number of 4 KB pages read and written per second. This includes paging activity associated with reading or writing from file systems. By subtracting PgspIn or PgspOut from this value, you get the number of 4 KB pages read or written from file systems per second.

11.5.7 The monitor tool

The **monitor** tool is an AIX System performance monitor program. **monitor** can be used to display system statistics of various short time performance values on a full screen terminal for AIX releases 3.1 up to 4.3. It exceeds the capabilities of **topas** in a few areas.

Attention: This tool is not officially supported by IBM. Use it at your own risk.

The latest version of **monitor** is available from:

<http://www.mesa.nl/pub/monitor>

You can also find the license agreement for **monitor** on that server.

While **monitor** is running, you can press h or ? to see its interactive commands.

An output of **monitor** is shown in Example 11-14.

Example 11-14 Output of monitor

```

AIX System monitor v2.1.7PRE6 07apr2000: saphost Thu Aug 9 16:05:27 2001
Uptime: 1 days, 21:56 Users: 0 of 0 active 0 remote 00:00 sleep time
CPU: User 17.0% Sys 25.5% Wait 0.0% Idle 57.5% Refresh: 1.00 s
0% 25% 50% 75% 100%
>>>>>>>>>=====

Runnable (Swap-in) processes 1.00 (1.00) load average: 1.35, 0.63, 0.25

Memory      Real      Virtual    Paging (4kB)    Process events    File/TTY-I/O
free        520 MB    2046 MB    3610.1 pgfaults  46899 pswitch     0 iget
procs      125 MB     1 MB      0.0 pgin        173690syscall     3 namei
files      378 MB     3640.1 pgout   866790read      0 dirblk
total     1024 MB    2048 MB    0.0 pgsin       866631write      15140374readch
IO (kB/s)  read    write busy%    0.0 pgsout       0 fork           15137369writetech
hdisk0     0.0     0.0     0               0 exec           0 ttyrawch
hdisk1     0.0 14560.4 100 Client Server NFS/s 0 rcvint         0 ttycanch
hdisk3     0.0     0.0     0 0.0 0.0 calls 0 xmtint         175 ttyoutch
hdisk2     0.0     0.0     0 0.0 0.0 retry 0 mdmint
hdisk4     0.0     0.0     0 0.0 0.0 getattr
hdisk5     0.0     0.0     0 0.0 0.0 lookup Netw read write kB/s
cd0        0.0     0.0     0 0.0 0.0 read tr0 0.0 0.2
           0.0 0.0 write lo0 0.0 0.0
           0.0 0.0 other

```

The starting screen of **monitor** shows, to a great extent, the same information as **topas**. We will not go into details for the displayed parameters.

The **monitor** command has some nice features, like the alternative screen. If you press the “a” key while **monitor** is running, you will see a different screen, in which the main part of the screen is filled with a list of processes that are sorted according to their CPU usage. If you enable the *SMP multiprocessor cpuinfo* option by pressing the “s” key, you will see a screen that is similar to the screen shown in Example 11-15.

Example 11-15 Alternative screen of monitor with SMP multiprocessor cpuinfo

```

Load averages: 0.18, 0.16, 0.21          tequila  Thu Aug 16 21:19:47 2001
Cpu states: 0.1% user 0.1% system 0.0% wait 99.7% idle
Logged on: 0 users 0 active 0 remote 00:00 sleep time
Real memory: 1191.5M procs 847.1M files 8.7M free 2047.4M total
Virtual memory: 1762.7M used 1213.3M free 2976.0M total
CPU USER KERN WAIT IDLE% PSW SYSCALL WRITE READ WRITEkb READkb
#0 1 1 0 98 4 205 0 12 0.20 9.54
#1 0 0 0 100 27 30 0 13 0.00 0.00
#2 0 0 0 100 4 0 0 0 0.00 0.00
#3 0 0 0 100 6 12 0 0 0.00 0.00
#4 0 0 0 100 2 3 0 1 0.00 0.17
#5 0 0 0 100 0 0 0 0 0.00 0.00
#6 0 0 0 100 0 0 0 0 0.00 0.00
#7 0 0 0 100 0 0 0 0 0.00 0.00
SUM 0 0 0 100 43 250 0 26 0.20 9.71

  PID USER  PRI NICE SIZE RES STAT TIME CPU% COMMAND
 2064 root  127 21 272k 8k run 22+15:18 12.5/97.1 Kernel (wait)
 2322 root  127 21 272k 8k run 22+15:04 12.5/97.1 Kernel (wait)
 1806 root  127 21 272k 8k run 22+14:49 12.5/97.0 Kernel (wait)
 1032 root  127 21 272k 8k run 22+04:23 12.5/95.2 Kernel (wait)
 1548 root  127 21 272k 8k run 22+13:37 12.4/96.8 Kernel (wait)
 774 root  127 21 272k 8k run 21+18:56 12.4/93.5 Kernel (wait)
 1290 root  127 21 272k 12k run 22+10:00 12.4/96.2 Kernel (wait)
 516 root  127 21 272k 8k run 21+06:13 12.3/91.2 Kernel (wait)
37946 root  60 0 668k 764k Frun 0:00 0.1/ 2.3 monitor
15494 root  60 0 2852k 424k slp 32:46 0.0/ 0.1 i41lmd
16784 slzadm 60 0 1121k 716k slp 10:01 0.0/ 0.1 rslgcoll
29732 slzadm 60 0 69M 56M slp 6:42:24 0.0/ 8.3 disp+work
5216 root  60 0 365k 36k Fslp 6:34:50 0.0/ 1.2 syncd
39480 slzadm 60 0 76M 64M slp 2:16:16 0.0/ 2.8 disp+work
35780 slzadm 61 0 1112k 1280k slp 2:08:35 0.0/ 1.3 saposcol
5428 slzadm 60 0 76M 62M slp 1:53:59 0.0/ 2.3 disp+work
9670 db2slz 60 0 14M 14M slp 1:41:57 0.0/ 2.1 db2sysc
27088 db2slz 60 0 13M 13M slp 1:28:50 0.0/ 1.8 db2sysc
2580 root  16 21 276k 8k slp 48:03 0.0/ 0.1 Kernel (lrud)
31418 db2slz 60 0 6395k 5992k slp 42:53 0.0/ 0.9 db2sysc

```

In this screen, the workload of each processor of the SMP system is shown individually. The eight processors in this example are numbered from #0 to #7. In our example, all eight processors are almost idle. Thus, you can see eight kernel wait processes in the process list.

11.5.8 The SAP R/3 transaction ST06

Using transaction ST06, you can monitor the operating system. There are two types of analysis available, a snapshot analysis, where the user can see the actual values, and an analysis based on the values of the previous hours. The following individual analysis can be performed:

- ▶ Snapshot analysis of:
 - CPU
 - Memory
 - Swap space
 - Disks
 - Network connections
 - File systems
- ▶ Integrated analysis of:
 - CPU
 - Memory
 - Swap space
 - Disks
 - Network connections
 - File systems
 - Operating system log
 - Hardware info

For more information on the transaction ST06, refer to the SAP R/3 documentation.

11.6 AIX performance hints

A performance degradation in your SAP R/3 system can have its cause in many components of the system. You can have a bottleneck in the SAP R/3 system, in the database, in the network connections, in the disk I/O subsystem, in the hardware of your server, or in the operating system.

In this section, we give some hints for analyzing the workload in AIX and discuss some common problems. Because the workload of each system is different, we cannot present a primer on how to get the best performance for an SAP R/3 system.

You can find more information on how to understand the performance of your AIX system in the redbook *AIX 5L Performance Tools Handbook*, SG24-6039 and in the *Performance Management Guide, AIX 5L Version 5.1*.

If you experience bad performance just after having installed SAP R/3 on your system, you should first check your settings of the memory management of SAP R/3 (refer to Section 11.3, “Models for SAP R/3 Extended Memory” on page 318) and of the AIX operating system (refer to Section 11.4, “AIX concepts” on page 329).

If a sudden performance degradation is seen, you should first try to find out the type of degradation and the date and time of its first occurrence. Perhaps you can also record a specific pattern of this performance degradation. You should check all error logs and the server journal (see Section 10.4.1, “Server journal” on page 301). Additionally, you should consult the journal where the changes in the network environment are endorsed and check for actions that may have lead to this situation. If you verify the configuration changes, you often find the change of some settings that is the reason for the performance degradation. After you have identified the cause of the problem, you should solve it; otherwise, you have to do some performance analysis.

To start the performance analysis, you should first check to see if the operating system has recorded some errors or warnings. For this purpose, you can use the **errpt** command.

If no problems are recorded in the error report, a good starting point for a performance analysis are the tools **topas** and **monitor**. Using these tools, you get an overview of the workload of the AIX system:

- ▶ You can check the saturation of the processors. A workload is called CPU-bound if the utilization of the CPUs is near 100 percent, adding the user and the system time. If the CPU spends a considerable percentage of time waiting for I/O, the execution of processes is blocked because of the I/O subsystem. For further investigation, you should use the **vmstat** command. You should also not forget to check the disk I/O pacing settings.
- ▶ You can check the status of the disk subsystem. If you see that many hdisks are busy for a long time, you should find out which processes are writing to the disks. First, you can check if the system is swapping. You can use the information on the amount of free memory and on the characteristics of paging to check for swapping. For further investigation, you should check the system using the **filemon** and **iostat** commands.

- ▶ You can check the activity of the network interfaces. Compare the given values for the amount of data transferred through each of the network interface with your experience of normal operations. You can check the use of the mbufs using the `netstat -m` command.
- ▶ You can check the processes that are consuming the CPU time. Check if the distribution of the processes correspond to your experience for that system.
- ▶ You can check how many processes are located in the run queue and in the wait queue and compare these values with your experience of normal operation.

In the following sections, we discuss some examples of observed work loads.

11.6.1 High value of disk I/O wait

If your system is spending most of the time in user and system mode and has a high value for disk I/O wait, then normally there is nothing wrong with the computing power or the amount of real memory of your system. The processors just have to wait for disk I/O requests to finish. Therefore, the bottleneck can be in the disk I/O subsystem.

If you use Oracle with asynchronous disk I/O enabled, you can first check the number of running aioservers. In case the number of running aioservers is equal to the value of maxservers, you can think of increasing the value of maxservers. Refer to Section 11.4.3, “Asynchronous disk I/O” on page 338 for a detailed discussion of asynchronous disk I/O.

11.6.2 High disk activity

If you see a huge activity on one of the hdisks, you should first check which file systems are located on that hdisk.

To check, for example, hdisk0, which normally is included in the rootvg, you can use the command:

```
lspv -l hdisk0
```

In Example 11-16, a sample output of this command can be seen.

Example 11-16 Output of the lspv -l hdisk0 command

hdisk0:				
LV NAME	LPs	PPs	DISTRIBUTION	MOUNT POINT
hd5	1	1	01..00..00..00..00	N/A
hd6	128	128	00..102..26..00..00	N/A
hd8	1	1	00..00..01..00..00	N/A
hd4	3	3	00..00..03..00..00	/

hd2	30	30	00..00..30..00..00	/usr
hd9var	1	1	00..00..01..00..00	/var
hd3	2	2	00..00..02..00..00	/tmp
hd1	1	1	00..00..01..00..00	/home

The next step would be to use the **filemon** command to check which of these file systems is the most active on the hdisk. You could find out, for example, that most of the disk activity comes from a paging space on that disk. If you have located the file system in which the main activity is located, you should try to find out which processes are writing to that file system. If you find out that these I/O operations cannot be avoided, you can think of redistributing some of the I/O operations to other hdisks.

11.6.3 Paging

Paging degrades the performance of your system, so you should try to avoid paging whenever you can.

The first thing to check if a system is paging are the programs that are using memory. You should use your experience of normal operations and check if there are any programs that behave strangely in using more memory than they normally do. If you configure, for example, Oracle to use a bigger SGA, then it might be that the system has to start paging to fulfill the changed memory needs of Oracle. It could also be that some programs run on the server that are normally not active on that system.

If you have just installed SAP R/3 and you experience paging, you should check that the memory sizing for SAP R/3 has been performed correctly. If you have a central system, the memory sizing has to include the memory requirements of all processes of SAP R/3 *and* the database.

On a database server, you should check to see if the configuration of the database and the used file system cache are set correctly. Refer to “File system caching” on page 335 for a detailed discussion about the file system caching.

The second thing to check is the pattern of the paging. The amount of paging that is tolerable depends on the hardware you use. For example, a number of 20 pages/s is no problem for an IBM @server pSeries model p680 with a balanced I/O subsystem, but it is a problem for an RS/6000 model 43P. You also should check if the paging only occurs if seldom used programs are running.

11.6.4 The run queue

The value for the numbers of processes located in the run queue shows the utilization of your CPUs. If there are, for example, 10 processes in the run queue and you have a 12 CPU machine, then this is perfect. But if you have only a 2 CPU machine, then your CPUs are saturated.

11.6.5 Long running batch jobs

A batch process uses only one work process. Because the work processes are not multithreaded, this process can only run on one processor. So adding CPUs will not improve the run time of a batch process.

It is important that you check what is the bottleneck that limits the run time.

Some possible bottlenecks could be:

- ▶ The usage of expensive SQL statements
- ▶ Paging on the application server where the batch job is running
- ▶ A limited disk I/O subsystem on the database server
- ▶ A limiting bandwidth or latency of the front-end network

Depending on the result of the analysis, you can try to optimize the ABAP program, customize the memory management on the application server, tune the disk I/O subsystem on the database server, or schedule the batch process so that it runs on the database server.

11.6.6 Related information

These publications are also relevant as further information sources:

- ▶ *AIX 5L Performance Tools Handbook*, SG24-6039
- ▶ *Performance Management Guide, AIX 5L Version 5.1, Second Edition (April 2001)*
- ▶ *RS/6000 SP System Performance Tuning Update*, SG24-5340
- ▶ *Database Performance on AIX in DB2 UDB and Oracle Environments*, SG24-5511
- ▶ *AIX Version 4.3 Differences Guide*, SG24-2014
- ▶ IBM SAP Marketing White Paper: 64 Bit SAP R/3 on the RS/6000, ISICC, Version 1.3, April 2001

- ▶ Configuring and Tuning IBM Storage Systems In an Oracle Environment, IBM/Oracle International Competency Center, Version 1.0, published May 25, 1999

These Web sites are also relevant as further information sources:

- ▶ <http://service.sap.com/>



Part 3

Operation

SAP R/3 system copy

In a nutshell:

- ▶ Learn about the different options for creating copies of a client or a whole SAP R/3 system.
- ▶ Determine the business requirements and the technical boundaries for the SAP R/3 system copy.
- ▶ Choose an appropriate method for the SAP R/3 system copy.
- ▶ Perform all required technical actions after the SAP R/3 system copy.

Performing an SAP R/3 system copy is a task that an SAP R/3 system administrator has to fulfill several times at different project stages. This chapter describes the reasons and supported methods for an SAP R/3 system copy.

A list of necessary actions is provided to help you deliver a new SAP R/3 system in a consistent state. It includes coverage of setting up the three layers (operating system, database, and SAP R/3 system) to ensure that the copied SAP R/3 system has the same behavior as the original SAP R/3 system. Special requirements resulting from duplicate systems in the same landscape or from overwriting an SAP R/3 development system are taken into account.

The information in this chapter reflects the authoring team's experiences with SAP R/3 Release 4.6 (on which this chapter is based), and it is a supplement to existing SAP R/3 documents summarized in the reference section.

This chapter covers the highlighted area in Figure 12-1, which is derived from the SAP R/3 infrastructure model that is used throughout the book.

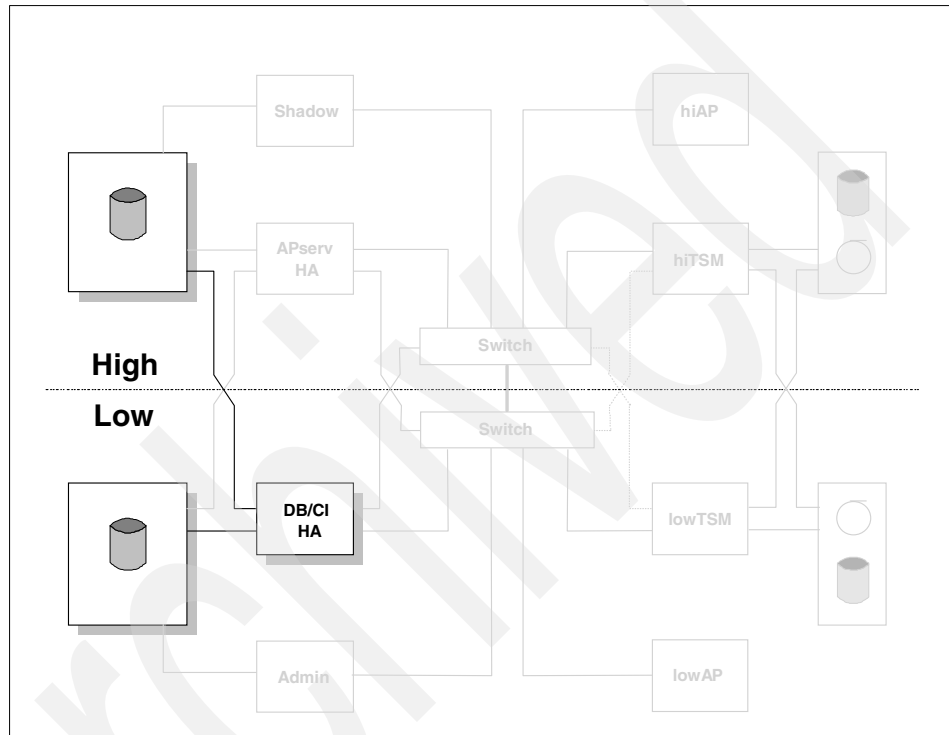


Figure 12-1 System copy

12.1 Introduction

An SAP R/3 system landscape is never static over a long period of time. Upgrades, projects to implement new functionality, corporate mergers, or consolidation of IT infrastructure can influence the requirements for an SAP R/3 landscape. These changes can lead to the need to copy an existing system, which is described in the next sections together with the possible methods.

12.1.1 Reasons for copying an SAP R/3 system

During the life cycle of an SAP R/3 system landscape, you will often receive the requirement/request to generate an SAP R/3 system based on an existing one. The following activities are generally the reasons to perform an SAP R/3 system copy:

- ▶ Initialize new SAP R/3 environments
- ▶ Build a demo, training, or test system, for example, for verification of SAP R/3 upgrades
- ▶ Refresh quality assurance systems with current productive data
- ▶ Create special systems for reporting or data mining on a regular basis
- ▶ Synchronize and consolidate systems in an SAP R/3 landscape
- ▶ Move the SAP R/3 system to new hardware after increased system requirements or workload
- ▶ Rename SAP R/3 SID or host name due to company related reasons

The tight interconnection between SAP R/3 customizing and application data makes it hard to copy only parts of an SAP R/3 system in a consistent manner. Because of the available time frame, the not completely documented history of the system, and other conditions the SAP R/3 system copy is often the most applicable procedure.

A system copy should only be done by a person with experience in copying systems and with knowledge of the operating system, the database, and the ABAP Dictionary, for example, a Certified Basis Consultant.

12.1.2 Terminology

The architecture of an SAP R/3 system landscape and the related technical processes are very complex. People from different organizations use different terms to describe these processes, but often they do not have the same understanding for the same term. For a better understanding of this chapter and a better communication between all involved departments, we will define the following terms:

System copy

The whole process of duplicating and adjusting an SAP R/3 system is summarized herein. Apart from the database, this includes aspects such as creating users at operating system level and adaptations to SAP R/3 kernel and profiles. SAP differentiates between homogeneous system copy and heterogeneous system copy.

Homogeneous system copy	This means that the underlying operating system and database will stay unchanged between source and target system.
Heterogeneous system copy	This is also called OS/DB migration, because the creation of the new system is associated with a change of the operating system and/or the database.
Database copy	This is the database-dependent part of the SAP R/3 system copy. It comprises the whole underlying database, not only parts of it.
Source system	This is the SAP R/3 system containing the original database used as the input for the database copy. The collectivity of the whole customized SAP R/3 system, including all files, interfaces, system settings, and others, represents the source system.
Target system	This is the SAP R/3 system to which the database copy is to be applied. The target system can be a new or an existing SAP R/3 system.
Placeholders	Placeholders such as <SID>, <sid>adm, or <instance_id> are used in the same way as in the SAP R/3 system installation documentation, and must be replaced with the values valid for your site.

Heterogeneous system copy is a more complex process. You have to ensure the approval for the new environment by SAP and you need a Basis Consultant with special certification for OS/DB migration. SAP supplies an OS/DB migration service, including special migration tools to support this transition. This is necessary to minimize the risk with regards to the availability and performance, especially for productive environments. We do not describe heterogeneous system copy, as it is outside the scope of this redbook. For further information, refer to SAP Note 82478.

12.2 Methods for SAP R/3 system copy

In general you have different possibilities to copy data between two SAP R/3 systems:

- ▶ SAP R/3 Transport system

The SAP R/3 Transport system allows you to transport client-dependent and client-independent data to other systems in a defined environment. This concept can be used to set up a new SAP R/3 system landscape right from the start and to keep the repository and customizing synchronized in this landscape. However, SAP R/3 transports are not suitable for moving a huge amount of business application data between the systems.

- ▶ SAP R/3 client copies

The SAP R/3 functions for client copy or client transport are not designed to copy clients of a size of several GB. The remote copy function cannot transport all the system settings and data to the target system and is not supported by SAP as a system copy method for productive clients (SAP Note 96866 and 89188). Apart from experiencing performance problems, you would, for example, no longer be able to access change documents and archived data. Furthermore, the client transport does not consider the client independent repository changes; additional transports are necessary. For further information, refer to the online documentation of client copy and to the following SAP Notes: 19574, 24853, 67205, and 70547.

- ▶ Batch output and batch input sessions

If your production system (being the source system) is too large to be copied within a reasonable amount of time, or if your hardware resources are not sufficient, consider transferring a subset of the data to the target system via batch output/input. This is also advisable if the source system contains confidential data that should not be stored in a test system. The drawback of this technique is the high programming and test effort. There are no prepared and maintained tools available, and adjustments are probably necessary with every SAP R/3 upgrade. SAP assumes no liability for the results and does not give support for this procedure!

- ▶ SAP R/3 system copy

SAP supports different methods to perform an SAP R/3 system copy:

- R3load procedure with SAP R/3 installation or migration tools
- Database dependent backup and restore procedures
- Database dependent export and import procedures

The following sections give a short overview of the different methods (described also in SAP Note 89188) and their advantages and disadvantages.

Attention: Make sure that the option you have selected for your data transfer fulfills your requirements and is approved by SAP!

12.2.1 R3load procedure

The utility R3load enables you to export and import data in an SAP defined, database independent format. This utility is used by the SAP R/3 installation tool *R3SETUP* (as of Release 3.1I_SR1) to perform system copies. The OS/DB migration process is based on these tools along with customized control files for R3SETUP delivered by SAP.

For very large databases, and to minimize downtime, an incremental migration procedure is under development and available for some database platforms (see SAP Note 353558).

Advantages:

- ▶ Export is in a platform/database independent format.
- ▶ Several reloads with exported data possible.
- ▶ Automatic distribution to parallel load processes.
- ▶ Implicit database reorganization of target database.

Disadvantages:

- ▶ Additional disk space for the unloaded data necessary
- ▶ Additional space necessary for sort during unload (up to several GB)
- ▶ Additional effort and downtime for unloading the source database
- ▶ Higher time exposure for load than for the restore procedure (see Section 12.2.2, “Backup and restore procedures” on page 370)

12.2.2 Backup and restore procedures

You can use backups created on the source system for restoring the database on the target system. The restore procedures are database dependent and can be installation specific. You have to provide an environment on the target system that enables you to restore backup data of another system. After the restore, you have to perform a change of the database name (SID) and some other adjustments. Section 12.5, “Performing an SAP R/3 database copy” on page 378 and Section 12.6, “Subsequent technical actions” on page 387 discuss this procedure in more detail.

Advantages:

- ▶ Short or no additional downtime of the source system
- ▶ Use of regular backups possible
- ▶ Several repeats for restore possible
- ▶ Very fast

Disadvantages:

- ▶ No implicit database reorganization of target database
- ▶ Migration of operating system not possible

There are special occurrences of the mentioned method:

- ▶ ESS FlashCopy

The ESS FlashCopy functionality can be used to create a system copy, but will not be covered in this chapter. More information can be found in Chapter 9, “Shadow database” on page 259, or in the redbook *R/3 Data Management Techniques Using Tivoli Storage Manager*, SG24-5743.

- ▶ Copy with operating system commands

Another possibility to copy a whole system or parts of it are, in general, the operating system commands, such as **tar**, **cpio**, **backup**, **restore**, or **rcp**. Apart from a longer downtime of the source system, you have to preconceive the effects of these commands to sparse files and the limitation by maximum file sizes. After errors, you have normally to repeat the whole procedure.

- ▶ Procedures to change <SID>, <instance_id>, or host name

During the consolidation of different SAP R/3 systems, it can be necessary to change “only” the <SID>, <instance_id>, or host name for a system to achieve unique system characteristics. The necessary adjustments are a subset of activities covered in Section 12.6, “Subsequent technical actions” on page 387.

12.2.3 Export and import procedures

The export and import procedure was integrated in the SAP R/3 installation program *R3INST* up to SAP R/3 Release 3.1, based on the Oracle tools *exp* and *imp*. Similar solutions or tools were available for other databases as well, such as *xload* for SAPDB (Adabas D). However, these procedures are no longer recommended and officially supported by SAP for SAP R/3 4.x releases; the *R3load* procedure should be used instead.

12.3 Organizational preparations

The process of SAP R/3 system copy is not only a technical process. The request for copying an SAP R/3 system is usually not due to technical reasons, but is part of a development project or other business related alterations. Therefore, it is necessary to determine the real requirements and the conditions under which the system copy can take place.

Pitfall ahead!

We recommend defining a project with all involved departments for planning this purpose (see Section 2.5.1, “Roles in a productive environment” on page 21).

Keep the following considerations in mind to prevent user dissatisfaction, inadequate performance, and system availability:

- ▶ Determine the required data quality of the target system
- ▶ Determine the maximum allowed downtime of source and target system
- ▶ Choose the right time for the system copy as a conclusion of the first two points (for example, month-end closing in the source system)
- ▶ Select and verify hardware resources for the target system
- ▶ Ensure the availability of resources and (certified) support for implementation and verification
- ▶ Review the installation specific inbound and outbound interfaces of the source system
- ▶ Specify the SAP R/3 SID, DB SID, and instance number for the target system, if it is a new system
- ▶ Select the source system and the copy method
- ▶ Examine your SAP contract if you plan to install an additional system

General requirements and restrictions are described in the SAP manuals.

12.4 Technical preparations

The technical procedure for an SAP R/3 system copy is described in detail in the guides *R/3 Homogeneous System Copy* and *R/3 Heterogeneous System Copy* of the respective update level. They can be downloaded from the SAP Service Marketplace at <http://service.sap.com/instguides>. This section describes the necessary preparation activities in a more condensed format.

To perform an SAP R/3 system copy, the versions of the SAP R/3 system and underlying database and operating system must be the same on the target and source system. Existing SAP R/3 systems often reached their current SAP R/3 release level by upgrades. If you plan to use the R3load procedure, you have to order the right version and platform combination of the SAP R/3 installation kit. The copy tools are located on the SAP Kernel CD-ROM of the installation kit (the SAP R/3 migration tools are only necessary for OS/DB migrations). Check also the availability of installation software and procedures for additional components that are necessary to install a consistent target system, such as specific printing solutions.

To use the copied SAP R/3 system, a new SAP R/3 license key for the target system is required. The license key of the source system is not valid for this system.

Bright idea!

You have, however, the possibility to provide the license key for the target system with the database of the source system. The procedure for providing the license key with the source system database is as follows:

1. Execute the **saplicense -get** command on the target system to get the so-called customer or hardware key.
2. Order a new license key via SAPNet with this customer key and the relevant installation number.
3. Insert the new license into the license table of the source database with the **saplicense -install** command as user <sid>adm.
4. All known licenses can be listed with the **saplicense -show** command.

This procedure is recommended for regular copies of the source system. The SAP R/3 license key is available to the target system immediately after the database copy. SAP Note 174911 contains additional information about customer key generation and temporary license generation.

12.4.1 Preparing the source system

In order to create a consistent copy of the database, it is necessary to prepare the source system accordingly. The importance of these steps depends on the purpose of your target system. For example, some of these are very important if you migrate your production system, but less important if you create only a demo or training system.

For better readability, we address the SAP R/3 transactions directly with their transaction code, and we do not describe the selected menu entries. The word *transaction* is abbreviated to *TX*. All of the following actions can be executed by an SAP R/3 user with sufficient authorizations.

The following list contains preparatory actions on SAP R/3 system level:

TX SM13

Display update records

To generate a complete list, you have to purge the fields Client, User, and From date in the selection screen. If canceled update records exist, they must be processed again or deleted. If pending records exist, wait for the processing to finish.

Check whether this action was successful using transaction SE16 for table VBDATA, to verify that this table is empty. If you find inconsistencies between TX SM13 and the contents of table VBDATA you can reorg the table with report RSM13002 and parameter REORG set to 'X' (SAP Note 67014).

TX SM35

Batch Input Monitoring

Process all outstanding batch input sessions and reorganize them and the log file afterwards with ABAP report RSBDCREO. Depending on your SAP R/3 release, it may be necessary to export remaining sessions on the source system for later reimport into the target system.

TX ICNV

Incremental table conversion

One prerequisite for an export is that no incremental conversion is in progress. If tables are in the state For conversion or in the state Done, delete the entries from the list. For any other state, you have to finish the incremental conversion; otherwise, you have to defer the system copy. This is in the first place valid for the R3load procedure.

TX SE14

Utilities for dictionary tables

Check for open conversions of all types by selecting **DB requests -> (all menu entries)**. If all resulting lists are empty, you can delete all invalid temporary tables (QCM tables) from the list generated by **Extras -> Invalid temp. tables**. Please refer also to SAP Note 9385.

TX DB02

Analyze tables and indexes

These steps are only valid for the R3load procedure. Check the consistency of database dictionary and ABAP Dictionary by selecting **Goto -> Check -> Installation ->**

Database <-> ABAP Dictionary. There should be no tables in the list “Unknown objects in ABAP Dictionary - DB Tables” or only uncritical tables, which can easily be reimported by using transport requests. Otherwise, you will lose data.

You may also check the consistency between ABAP Dictionary and the nametab. Execute the report RUTCHKDD with variant SAP&CHKALL_NT.

TX SE03

Transport organizer tools

Select **Find requests** and **Execute** to search for modifiable requests and tasks. Release all tasks, repairs, and transports if the system that will be moved is your development system, and the SID will be changed (refer to Section 12.7, “Special treatment for the development system” on page 402 for further information).

TX SM37

Job selection

SAP recommends you switch all released jobs to status Scheduled to prevent the target system from running jobs in an invalid environment. Do not forget to include jobs that are event triggered when generating your list! This also applies to the jobs which must run periodically (see SAP Note 16083).

Often, there are hundreds of jobs in your production system. In this case, you can execute the ABAP report BTCTRNS1 (see SAP Note 37425) to change the status of all released jobs (except RDDIMPDP jobs). You have to release these jobs after the export or backup of the source system database with the ABAP report BTCTRNS2.

Another possibility, described in Section 12.6, “Subsequent technical actions” on page 387, is to prevent the target system from starting SAP R/3 background jobs directly after the system copy by reducing the number of SAP R/3 background work processes to 0.

Attention: These procedures demand great care when cleaning up the target system! Otherwise, you could accidentally damage your source system!

The following additional prerequisites have to be fulfilled to enable a migration of the source system:

- ▶ No PREPARE of an SAP R/3 Release Upgrade is performed. You have to finish the whole upgrade procedure before executing a system copy.
- ▶ If you have archived data in the source system, you must make this data accessible in the target system.

Bright idea!

To reduce the time required to unload and load the database, minimize the amount of data in the source system. You can reach this by deleting unnecessary data (spool requests, ABAP dumps, test clients, and so on) and by archiving data. This can be effective when using the R3load method.

Some preparations for application consistency checks are proposed in the guide *R/3 Homogeneous System Copy*. Generate an output from the transactions presented in the guide for comparing with the results in the target system.

12.4.2 Preparing the target system

The necessary activities on the target system depend on the current system status. If you create the system on a new server, you have to establish the SAP R/3 environment by performing the SAP R/3 installation procedure with R3SETUP or by copying and adjusting all non-database files from the source system.

The usage of the R3SETUP installation tool, together with the appropriate control files, such as DBR3CP.R3S, is described in the relevant SAP manuals. Using R3SETUP with this control file supports you in most of the preparation tasks and performs a lot of the subsequent technical actions described in Section 12.6.2, “Actions on database level” on page 390.

The second alternative requires a deep understanding of the SAP R/3 system architecture on your platform. This is not covered in detail in this redbook, as the files and post installation steps are subject to change with every new SAP R/3 release. Therefore, SAP does not support you in performing this alternative.

Another starting point is a target system holding already a functioning database. In this case, you perform only a database copy. To prepare a consistent environment for a database copy, you have to delete some data on the target system.

Pitfall ahead!

Depending on the customization of your existing target system, it can be useful/necessary to export parts of the database for later import, for example, SAP R/3 printer definitions or existing users with their authorizations.

If you restore the database to a system that was holding another database before, you have to stop all active SAP R/3 and database processes and completely clean up the following directories in a default SAP R/3 system installation before performing the database copy:

- ▶ /sapmnt/<SID>/global
- ▶ /usr/sap/<SID>/DVEBMGS<instance_id>/data
- ▶ /usr/sap/<SID>/DVEBMGS<instance_id>/log
- ▶ /usr/sap/<SID>/DVEBMGS<instance_id>/work
- ▶ Database related files - DB2 databases
 - Logon as user db2<sid> to delete the databases and according files
 - \$> db2start
 - \$> db2 drop database <SID>
 - \$> db2 drop database ADM<SID> (If available)
 - /db2/<SID>/db2dump
 - /db2/<SID>/saparch
 - /db2/<SID>/sapdatat
 - /db2/<SID>/saprest
- ▶ Database related files - Oracle databases
 - /oracle/<SID>/mirrlog*
 - /oracle/<SID>/origlog*
 - /oracle/<SID>/saparch
 - /oracle/<SID>/sapbackup
 - /oracle/<SID>/sapcheck
 - /oracle/<SID>/sapdata*
 - /oracle/<SID>/sapreorg
 - /oracle/<SID>/saptrace

Starting with the new releases, such as SAP R/3 4.6C_SR2 or higher, the SAP R/3 SID and the database SID can differ. Current installations use the same name for both. Therefore, we refer to the abbreviation SID only.

Compare the available file systems with the source system. Create and mount additional file systems or increase their size if necessary.

If you want to change your SAP R/3 SID only, create or rename groups and users according to SAP R/3 requirements. User IDs, group IDs, and all characteristics should be synchronized between all systems of the SAP R/3 landscape (see Chapter 10, “Hints and tips” on page 275 for more information).

Additional preparations depend on your selected copy method.

12.5 Performing an SAP R/3 database copy

For a better understanding of the following sections, we assume the SAP R/3 system landscape shown in Figure 12-2.

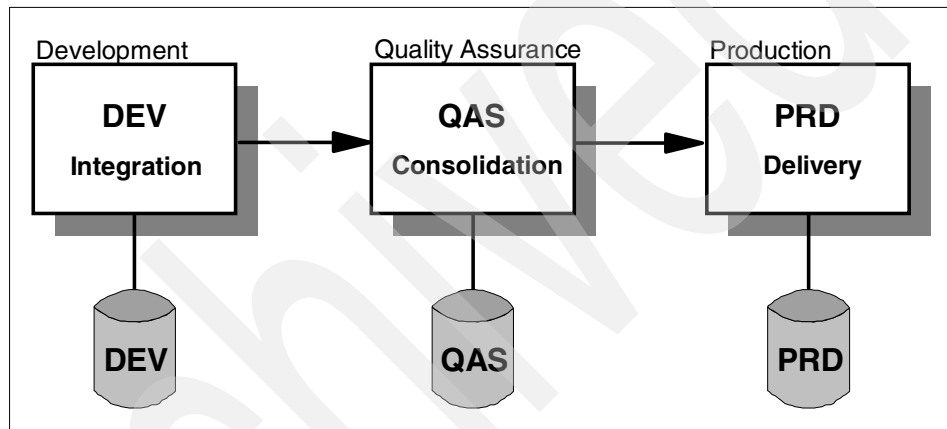


Figure 12-2 Sample SAP R/3 system landscape

- ▶ SAP R/3 system DEV - Development system and TMS domain controller.
- ▶ SAP R/3 system QAS - Quality assurance system (on server *target*).
- ▶ SAP R/3 system PRD - Production system (on server *source*).
- ▶ The <SID>s, such as in user name qasadm, are treated accordingly.
- ▶ Software versions:
 - AIX 4.3.3
 - DB2 UDB 6.1
 - Oracle 8.1.7 - 64bit (Oracle tablespaces on JFS)
 - SAP R/3 4.6C with SAP kernel 4.6D
- ▶ Backup infrastructure (see Chapter 7, “Backup and recovery” on page 185 for more information).

For the description of the database copy process we assume that the database is copied onto a prepared system (see Section 12.4.2, “Preparing the target system” on page 376) with an SAP R/3 instance and the database software installed. In our example, we want to update the quality assurance system QAS with current production data of the PRD database.

We cover only the backup and restore method for Oracle and DB2 databases using TSM. The technical approach for the SAP R/3 migration procedure and tools is described in detail in the SAP manuals *R/3 Homogeneous System Copy* and *R/3 Heterogeneous System Copy* of the respective update level. They can be downloaded from the SAP Service Marketplace at:

<http://service.sap.com/instguides>

All commands in this section are examples for the selected environment. If a user is not explicitly named with the described command, the actions are done by user root.

12.5.1 Example of backup and restore procedure of a DB2 database

In this chapter, we describe the necessary steps to create a consistent copy of the DB2 database on your target system using database backups generated with Tivoli Storage Manager (TSM) and the redirected restore function of DB2. We provide you with a verified sample procedure for renaming the database system to the new DB2 database name.

Actions on the source system

In addition to the preparations described in Section 12.4.1, “Preparing the source system” on page 373, you have to back up the database at the defined point in time. You can also use one of the regular backups, depending on your defined requirements. Perform the backup with the SAP R/3 utility **brdb6brt** (see SAP Note 326691) and the following options to create a backup of database PRD and a DB2 Command Line Processor (CLP) script to restore the database:

```
su - db2prd
db2start
brdb6brt -s PRD -bm BOTH -bpt ADSM 1 (for offline backups)
brdb6brt -s PRD -bm BOTH -bpt ADSM 1 -ol (for online backups)
```

Using an online backup requires more attention to transaction log files and is not covered in this chapter.

The CLP script to restore the database is named PRD.scr in our case and will be created in the local directory. You can redirect it to another directory with the option **-ip**. When you specify the option **-es**, comment lines are omitted in the script and makes it easier to read.

Actions on the target system

Perform the following steps on your target system:

1. Prepare the restore of the database
 - For the restore of the PRD database, you have to copy some configuration files for the TSM client from the source system to the target system (see also Chapter 7, “Backup and recovery” on page 185). The file `/usr/tivoli/tsm/client/ba/dsm.prddb.opt` (with correct ownership) is needed in our example.
2. Adjust the script file for redirected restore

Copy the script file `PRD.scr` from the source system to `/db2/QAS/QAS.scr` on the target system and change ownership to `db2qas:dbqasadm`. Substitute all occurrences of the string `PRD` to `QAS` in the script, except for the clauses `RESTORE DATABASE PRD` and `RESTORE DATABASE PRD CONTINUE`. If necessary, you can redefine the complete disk layout of your target database, including the location, size, and number of containers.

Remove the comment signs (`--`) in front of the following lines:

- `TAKEN AT` and the next line, if you do not want to use the last backup
- `NEWLOGPATH` and the next line, to change the log configuration
- `WITHOUT ROLLING FORWARD`, for offline backups

Additional information is available in SAP Note 122222.

Example 12-1 shows the `OAS.scr` script.

Example 12-1 Resulting DB2 CLP script QAS.scr

```
UPDATE COMMAND OPTIONS USING S ON Z ON QAS.out V ON;  
ECHO @./QAS.scr@;
```

```
RESTORE DATABASE PRD  
USE ADISM OPEN  
1  
SESSIONS  
TAKEN AT  
20010731001340  
INTO  
QAS  
NEWLOGPATH  
/db2/QAS/log_dir/  
WITH  
2  
BUFFERS BUFFER  
1024  
REDIRECT  
WITHOUT ROLLING FORWARD  
;
```

```

SET TABLESPACE CONTAINERS FOR 0 USING (
  FILE      /db2/QAS/sapdata1/SYSCATSPACE.container000 50000
  FILE      /db2/QAS/sapdata2/SYSCATSPACE.container001 50000
) ;

SET TABLESPACE CONTAINERS FOR 3 USING (
  PATH      /db2/QAS/sapdata4/TEMP4
) ;

...

SET TABLESPACE CONTAINERS FOR 27 USING (
  FILE      /db2/QAS/sapdata2/PSAPUSER11.container000 2500
) ;

RESTORE DATABASE PRD CONTINUE ;

```

3. A redirected restore of the database backup is created with **brdb6brt**.

Log on as user db2qas and adapt DSM_CONFIG and DSMI_CONFIG in the user environment using the following commands:

```

[target]> su - db2qas
db2qas > setenv DSMI_CONFIG /usr/tivoli/tsm/client/ba/dsm.prddb.opt
db2qas > setenv DSM_CONFIG /usr/tivoli/tsm/client/ba/dsm.prddb.opt

```

In this way, you can access the backups of the database PRD via the virtual node name of the source system. According to SAP Note 91976, create a database QAS and adjust the configuration regarding the TSM connection of the database with the following commands:

```

db2qas > db2 create database QAS
db2qas > db2 update db cfg for QAS using tsm_nodename sourceprddb
db2qas > db2 update db cfg for QAS using tsm_password sourcepw
db2qas > db2 update db cfg for QAS using tsm_mgmtclass sourcemc

```

(For DB2 UDB Version 6, the parameters were adsm_node, adsm_password, and adsm_mgmtclass.)

A quick alternative is to temporarily switch off the authentication of the TSM server (if possible).

To avoid an abort of the restore after tablespace creation, verify the TSM parameter COMMTIMEOUT, depending on your environment.

Start the redirected restore of database PRD into the new database QAS with the following commands:

```

db2qas > db2start
db2qas > db2 -tvf /db2/QAS/QAS.scr

```

The offline backup is identified by the timestamp in the script for the DB2 CLP.

4. Clean up the environment

Delete the configuration files that still refer to the old SID. The file `/usr/tivoli/tsm/client/ba/dsm.prddb.opt` was copied at the beginning of the procedure.

Redefine the configuration information regarding the TSM connection of the database with the following commands:

```
db2qas >db2 update db cfg for QAS using tsm_nodename targetqasdb
db2qas >db2 update db cfg for QAS using tsm_password targetpw
db2qas >db2 update db cfg for QAS using tsm_mgmtclass targetmc
```

(For DB2 UDB Version 6, the parameters were `adsm_node`, `adsm_password`, and `adsm_mgmtclass`.)

The necessary adjustments to use the restored database on your target server are described in Section 12.6.1, “Actions on operating system level” on page 387 and Section 12.6.2, “Actions on database level” on page 390. Some of the steps described can be performed by restarting R3SETUP after the database copy as user root, as described in the installation manual. As a prerequisite, set the environment variable LIBPATH to `/usr/sap/QAS/SYS/exe/run:/usr/lpp/db2_06_01/lib`.

Please be aware that R3SETUP will start your SAP R/3 system. Depending on your handling of released background processes (see Section 12.4.1, “Preparing the source system” on page 373), you have to prepare your environment before the SAP R/3 system is started for the first time with the newly copied database. If you perform regular database copies to an existing system, we recommend you save a copy of your `DBR3CP.R3S` before restarting R3SETUP. This copy is customized with your system specific parameters values and can be used as a template to perform only the subsequent technical actions covered by R3SETUP after the next database copy.

12.5.2 Example for backup/restore procedure of an Oracle database

In this chapter, we describe the necessary steps to create a consistent copy of the Oracle database on your target system using database backups generated with Tivoli Storage Manager (TSM) and Tivoli Data Protection (TDP) for R/3 on the source system. TDP for R/3 is based on the SAP R/3 backint interface and consists mainly of the executables `backint` and `backfm`. We provide a verified sample procedure for renaming the database system to the new Oracle SID.

Actions on the source system

There are no specific actions to be done on your source system except for the tasks described in Section 12.4.1, “Preparing the source system” on page 373. You have to back up the database at the defined point in time or you can perform a regular backup. (This depends on the defined requirements.) During the restore, you have to retrieve some configuration information from the source system. Further details are explained in the next section Actions on the target system.

Actions on the target system

Perform the following steps on your target system

1. Prepare the restore of the database.

For the restore of the PRD database, you have to copy some configuration files for the TSM client and TDP for R/3 from the source system to the target system (see Chapter 7, “Backup and recovery” on page 185). The following files (with correct ownership) are needed in our example:

- /usr/tivoli/tsm/client/ba/dsm.prddb.opt
- /oracle/PRD/817_64/dbs/initPRD.bki
- /oracle/PRD/817_64/dbs/initPRD.utl

Adapt the path information for BACKAGENT to the SID of the target system in the configuration file initPRD.utl.

The restore requires the same directory paths as the ones that existed on the source system at the time of the backup. Therefore, execute the following command to create a temporary symbolic link for PRD:

```
ln -s /oracle/QAS /oracle/PRD
```

The restore utilities are not able to create any subdirectories necessary for the data files. Data files for missing directories will be skipped during restore. You have to compare your systems and establish the same directory structure in your sapdata file systems on the target system as existing on the source system. Adapt the ownership afterwards with the following command:

```
chown -R oraqas:dba /oracle/QAS/sapdata*
```

2. Restore of the database backup created with TSM and TDP for R/3.

Log on as user oraqas and adapt DSM_CONFIG and DSMI_CONFIG in the user environment using the following commands:

```
[target]> su - oraqas
oraqas > setenv DSMI_CONFIG /usr/tivoli/tsm/client/ba/dsm.prddb.opt
oraqas > setenv DSM_CONFIG /usr/tivoli/tsm/client/ba/dsm.prddb.opt
```

The following command sets the TSM node password in the file initPRD.bki:

```
oraqas > backint -p /oracle/QAS/dbs/initPRD.utl -f password
```

Select the offline backup you want to restore in the backint file manager started with the following command:

```
oraqas > backfm -p /oracle/QAS/dbs/initPRD.utl
```

You can also use an online backup, but this requires the restore of all offline redo logs that are necessary for the database recovery to the current log sequence number. This procedure, of an standard database recovery, is not covered in this example.

3. Adapt the restored database files.

Change the ownership of the restored files from user oraprd to user oraqas with the following command:

```
find /oracle/QAS -user oraprd -exec chown oraqas {} \;
```

During a backup of the Oracle database, only one member of every online redo log group is saved. Therefore, you have to create the second member by copying the first one. Select a command depending on which kind of redo log was backed up and restored. The following command shows an example after a restore of /oracle/QAS/origlog*:

```
/oracle/QAS > cp -p origlogA/log_g11m1.dbf mirrlogA/log_g11m2.dbf
```

Execute this command, in a similar fashion, for all your members in the redo log groups.

4. Prepare the startup of database with SID PRD to force a database recovery.

For the database backup, the **sapdba** tool stops the database with **shutdown immediate**. For that reason, during the startup of the database, the online redo logs can be necessary (see also SAP Note 157127). However, these redo logs can only be used before renaming the database SID and starting the database with option RESETLOGS.

Perform the startup of the database with SID PRD using the configuration file initPRD.ora, which has to be copied from the source system to /oracle/QAS/817_64/dbs on the target system.

To start up the database successfully, you need all the copies of your Oracle control files at the locations specified in initPRD.ora. Copy the restored control file to the specified locations.

5. Start up the database with SID PRD and clear up online redo logs.

Log on as user oraqas, adapt the user environment to use SID PRD, and start the database by executing the following commands:

```
[target]> su - oraqas  
oraqas > setenv ORACLE_SID PRD
```

```

oraqas > setenv ORACLE_PSRV PRD
oraqas > svrmgrl
SVRMGR> connect internal
SVRMGR> startup mount;
SVRMGR> alter database open;

```

You have to switch the online redo logs as often as redo log groups exist to clean up all online redo logs and to ensure a consistent state of the database data files with the following command:

```
SVRMGR> alter system switch logfile;
```

6. Recreate Oracle control files with new SID.

Create a trace file for renaming the Oracle SID with the following Server Manager command and shutdown the database afterwards:

```

SVRMGR> alter database backup controlfile to trace;
SVRMGR> shutdown;
SVRMGR> exit;

```

The resulting trace file ora_<process_number>_prd.trc is located in /oracle/QAS/saptrace/usertrace. Copy this trace file with a new name, for example, createcntrlQAS.sql to /oracle/QAS. Edit the file and substitute all occurrences of PRD with QAS in the file and substitute REUSE with SET in the CREATE CONTROLFILE statement. You should get a script similar to the one shown in Example 12-2.

Example 12-2 Resulting SQL script createcntrlQAS.sql

```

STARTUP NOMOUNT
CREATE CONTROLFILE SET DATABASE "QAS" RESETLOGS ARCHIVELOG
    MAXLOGFILES 24
    MAXLOGMEMBERS 3
    MAXDATAFILES 1022
    MAXINSTANCES 50
    MAXLOGHISTORY 7941
LOGFILE
GROUP 11 (
    '/oracle/QAS/mirrlogA/log_g11m2.dbf',
    '/oracle/QAS/origlogA/log_g11m1.dbf'
) SIZE 50M,
GROUP 12 (
    '/oracle/QAS/mirrlogB/log_g12m2.dbf',
    '/oracle/QAS/origlogB/log_g12m1.dbf'
) SIZE 50M,
GROUP 13 (
    '/oracle/QAS/mirrlogA/log_g13m2.dbf',
    '/oracle/QAS/origlogA/log_g13m1.dbf'
) SIZE 50M,
GROUP 14 (
    '/oracle/QAS/mirrlogB/log_g14m2.dbf',

```

```

        '/oracle/QAS/origlogB/log_g14m1.dbf'
    ) SIZE 50M
DATAFILE
    '/oracle/QAS/sapdata1/btabd_1/btabd.data1',
    '/oracle/QAS/sapdata1/btabd_10/btabd.data10',
    '/oracle/QAS/sapdata1/btabd_11/btabd.data11',
    ...
    '/oracle/QAS/sapdata18/vbukd_2/vbukd.data2',
    '/oracle/QAS/sapdata18/vbukd_3/vbukd.data3'
;
# Recovery is required if any of the data files are restored backups,
# or if the last shutdown was not normal or immediate.
# RECOVER DATABASE
# All logs need archiving and a log switch is needed.
ALTER SYSTEM ARCHIVE LOG ALL;
# Database can now be opened normally.
ALTER DATABASE OPEN RESETLOGS;
# No tempfile entries found to add.

```

Execute the modified SQL script as user oraqa to recreate the control files with new Oracle SID using the following commands:

```

[target]> su - oraqa
oraqa > svrmgrl
SVRMGR> @/oracle/QAS/createcntrlQAS.sql
SVRMGR> exit

```

The step is finished with the successful start of the database.

7. Clean up the environment.

Delete the configuration files that still refer to the old SID. The following files were copied at the beginning of the procedure:

- /usr/tivoli/tsm/client/ba/dsm.prddb.opt
- /oracle/PRD/817_64/dbs/initPRD.bki
- /oracle/PRD/817_64/dbs/initPRD.ora
- /oracle/PRD/817_64/dbs/initPRD.utl

The necessary adjustments to use the restored database on your target server are described in Section 12.6.1, “Actions on operating system level” on page 387 and Section 12.6.2, “Actions on database level” on page 390. Some of the steps described there can be performed by restarting R3SETUP after the database copy as user root, as described in the installation manual. As a prerequisite, set the environment variable LIBPATH to /usr/sap/QAS/SYS/exe/run:/oracle/QAS/817_64/lib64 and change the passwords of the database users to the default. This procedure is covered in Section 12.6.2, “Actions on database level” on page 390.

Please be aware that R3SETUP will start your SAP R/3 system. Depending on your handling of released background processes (see Section 12.4.1, “Preparing the source system” on page 373), you have to prepare your environment before the SAP R/3 system is started for the first time with the newly copied database. If you perform regular database copies to an existing system, we recommend you save a copy of your DBR3CP.R3S before restarting R3SETUP. This copy is customized with your system specific parameters values and can be used as a template to perform only the subsequent technical actions covered by R3SETUP after the next database copy.

12.6 Subsequent technical actions

This section contains a summary of all published subsequent actions required on the target system, extended by the authoring team’s own experience. The actions required for your installation can be different, depending on your specific environment. If your system is installed on another platform, such as DB2/390 or Informix, additional actions may be necessary.

Due to company related reasons, it can become mandatory to rename the SAP R/3 SID, to change the SAP R/3 instance number, the host name of the database, or the application server. These tasks represent a subset of the activities described in the following sections.

12.6.1 Actions on operating system level

This section describes actions relevant to verifying the environment. If your target system is accurately installed, most of the following settings are already available as a result of system preparation (see Section 12.4.2, “Preparing the target system” on page 376 for more details). The fundamental principles for managing the environment are defined in Section 2.7, “Guidelines for operating an SAP R/3 environment” on page 30. The following activities can be relevant for your database server and your application server as well:

- ▶ Check the necessary operating system settings with the following commands:
 - Number of processes per user (`lsattr -El sys0 -a maxuproc`)
 - Settings for asynchronous IO (`lsattr -El aio0`)
 - Size of paging space (`lsp -a`)
 - Installed language environments (`locale -a`)
 - Installed LPPs and drivers for instance printers (`ls1pp -l`)
- ▶ Verify the settings and order for host name resolution:
 - Domain Name Service (DNS) and `/etc/hosts`

- Environment variable NSORDER and file /etc/netshvc.conf
- Verify other host name related communication settings:
- Adapt entries in the .rhosts files of different users or other authentication methods like Kerberos.
- Check the entry in the saprouter table for access to the SAP online support system (SAPNet) and restart the saprouter if applicable.
- Update the connect string for the SAP R/3 front ends of the end users, for example, by distributing a new SAPLOGON.INI configuration file.
- ▶ Verify and adapt the socket port definitions if you have changed <SID> or <instances_id>. The following service names can be affected:
 - sapdp<instance_id> and sapdp<instance_id>s
 - sapgw<instance_id> and sapgw<instance_id>s
 - sapms<SID>
 - sapdb2<SID> and sapdb2<SID>i for DB2 databases
 - Other entries for other platforms
- ▶ Establish all necessary AIX printer queues.
- ▶ Install and customize additional software components, for example, SAP ArchiveLink, if required.
- ▶ Include the system in the system management and support structures specific to your environment, for example, implement the backup procedures or installation specific scripts.
- ▶ If you have changed your SAP R/3 SID, create or rename groups and users according to SAP R/3 requirements. User IDs and group IDs and all characteristics should be synchronized between all systems of the SAP R/3 landscape (see Section 10.1.2, “Importance of synchronizing” on page 279 for more information).
- ▶ Maintain the crontab entries for all corresponding users.
- ▶ Depending on the cause of your copy, it can be useful to restore the contents of the following directories or parts of them from the source system to your target system:
 - /sapmnt/<SID>
 - /home/<sid>adm
 - /usr/sap/<SID>
 - Any installation specific file systems

Change the ownership of the restored files according to your SAP R/3 SID.

- ▶ The file system structure of an SAP R/3 system consists of a lot of symbolic links to map the same structure to different platforms. After a change of the SID of an existing system, you have to adapt these links. The following command can help you to find all affected files:

```
find / -type l | xargs -i ls -l {} | grep PRD
```

- ▶ The information about <SID>, <instance_id> and host name is part of the file name *or* file contents of different configuration files. You have to adjust the following files and directories:

- /sapmnt/<SID>/profile/*
- /usr/sap/<SID>/D*<instance_id>
- /home/<sid>adm/*.env*
- /home/<sid>adm/*

DB2 specific files:

- /var/db2/v61/default.env
- /var/db2/v61/profiles.reg

Oracle specific files:

- /oracle/<SID>/<rel>_<bit>/*.env*
- /oracle/<SID>/<rel>_<bit>/dbs/init<SID>.???
- /oracle/<SID>/<rel>_<bit>/network/admin/listener.ora
- /oracle/<SID>/<rel>_<bit>/network/admin/tnsnames.ora
- /etc/oratab

- ▶ The old files for the SAP roll area, the SAP paging area, and the SAP system log can be deleted, if the <instance_id> was changed. They are located in subdirectories data and log of the file system /usr/sap/<SID>/D*<instance_id>.
- ▶ Adjust the mount points of the SAP R/3 file systems after a <SID> change. If you transferred the SAP R/3 volume groups from one server to another using the **importvg** command, verify the sequence of the file systems in /etc/filesystems afterwards.
- ▶ Adapt the SAP R/3 profiles to represent the requirements of the target system, for example, default login client and table logging, resource allocation and shared memory definitions, and so on.
- ▶ Back up the whole environment after finishing all adjustments, including the adjustments on the database and SAP R/3 levels described in “Actions on database level” on page 390 and “Actions on SAP R/3 system level” on page 393.

12.6.2 Actions on database level

This section describes actions relevant to verify and adjust the environment after a database copy or change of the SID. The following activities are valid for all database platforms:

- ▶ Activate the logging mechanism of the database.
- ▶ Verify kernel extensions in /etc/inittab if required.

DB2 specific actions

Some activities have to be used in DB2 based environments only.

- ▶ Verify the settings of the TSM environment variables DSM_CONFIG and DSMI_CONFIG for the users db2<sid> and <sid>adm.
- ▶ Install the current version of the SAP DB2 admin tools using the sddb6ins installation program. Depending on your configuration, this step also creates the admin database ADM<SID>. SAP Note 410252 gives you more information.
- ▶ The passwords of the users <sid>adm, and sapr3 or sap<sid> are encrypted with the key defined in their environment variable DB2DB6EKEY and stored in the file /sapmnt/<SID>/global/dscdb6.conf. If you have copied the system, changed the SID, the host name, or the variable DB2DB6EKEY, then you have to update the entries in the file dscdb6.conf with the following commands:

```
su - <sid>adm
dscdb6up <sid>adm <password>
dscdb6up sapr3 <password> or dscdb6up sap<sid> <password>
```

Starting with the new SAP R/3 releases, such as SAP R/3 4.6C_SR2 or higher, the databases can be created with a new user named sap<sid> instead of sapr3 and a changed tablespace set. Please consider these changes if you have a newer start release.

Important: The environment variable DB2DB6EKEY is very important for password management and must be kept synchronized on both database and application servers.

Oracle specific actions

Some activities have to be used in Oracle based environments only:

- During the Oracle installation, the SID is linked with the software. After a change of the SID, you have to relink your Oracle software (SAP Note 97953) or establish a symbolic link with a command, such as the following, which is valid for our example:

```
ln -s /oracle/PRD /oracle/QAS
```

- If you want to use R3SETUP for finishing the installation after a database copy, it is necessary to reset the passwords of the database users back to default. You can carry out this activity with the following SQL commands:

```
su - ora<sid>
svrmgrl
SVRMGR > connect internal;
SVRMGR > alter user SAPR3 identified by sap;
SVRMGR > alter user SYSTEM identified by manager;
SVRMGR > alter user SYS identified by change_on_install;
```

- The OPS\$ user mechanism and a table storing the password of database user sapr3 (owner of all SAP R/3 tables in the database) is used for the database connect of the operating system user <sid>adm. In our sample configuration, you have to execute the following SQL commands on database level using, for example, Oracle Server Manager to enable the connect to user qasadm:

```
su - ora<sid>
svrmgrl
SVRMGR > connect internal;
```

Due to security reasons, you should delete the old user information regarding the source system to prevent unauthorized access to the database:

```
SVRMGR > drop table OPS$PRDADM.SAPUSER;
SVRMGR > drop user OPS$PRDADM cascade;
SVRMGR > create user OPS$QASADM identified externally default tablespace
psapuser1d temporary tablespace psaptemp profile default;
SVRMGR > grant connect, resource, dba to OPS$QASADM with admin option;
SVRMGR > create table OPS$QASADM.SAPUSER ( userid varchar2(256), passwd
varchar2(256) );
SVRMGR > insert into OPS$QASADM.SAPUSER values ('SAPR3', 'sap');
```

You can add additional OPS\$ users, for example, OPS\$DEVADM in our example, if the use of the testimport capability of the SAP R/3 transport program *tp* is required. See SAP Notes 175627, 29726, and 400241 for further information.

For the creation of the OPS\$ user and the password change for database users you can also use the SAP R/3 utilities CHDBPASS (see SAP Notes 319211 and 361641), **sapdba** and **brconnect** (see SAP Notes 150790 and 403704). Therefore, it can be necessary to execute the latest version of the sapdba_role.sql script. You will find more information on executing the script in SAP Note 134592.

Since SAP R/3 release 4.5B, the password of the database user sapr3 is stored encrypted in the table SAPUSER. Encrypted entries can only be maintained with the tools specified above.

Starting with the new SAP R/3 releases, such as SAP R/3 4.6C_SR2 or higher, the databases can be created with a new user named sap<sid> instead of sapr3 and a changed tablespace set. Please consider these changes if you have a newer start release.

Database independent actions

Executing the actions on the SAP R/3 system level, as described in “Actions on SAP R/3 system level” on page 393, often means that you delete information about the environment of the source system. To save yourself some effort, you can use the SQL command **delete from sapr3.<table>** to remove the data from the following affected tables:

DDLOG	Buffer synchronization between application servers (see SAP Note 25380)
TPFET	Table of profile parameters
TPFHT	Profile header, and administration data for profiles in DB
TLOCK	Lock table of the SAP R/3 Change and Transport System
MONI	SAP R/3 monitoring data
PAHI	History of system, database, and SAP R/3 parameters
OSMON	Operating system monitoring data
DBSNP	Database snapshots
SDBAH	Header Table for DBA Logs
SDBAD	Detail Table for DBA Logs
TSLE4	SAP R/3 instances and their operating systems
DBSTA*<db>	Database statistics (the extension <db> is database dependent - ORA, DB6, and so on)

Bright idea!

We recommend that you assemble all commands in an SQL script for additional database copies. For Oracle databases, you can also apply the **truncate table sapr3.<table>** command.

Additionally, schedule daily tasks, such as update statistics, for the database (if implemented with crontab); otherwise, it will be covered in the next section with SAP R/3 transaction DB13.

12.6.3 Actions on SAP R/3 system level

This section contains a list of SAP R/3 transactions you can use to adapt database information regarding the old SAP R/3 SID, instance number, or host name. Perform the steps in the list in the given order to avoid duplicate work.

The list is valid for different kinds of SAP R/3 system copies. Depending on your task (for example, system rebuild or only a host name change), you can skip some of these steps.

For better readability, we address the SAP R/3 transactions directly with their transaction code and we do not describe all the selected menu entries. The word *transaction* is abbreviated to *TX*. If no logon client and no user is specified, the action can be executed by an SAP R/3 user with sufficient authorizations. For more informations on using the specified transactions, see the SAP Library.

Before you start the copied SAP R/3 system for the first time, prepare the environment in the following manner:

1. Modify the SAP R/3 instance profiles to avoid the start of released background jobs and undesirable output:
 - `rdisp/wp_no_btc = 0`
 - `rdisp/wp_no_spo = 0`

The resulting number of work processes has to be different from the number defined in the instances of your operation modes to prevent the start of background work processes by a switch to another operation mode. This is only valid if the SID and host name will not be changed; otherwise, all instance definitions of the operation modes are unusable!

2. Start only the central instance of your SAP R/3 system.
3. Lock the system to prevent users from logging on to the system with the following commands:

```
su - <sid>adm
cd /usr/sap/trans/bin
tp locksys <SID>
```

The SAP R/3 user SAP* and DDIC are still able to log on. If you need to logon with another user during the activities of the next steps, you can temporarily open the system up with the command `tp unlocksys <SID>`.

Alternatively, lock all SAP R/3 users not necessary during this cleanup phase and unlock them afterwards. This can be performed with the function Mass changes of SAP TX SU01.

Start with the integration of the new system in your landscape.

Log on to your domain controller of the SAP Transport Management System (SAP R/3 system DEV in our SAP R/3 system landscape), execute the following transaction, and stay logged on:

TX STMS Maintain the SAP Transport Management System
Select **Overview -> System** and delete the newly created system from the list, if it was existent before.

Log on to client 000 of the created SAP R/3 system (system QAS in our SAP R/3 system landscape) and execute the following transactions:

TX SE06 Post-installation methods for Transport Organizer
Configure the Correction and Transport System (CTS) with the option "Database copy or migration". Answer the question "Do you want to reinstall the CTS?" with Yes. Select the source system of the database copy. Answer the questions "Delete TMS configuration?" and "Delete old version of transport routes?" with No, if the target system was already part of the same SAP R/3 transport domain as the source system.

If originals of SAP R/3 development objects were located in the source system, for example, the development system was the source system, the following additional question will occur: "Change originals from DEV to QAS?". Normally there is no change necessary, except for renaming the development system.

This transaction also releases all transport, repair, and customizing requests that have not been released in the source system.

You can verify the successful adjustment of CTS with the **tp connect QAS** command as user qasadm in directory /usr/sap/trans/bin. The return code should be 0.

TX STMS Maintain the SAP Transport Management System
A dialog box will appear. Insert a short description for the target system and press Enter. The system is now waiting for inclusion into the transport domain.

On DEV, that means the domain controller of the SAP Transport Management System executes the following transaction:

TX STMS Maintain the SAP Transport Management System

Select **Overview -> System**, approve the inclusion of the newly created system, and update the configuration, if necessary.

One part of the system configuration is the *Transport tool* record. Relevant changes to the tp parameters are reflected in the configuration files of the SAP Transport and Management System (TPPARAM and TP_DOMAIN_DEV.PFL in our example).

Distribute and activate the configuration with **Extras -> Distribute and activate configuration**.

Log on to the created QAS SAP R/3 system and execute the following transactions:

TX STMS Maintain the SAP Transport Management System

Select **Overview -> Transport routes**, and then **Configuration -> Adjust with controller**, to make the configuration consistent. When you re-enter transaction STMS afterwards, no problems should be reported.

TX SE16 Maintain entry of table INSTVERS

After you log on to SAP R/3, the system may display the following message: "Your system has not been installed correctly. (message 735)". In this case, change the last record that has value 0 in the field STATUS in table INSTVERS. Enter the new host name of the database server in the field DBHOSTNAME, as described in SAP Note 144978.

TX SM37 Job selection

Your approach depends on how you have changed the job's status during the preparation of the source system (see Section 12.4.1, "Preparing the source system" on page 373).

Check and adapt all existing SAP R/3 background jobs. We recommend that you switch all released jobs to status *scheduled* to prevent the system from running jobs in an invalid environment. Do not forget to consider event driven jobs and jobs released with start dates in the future!

Adapt and release only jobs needed in the target system. Check also for dedicated *target servers* and maintain print specifications of released jobs to prevent errors or misleading output.

After a change of the SAP R/3 SID, it may be necessary to adjust the paths of jobs calling an external command or program.

Verify jobs with status *active* or *ready* by selecting the job and selecting **Job -> Check status**.

TX VP01

Maintain the print parameters

Adapt the output devices to prevent misleading output from a test system generated by a database copy from a production system. This task is dependent on your environment and application specific.

TX SPAD

Spool administration

Adjust SAP R/3 printer definitions with **Utilities -> For output devices -> Assign server**. You can define or adapt virtual print servers for load balancing if requested by maintaining the list of spool servers by selecting **Configuration -> Spool server**.

To prevent misleading output after copying the production system, you can lock unnecessary printers. Verify the current settings by selecting **Configuration -> Check installation**.

If you exported the SPAD data during the preparation of the target system, you can import the data now.

TX SP12

Administration of Temporary Sequential objects (TemSe)

Relevant data is stored in files on the operating system level and in the SAP R/3 database. Detect and delete inconsistent entries by selecting **TemSe database -> Consistency check**. During this check, no background jobs should be active! See SAP Notes 16875 and 48400 for further information.

Log files of background jobs can only be accessed if you have restored the related files from the source system to /usr/sap/QAS/SYS/global (corresponds to /sapmnt/QAS/global), as mentioned in Section 12.6.1, "Actions on operating system level" on page 387.

Additionally, you have to establish a symbolic link with the

SID of the source system with the command `ln -s /usr/sap/QAS /usr/sap/PRD`.

Now you can restart the SAP R/3 system with the required number of spool and background work processes configured. Please restart *all* available instances.

- TX RZ12** RFC server group maintenance
- Delete the group of source systems and create a new group for RFC access, according to the naming conventions, and insert available instances.
- TX SMLG** Maintain assignments of instances to logon groups
- Delete the group of source systems and create a new dialog logon group, according to the naming conventions, and insert available instances.
- TX RZ10** Maintenance of profile parameters
- Delete old profiles if you have not deleted the table contents during the database activities (see Section 12.6.2, “Actions on database level” on page 390). Afterwards, you can import the profiles of all new instances by selecting **Utilities -> Import profiles -> Of active servers**.
- TX RZ04** Maintain operation modes and instance definitions
- You can easily insert all new instance definitions by selecting **Operation mode -> Maintain instances -> Operation mode view** and then selecting **Settings -> Based on current status -> New instances -> Generate**. Now you can create new operation modes if necessary and delete the previous ones.
- Check the consistency with profiles imported in the step before by selecting **Instance -> Consistency check**.
- TX SM63** Maintain operation mode timetable
- Adapt the timetable to the requested status if you have defined operation modes in the step before.
- TX SM13** Display update records
- Delete open and failed update records of the source system if you were not able to do so during the preparation phase. To generate a complete list, you have to purge the fields Client, User, and From date in the selection screen.

TX SM58

Asynchronous RFC Error Log

Depending on your environment, check and clean the transactional RFC. Select all users with a star and an adequate time frame. Select **Log file -> Reorganize** and specify the data you want to delete. For further information, refer to the SAP Library.

TX SM61

Maintain background control objects

Delete all invalid control objects. All control objects necessary for your environment should be activated. Missing entries for your server will be automatically created whenever necessary.

TX ST03

Performance database

Delete old content of the performance database if you have not deleted the relevant table contents during database activities (see Section 12.6.2, “Actions on database level” on page 390) by selecting **Goto -> Performance database -> Contents of database**. The rows of the resulting list can be cleaned up, for example, by server or by group (ID).

Afterwards, it is useful to adapt the reorganization parameters for the database by selecting **Goto -> Parameters -> Performance database**.

Check the access to actual statistic data and, if necessary, delete the statistic file according to SAP Note 6833 by selecting **Workload -> Reorganize -> Delete seq.stat.file**.

TX SM65

Analysis tool for background processing

Select **Goto -> Additional test** to perform an extensive test of the background processing environment in the new system. Select the appropriate check boxes in the selection screen to detect problems and inconsistencies. Clean up all inconsistent entries because they can prevent normal background processing. No jobs should be active at this time!

TX SM35

Batch Input Monitoring

Delete all information about batch input sessions of the source system.

If it is required to access the log file of the source system, you have to rename the copy of the source system /sapmnt/QAS/global to BItarget00

(BI<hostname><instance_id>). This also enables you to reorganize the batch input sessions and the log file afterwards with ABAP report RSBDCREO.

Depending on your release, it can be necessary to import remaining sessions exported from the source system.

TX SM69

Maintain external commands

Delete or adapt entries which are only valid, allowed, or available in the environment of the source system (for example, individually defined scripts or changes of paths).

TX SE03

Transport organizer tools

Select **Set System Change Option** and Execute (F8). If you want to allow or prevent changes to the SAP R/3 repository and cross-client customizing, then you can use this activity to adapt the *Global setting* and the settings for software components and name spaces according to the system concept. The *Global setting* specifies whether objects in the repository and in the cross-client customizing are modifiable or not. Additionally, each repository object is a member in a name space or name range, and in a software component.

TX SCC4

Client administration

In the client table, you can set whether changes to Repository and Customizing are allowed and recorded for each specific client.

Adapt the client role and which kind of changes are allowed and recorded in the specific clients. Some modification activities are prohibited for clients with role *Production*.

Adapt or delete the logical system (relevant for ALE, archiving, and IDOC processing) to prevent wrong communication links. This is only possible with a client role that is not *Production*.

TX SM54

TXCOM maintenance

Maintain the entries in the table according to the external connections that are permitted to the new system. Refer to the SAP Library for further information.

TX SM55

THOST maintenance

Verify the symbolic host names defined in this table to represent very long host names. Use this feature only if it

is absolutely necessary, in addition to the capabilities at operating system level.

TX SM59

Maintain RFC destinations

Verify and adapt RFC destinations to other SAP R/3 or non-SAP R/3 systems to enable or disable communication. As a rule, all RFC destinations to systems of the production environment should be disabled to prevent failures that are caused by the target system when processing incorrect data.

After a host name change, you have to also adjust the RFC destinations in other SAP R/3 systems of your SAP R/3 system landscape.

TX OSS1

Logon to SAPNet

Verify that the logon to SAPNet is possible if required. Update system data in SAPNet if the characteristics of the target system have changed.

TX SCOT

SAPconnect administration

Adapt the RFC destination for your fax and other defined external connections, for example, to prevent sending production output from a test system.

TX SU01

User maintenance (client dependent)

Lock or remove users that are not allowed to use this system according to the implemented authorization concept.

Depending on your administration concept, it is necessary to update information about passwords of default users (SAP*, DDIC, and so on), for example, in a spreadsheet, with values from the source system by the date of the database backup or export.

If you have exported the users, user profiles, and authorizations from the target system before database copy, you can import them now. Clean up the resulting inconsistencies in SAPOffice with SAP R/3 report RSSOUSCO.

TX DB13

DBA planning calendar

Schedule necessary database maintenance actions coordinated with other schedules.

TX SDCC	<p>Service Data Control Center</p> <p>Verify and update the scheduling of sessions for SAP EarlyWatch Alert (EWA) if applicable. See SAP Note 178631 for further information.</p>
TX SCC5	<p>Client deletion</p> <p>Delete unnecessary clients. Remove the relevant entries from client table T000; otherwise, the unauthorized logon to the system as user SAP* with password PASS (coded in SAP R/3 kernel) is possible!</p>
TX SA38	<p>Reports for reorganization jobs</p> <p>Execute the following reports in the background with suitable parameters (according to SAP Note 16083) to delete information only relevant to the source system:</p> <ul style="list-style-type: none"> ▶ Report RSBTCDEL and RSBTCPRIDEL Clean up information about old background jobs. ▶ Report RSPO0041 or RSPO1041 Clean up spool information (SAP Note 48400 and 41547). ▶ Report RSBDCREO Clean up information of batch input sessions (see TXSM35). ▶ Report RSSNAPDL Clean up information of old ABAP dumps.
TX SGEN	<p>Load generator</p> <p>Use this transaction to (re-)generate ABAP program loads to speed up first program execution. This is only necessary after a change of application server's operating system (including the change from 32-bit to 64-bit) or the usage of the R3load procedure. Schedule this action as a background job during the night because of high system load. In addition, SAP R/3 buffers will be destroyed, which can cause problems with dialog tasks at this time.</p>
TX SM02	<p>System messages</p> <p>Generate a system message to inform users about the system status.</p>
TX RZ20 / RZ21	<p>CCMS monitoring</p> <p>Adapt CCMS monitoring settings (for example, alert thresholds or central monitoring capabilities) if required.</p>

Additional actions may be necessary depending on your database platform and your specific environment, such as:

- Make archived data from the source system (data that does not reside in the database but was moved to a different storage location using SAP R/3 Archive Management) accessible to the target system. Adapt the file residence information in the target system. Refer to the SAP Library for additional help.
- Delete confidential or unnecessary application data from the system.

The following actions are suitable for checking the consistency of the target system (Refer also to Section 13.4, “SAP R/3” on page 422):

TX SICK	Installation check Perform an initial consistency check of the system.
TX SM51	List of SAP R/3 systems Perform a server check for all available instances.
TX SM21	System log analysis Check the system log (local and central) for relevant error messages.
TX DB02	Database tables and indexes Refresh the available information and check the consistency of the database.
TX SPAD	Spool administration Check the installation as described and verify output of your most important printers.

Pitfall ahead!

Depending on the importance of your target system, the staff of the user departments should test their most frequently used transactions. Some application consistency checks are proposed in the guide *R/3 Homogeneous System Copy*. Please check this guide also for complementary information delivered by SAP for subsequent releases.

12.7 Special treatment for the development system

Special considerations are necessary if you plan to copy your production system to your development system. Such a procedure does not conform to the strategy recommended by SAP to retain the development system. Only this system encloses all the history information of your implementation projects and modified repository objects.

The following important information might be lost:

- ▶ Version history of the repository objects
- ▶ Information about all transports generated in the development system
- ▶ Access keys for all registered developers and modified SAP R/3 objects
- ▶ Projects of the SAP Implementation Guide for R/3 Customizing (IMG) and all related notes
- ▶ Special settings of user characteristics

You have to perform special tasks to prevent the loss of this information.

Pitfall ahead!

We recommend you keep the old system as long as possible for comparisons, for example, until the next SAP R/3 upgrade.

12.7.1 Version history

The complete version history of all repository objects is part of the development system. Because this version history is particularly important for later SAP R/3 upgrades (modification adjustment), the following workaround can be used to export the complete version database from the old development system and import it again after the system copy to the newly created system.

The procedure is release dependent and described in SAP Note 130906. Do not forget to check the log file! Versions generated in the source system of the database copy are completely overwritten in the newly created system with this procedure.

Attention: Please note that no change request data is copied. References from the version database to change requests will only remain intact if the corresponding change request was previously imported in the source system of the database copy.

12.7.2 Repair and transport requests

Only repair and transport requests imported in the source system of the database copy are known to the new development database and can be used as a pattern for regeneration or for re-export, and can be accessed for references by the version database.

Bright idea!

To prevent the loss of actual modifications, all repair and transport requests should be released and transported to all systems, if possible. Open tasks currently under work in the development system can be exported from the old development system and afterwards imported in the newly generated development system only. The developer has to use the imported transport request as a pattern for creating the new transport request before proceeding with his work.

With this approach, you ensure the complete transport of all involved development objects when the task is finished. Defining a code freeze for the period of recreating the new development system is very useful to avoid inconsistencies in your SAP R/3 landscape.

The table E070L contains the last number used for repair or transport requests. After initializing the system with SAP R/3 transaction SE06, adjust the counter in this table to prevent an overwrite of existing repair and transport requests.

12.7.3 Registration of developers and objects

The SAP Software Change Registration (SSCR) is a procedure, valid from SAP R/3 Release 3.0A onwards, which registers all manual changes to SAP sources and SAP Dictionary objects. The following information is registered:

- ▶ Every development user who undertakes changes to objects in your SAP R/3 system or creates new objects
- ▶ Every changed object owned by SAP

Normally, the keys are requested and stored in the development system only. With the database copy you will lose all the keys. You have two possibilities:

- ▶ You can re-enter the developer and object keys when you need them. A list of requested keys can be generated in the SAPNet.
- ▶ Export the tables DEVACCESS and ADIRACCESS from the original development system and import the tables after the database copy into the new development system.

A prerequisite for this procedure is the equal SAP Installation Number for your source and your target system.

12.7.4 IMG projects

All generated projects of the SAP Implementation Guide for R/3 Customizing (IMG) and all related notes you have added to your projects are also part of the development system database and will be lost with the recreation. To import this information to your new development system, you can export transport requests for all necessary projects. Execute SAP transaction SPRO and select **Goto ->**

Project Management. Select your projects from the list and assign them to a transport request. You can receive more information about known problems from SAP Notes by referring to the ABAP report RSTRANSPROJECT (or RSTRAPRO for older releases).

12.7.5 SAP R/3 user

Normally, your development environment is a special environment in your SAP R/3 system landscape. Therefore, you have often users defined with other user profiles and authorizations (such as ABAP development) than, for example, in your production system. To preserve these settings of your old development system, you can export the relevant tables for later import by a SAP transport or client export with profile SAP_USER. This procedure is valid for all systems in general, not only for development systems.

12.8 Reference information

This section summarizes the sources where you can obtain additional information. The following Internet addresses are part of the SAP Service Marketplace, also known as SAPNet (formerly Online Service System - OSS). You can access the information only with a valid SAPNet user ID and password.

- ▶ <http://service.sap.com/instguides>

Here you can view and download all installation and migration guides, such as the manual *R/3 Homogeneous System Copy*.

- ▶ <http://service.sap.com/osdbmigration>

Here you find additional information about the available migration procedures and services if you have to change your platform.

- ▶ <http://service.sap.com/notes>

With this link you can access all the SAP Notes mentioned in this chapter and all release-specific notes according to the installation guides. You can search in the components BC-INS and BC-INS-MIGR3 for additional information regarding system copy and migration.

SAP also offers the course TABC92 *Operating System and Database Migration* for experienced Basis Consultants to become certified for Heterogeneous System Copies.

Additional reference information depends from the selected method, such as TSM or database specific documentation.

Daily tasks to prevent error situations

In a nutshell:

- ▶ Maintain your environment in a proactive way.
- ▶ Investigate the reason for every error and develop a plan to avoid a repeat situation.
- ▶ Perform extensive housekeeping to avoid bothersome objects.

This chapter describes monitoring tasks that can be used to detect and avoid errors in an SAP R/3 system landscape. The tasks should be performed regularly and can be divided into the following topics:

- ▶ Monitoring methods
- ▶ AIX operating system (including SP and HACMP specific tasks)
- ▶ Database system
- ▶ SAP R/3
- ▶ Backup
- ▶ Cleaning the log files

Each topic will be described in one of the following sections.

This chapter covers the highlighted area in Figure 13-1, which is derived from the SAP R/3 infrastructure model that is used throughout the book.

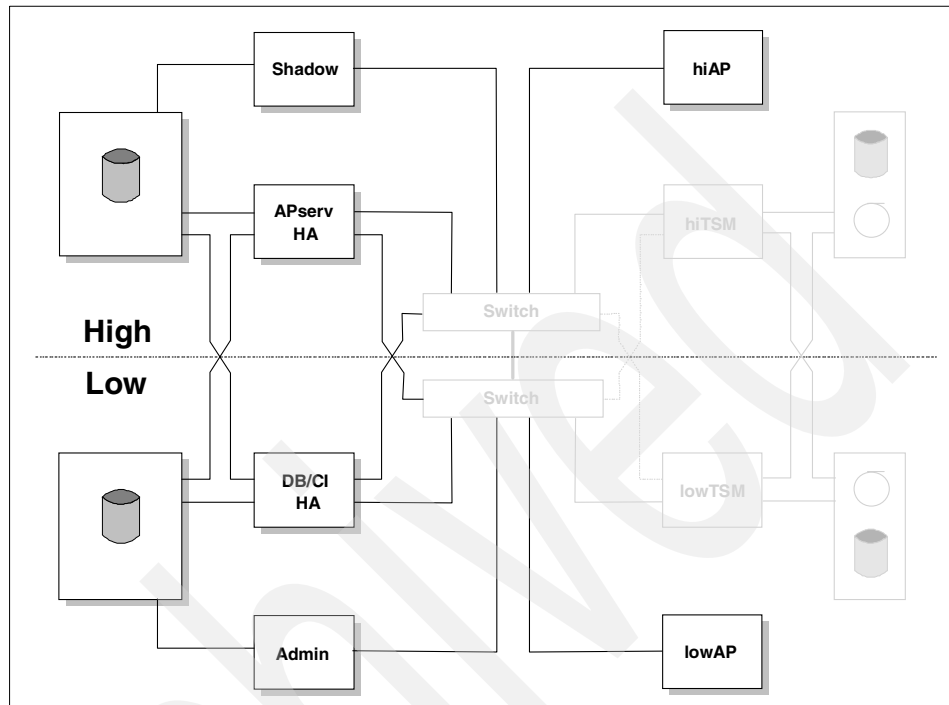


Figure 13-1 Daily tasks to prevent error situations

13.1 Monitoring methods

There is an ample variety of system management products from software vendors, such as Tivoli, BMC Software Inc., Computer Associates, and others. System management tools effectively assist system administrators with their daily work. It is not within the scope of this book to show how to design and set up a complete system management environment. The following sections describe the important checks which have to be performed and gives examples for automating recurrent actions.

Generally, there are two possibilities for how to take care of your systems. Either you check and monitor your systems regularly, which is *proactive*, or you start investigation when a failure has already occurred, which is *reactive*. Working in a reactive way implies system down time, as there are no regular tasks performed to avoid failures.

Bright idea!

Therefore, we strongly recommend that you monitor your systems in a proactive way. This approach supports you in avoiding system down time and helps to maintain a reliable infrastructure. In this chapter, we provide a collection of regular tasks that should be used to monitor your environment proactively.

13.2 AIX operating system

Daily checks need to be run on all AIX systems in the SAP R/3 environment, including application server, database server, TSM server, admin server, and other servers running an AIX operating system.

The following examples assume a configuration as described in Section 10.1.4, “Principles of the SP system” on page 280.

13.2.1 Checking the AIX error log

The AIX error log contains hardware and software error messages. Use the **errpt** command to show the error messages. For example, to show all messages starting on 17 July 2001 at 0:00 hours for all servers, use the following commands:

```
export WCOLL="/wcoll_all"
dsh 'errpt -s 1707000001'
```

Bright idea!

Persistent errors easily fill the error report and annoy system administrators with logging the same event innumerable times. Condensing the error report, sorted by identifiers, may help system administrators to keep an overview. The script in Example 13-1 can help you in summarizing the errors. The summary produced by the script shows the number of occurrences of each error identifier in the error report.

Example 13-1 Script to summarize AIX error report by identifiers

```
#!/bin/ksh

TEMPFILE="/tmp/$$.tmp"

/usr/bin/errpt |
/usr/bin/tee ${TEMPFILE} |
/usr/bin/sed "/^IDENTIFIER/d" |
/usr/bin/tr -s " " " " |
/usr/bin/cut -d" " -f1,5,6- |
/usr/bin/sort |
/usr/bin/uniq |
while read FIELD1 FIELD2 DESCRIPTION ; do
    /usr/bin/grep "${FIELD1}.${FIELD2}" ${TEMPFILE} |
    /usr/bin/echo "$( /usr/bin/wc -l ) x \c"
```

```

        /usr/bin/echo "${FIELD1} ${FIELD2} ${DESCRIPTION}"
done
        |
/usr/bin/sort -nr

/usr/bin/rm ${TEMPFILE}

```

Monitoring the error report each hour can be automated in the following way:

- ▶ Write the output of **errpt** to a temporary file each hour.
- ▶ Compare the file with the output written one hour ago.
- ▶ Then send the differences between the two to a mail address of your choice, generally to the system administrator.

13.2.2 Checking file systems

To ensure a smooth operation of your AIX systems, there should always be enough space available in all file systems. The following file systems are most likely to fill up fast:

- ▶ /
- ▶ /tmp
- ▶ /var (and other file systems with transactional data or log files)
- ▶ /usr/sap/<SID>
- ▶ Archive directories, such as:
 - /oracle/<SID>/saparch (using an Oracle DB)
 - /db2/SID/log_archive (using a DB2)

Check the percentage of used space and used inodes of all file systems on a daily basis. For example, you can use the following command to show all file systems that reach a saturation point of 95 percent in either space or inode usage:

```

/usr/bin/df -k | \
/usr/bin/awk '{if ($4 > 95 || $6 > 95) {print $1 " " $4 " " $6} }'

```

Pitfall ahead!

Depending on the method and the frequency with which the redo log files in the archive directories are archived, it may make sense to set up daemons to monitor the saturation of your archive directories. For example, archive the redo log files automatically when an archive directory reaches a saturation point of 80 percent.

The `/usr/sap/<SID>` directory contains, among other files, SAP R/3 trace files, log files, statistics and temporary sort files. Huge sorts or a high trace level can fill up the file system, leading to the termination of a process. Therefore, we suggest that `/usr/sap/<SID>` should have 2 GB of free space in addition to the space needed for your daily business.

13.2.3 Checking mirrors of logical volumes (LV)

Losing a mirror of an LV due to a problem is generally indicated by an entry in the AIX error report. If such an error message has not been noticed, there will be no further warnings. In this case, you might lose the fault- and disaster-tolerance. To detect if any of the systems are running without all their defined LV mirrors, the following check can be performed:

```
export WCOLL="/wcoll_all"
dsh 'lsvg | lsvg -il | grep stale'
```

13.2.4 Checking mailboxes

Some messages are neither directed to the error report nor to the system console. Such messages are, for example, standard outputs of **cron** commands, daemons, and other processes, which have no output device defined. Some outputs are sent as mails to a user's mail box.

In an SAP R/3 environment, the important users, to which mails could be sent, are `root`, `ora<sid>` or `db2<sid>`, and `<sid>adm`. Check these user's mail boxes with the **mail** command.

New mails are held in the file `/var/spool/mail/<user>`. Old mails, that have already been read, can be found in the user's home directory in the file `mbox`.

Bright idea!

To reduce the amount of mail sent to a certain user, there should always be an output device specified when using cron jobs, daemons, and other processes. The console device `/dev/console` is most commonly used as a standard out and standard error device for cron jobs and daemons.

If requested, the mails can be forwarded to any other user or mail address by creating a `.forward` file in the forwarding user's home directory. The file contains the mail address to which the mails shall be forwarded. For example, create a `.forward` file in `/oracle/<SID>` containing the administrators mail address. All mails from user `ora<sid>` will then be forwarded to the administrators mail box. To send mails over a gateway to other mail systems, such as Lotus Notes or MS Outlook, you have to configure and activate the **sendmail** program. For a complete description of sendmail consult, for example, *sendmail* by Bryan Costales, et al.

13.2.5 Checking printer queues

Printer queues may be frequently used by the SAP R/3 system. Therefore, they should be monitored regularly. To show the status of all printer queues, you can use the following command:

```
smitty qstatus
```

For various reasons, printer queues can occasionally become inactive. If this happens, they need to be enabled again to resume printing. This action can be automated with a little script, as shown in Example 13-2, and may be scheduled daily or hourly in the crontab.

Example 13-2 Script to set printer queues up

```
for PRINTER in `lsallq`  
do  
    /usr/bin/qadm -U $PRINTER  
done
```

13.2.6 Checking network routes

Lost route definitions, or routes that have been improperly changed can result in a loss of performance. Therefore, we suggest you check your route definitions regularly. You may use the following command:

```
netstat -r
```

The following algorithm may support you in automating the check for network route changes:

- ▶ Write the output of **netstat -r** to a temporary file daily.
- ▶ Compare the file with the output written one day ago.
- ▶ Then send the differences to a mail address of your choice, generally to the system administrator.

13.2.7 Checking the AIX system console log

The AIX console displays messages of system boot/shutdown, standard output of processes started by the inittab and other processes or services, that use the console as an output device for stdout or stderr. An AIX default installation displays console messages on the device which has been defined as the AIX console. In this case, there is no possibility to analyze errors once they have been overwritten on the console.

Pitfall ahead!

It is strongly recommended that you write the console output to a file for further investigation; refer to Section 10.2.5, “Redirect the console” on page 286 for more information.

13.2.8 Differences for an SP environment

Beside all checks for the AIX operating system (as described in the preceding sections), there are additional checks that have to be performed when using an SP system.

Checking HW conditions for frame, switch, node and CWS

Use the graphical utility Perspectives to check the SP hardware (HW) conditions, such as:

- ▶ Power on/off
- ▶ LED/LCD flashing and contents
- ▶ Node and switch response

Perspectives allows you to configure your personal view of monitoring areas. We suggest that you keep Perspectives visible on your display for continuous control.

For a detailed explanation of monitoring and diagnosing SP error conditions, consult *PSSP: Diagnosis Guide*, GA22-7350.

Checking SP daemons

There are a couple of SP specific daemons that should be running for a smooth operation of your SP.

Check that daemons listed in Table 13-1 are running on your SP nodes.

Table 13-1 Important SP specific daemons for nodes and CWS

SRC Group	Daemon name ^a	Running on Nodes/CWS
hags	hags, hagsglsm	Nodes + CWS
haem	haem, haemaixos	Nodes + CWS
hats	hats	Nodes + CWS
pman	pman, pmanrm	Nodes - CWS
	spconfigd	Nodes + CWS
sdr	sdr	CWS
hr	hr	CWS

SRC Group	Daemon name ^a	Running on Nodes/CWS
	hardmon	CWS
	splogd	CWS
	kadmind	CWS
	kerberos	CWS
	supfilesrv ^b	CWS

a. On the CWS, daemons are listed as: <daemon_name>.<cws_name>

b. If using file collection

You can use the following command to search for daemons that are active:

```
dsh '/usr/bin/ps -ef | /usr/bin/grep <daemon_name>'
```

Checking the Kerberos ticket lifetime

On the control workstation (CWS), issue the **klist** command to show the lifetime of your actual Kerberos ticket. If the ticket has expired or will expire soon, generate a new ticket with the **kinit** command.

Checking SP log files

In an SP environment, there are some slight differences concerning the location of log files. The SP nodes' console logs can only be redirected to the SP log directory /var/adm/SPlogs/sysman.

Additional SP logs can be found in the following directories:

- ▶ /var/adm/SPlogs
- ▶ /var/adm/SPlogs/sysman
- ▶ /var/adm/SPlogs/css
- ▶ /var/adm/SPlogs/kerberos
- ▶ /var/ha/log

Additionally, you may find valuable information in one of the following Redbooks:

- ▶ *Inside the RS/6000 SP*, SG24-5145
- ▶ *PSSP Version 3 Survival Guide*, SG24-5344
- ▶ *RS/6000 SP Cluster: The Path to Universal Clustering*, SG24-5374
- ▶ *RS/6000 SP Software Maintenance*, SG24-5160

13.2.9 Monitoring HACMP

If working in a fault- or disaster-tolerant configuration, the administrator should check the cluster history log daily. A separate log file is created every day, when at least one cluster event is raised. These log files record all actions that take place within the cluster. The location of the log files is `/usr/sbin/cluster/history/cluster.<mmddyyyy>`.

If the cluster automatically performs an action, it has been driven by an event, such as a network failure, a loss of electrical power, or a system breakdown. Therefore, the administrator always has to find out the reason for the events.

When a cluster event occurs, lots of “operator notifications” are logged in the AIX error report. Therefore, we recommend using a condensed form of the AIX error reporter (see Section 13.2.1, “Checking the AIX error log” on page 409), when working in a HACMP environment.

13.2.10 Summary for AIX, SP, and HACMP daily checks

Table 13-2 shows a summary of monitoring tasks concerning AIX, SP and HACMP.

Table 13-2 AIX, SP and HACMP checks

What to check	Command or Log file	Frequency
AIX error log	<code>errpt</code>	Daily/hourly
Saturation of file systems	<code>df -k</code>	Daily/hourly
Mirrors	<code>lsvg lsvg -i1 grep stale</code>	Daily
Mailboxes	<code>mail</code>	Daily
Printer queues	<code>smitty qstatus</code>	Daily/hourly
Network routes	<code>netstat -r</code>	Daily
AIX Console log	<code>/var/domain/SAP/logs/console.log</code>	On demand
SP HW conditions	Perspectives	Permanently
SP daemons	<code>dsh 'ps -ef grep <daemon_name>'</code>	Daily
SP Kerberos ticket lifetime	<code>klist</code>	Weekly
SP log files	<code>/var/adm/SPlogs/...</code> <code>/var/ha/log</code>	On demand

What to check	Command or Log file	Frequency
HACMP Cluster history log	/usr/sbin/cluster/history/cluster.<timestamp>	Daily

13.3 Database system

The database is one of the core elements of an SAP R/3 environment. For this reason, avoiding or early detection of database errors is one of the key issues when performing monitoring and maintenance tasks. The tasks in the following sections should be performed regularly.

13.3.1 Checking database free space

Use SAP R/3 transaction code (TX) DB02 to get an overview of tablespace usage and changes.

► TX DB02

In the tablespace box click on space statistics

The columns %-Used and Chg/day help you to calculate, how fast your tablespaces grow and which tablespaces may need an extension. As a rule all tablespaces over 87 percent have to be extended. Extend tablespaces early in order to avoid stagnation of your production system. The four most important tablespaces are

- PSAPSTABI
- PSAPSTABD
- PSAPBTABI
- PSAPBTABD

Pitfall ahead!

Implement a concept for archiving SAP R/3 data right from the start in order to prevent SAP R/3 systems from growing too large too quickly.

13.3.2 Update optimizer statistics

The database optimizer decides how to run database queries in the most efficient way. This includes questions such as which processing sequence is the most powerful, or which index should be used for scanning the tables. A cost based optimizer can only provide good access paths if supplied with meaningful statistics. Therefore, we suggest that you update your database statistics as follows:

- ▶ Using Oracle
 - Check optimizer statistics for all tablespaces weekly.
 - Update optimizer statistics for all entries in table DBSTATC weekly.
- ▶ Using DB2
 - Run update statistics for all entries in table DBSTATC daily.
 - Run update statistics for all tables weekly.
- ▶ Using Oracle
 - Select **TX DB13 -> Edit -> create action.**
 - Enter a time of your choice (for example, 04:00:00).
 - Enter period weeks (1=every week).
 - Select Check optimizer statistics.
 - Continue.
 - Select PSAP% All SAP tablespaces.
 - Continue.
 - Select E (estimate table).
 - Select **TX DB13 -> Edit -> create action.**
 - Enter a time of your choice (for example, 04:00:00).
 - Enter period weeks (1=every week).
 - Select Update optimizer statistics.
 - Continue.
 - Select DBSTATCO: All tables marked in DBSTATC.
- ▶ Using DB2
 - Select **TX DB13 -> Edit -> schedule action.**
 - Enter a time of your choice (for example, 04:00:00).
 - Enter period weeks (1=every week).
 - Select Upd. Statistics + Reorgcheck all tables.

- Repeat the following action for every weekday:
 - Select **TX DB13 -> Edit -> schedule action**.
 - Enter a time of your choice (for example, 04:00:00).
 - Enter period weeks (1=every week).
 - Select Update Statistics + Reorgcheck (DBSTATC).

Pitfall ahead!

While the optimizer's statistics are being updated, the performance of your SAP R/3 system is impacted. Therefore, we suggest that you run the update of the statistics at night or during weekends.

If needed, you also may run update statistics for single tables using transaction DB20.

13.3.3 Running Oracle database system check

This section only applies to Oracle databases and not to DB2 databases. For DB2 there is no equivalent tool to perform an SAP R/3 database check.

The following issues are checked:

- ▶ Physical consistency of control, redo log, and data files. For example
 - Missing control files
 - Missing or corrupted data files
 - Tablespaces offline or in backup mode
- ▶ Severe error messages in the database alert file, such as:
 - ORA-1113, ORA-1122, ORA-1555, and others
- ▶ Database space concerning fragmentation and point of saturation. For example:
 - Segments with more than 80 percent of extents used
 - Segments leading to tablespaces overflow by allocating one extent
 - Tablespaces that are more than 90 percent full
- ▶ Oracle parameters in the init<sid>.ora file. For example:
 - DB_BLOCK_BUFFERS < 8960, DB_FILES > 254, and others

Schedule a check of your database once a week. You may use the following command:

```
sapdba -check
```

Another possibility to schedule the check is using transaction DB13, as follows:

1. Select **TX DB13 -> Edit -> Create action**.
2. Enter a time of your choice (for example, 02:00:00)
3. Enter period weeks (1=every week)
4. Select Check database
5. Continue

A database system check slows down performance of your SAP R/3 system. We recommend that you schedule the check for when the SAP R/3 system is likely to be unused or lightly loaded.

13.3.4 Checking for database errors

The database log file contains all actions and errors that take place in the database system. Database errors are also logged in the SAP R/3 system log. Check the SAP R/3 system log daily (as described in Section 13.4.3, “Checking the SAP R/3 system log” on page 423). For continuous information of database actions and errors, consult the database diag file, which is:

- ▶ /db2/<SID>/db2dump/db2diag.log for DB2
- ▶ /oracle/<SID>/saptrace/background/alert_<SID>.log for Oracle

13.3.5 Searching for missing indices

Missing indices may reduce the performance of your database queries. Use the following transaction in order to find missing indices:

TX DB02 -> Goto -> Tables and indices -> Missing indices

If an index is missing in the data dictionary, recreate the index using transaction SE11; if it is missing in the database, recreate the index using transaction SE14.

13.3.6 Monitoring performance

In order to detect problems at an early stage, it is important that you monitor the performance of your database system. Perform the following steps for monitoring database performance and ensure that the listed values fulfill the following recommendations:

For DB2:

TX ST04	Overall buffer quality > 96 percent
	Data hit ratio > 95 percent

Index hit ratio > 98 percent
Average physical read time < 10 ms
Average physical write time < 50 ms

For Oracle:

TX ST04

Data buffer quality > 94 percent
Reads/user call < 20
Time/user call < 20 ms
SQL area pinratio > 97 percent

13.3.7 Reorganization

Coping with database growth and dealing with large numbers of updates, deletes, or inserts can cause tables to become fragmented. Fragmentation is a waste of storage and may impact performance. Therefore, data and index reorganization represents an important task.

Although reorganization usually leads to a performance improvement, it is an expensive task. While running an online reorganization, users are affected by a performance decrease. Oracle tablespace reorganization can only be performed when the database is offline, while Oracle table reorganization may be performed online as of SAP R/3 Version 4.6.

To reduce fragmentation, you should always know your future needs for table growth and allocate space in advance. For a closer look at how to select database parameters to reduce fragmentation, consult the redbook *Database Performance on AIX in DB2 UDB and Oracle Environments*, SG24-5511.

In the following two sections, we describe how to find out which DB2 and Oracle tables need to be reorganized.

Reorganizing DB2 tables

Running the job update statistics for all tables, as described in Section 13.3.2, “Update optimizer statistics” on page 417, includes a check for reorganization on all tables. The DBA planning calendar provides the following job for table reorganization:

1. Select **TX DB13 -> Edit -> Schedule action**.
2. Select **Reorganize flagged tables & update stats**.

Selecting this job provides you with a list of tables for which reorganization is recommended. You can choose which of the tables you want to be reorganized. Temporary space is needed for reorganization. You can either enter the name of a temporary table space or, otherwise, a working copy of the original table is created in the same table space where the original table resides. This requires sufficient free space in the table space.

Reorganizing Oracle tables

We recommend that you use the **sapdba** utility for reorganization. **sapdba** supports you by performing the required actions for a reorganization. The syntax of the command is:

```
sapdba -> d - Reorganization
```

If you do not know which tables need to be reorganized, start with a check to find out reorganization candidates. There are different levels of reorganization. You either can reorganize a table or a complete tablespace. Select the desired method and follow the steps that are proposed by **sapdba**. For a detailed description, refer to the SAP Library.

13.3.8 Summary for DB regular checks and tasks

Table 13-3 shows a summary of DB administrative tasks.

Table 13-3 Regular DB checks and tasks

What to do	Command or Log file	Frequency
Check database free space	TX DB02	Weekly
Update optimizer statistics	TX DB13	Schedule once check weekly
Check Oracle DB	sapdba -check or TX DB13	Schedule once check weekly
Check DB errors	TX SM21 /db2/<SID>/db2dump/db2diag.log or /oracle/<SID>/saptrace/background/ alert_<SID>.log	Daily
Check for missing indexes	TX DB02	Weekly
Monitor performance	TX ST04	Weekly
Run DB reorg	TX DB13 sapdba	On demand

13.4 SAP R/3

Daily checks need to be run on all SAP R/3 systems. Checks are performed via the SAP GUI. The following sections describe all important checks and the affiliated SAP R/3 transactions (TX). It is presumed that the administrator has the required SAP R/3 authorizations for running the particular transactions.

13.4.1 Housekeeping batch jobs

In a productive SAP R/3 environment, a couple of background jobs should always be configured. These jobs perform regular tasks in order to prevent an uncontrolled growth of tables and also deliver statistical data for performance analysis. Examples of such tasks are the deletion of old print requests and the deletion of old ABAP short dumps. Table 13-4 shows the most important standard batch jobs that should be scheduled by an administrator (see also SAP Note 16083).

Table 13-4 *Suggested house keeping jobs*

SAP R/3 naming conventions	ABAP program	Frequency	Description
SAP_REORG_JOBS	RSBTCDEL	Daily	Deletes expired background jobs.
SAP_REORG_SPOOL	RSPO0041	Daily	Deletes old spool requests.
SAP_REORG_BATCHINPUT	RSBDCREO	Daily	Deletes old batch input folders.
SAP_REORG_ABAPDUMPS	RSSNAPDL	Daily	Deletes old ABAP short dumps.
SAP_REORG_JOBSTATISTIC	RSBPSTDE	Daily	Deletes old job statistics that are not required anymore.
SAP_REORG_UPDATERECORDS	RSM13002	Daily	Deletes incomplete updates.
SAP_COLLECTOR_FOR_JOBSTATISTIC	RSBPCOLL	Monthly	Generates run-time statistics for background jobs.
SAP_COLLECTOR_FOR_PERFMONITOR	RSCOLL00	Hourly	Collects statistics for system performance.

Since SAP R/3 Release 4.6C, SAP provides a transaction that automatically schedules background jobs suggested by SAP. It is selected as follows:

TX SM36 -> Goto -> Standard jobs -> Reorg. job -> Default scheduling

After having scheduled the batch jobs, check the status of the jobs regularly (see Section 13.4.4, “Checking background jobs” on page 424).

13.4.2 Checking SAP R/3 instances and application servers

Use transaction code SM51 to check that the central instance and all application servers are running. For each server name, perform the following steps:

- ▶ Double click server name to show work processes for each server.
- ▶ Verify that all configured and required work processes are running.
- ▶ Verify that the status of waiting or running is normal.
- ▶ Keep an eye on long running dialog processes.
- ▶ Keep an eye on processes waiting a long time for a semaphore (see SAP Note 33873 for a description of the different semaphores).
- ▶ A status of Hold or Killed/Completed should be investigated in the SAP R/3 system log.
- ▶ Restart killed/completed processes.

13.4.3 Checking the SAP R/3 system log

The SAP R/3 system log contains messages describing actions that take place in the SAP R/3 system. The messages are categorized into problem classes. You should check at least once a day all messages of classes problem and warning. Follow these steps to do so:

1. Select **TX SM21 -> System log -> Choose -> Central system log**.
2. Select time frame of your interest.
3. Select **Problem classes: Problems and warnings-> System log -> Reread system log**.

Bright idea!

If the central instance (CI) is not on the same host as the database, you should adjust the profile parameter `rslg/collect_daemon/host`, which points to the `SAPDBHOST` as a default. Set the parameter in the default profile as `rslg/collect_daemon/host = <CI_host_name>`.

Additionally, we suggest that you change the following parameters. This increases the amount of space of the SAP R/3 the system log, which can grow quickly in a production environment. For example, double the default sizes, as shown in the following:

- ▶ `rslg/max_diskspace/central = 4000000`
- ▶ `rslg/max_diskspace/local = 1000000`

13.4.4 Checking background jobs

We have discussed the housekeeping batch jobs in Section 13.4.1, “Housekeeping batch jobs” on page 422. The remaining batch jobs in an SAP R/3 system are generally application specific. Depending on the business needs of your SAP R/3 system, you should check all background jobs regularly, at least on a daily basis. To search for canceled background jobs, run the following transaction:

- ▶ TX SM37
 - Enter Job name: *.
 - Enter User name: *.
 - Select Job status: Canceled.
 - Select the timeframe of your interest.
 - Select **Job** -> **Execute**.

Mark each canceled job and click on the job log button for further investigation.

13.4.5 Checking update records

To check for incomplete update records, run the following transaction:

- ▶ TX SM13
 - Enter Client: *.
 - Enter User: *.
 - Select time frame of your interest.

Investigate any incomplete updates and see if there is any reason for the problem.

13.4.6 Checking ABAP dumps

An ABAP short dump is created whenever an ABAP program fails for any reason. To list ABAP short dumps, run:

- ▶ Select **TX ST22 -> Goto -> select short dump**.
- ▶ Enter a time frame of your interest.
- ▶ Select **Program -> Execute**.

Double-click on each ABAP dump for a detailed description of the error.

Normally, the batch job SAP_REORG_ABAPDUMPS is scheduled, so old ABAP dumps will be deleted regularly and automatically. There may be a need to keep particular short dumps for later analysis. Mark these jobs in the following way:

- ▶ Select a dump.
- ▶ Select **Short dump -> Keep/release**.

13.4.7 Checking TemSe consistency

Problems sometimes occur with TemSe because of inconsistent table entries. This can happen especially if you restore from the database and perform a database copy or delete a client without deleting its objects first. Run the following transaction to check TemSe consistency:

TX SP12 -> TemSe database -> Consistency check

The consistency check forces the deletion of orphaned TemSe objects (see also SAP Note 16875).

13.4.8 Checking for lock entries

When performing data changes, rows or tables have to be locked in order to ensure consistency. This happens, for example, when records are updated. Generally, a lock is released as soon as the transaction is committed or rolled back. If a lock entry is kept persistent, you should check for the underlying reason.

To check if system lock entries are present, run:

- ▶ TX SM12
 - Enter Table name: *.
 - Enter Client: *.
 - Enter user name: *.

- Select -> **Lock entry** -> **List**.

Pitfall ahead!

We suggest that you investigate all locks that are older than five days. Only delete locks if it is obvious that they are not being used anymore. Before you delete locks, check for processes using TX SM51 and for users logged on using TX AL08 and finally contact the user in question.

Attention: Reckless deletion can lead to a corrupt and inconsistent database.

13.4.9 Monitoring performance and workload

It is important to monitor the performance of your SAP R/3 systems in order to detect bottlenecks or increasing load at an early stage.

Before starting to monitor performance and workload, we suggest that you adjust values that control the residence time of statistical data. Perform the following steps:

1. **TX ST03 -> Goto -> Parameters -> Performance database -> Edit -> Modify parameters**
2. In the box “Residence time of statistical data,” enter a residence time frame of your interest. We suggest:

Standard statistics:

Days: 10

Weeks: 5

Month: 3

3. Use the following transactions to monitor performance and workload:

System buffers

TX ST02

Buffers should have a hit ratio of 98 percent and swaps of zero. One exception is the program buffer, this buffer may; have swaps. You can use the history button to show buffer history.

System memory

TX ST02

Ensure that the current usage of the following items is below 60 percent:

- Roll area
- Paging area
- Extended Memory
- Heap Memory

Workload analysis

TX ST03 -> Goto -> Survey graphics

Determine if the workload remains constant or grows continuously.

An increasing workload or decreasing performance may require an expansion of your hardware (see Chapter 11, “Performance” on page 305 for more information).

13.4.10 Analyzing trace files of work processes

In a few situations, you may need to investigate a work process' trace file. They can be found in the `/usr/sap/<SID>/<instance>/work` directory.

Work process trace files generally contain lots of messages in the first part. For error analysis, we suggest that you read the trace file from bottom to top. It may also be useful to store trace files of work processes once, when your system runs properly. In case of a failure, you can compare traces of erroneous work processes with those traces you have stored while the system was running correctly.

13.4.11 Summary for SAP R/3 checks

Table 13-5 shows a summary of checks an SAP R/3 administrator should perform regularly.

Table 13-5 SAP R/3 regular checks

What to check/do	Command or Log file	Frequency
Housekeeping batch jobs	TX SM36	Schedule once check daily
Instances and application servers	TX SM51 TX SM50	Daily
System log	TX SM21	Daily
Background jobs	TX SM37	Daily
Update records	TX SM13	Daily
ABAP dumps	TX ST22	Daily
TemSe consistency	TX SP12	Weekly
Lock entries	TX SM12	Weekly
Performance and workload	TX ST02 TX ST03	Daily Weekly
Trace files of work processes	<code>/usr/sap/<SID>/<instance>/work</code>	On demand

13.4.12 Transactions for administration assistance

Instead of performing single checks for your SAP R/3 systems, as described in the preceding sections, you may use overlaid transactions that guide and support you in performing all necessary steps. Such transactions are:

- ▶ SSAA: System administration assistant
- ▶ RZ20: CCMS monitor sets

SSAA

You can access the System Administration Assistant (SAA) by calling transaction SSAA, which has been delivered in Version 4.5B for the first time. The SAA is designed to compose administration tasks at a central point. It supports important and frequent tasks and organizes them by subject and by frequency. The SAA changes the frequency of tasks depending on whether the system is a development, quality assurance, or production system. For detailed information on using and configuring SAA, the following information may be useful:

- ▶ Documents at <http://service.sap.com/rrr> (follow the menu **Ready-to-Run R/3 -> Media Center -> Consulting information**)
 - System Administration Assistant: Benefits
 - System Administration Assistant: Technical Background
 - System Administration Assistant: DB2
 - System Administration Assistant: Oracle
- ▶ SAP Note 104019

Transaction RZ20

Transaction RZ20 has an object based monitoring architecture that simplifies the tasks involved in monitoring a system. It integrates information from the entire SAP R/3 environment and presents an overview of the status of the SAP R/3 systems. Alerts are raised and displayed for each monitored attribute if threshold conditions are met. Therefore, you should configure your thresholds accurately before working with transaction RZ20.

The transaction RZ20 is also used by the SAA as an underlying part of monitoring different areas.

13.5 Backup

Pitfall ahead!

Maintaining an SAP R/3 environment includes regular system backups and, of course, regular checks to determine whether the backups have completed successfully or not. Besides monitoring your backup operations, we strongly recommend that you perform restore tests regularly.

Assuming a backup and recovery concept using TSM (see Chapter 7, “Backup and recovery” on page 185), the queries described in the following sections have to be performed.

13.5.1 Checking the TSM client schedules

TSM client schedules include:

- ▶ Incremental backups of AIX, SAP R/3, and DB executables and interfaces
- ▶ SAP R/3 DB backups
- ▶ SAP R/3 DB redo log backups (often called archives)
- ▶ NIM mksysb backups
- ▶ Other backups or archives scheduled by TSM

To list all TSM client schedules that experienced an error condition, use the following command within the TSM admin command line interface. By using the parameters `begindate/enddate` and `begintime/endtime`, you can choose the time frame of your interest:

```
query event SAP * begindate=today-1 begintime=18:00 \  
enddate=today endtime=18:00 exceptiononly=yes
```

The status column of the output indicates the reason for unsuccessful schedules:

- | | |
|---------------|---|
| Missed | The backup did not start, because it was not possible to establish a connection between TSM server scheduler and TSM client scheduler daemon. The reason may be an inactive scheduler daemon on the TSM client or a problem with the network or network protocol. |
| Failed | Status failed indicates that the schedule returned with a return code (RC) not equal to zero. Be aware of the following exception: When a running schedule exceeds its schedule startup window, TSM shows the status failed, but the schedule actually completed successfully without any errors or warnings. |

If a schedule fails, use the log files for investigation. Table 13-6 lists all TSM log files and database backup protocols that may be helpful. The following environment is assumed:

DSM_LOG=/var/domain/SAP/logs

DSMI_LOG=/var/domain/SAP/logs

For DB2 databases:

- ▶ INSTHOME=/db2/<SID>
- ▶ DB2DB6_ARCHIVE_PATH=/db2/<SID>/log_archive
- ▶ DIAGPATH=/db2/<SID>/db2dump

For Oracle databases:

- ▶ ORACLE_HOME=/oracle/<SID>
- ▶ SAPARCH¹=/oracle/<SID>/sapbackup

Table 13-6 Backup log files and protocols

Description	Log file or query	Useful for investigation of
TSM client error log	/var/domain/SAP/logs/dsmerror.log	All schedules
TSM client API error log	/var/domain/SAP/logs/dsierror.log	Schedules using TSM API, such as DB and redo log backups
TSM client scheduler log	As configured in dsm.sys, for example, /var/domain/SAP/logs/scheduler.aix.log	All schedules
TSM server activity log	q actlog [parameters]	
Summary of brbackup protocols	/oracle/<SID>/sapbackup/backSID.log or TX DB12 -> Overview of all database backups	Oracle db backups
Detailed brbackup protocol	in /oracle/<SID>/sapbackup <coded timestamp>.<extension>	
Summary of brarchive protocols	/oracle/<SID>/sapbackup/arch<SID>.log or TX DB12 -> Overview of redo log backup logs	Oracle redo log archive
Detailed brarchive protocol	in /oracle/<SID>/sapbackup <coded timestamp>.<extension>	

¹ The SAPARCH environment variable is optional; if not, set the default points to /oracle/<SID>/saparch.

Description	Log file or query	Useful for investigation of
User exit run logs	/db2/<SID>/db2dump/ db2uext2.log.NODE000 or TX DB12 -> User exit logs	DB2 log archive
User exit error log	/db2/<SID>/db2dump/ db2uext2.err.NODE000	
brarchive protocol	/db2/<SID>/saparch/ brarchive.<timestamp>.<extension> or TX DB12 -> brarchive logs	DB2 log brarchive
DB backup log	TX DB12 -> Overview of all database backups or db2adutl query full [parameters]	DB2 backup log

13.5.2 Checking the TSM administrative schedules

TSM administrative schedules include the following tasks:

- ▶ Backup of the TSM database
- ▶ Migration and copy of storage pools
- ▶ Setting of reclamation thresholds
- ▶ Expiring of inventory

You may use the TSM admin command line interface and enter the following command to list failed administrative schedules:

```
query event * type=admin begindate=today-1 begintime=18:00 \ enddate=today
endtime=18:00 exceptiononly=yes
```

The status column shows the reasons for unsuccessful schedules:

- Missed** The TSM server could not start the schedule within its scheduled startup window.
- Failed** An error occurred; therefore, the schedule did not complete successfully.

In both cases, consult the TSM activity log for a more detailed analysis.

Attention: Be careful with return codes of administrative schedules! Only scheduled commands with parameter wait=yes guarantee a meaningful status. All other administrative schedules do not wait for completion of the schedule and may show a status of completed, but actually may have received an error.

13.5.3 Checking the TSM server activity log

Bright idea! The TSM activity log is located in the TSM server's database and has the function of logging all TSM server messages. In addition, it is possible to send TSM client messages to the central TSM activity log. You may run the macro shown in Example 13-3 in order to send all TSM client messages of type error and severe to the TSM activity log.

Example 13-3 TSM macro for sending client error messages to the server actlog

disable	Events	Actlog	all	nodename=*
enable	Events	Actlog	severe	nodename=*
enable	Events	Actlog	error	nodename=*
begin	EventLoggin	Actlog		
commit				

To query all server messages in the TSM server activity log that indicate an error or severe error, you should run the following TSM commands:

```
q actlog search=ANR????E
q actlog search=ANR????S
```

To query all client messages in the TSM server activity log, search for ANE instead of ANR, for example:

```
q actlog search=ANE????E
q actlog search=ANE????S
```

Use the parameters begindate/enddate and begintime/endtime to choose the time frame of your interest, such as:

```
beginday=today-1 begintime=18:00
enddate=today endtime=18:00 exceptiononly=yes
```

13.5.4 Checking TSM scheduler daemons

Each TSM virtual node needs an appended scheduler daemon in order to be contacted by the TSM server to run a schedule. If a TSM scheduler daemon is missing on one virtual node, schedules defined for that node cannot be executed. You may restart lost scheduler daemons manually. Another approach is to use a script, which is executed regularly by cron to restart missing daemons. A third possibility is to initiate scheduler daemons in the `/etc/inittab` using the `respawn` option.

13.5.5 Checking SP CWS mksysb

If working in an SP environment, the AIX `mksysb` of the control workstation (CWS) needs to be run manually, because it is written to the local tape device. Therefore, it is normally not started using a TSM schedule. Check the return code of this manual operation each time after it has finished. Generally, this action should be performed weekly.

13.5.6 Summary for backup operations daily checks

Table 13-7 shows a summary of all daily tasks concerning backup operations and the TSM server.

Table 13-7 Backup operations daily checks

What to check	Tool or log file	Frequency
TSM client schedules	<code>q event SAP * beginnd=-1 endd=today ex=yes</code>	Daily
TSM admin schedules	<code>q event * t=a beginnd=-1 endd=today ex=yes</code>	Daily
TSM activity log	<code>q actlog [parameters]</code>	Daily
TSM scheduler daemons	<code>ps -ef grep "dsmc sched"</code>	On demand
SP CWS mksyb	smitty return code or log file of a script	Weekly

13.6 Cleaning the log files

In a vital SAP R/3 environment, there are plenty of log files written. If these files are not deleted regularly, they can fill file systems and lead to stagnation. We suggest that you implement an algorithm that limits log file growth. This algorithm distinguishes between the following types of log files and uses an appropriate method for deletion:

- ▶ Log files that grow constantly, for example, smit.log. Each month, move the content of this log to a file with the same name and add a timestamp extension. Delete these moved log files after three months.
- ▶ Log files that are created once, but will not be deleted automatically, for example, SAP R/3 trace files. Delete all files older than 1 month.
- ▶ Log files that are maintained by a special command or utility, for example, AIX error report. Use the appropriate utility to delete old entries or old log files.

Table 13-8 shows a collection of log files. Each log file is assigned to one of the previous defined types and lists an appropriate method for deletion of old log entries, if available.

Table 13-8 Log files

Description	Log file	Type	Method or utility for deletion
AIX error report	/var/adm/ras/errlog	3	errclear
SMIT log and script log	\$HOME/smit.log \$HOME/smit.script	1	
Mailbox	\$HOME/mbox	1	
Switch user log	/var/adm/sulog	1	
Log of successful logins	/var/adm/wtmp	1	
cron log	/var/adm/cron/log	1	
SNMP log	/var/tmp/snmpd.log	1	
	/var/tmp/dpid2.log		
Old mails	/var/spool/mail/root	1	
Failed login log	/etc/security/failedlogin	1	

Description	Log file	Type	Method or utility for deletion
HACMP cluster log	/var/adm/cluster.log	1	
HACMP cluster manager log	/tmp/cm.log	1	
DB backup	Your script logs for database backup	2	
DB log archive	Your script logs for database log archive	2	
brarchive protocols	\$SAPARCH/<coded timestamp>.<ext>	3	sapdba -cleanup^a
brbackup protocols	/oracle/<SID>/sapbackup/<coded timestamp>.<ext>	3	
Oracle alert log	/oracle/???/saptrace/background/alert_???.log	1	
Oracle audit files	/oracle/???/rdbms/audit/ora_*.aud	2	
DB2 diag log	/db2/???/db2dump/db2diag.log	1	
DB2 userexit run log	/db2/<SID>/db2dump/db2uext2.log.NODE000	1	
DB2 user exit error log	/db2/<SID>/db2dump/db2uext2.err.NODE000	1	
TSM error log	\$DSM_LOG/dsmerror.log	3	Automatic log pruning if configured in dsm.sys
TSM API error log	\$DSM_LOG/dsierror.log	3	
TSM scheduler logs	Name as configured in dsm.sys		

a. Additionally, **sapdba -cleanup** deletes all log files as configured in your sapdba configuration file (sapdb<SID>.dba)

Troubleshooting and first aid after system failure

In a nutshell:

- ▶ Use a structured problem determination procedure; never use trial-and-error.
- ▶ Know the contact persons for technical and organizational problems.
- ▶ Have the current system documentation for all components in place.
- ▶ Plan to restore the original state before implementing temporary fixes.

If error situations occur, a structured and planned troubleshooting procedure increases the speed in which a system can be brought back up to full operation and avoids accidental and destructive changes on the system.

This chapter outlines such a structured path for problem determination in an SAP R/3 infrastructure. It describes activities on operating system and database level, as well as SAP R/3 commands that allow you to determine the cause of a problem. For some common errors, we provide procedures to fix the problem. However, after determining the cause of a problem, you may either be able to fix the problem yourself or to contact the support line with valuable background information.

This chapter covers the highlighted area in Figure 14-1, which is derived from the SAP R/3 infrastructure model that is used throughout the book.

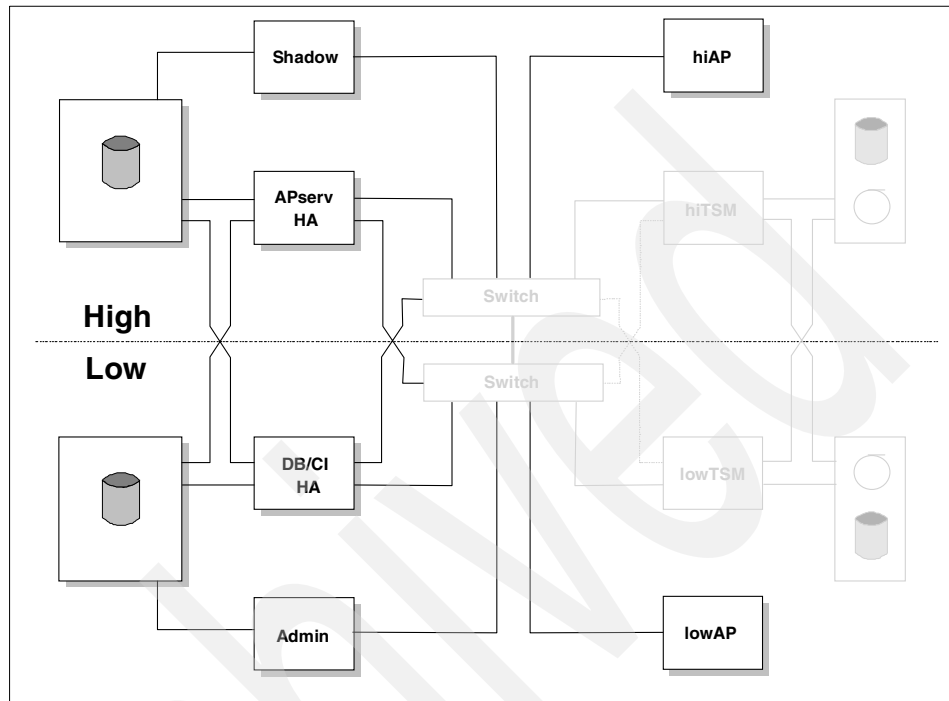


Figure 14-1 Troubleshooting and first aid after system failure

14.1 Steps to locate and analyze a faulty component

This section describes a method to determine the layer of a faulty component within an SAP R/3 system. This method covers the layers:

- ▶ Network
- ▶ Operating system
- ▶ Database system
- ▶ SAP R/3
- ▶ Presentation layer (SAP GUI)
- ▶ Backup system
- ▶ Hardware

It is not a trivial task to locate a faulty component within a relative complex system infrastructure for SAP R/3. An uncoordinated procedure, such as trial-and-error, can take either a long time or a side effect can damage the system accidentally. However, a more detailed knowledge about the location of the faulty component gives you the chance to either fix the problem yourself or to contact the support line with valuable background information. The described method is a suggestion based on many years of practical experience, but might not cover all specific environments.

Bright idea!

The layer with the faulty component should be detected first independent from the situation. You can follow the paths given in the flowcharts shown in Figure 14-2 on page 441 and Figure 14-3 on page 442.

First, select the suitable flowchart relevant to the error situation. The following list of some common situations helps you to choose the correct chart:

- ▶ Some transactions abort with a dump (see Figure 14-2 on page 441)
- ▶ Transports fail or TX STMS aborts (see Figure 14-2 on page 441)
- ▶ Any SAP R/3 user cannot connect (see Figure 14-3 on page 442)
- ▶ Instance or database does not start (see Figure 14-3 on page 442)
- ▶ Backup or archive operation fails (see Figure 14-3 on page 442)

To read the flowcharts, start at the top of the chart and choose the arrow that describes the current error situation. Then follow this arrow and execute the action described in the next box. After executing some actions along the path you have determined or solved the problem or reached one of the endpoints. They describe the layer where the faulty component is most probably located and are marked with a thick line above them in the flowchart. Contact the person which is responsible for the administration of the relevant layer. The end points in the flow charts refer to the following contact points:

PC	Contact the PC administrator.
Network	Contact the network administrator.
SAP	Contact the SAP administrator or an SAP R/3 Basis Consultant.
DB	Contact the DB administrator or an SAP R/3 Basis Consultant.
AIX	Contact the AIX administrator or an SAP R/3 Basis Consultant.
TSM	Contact the TSM administrator.

All the activities shown in the flowcharts are described, starting in Section 14.2, "Transactions" on page 443. The activities contain SAP R/3 transactions (rectangle boxes), combined tasks (circles), SAP R/3 or database tools (hexagon), AIX commands (rhombus), and log files named in the activity descriptions of every layer. A short description of what you should expect and

how you can get the information is provided in the individual sections. On the edge of every activity element, there is a small circle containing a number. This number refers to the appropriate section with further explanations. For example, number 2.2 refers to Section 14.2.2, “Transaction SM21” on page 443.

All command line utilities must be executed as AIX user root, if not stated otherwise.

Attention: The descriptions given for activities like tasks, tools, and commands are only valid in the context of the relevant flow chart. They do not cover *all* possible situations. The descriptions are tailored to the situations that are outlined in the flow charts.

Pitfall ahead!

A complex system like SAP R/3, which delivers a wide spectrum of services, has a lot of components. In every one of these components, an error might occur. To avoid unnecessarily long down times for maintenance tasks and accidental destructions during maintenance tasks, follow the rules listed below:

- ▶ Apply only changes with well-known effects and results.
- ▶ Plan to restore the original state before implementing temporary fixes.
- ▶ Check possible influences on the systems fault- or disaster-tolerance.
- ▶ A dedicated person must be responsible for coordinating all repair tasks.

Keep in mind that this chapter only discusses selected error situations and gives suggestions to help you start your investigations and to fix minor problems. Detailed activities are not discussed here. For detailed information, refer to the SAPNet or the appropriate product manuals.

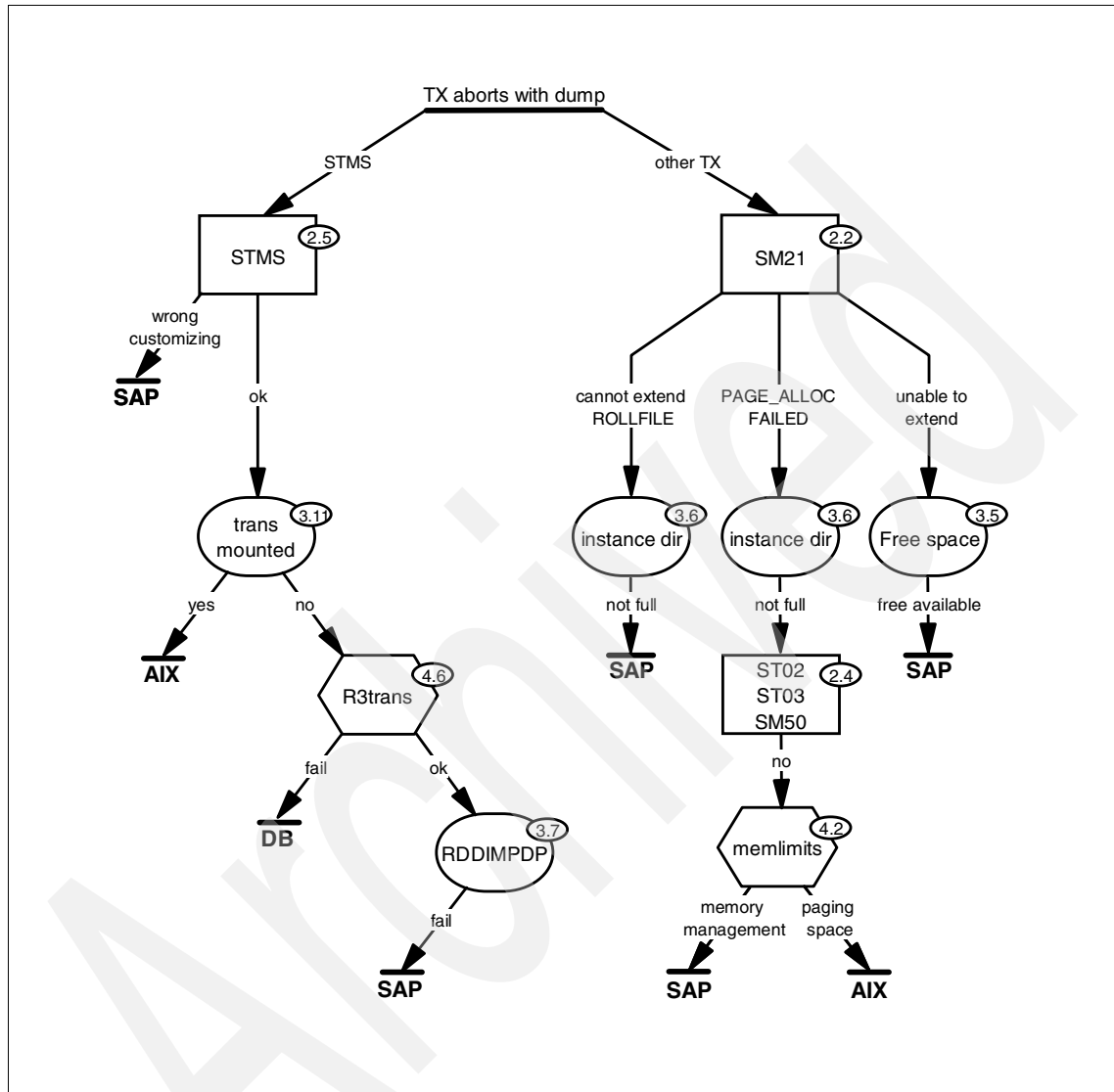


Figure 14-2 Determine a faulty component if SAP R/3 transactions abort / dump

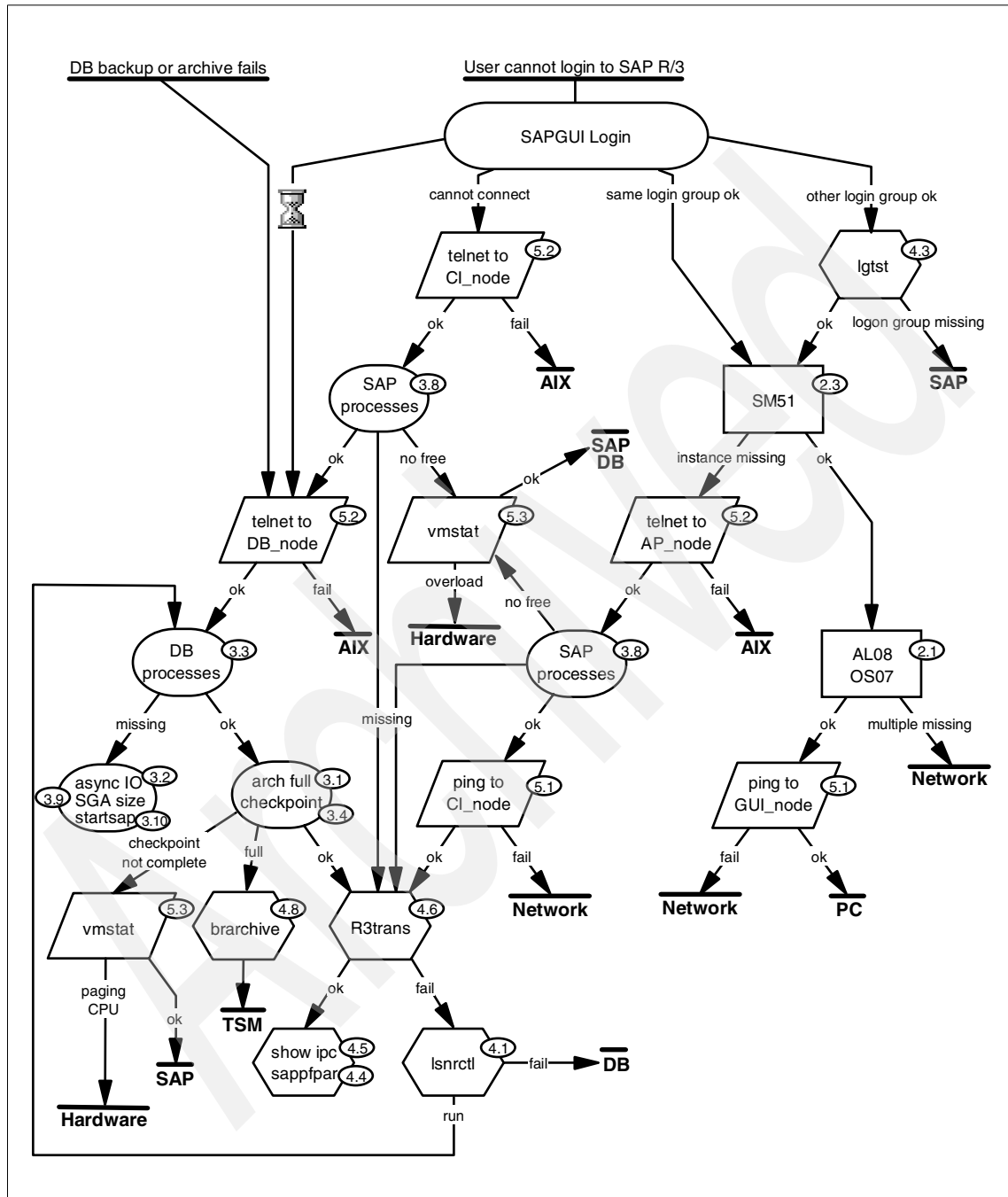


Figure 14-3 Determine a faulty component if a user cannot connect to SAP R/3

14.2 Transactions

This section contains SAP R/3 transactions, in alphabetical order, which are useful for locating faulty components.

Attention: The description given for each transaction is only valid in the context of the relevant flow chart. It does not cover *all* possible situations. The descriptions are tailored to the situations that are outlined in the flow charts (see Figure 14-2 on page 441 and Figure 14-3 on page 442).

14.2.1 Transactions AL08 and OS07

This task checks whether only a single or more users are unable to connect to SAP R/3. The following list is a guide to checking which users are currently logged on:

- Use transaction AL08 to display the host name of the PCs, where the SAPGUI of the active SAP R/3 users run. A meaningful naming concept for these terminals may give you a chance to detect an unusual pattern of active users. For example, if there are no active users of a certain subsidiary, there might be a problem with the corporate wide network. If a network problem arises when users are already connected, they will be still shown as connected within AL08 as long as they are not inputting or requesting new data from SAP R/3.
- Use transaction OS07 to check the network availability to already connected SAP R/3 users using the following menu path:

OS07 -> Detail analysis menu -> LAN Check by Ping -> Presentation server -> Select All -> 10 x Ping

14.2.2 Transaction SM21

This action determines the instance on which the SAP R/3 transaction was processed before it was terminated. Because SAP R/3 uses load balancing, this is not always obvious. The instance and work process number, the reason for abortion, and the error code are documented in the SAP R/3 central system log. On a system with several instances, the host name is also written into the system log (shown as bold marked columns in Example 14-1 on page 444). For further information, see Section 13.4.3, “Checking the SAP R/3 system log” on page 423.

Example 14-1 SAP R/3 central system log

Time	Instance	Ty.	Nr	Cl.	User	MNo	Text
12:28:16	terra_SLZ_10	BTC	15	800	USER01	BY0	> DbSlModifyDB6(SQLExecute) SQLS
12:28:16	terra_SLZ_10	BTC	15	800	USER01	BY0	> Driver][DB2/6000] SQL0289N Unabl
12:28:16	terra_SLZ_10	BTC	15	800	USER01	BY0	> table space "PSAPBTABI". SQLSTAT
12:28:16	terra_SLZ_10	BTC	15	800	USER01	EAZ	Failed to maintain status entry fo
12:32:35	terra_SLZ_10	DIA	0	000	SAPSYS	BY4	Database error -289 at INS access
12:32:35	terra_SLZ_10	DIA	0	000	SAPSYS	BY0	> DbSlExeModifyDB6(SQLExecute) S
12:32:35	terra_SLZ_10	DIA	0	000	SAPSYS	BY0	> Driver][DB2/6000] SQL0289N Unabl
12:32:35	terra_SLZ_10	DIA	0	000	SAPSYS	BY0	> table space "PSAPBTABI". SQLSTAT
12:32:35	terra_SLZ_10	DIA	0	000	SAPSYS	R68	Perform rollback

14.2.3 Transaction SM51

This task checks the state of the instances and work processes on these instances. The following list is a guide to performing the checks:

- The SAP R/3 transaction SM51 shows all active instances. Check whether all configured instances are listed and offer services, such as dialog or background processing. Example 14-2 shows a system with five instances on three nodes.

Example 14-2 Output of transaction SM51 (instances)

Server names	Host name	Type
tequila_SLZ_10	tequila	Dialog Update Enqueue Backg Spool Upd2
terra_SLZ_10	terra	Dialog Backg Spool
terra_SLZ_20	terra	Dialog
zombie_SLZ_10	zombie	Dialog Backg Spool
zombie_SLZ_20	zombie	Dialog

- Further detailed information about the work processes that belong to a specific instance, for example, work process types and states as well as error conditions, can be displayed with transaction SM51. Follow the menu path below to get an output like Example 14-3. The SAP R/3 Transaction SM50 delivers the same output without any menu selections:

SM51 -> select an instance with double-click

Example 14-3 Output of transaction SM51 (work processes)

No	Ty.	PID	Status	Reasn	Start	Err	Sem	CPU	Time	Report	Cl.	User
0	DIA	36482	Running		Yes					SAPLTHFB	000	SAP*
1	DIA	41378	waiting		Yes	2						
2	DIA	19980	waiting		Yes							
3	BGD	16236	waiting		Yes							
4	BGD	30526	waiting		Yes							
5	BGD	20434	waiting		Yes							
6	BGD	14858	waiting		Yes							

7	DIA	34802	waiting	Yes
8	DIA	39768	waiting	Yes
9	DIA	26396	waiting	Yes
10	UPD	22510	waiting	Yes
11	UPD	23378	waiting	Yes
12	UPD	6492	waiting	Yes
13	UPD	8992	waiting	Yes
14	ENQ	28532	waiting	Yes
15	BGD	29516	waiting	Yes

14.2.4 Transactions ST02 and ST03

The reason for aborting a transaction is always documented in the SAP R/3 central system log (SM21). Further information can be obtained from there using the following steps:

1. If the reason relates to the memory consumption, then use ST03 and the following menu path to determine the amount of memory the transaction has used:

ST03 -> Detail analysis menu -> Workload -> Memory Profile -> Sort Memory

The term memory in this section means the sum of extended and private memory, which is limited by the sum of real memory plus paging space. For further information, see Section 11.3, “Models for SAP R/3 Extended Memory” on page 318.

In general, there are two error situations:

- a. A single transaction consumes all memory.
- b. The sum of used memory by all transactions exceeds the available memory.

The transaction with the biggest memory consumption can be found at the top of the output, as shown in Example 14-4.

Example 14-4 Output of SAP R/3 transaction ST03

Tcode or Report	Mem. total avg [Bytes]	Extended memory avg used	[Bytes] max used	Workproc. reservations	Private avg used
ZZRES01	287.695.820	211.468.783	250.880.687	4	72.086.619
SM36	6.836.761	5.640.357	7.419.461	0	0
SM37	3.566.951	2.421.644	6.768.447	0	0
RSABAPPROGRA	2.955.511	1.664.812	250.880.687	1	0
ABAP_DOCU_SH	2.382.405	1.271.146	1.832.846	0	0
DB6COCKPIT	2.145.350	961.553	6.768.447	0	0
SM50	2.050.378	861.810	7.419.461	0	0
ST02	1.989.373	911.419	34.161.856	0	0

2. In case “a”, split the input data into smaller parts or choose a more detailed selection criteria to prevent future error situations.
3. In case “b”, monitor the memory usage of all transactions and programs. A usage history can be found with transaction ST02 by following the menu path below:

ST02 -> select line Extended Memory -> Mode list

Example 14-5 shows the transactions that consume the largest amount of extended memory as a bold printed column. These transactions or reports are subject for further investigations.

Example 14-5 Output of SAP R/3 transaction ST02

History									
No.	Name	Attchd	Ext Mem [kB]	Heap [kB]	I mode Globl/kB	E mode Globl/kB	Imode0 [kB]	Imode1 [kB]	Imode2 [kB]
1	RESIDENCY	X	179539	0	0	0	1.441	1.382	176717
2	SAP*	X	178458	0	0	0	1.441	300	176717
3	SAP*	X	244900	0	0	0	1.441	300	243159
4	USER03	X	23014	0	0	0	339	0	141
5	USER03	X	23014	0	0	0	339	0	141
6	USER35	X	28649	0	0	0	1.443	0	0

4. The memory consumption can also be monitored during the run time of a transaction or program with transaction SM50. The values of the actual memory consumption are available as long as the report or transaction has not finished and the work process state is in PRIV mode. This is shown in Example 14-6 as bold marked items.

Example 14-6 Output of SAP R/3 transaction SM50

	No	Ty.	PID	Status	Reasn	Start	Err	Sem	CPU	Time	Report	Cl.	User
0	DIA		36480	Running	Yes				30	ZZRES01	000	SAP*	
1	DIA		41368	stopped	PRIV	Yes	2		147		000	SAP*	
2	DIA		35524	stopped	PRIV	Yes			1425		000	SAP*	
3	DIA		16234	Running	Yes				63	ZZRES01	000	SAP*	
4	DIA		30526	Running	Yes					SAPLTHFB	000	SAP*	
5	DIA		20434	waiting	Yes								
6	DIA		14858	waiting	Yes								
7	DIA		34802	waiting	Yes								
8	DIA		39768	waiting	Yes								
9	DIA		26396	waiting	Yes								
10	UPD		22510	waiting	Yes								
11	UPD		23378	waiting	Yes								
12	UPD		6492	waiting	Yes								
13	UPD		8992	waiting	Yes								
14	ENQ		28532	waiting	Yes								
15	BGD		29516	waiting	Yes								

- If it is not possible to log in to SAP R/3, use the command line utility **dpmon**, which delivers an output format similar to transaction SM50 (see Example 14-7):

```
cd /sapmnt/<SID>/exe
./dpmon pf=../profile/<SID>_DVEBMGS<instance_number>_<host> 1
```

Example 14-7 Output of the dpmon command

No	Ty.	Pid	Status	Cause	Start	Err	Sem	CPU	Time	Program	C1	User
0	DIA	36482	Wait		yes	0	0		0			
1	DIA	41368	Stop	PRIV	yes	2	0		421		000	SAP*
2	DIA	35524	Stop	PRIV	yes	0	0		58		000	SAP*
3	DIA	16236	Wait		yes	0	0		0			
4	DIA	30526	Wait		yes	0	0		0			
5	DIA	20434	Wait		yes	0	0		0			
6	DIA	14858	Wait		yes	0	0		0			
7	DIA	34802	Wait		yes	0	0		0			
8	DIA	39768	Wait		yes	0	0		0			
9	DIA	26396	Wait		yes	0	0		0			
10	UPD	22510	Wait		yes	0	0		0			
11	UPD	23378	Wait		yes	0	0		0			
12	UPD	6492	Wait		yes	0	0		0			
13	UPD	8992	Wait		yes	0	0		0			
14	ENQ	28532	Wait		yes	0	0		0			
15	BTC	29516	Wait		yes	0	0		0			

- A suspicious program can be monitored during its run time with transaction SM50 as long as the report or transaction has not finished and the work process state is in PRIV mode. Follow the menu path below:

SM50 -> select Workprocess -> Details

Example 14-8 shows a program that has consumed a large amount of memory and roll areas. The work process state is in PRIV mode, because the extended memory is already exhausted.

Example 14-8 Output of transaction SM50

Roles / memory	Number	Time (usec
Roll In	1	889
Roll Out	2	128.822
Roll	1.982.464	(Bytes)
Page	32.768	(Bytes)
Memory (sum)	250856450	(Bytes)
Memory (sum private)	110946352	(Bytes)
Memory (used)	248994542	(Bytes)
Memory (max. transaction)	250856450	(Bytes)
Memory (max. dialog step)	250856450	(Bytes)

7. If not a single transaction but, rather, the sum of all transactions consumes too much memory, check the settings of the SAP R/3 memory management. For detailed information, see Section 11.3, “Models for SAP R/3 Extended Memory” on page 318.

14.2.5 Transaction STMS

This task uses the transaction STMS, which is the interface within SAP R/3 to the SAP R/3 *Transport Management System (TMS)*. The following list is a guide to checking, with transaction STMS, whether TMS is customized correctly:

- ▶ All expected source and target systems should be defined and connected by at least one transport layer. Trigger the distribution of TMS definitions to ensure all systems have the same centrally managed settings. Use the following menu path:

STMS -> Overview -> Transport routes -> Configuration -> Check -> Transport routes

- ▶ The transport domain controller must be able to reach every system inside the SAP R/3 transport domain and execute the program tp there. This can be checked using the following menu path:

STMS -> Overview -> Imports -> Import Queue -> Check -> Transport tool -> Check transport tool on all systems? -> Yes

14.3 Tasks

This section contains tasks in alphabetical order that are useful for locating faulty components.

Attention: The description given for each task is only valid in the context of the relevant flow chart. It does not cover *all* possible situations. The descriptions are tailored to the situations that are outlined in the flow charts (see Figure 14-2 on page 441 and Figure 14-3 on page 442).

14.3.1 Archive file system full

Database logs are essential for a recovery of the database. Therefore, it is crucial to keep all of them in a safe place, for example, in a separate file system on disk.

Archive file system full (DB2)

Under normal circumstances, online active log files are created in the log_dir directory. When they no longer contain active pages, a user exit transfers them to the log_archive directory.

Important: Never move DB2 log files manually from log_archive into the log_dir directory!

- ▶ If there is no free space in the archive file system to store further offline retained logs, the user exit aborts with an error code greater than zero. This condition can be shown in the database diag log as well as in the user exit error log (Example 14-9). In that case, the database allocates extra online logs, if necessary. Eliminate the cause of the error to avoid a database hang.

```
/usr/bin/df /db2/<SID>/log_archive
```

Further information can be gathered from the following log files:

- ▶ /db2/<SID>/db2dump/db2uext.log.NODE000
- ▶ /db2/<SID>/db2dump/db2uext.err.NODE000
- ▶ /db2/<SID>/db2dump/db2diag.log

Example 14-9 Log file db2uext.err and db2diag.log

```
Time of Error:   Thu Aug  9 14:30:08 2001
Action:         ARCHIVE
Database Name:  SLZ
Log File Path:  /db2/SLZ/log_dir/
Log File Name:  S0000015.LOG
Audit Log File: /db2/SLZ/db2dump/db2uext2.log.NODE0000
Media Type:     Disk
User Exit RC:   39
> Error isolation: Error writing file:
/db2/SLZ/log_archive/SLZ/S0000015.LOG.200
10809121108.NODE0000
Error code: 0
Error message: Error 0
```

User Exit returned error on ARCHIVE log file S0000015.LOG from
/db2/SLZ/log_dir/ for database SLZ, **error code 8**

- ▶ If there is no free space in the log_dir directory to allocate new online log files, the database hangs. In this situation, all database processes are active, but neither an SAP R/3 user nor a direct connected session to DB2 is able to finish any update operation. In this case, extend the log_dir directory or free some space in the log_archive directory using your archiving procedure.

- ▶ The DB2 user exit will be activated every five minutes or if the database starts to move the DB2 offline logs from the directory log_dir to log_archive:

```
/usr/bin/df /db2/<SID>/log_dir
```

Further information can be gathered from the following log files:

- ▶ /db2/<SID>/db2dump/db2uext.log.NODE000
- ▶ /db2/<SID>/db2dump/db2uext.err.NODE000

Archive file system full (Oracle)

Under normal circumstances, online redo log files exist in the origlog and mirrlog directories. If they contain no active transactions, the Oracle archiver process copies them to the saparch directory.

- ▶ If there is no free space in the archive file system to store further offline redo logs, the archiver process stops copying. This condition can be shown in the Oracle alert log (Example 14-10). In this case, the database uses all available online redo logs as long as they are all used. Eliminate the cause of the error to avoid a database hang.

Further information can be gathered from the
/oracle/<SID>/saptrace/background/alert_<SID>.log log file.

Example 14-10 Oracle alert log

```
Thread 1 advanced to log sequence 2148
  Current log# 14 seq# 2148 mem# 0: /oracle/MUC/origlogB/log_g14m1.dbf
  Current log# 14 seq# 2148 mem# 1: /oracle/MUC/mirrlogB/log_g14m2.dbf
Wed Aug 1 17:01:42 2001
ORACLE Instance MUC - Can not allocate log, archival required
Wed Aug 1 17:01:42 2001
Thread 1 cannot allocate new log, sequence 2149
All online logs needed archiving
  Current log# 14 seq# 2148 mem# 0: /oracle/MUC/origlogB/log_g14m1.dbf
  Current log# 14 seq# 2148 mem# 1: /oracle/MUC/mirrlogB/log_g14m2.dbf
```

- ▶ The database hangs if there are no unused online redo logs available. All database processes are active, but neither any SAP R/3 user or a direct connection with the **svrmgr1** command is able to finish any update operation. In this situation, check the free space at first. There must be at least one redo log file of free space. Extend the saparch directory or free some space using your archiving procedure:

```
/usr/bin/df /oracle/<SID>/saparch
```

14.3.2 Asynchronous I/O failed (Oracle)

An Oracle database needs an active asynchronous input and output subsystem (async I/O) and appropriate privileges for access to the data files. Otherwise, the database cannot be mounted or opened. The following list is a guide to finding the cause of the error:

- In some cases, missing file access permissions at the operating system level are masked by misleading asynchronous I/O errors, as shown in Example 14-11. Therefore, it is useful to put the database into the mount state and adjust the database's file permissions. Use the following command sequence to perform these actions:

```
/usr/bin/su - ora<sid>
/oracle/<SID>/<version>/svrmgrl
SVRMGR> connect internal
SVRMGR> startup mount
/usr/bin/chown -R ora<sid> /oracle/<SID>/sapdata*
/usr/bin/chmod 600 /oracle/<SID>/sapdata*/*/*data*
```

Further information can be gathered from the
/oracle/<SID>/saptrace/background/alert_<SID>.log log file.

Example 14-11 Output of svrmgrl command

```
ORA-01110: data file 3: '/oracle/I11/sapdata2/roll_1/roll.data1'
ORA-27061: skgfospo: waiting for async I/Os failed
IBM AIX RISC System/6000 Error: 22: Invalid argument
ORA-01110: data file 2: '/oracle/I11/sapdata1/temp_1/temp.data1'
ORA-27061: skgfospo: waiting for async I/Os failed
IBM AIX RISC System/6000 Error: 22: Invalid argument
SVRMGR> shutdown immediate
ORA-01109: database not open
Database dismounted.
ORA-27061: skgfospo: waiting for async I/Os failed
BM AIX RISC System/6000 Error: 22: Invalid argument
SVRMGR> startup mount
ORA-01031: insufficient privileges
```

- If the previous step fails, check whether asynchronous I/O is active. The appropriate number of asynchronous I/O servers depends on several criteria. For a suitable of the maximum number of I/O servers, please refer to “Recommendations for an SAP R/3 system” on page 340. Changes to the asynchronous I/O parameters require a reboot to become effective. Example 14-12 on page 452 shows a node with the minimal required settings to operate asynchronous I/O. Use the following command to display the asynchronous I/O parameter settings:

```
/usr/sbin/lstatr -EHl aio0
```

Example 14-12 Output of lsattr command

attribute	value	description	user_settable
minservers	1	MINIMUM number of servers	True
manservers	10	MAXIMUM number of servers	True
maxreqs	4096	Maximum number of REQUESTS	True
kprocprio	39	Server PRIORITY	True
autoconfig	available	STATE to be configured at system restart	True

14.3.3 Database processes

A database system, such as DB2 or Oracle, consists of the database's core processes, the client processes, and some tablespace files.

Database processes (DB2)

This task checks whether all required database processes are active. The number of these processes varies depending on their type, for example, database writer or agent processes, and the number of active SAP R/3 work processes. For future comparison, it is helpful to document the correct numbers for every database. The following list is a guidance to find the error reason:

- Every database process can be displayed with the **ps** command, as shown in Example 14-13. At least the following processes should be running:
 - Some DB2 agent processes (db2agent)
 - Some DB2 page cleaner processes (db2pclnr)
 - Some DB2 pre fetcher processes (db2pfchr)
 - Some miscellaneous DB2 core processes (db2*) started by user db2as
 - Some miscellaneous DB2 core processes (db2*) started by user db2<sid>
 - One DB2 watch dog process (db2wdg)

Enter the following command to obtain a list of running DB2 processes:

```
/usr/bin/ps -ef | /usr/bin/grep db2
```

Example 14-13 DB2 processes

root	16254	1	0	Jul 24	-	0:00	db2wdog
db2as	16512	16254	0	Jul 24	-	0:01	db2sysc
db2as	16770	16512	0	Jul 24	-	0:00	db2gds
db2as	17028	16512	0	Jul 24	-	0:00	db2ipccm
db2as	17286	16770	0	Jul 24	-	0:00	Scheduler
db2as	17544	16512	0	Jul 24	-	0:00	db2tcpcm
db2as	17802	16512	0	Jul 24	-	0:00	db2tcpdm
db2slz	5456	11618	0	Jul 24	-	0:01	db2sysc
db2slz	6478	5456	0	Jul 24	-	0:00	db2ipccm

db2slz	8522	5456	0	Jul 24	-	0:00	db2gds
db2slz	9552	5456	0	Jul 24	-	0:00	db2tcpcm
db2slz	14242	5456	0	Jul 24	-	0:00	db2spprm
db2slz	14750	8522	0	Jul 24	-	0:00	db2resyn
db2slz	15284	8522	0	Jul 24	-	0:00	db2srvlst
db2slz	16034	8522	0	Jul 24	-	0:00	db2spm1w
db2slz	19354	8522	0	Jul 24	-	0:43	db2loggr
db2slz	19608	8522	0	Jul 24	-	0:00	db2dlock
db2slz	22988	8522	0	Jul 24	-	0:00	db2event
db2slz	20382	8522	0	Jul 24	-	0:35	db2pfchr
db2slz	20640	8522	0	Jul 24	-	0:30	db2pfchr
db2slz	21156	8522	0	Jul 24	-	0:26	db2pfchr
db2slz	21930	8522	0	Jul 24	-	0:09	db2pc1nr
db2slz	22188	8522	0	Jul 24	-	0:09	db2pc1nr
db2slz	22446	8522	0	Jul 24	-	0:09	db2pc1nr
db2slz	26062	6478	0	14:57:32	-	0:05	db2agent (SLZ)
db2slz	29672	6478	0	14:57:32	-	0:01	db2agent (SLZ)
db2slz	30188	6478	0	14:57:32	-	0:00	db2agent (SLZ)
db2slz	30702	6478	0	14:57:32	-	0:00	db2agent (SLZ)

- If some of the processes listed above are missing, the database might be inactive or may have crashed before. In that case, you can try to restart the database, which automatically performs a crash recovery if needed. Use the following commands to restart the database:

```
/usr/bin/su - db2<sid>
/db2/<SID>/sqllib/bin/db2 db2start
```

Further information can be gathered from the
/db2/<SID>/db2dump/db2diag.log log file.

Database processes (Oracle)

This task checks whether all required database processes are active. The number of these processes varies depending on their type, for example, database writer or shadow processes, and the number of active SAP R/3 work processes. For future comparison, it is helpful to document the correct numbers for every database. The following list is a guidance to find the error reason:

- Every database process can be displayed with the **ps** command, as shown in Example 14-14 on page 454. At least the following processes should be running:
 - One listener process (tnslsnr)
 - Some Oracle shadow processes (oracle<SID>)
 - One Oracle database writer (ora_dbw)
 - One Oracle log writer (ora_lgwr)

- One Oracle archiver process (ora_arch)
- One Oracle check point process (ora_ckpt)
- One Oracle reconnect process (ora_reco)
- One Oracle system monitor (ora_smon)
- One Oracle process monitor (ora_pmon)

Enter the following command to obtain a list of running Oracle processes:

```
/usr/bin/ps -ef | /usr/bin/grep ora
```

Example 14-14 Oracle processes

orazmb	10092	1	0	Jan 03	- 0:50	/oracle/ZMB/bin/tnslsnr	LISTEN
zmbadm	10598	1	0	Feb 02	- 0:06	ora_arch_ZMB	
zmbadm	11598	1	0	Feb 02	- 58:02	ora_dbw0_ZMB	
zmbadm	15646	1	0	Feb 02	- 14:37	ora_lgwr_ZMB	
zmbadm	17010	1	0	Feb 02	- 17:22	ora_ckpt_ZMB	
zmbadm	19694	1	0	Feb 02	- 0:06	ora_reco_ZMB	
zmbadm	20766	1	0	Feb 02	- 87:09	ora_smon_ZMB	
zmbadm	23336	1	0	Feb 02	- 5:42	ora_pmon_ZMB	
orazmb	12528	1	0	Feb 02	- 6:04	oracleZMB	(DESCRIPTION=(LOCAL=
orazmb	12816	1	0	Feb 02	- 6:35	oracleZMB	(DESCRIPTION=(LOCAL=
orazmb	13200	1	0	Feb 02	- 0:00	oracleZMB	(DESCRIPTION=(LOCAL=
orazmb	16298	1	0	Feb 02	- 0:00	oracleZMB	(DESCRIPTION=(LOCAL=
orazmb	18300	1	0	Feb 02	- 2:03	oracleZMB	(DESCRIPTION=(LOCAL=

- If some of the processes listed above are missing, the database might be inactive or may have crashed before. In that case, you can try to restart the database, which automatically performs a crash recovery if needed. Use the following commands to restart the database:

```
/usr/bin/su - ora<sid>
/oracle/<SID>/<version>/svrmgrl SVRMGR> connect internal
SVRMGR> startup mount
SVRMGR> alter database open
```

Further information can be gathered from the
/oracle/<SID>/<version>/dbs/alert_<SID>.log log file.

14.3.4 Event ‘Checkpoint not complete’ (Oracle)

This action determines the number of occurred “Checkpoint not complete” events. Checkpoints are regularly triggered within Oracle to write all dirty database blocks from the database block buffer to the disks. This event usually occurs during times of heavy database write activity when the database writer processes are not able to complete one checkpoint before the next checkpoint is triggered.

The following list is a guide to finding the cause of the error:

- Check the current setting for log items, because “Checkpoint no complete” events are only documented within the database alert log file if the according parameter is set:

```
/usr/bin/vi /oracle/<SID>/<version>/dbs/init<SID>.ora
log_checkpoints_to_alert = true
```

- Determine how often the event “Checkpoint not complete” occurs during one day by analyzing the database alert log. Example 14-15 shows such a file. The event indicates more of a performance problem than an error. All further database operations are delayed until all checkpoints are completed. Use the following command to extract these events from the alert log:

```
/usr/bin/grep -E 'Jun 3|Checkpoint not complete' \
/oracle/<SID>/saptrace/background/alert_<SID>.log
```

Further information can be gathered from the
/oracle/<SID>/saptrace/background/alert_<SID>.log log file.

Example 14-15 Oracle alert log

```
Sun Jun  3 16:12:14 2001
Beginning log switch checkpoint up to RBA [0x20cd.2.10], SCN: 0x0000.02986028
Thread 1 advanced to log sequence 8397
Current log# 14 seq# 8397 mem# 0: /oracle/ZMB/origlogB/log_g14m1.dbf
Current log# 14 seq# 8397 mem# 1: /oracle/ZMB/mirrlogB/log_g14m2.dbf
Sun Jun  3 16:12:38 2001
Thread 1 cannot allocate new log, sequence 8398
Checkpoint not complete
Current log# 14 seq# 8397 mem# 0: /oracle/ZMB/origlogB/log_g14m1.dbf
Current log# 14 seq# 8397 mem# 1: /oracle/ZMB/mirrlogB/log_g14m2.dbf
Sun Jun  3 16:13:37 2001
Completed checkpoint up to RBA [0x20cb.2.10], SCN: 0x0000.02985ffdf
```

14.3.5 Free space problem

A database used for SAP R/3 always needs sufficient free space, because SAP R/3 inserts new data records into the underlying database during normal operation.

Free space problem (DB2)

A very common cause of error is the abortion of transactions due to a lack of database space. The following list is a guidance to finding the specific cause of error:

- The first action during error determination that affects only certain transactions is to analyze the SAP R/3 central system log using transaction SM21. The system log entry gives you the precise information about which work process log file of a certain instance has to be checked. Besides this, sometimes the cause of the error is directly displayed, as can be seen in Example 14-16.

Example 14-16 Output of SAP R/3 transaction SM21

Time	Instance	Ty.	Nr	Cl.	User	MNo	Text
12:12:58	sky_SLZ_10	DIA	0	000	SAPSYS	BY4	Database error -289 at INS access
12:12:58	sky_SLZ_10	DIA	0	000	SAPSYS	BY0	> DbS1ExeModifyDB6(SQLExecute)
12:12:58	sky_SLZ_10	DIA	0	000	SAPSYS	BY0	> SQL0289N Unable to allocate new
12:12:58	sky_SLZ_10	DIA	0	000	SAPSYS	BY0	> table space "PSAPBTABD". SQLSTA
12:12:58	sky_SLZ_10	DIA	0	000	SAPSYS	R68	Perform rollback
12:12:59	sky_SLZ_10	DIA	0	000	SAPSYS	AB0	Run-time error "DBIF_RSQ_L_SQL_ERR
12:13:00	sky_SLZ_10	DIA	0	000	SAPSYS	AB1	> Short dump "010718 121259 sky S
12:13:00	sky_SLZ_10	DIA	0	000	SAPSYS	D01	Transaction termination 00 (DBIF_
12:13:00	sky_SLZ_10	DIA	0	000	SAPSYS	R68	Perform rollback

- If the transaction has been terminated with an error message, such as 'Unable to extend table', the next steps are to check the database error log and the work process log files for further error entries (see Example 14-17). Use the following commands to list the contents of the work process log files:

```
/usr/bin/telnet <ap_or_ci_node>
/usr/bin/pg /usr/sap/<SID>/<instance>/work/dev_w<number>
```

Further information can be gathered from the
/usr/sap/<SID>/<instance>/work/dev_w<number> log file.

Example 14-17 Work process trace file

```
-----
trc file: "dev_w0", trc level: 1, release: "46C"
-----
*  ACTIVE TRACE LEVEL          1
...
C  *** ERROR in DB6Execute[dbdb6.c, 2682]
B  ***LOG BY4=> sql error -289   performing INS on table BALDAT      [dbtran
B  ***LOG BY0=> DbS1ExeModifyDB6( SQLExecute ) SQLSTATE=57011: [IBM][CLI
[DB2/6000] SQL0289N  Unable to allocate new pages in table space PSAPBTABD".
B  dbtran ERROR LOG (hdl_dbsl_error): DbS1 'INS'
B  STMT: {stmt:#=0, bndfld:#=0, prop=0x4, distinct=0, select*=1,
B          fld:#=7, alias:p=0, fupd:#=0, tab:#=0, where:#=0,
```

```

B      groupby:#=0, having:#=0, order:#=0, primary=0, hint:#=0}
B  CRSR: {tab='', id=0, hold=0, prop=0, max.in@=0, fae:blk=0,
B      con:nm='R/3', con:id=0, con:vndr=7, val=2,
B      key:l=0, key:#=5, xfer=1, xin:#=1, xout:#=0, upto=0,
B      wa:p=70bd6158, init:p=0, init:#=0, init:b=0}
M  ***LOG R68=> ThIRollBack, roll back () [thxxhead 9500]
A  Wed Jul 18 13:36:07 2001
A  ABAP/4 Program SAPLSBAL_DB_INTERNAL .
A  Source LSBAL_DB_INTERNALU02 Line 58.
A  Error Code DBIF_RSQ_L_SQL_ERROR.
A  Module $Id: //bas/46C/src/krn/runt/absapsql.c#3 $ SAP.
A  Function ExecuteCall Line 5832.
A  SQL error -289 occurred when accessing table BALDAT

```

- If the database cannot allocate free space to fulfill a write operation, for example, to insert a row, extend the table space by adding a container. The database diag log contains detailed error messages, as shown in Example 14-18. Use the following command to list the contents of the diag log file:

```
/usr/bin/pg /db2/<SID>/db2dump/db2diag.log
```

Further information can be gathered from the
/db2/<SID>/db2dump/db2diag.log log file.

Example 14-18 Diag log db2diag.log

```

Data Title:section stmt PID:16802 Node:000
494e 5345 5254 2020 494e 544f 2020 2242      INSERT INTO "B
414c 4441 5422 2028 2022 4d41 4e44 414e      ALDAT" ( "MANDAN
5422 202c 2022 5245 4c49 4422 202c 2022      T" , "RELID" , "
4c4f 475f 4841 4e44 4c45 2220 2c20 2242      LOG_HANDLE" , "B
4c4f 434b 2220 2c20 2253 5254 4632 2220      LOCK" , "SRTF2"
2c20 2243 4c55 5354 5222 202c 2022 434c      , "CLUSTER" , "CL
5553 5444 2220 2920 2056 414c 5545 5328      USTD" ) VALUES(
203f 202c 203f 202c 203f 202c 203f 202c      ? , ? , ? , ? ,
203f 202c 203f 202c 203f 2029 20          ? , ? , ? )
...
2001-07-18-12.22.48.832896 Instance:db2slz Node:000
PID:16802(db2agent (SLZ)) Appid:*LOCAL.db2slz.010627133149
buffer_pool_services sqlbAllocateExtent Probe:830 Database:SLZ
Tablespace 4(PSAPBTABD) is full

```

Free space problem (Oracle)

A very common cause of error is the abortion of transactions because of a lack of database space. The following list is a guide to finding the specific cause of error:

- The first action during error determination, which only affects certain transactions, is to analyze the SAP R/3 central system log using transaction

SM21. The system log entry gives you precise information on which work process log file of a certain instance has to be checked. Beside this, sometimes the error reason's directly displayed, as can be seen in Example 14-19.

Example 14-19 Output of SAP R/3 transaction SM21

Time	Instance	Ty.	Nr	Cl.	User	MNo	Text
12:44:05	terra_SLZ_10	DIA	7	000	SAPSYS	BYL	Database error 1631 requires in
12:44:05	terra_SLZ_10	DIA	7	000	SAPSYS	BY4	Database error 1631 at SEL acce
12:44:05	terra_SLZ_10	DIA	7	000	SAPSYS	BY0	> ORA-01631: max # extents
12:44:05	terra_SLZ_10	DIA	7	000	SAPSYS	BY0	> (300) reached in table
12:44:05	terra_SLZ_10	DIA	7	000	SAPSYS	BY0	> SAPR3.DRA0#

- If the transaction has been terminated with an error message, such as “Unable to extend table”, the next steps are to check the database error log and the work process log files for further error entries (see Example 14-20). Use the following commands to list the contents of the work process log files:

```
/usr/bin/telnet <ap_or_ci_node>
/usr/bin/pg /usr/sap/<SID>/<Instance>/work/dev_w*
```

Further information can be gathered from the
/usr/sap/<SID>/<instance>/work/dev_w* log file.

Example 14-20 Work process trace file dev_w7

```
-----
trc file: "dev_w7", trc level: 1, release: "46C"
-----
* ACTIVE TRACE LEVEL          1
...
M ***LOG Q01=> tskh_init, WPStart (Workproc 7 1 15704) [thxxhead 0847]
M calling db_connect ...
B Try to connect as default user
B Using SQL-Net V2 with tnsname = 'MUC'
B Got '/usr/sap/trans' for TNS_ADMIN from environment
B Got 'MUC' for SID from profile-parameter 'rsdb/oracle_sid'
B Connecting via TNS_ADMIN=/usr/sap/trans, tnsname=MUC
B Got NLS_LANG=AMERICAN_AMERICA.US7ASCII from environment
B Thu Aug  2 12:06:35 2001
B ***LOG BXF=> table logging switched off in program RSCLXCOP by user SAP*
B Thu Aug  2 12:08:31 2001
B ***LOG B6B=> syn. mc switched off completely from RSCLXCOP by SAP*
B ***LOG BB1=> reset buffer TABL          [dbtbxbuf 1439]
B Thu Aug  2 12:08:33 2001
B ***LOG BB1=> reset buffer TABLP         [dbbfe    3957]
B Thu Aug  2 12:43:50 2001
B ***LOG BYL=> dba action required because of db error          [dbsh
B ***LOG BY4=> sql error 1631 performing SEL on table DRA0      [dbtran
B ***LOG BY0=> ORA-01631: max # extents (300) reached in table SAPR3.DRA0
```

- ▶ If the database cannot allocate free space to fulfill an operation, for example, to insert a row, extend the tablespace by adding a data file. Regular tablespace reorganizations allow the re-use of unused data areas and may therefore avoid the need to add more data files.

14.3.6 Instance file system full

Every SAP R/3 instance has a file system that contains some important files, for example, the SAP R/3 paging and roll file. Furthermore, there are some trace files. It makes sense to regularly check the remaining free space with the following command:

```
/usr/bin/df /usr/sap/<SID>/<Instance>/data
```

The following reasons may be causing file system space problems:

- ▶ If old trace files consume too much space, move or delete these files. If necessary, adjust the work process trace level with SAP R/3 transaction SM50.
- ▶ The statistic file /usr/sap/<SID>/<Instance>/data/stat may have grown extremely. To delete this file, but save current statistic data, use SAP R/3 transaction ST03 and follow the menu path:

ST03 -> Workload -> Reorganization -> Delete stats. File -> Update performance database from statistics file before delete? -> Yes

To avoid an very large statistic file, shorten the period of time between two reorganizations with SAP R/3 transaction ST03, as discussed in Chapter 13, “Daily tasks to prevent error situations” on page 407.

- ▶ Sort operations on large data areas can also fill up the instance file system completely. These sort files are temporary and disappear after the sort operation has been finished or terminated.

14.3.7 Report RDDIMPDP

This action checks if the SAP R/3 batch job RDDIMPDB is active. After the **tp** tool has inserted all information about the objects in a correction request, it triggers the event driven SAP R/3 batch job. The **tp** tool checks whether this job is scheduled correctly. If this job is missing for any reason, reschedule it with the report RDDNEWPP as SAP R/3 user DDIC in client 000. A **tp** output of an execution on system with the correct planned job is shown in Example 14-21 on page 460.

Run **tp** using the following command sequence:

```
/usr/bin/su - <sid>adm
cd /usr/sap/trans/bin
/sapmnt/<SID>/exe/tp checkimpdp <SID>
```

Example 14-21 Output of *tp* command

System SLZ: Background job RDDIMPDP is **scheduled event periodic**.

14.3.8 SAP R/3 processes

This action checks whether all required SAP R/3 processes are active. The number of work processes varies depending on their type (for example, dialog or message server) and the parameters that were defined in the SAP R/3 instance profiles. For future comparison, it is helpful to document correct numbers for every SAP R/3 instance. The following list is a guide to finding the cause of the error:

- ▶ SAP R/3 transaction SM50 shows all active work processes. At least one available dialog work processes in status *waiting* should be in the output of transaction SM50, as shown in Example 14-22. Busy work processes are in states PRIV, hold mode, or running.

Example 14-22 Transaction SM50

No.Ty.	PID	Status	Reason	Start	Err	Sem	CPU	Time	Program	ClieUser
0	DIA 14788	waiting		Yes						
1	DIA 21948	running		Yes					RSMON000 000	SAP*
2	DIA 28458	waiting		Yes						
3	DIA 21066	waiting		Yes						
4	DIA 24394	waiting		Yes						
5	DIA 20664	waiting		Yes						
6	DIA 9284	waiting		Yes						
7	DIA 25300	waiting		Yes						
8	DIA 26590	waiting		Yes						
9	UPD 27112	waiting		Yes						
10	UPD 26334	waiting		Yes						
11	UPD 24290	waiting		Yes						
12	ENQ 19116	waiting		Yes						
13	BTC 25038	waiting		Yes						
14	BTC 10088	waiting		Yes						
15	SPO 28162	waiting		Yes						
16	UP2 20396	waiting		Yes						

- If it is not possible to login to SAP R/3, use the command line utility **dpmon**, which uses an output format similar to transaction SM50 (see Example 14-23 for an output of the **dpmon** command):

```
cd /sapmnt/<SID>/exe
./dpmon pf=../profile/<SID>_DVEBMGS<instance_number>_<host> 1
```

Example 14-23 Output of the dpmon command

No	Ty.	Pid	Status	Cause	Start	Err	Sem	CPU	Time	Program	C1	User
0	DIA	27102	Wait		yes	0	0		0			
1	DIA	34338	Wait		yes	0	0		0			
2	DIA	33572	Wait		yes	0	0		0			
3	DIA	22748	Wait		yes	0	0		0			
4	DIA	30510	Wait		yes	0	0		0			
5	BTC	33302	Run		yes	0	0		48633	RSSNAPDL	000	SAP*
6	DIA	33054	Wait		yes	0	0		0			
7	DIA	32522	Wait		yes	0	0		0			
8	DIA	28146	Wait		yes	0	0		0			
9	DIA	25840	Wait		yes	0	0		0			
10	UPD	25570	Wait		yes	0	0		0			
11	UPD	18592	Wait		yes	0	0		0			
12	UPD	22198	Wait		yes	0	0		0			
13	UPD	24524	Wait		yes	0	0		0			
14	ENQ	21682	Wait		yes	0	0		0			
15	DIA	9180	Wait		yes	0	0		0			
16	BTC	19618	Wait		yes	0	0		0			
17	BTC	5470	Wait		yes	0	0		0			
18	SP0	16044	Wait		yes	0	0		0			
19	UP2	36192	Wait		yes	0	0		0			
20	UP2	26840	Wait		yes	0	0		0			

- Every work process has a corresponding process on operating system level. They can be listed with the **ps** command, as shown in Example 14-24 on page 462. At least the following processes should be running:
 - On central instance: One message server (ms)
 - On central instance: One syslog collector daemon (co)
 - On application servers: One syslog sender daemon (se)
 - Some work processes (dw)
 - One performance collector (saposcol)
 - One gateway reader (gwrdr)

Use the following command to display the currently running SAP R/3 processes:

```
/usr/bin/ps -ef | /usr/bin/grep sap
```

Example 14-24 SAP R/3 processes

zmbadm	10880	1	0	Jan 03	- 81:51	/usr/sap/ZMB/SYS/exe/run/sapos
zmbadm	18614	23894	0	Feb 02	- 17:19	ms.sapZMB_DVEBMGS00 pf=/usr/sa
zmbadm	20540	23894	0	Feb 02	- 24:12	co.sapZMB_DVEBMGS00 -F pf=/usr
zmbadm	24136	23894	0	Feb 02	- 10:13	se.sapZMB_DVEBMGS00 -F pf=/usr
zmbadm	16480	17964	0	Feb 02	- 19:12	gwrdd -dp pf=/usr/sap/ZMB/SYS/p
zmbadm	11872	17964	0	Feb 02	- 0:20	dw.sapZMB_DVEBMGS00 pf=/usr/sa
zmbadm	15488	17964	0	Nov 20	- 72:25	dw.sapZMB_DVEBMGS00 pf=/usr/sa
zmbadm	16658	17964	0	Nov 20	- 32:12	dw.sapZMB_DVEBMGS00 pf=/usr/sa
zmbadm	17312	17964	0	Feb 02	- 59:33	dw.sapZMB_DVEBMGS00 pf=/usr/sa
zmbadm	17608	17964	0	Feb 02	- 3:57	dw.sapZMB_DVEBMGS00 pf=/usr/sa

14.3.9 Shared memory segment too large (Oracle only)

If an Oracle database does not start, one reason could be that the configured SGA size exceeds the available main memory.

- An error message, as shown in Example 14-25, is misleading, because it is not the permissions that are wrong; it is Oracle that cannot allocate sufficient shared memory segments as defined in the profile. This error occurs especially after the system copies from a node with a lot of memory to a relatively small node with much less memory. In this case, the following command sequence triggers the error message, as shown in Example 14-25, by starting only the database processes:

```
/usr/bin/su - ora<sid>  
/oracle/<SID>/<version>/svrmgrl  
SVRMGR> connect internal  
SVRMGR> startup nomount
```

Further information can be gathered from the
/oracle/<SID>/saptrace/background/alert_<SID>.log log file.

Example 14-25 Output of svrmgrl command

```
SVRMGR> connect internal  
Connected.  
SVRMGR> startup mount  
ORA-27121: unable to determine size of shared memory segment  
IBM AIX RISC System/6000 Error: 22: Invalid argument  
SVRMGR> startup nomount  
ORA-27123: unable to attach to shared memory segment  
IBM AIX RISC System/6000 Error: 22: Invalid argument
```

- After decreasing the shared memory related parameter `db_block_buffer` within the file `init<SID>.ora`, the problem should disappear.

You can edit the `init<SID>.ora` file using the following command:

```
/usr/bin/vi /oracle/<SID>/<version>/dbs/init<SID>.ora
```

14.3.10 startsap fails (Oracle only)

The SAP R/3 command **startsap** is used to start the database and the appropriate instances. The following list is a guide to finding the cause of the error:

- The **startsap** command is executed as AIX user `<sid>adm`. However, the database has to be started as AIX user `ora<sid>`. Therefore, some Oracle executables must have special file access permissions (`rwsr-s--x`); otherwise, the instance startup fails, even though the database is up and running, as shown in Example 14-26.

Further information can be gathered from the `/home/<sid>adm/startdb.log` log file.

Example 14-26 Log file startdb.log

```
Attempting to contact (ADDRESS=(COMMUNITY=SAP.WORLD)(PROTOCOL=TCP)(HOST=sky)
T=1527))
OK (210 msec)
tnsping: V2 connect to SKY
----- Mon Jul 30 11:51:16 MSZ 2001
Connect to the database to check the database state:
R3trans: connect check finished with return code: 12
Database not available
```

- Starting the database as AIX user `<sid>adm` using the **svrmgr1** command can also fail due to missing permissions on the audit directory, as shown in Example 14-27. The **chmod** command below assigns the correct permissions:

```
/usr/bin/su - <sid>adm
/oracle/<SID>/<version>/svrmgr1
SVRMGR> connect internal
SVRMGR> startup mount
/usr/bin/chmod -R 775 /oracle/<SID>/<version>/rdbms
```

Example 14-27 Output of svrmgr1 command (missing file permission)

```
SVRMGR> connect internal
Password:
ORA-09925: Unable to create audit trail file
IBM AIX RISC System/6000 Error: 13: Permission denied
Additional information: 9925
```

- ▶ Starting the database as AIX user <sid>adm using the **svrmgr1** command can also fail due to missing permissions on certain executables, as shown in Example 14-28. The **chmod** command below assigns the correct permissions:

```
/usr/bin/su - <sid>adm
/oracle/<SID>/<version>/svrmgr1
SVRMGR> connect internal
SVRMGR> startup mount
/usr/bin/chmod 6751 /oracle/<SID>/<version>/bin/oracle*
```

Example 14-28 Output of svrmgr1 command (missing permission to SGA)

```
SVRMGR> connect internal
Connected.
SVRMGR> startup mount
ORA-07302: smscre: create failure in creating ?/dbs/sgadef@.dbf.
IBM AIX RISC System/6000 Error: 13: Permission denied
```

14.3.11 Transport directory

The transport directory is usually the central location for storing client transport requests and support packages. The following list is a guide to checking the correct state of this directory:

- ▶ The transport directory must be mounted and AIX user <sid>adm has to have read/write permissions to all relevant files in this file system. The permissions can be checked and adjusted if necessary using the commands below:

```
/usr/sbin/mount /usr/sap/trans
/usr/bin/su - <sid>adm
/usr/bin/touch /usr/sap/trans/log/TEST
/usr/bin/chmod -R 775 /usr/sap/trans
/usr/bin/chgrp -R sapsys /usr/sap/trans
```

- ▶ There must also be sufficient free space available for transport log files or new exports of SAP R/3 objects. SAP R/3 support packages especially are usually large compressed files. After decompression, they need considerably more space, but it saves a lot of space to delete the data files of successfully applied support packages. These files are also stored in the /usr/sap/trans/data directory. Use the following commands to check on the available space and to delete the appropriate data files:

```
/usr/bin/df /usr/sap/trans
/usr/bin/rm /usr/sap/trans/data/*<release><pkg_number>.SAP
```

14.4 Tools

This section contains SAP R/3, DB2, and Oracle tools in alphabetical order, which are useful for locating of faulty components.

Attention: The description given for each tool is only valid in the context of the relevant flow chart. It does not cover *all* possible situations. The descriptions are tailored to the situations that are outlined in the flow charts (see Figure 14-2 on page 441 and Figure 14-3 on page 442).

14.4.1 Active listener process [lsnrctl] (Oracle only)

In order to connect to the database, SAP R/3 work processes have to contact the Oracle listener first. If the listener is not active or does not provide a service handle for this SID, SAP R/3 work processes cannot connect to the database and therefore will not start. An output of a accurate working listener process is shown in Example 14-29. Use the following commands to display the status of the Oracle listener:

```
/usr/bin/su - ora<sid>  
/oracle/<SID>/<version>/bin/lsnrctl status
```

Further information can be gathered from the
/oracle/<SID>/<version>/network/log/listener.log log file.

Example 14-29 Output of lsnrctl command

```
Connecting to (ADDRESS=(PROTOCOL=IPC) (KEY=ZMB.WORLD))  
STATUS of the LISTENER  
-----  
Alias                LISTENER  
Version              TNSLSNR for IBM/AIX RISC System/6000: Version 8.0.5  
Start Date           03-JAN-00 12:10:39  
Uptime                72 days 0 hr. 52 min. 6 sec  
Trace Level          off  
Security              OFF  
SNMP                  ON  
Listener Parameter File /oracle/ZMB/network/admin/listener.ora  
Listener Log File     /oracle/ZMB/network/log/listener.log  
Services Summary...  
  ZMB                  has 1 service handler(s)  
The command completed successfully
```

14.4.2 Allocatable memory [memlimits]

Every SAP R/3 instance uses shared memory segments and allocates, if necessary, additional so called private memory from the heap. The sum of all shared memory segments and all private memory areas must not exceed the available operating system virtual memory. Virtual memory is limited to the size of real memory plus paging space.

Attention: Never use memlimits on a system where an SAP R/3 instance is active. It may crash the instance or produce wrong results.

- The `memlimits` tool checks the upper limit of usable main memory for the instance. If you ignore errors of `memlimits`, the instance may be unstable and abort transactions because of a lack of paging space (see Example 14-30). An increased AIX paging space avoids this error according to the SAP R/3 Installation Guide. To check the memory limits, run the `memlimits` command:

```
/sapmnt/<SID>/exe/memlimits
```

Example 14-30 Output of memlimits command

R/3 parameter em/initial_size_MB up to 584 permitted

```
Check the maximum address space per process usable  
both by process local memory and mapped file  
Maximum address space ( mmap(584 MB)+ malloc(4MB) ): 588MB
```

	Result
Maximum heap size per process.....:	256 MB
Maximum mapped file size (mmap).....:	584 MB
this value is probably limited by swap space	
Maximum protectable size (mprotect)..:	584 MB
em/initial_size_MB > 584 MB will not work	
Maximum address space per process....:	588 MB
this value is probably limited by swap space	
Total available swap space.....:	584 MB
*** ERROR =>	swap space too small, expect problems
	main memory size x 3 recommended , minimum 1 GB

- ▶ If **memlimits** does not return any error and transactions still abort, check the following process trace files for further information:
 - /usr/sap/<SID>/<instance>/work/dev_w*
 - /usr/sap/<SID>/<instance>/work/dev_ms
 - /usr/sap/<SID>/<instance>/work/dev_disp

14.4.3 Available logon groups [lgtst]

This task checks the availability of logon groups for a certain SAP R/3 system. The command output shows all currently active logon groups and all offered services, such as spool, update, or batch. All configured logon groups and services as well as all application servers should be active. These items are shown in Example 14-31 as bold marked items. The syntax of the relevant command is:

```
/sapmnt/<SID>/exe/lgtst -H <CI_node> -S sapms<SID>
```

Further information can be gathered from the /sapmnt/<SID>/exe/dev_lg log file.

Example 14-31 Output of the lgtst command

```
list of reachable application servers
-----
[node21_PRD_02] [node21] [9.23.20.21] [sapdp02] [3202] [DIA VB BTC SPO VB2 ]
[node22_PRD_02] [node22] [9.23.20.22] [sapdp02] [3202] [DIA VB BTC SPO VB2 ]
[node22_PRD_01] [node22] [9.23.20.22] [sapdp01] [3201] [DIA VB BTC SPO VB2 ]
[node21_PRD_01] [node21] [9.23.20.21] [sapdp01] [3201] [DIA VB BTC SPO VB2 ]
[node20_PRD_00] [node20] [9.23.20.20] [sapdp00] [3200] [DIA VB ENQ BTC SPO

list of selectable login-classes with favorites
-----
[Controlling] [9.23.20.21] [3201] [46C]
[MaterialMa] [9.23.20.21] [3201] [46C]
[SPACE]       [9.23.20.21] [3201] [46C]
```

14.4.4 Check SAP profiles and logs [sappfpar]

The period of time between a modification and the next start of the instance can be long. Modifications of SAP R/3 profiles only become effective at the next start of the instance. Incorrect profile changes might prevent a correct instance start. To avoid this case all profiles should be checked right after a modification with the **sappfpar** tool, which reports incorrect parameters, as shown in Example 14-32 on page 468.

- If you run into that problem, there are three ways to fix the problem. If possible, activate a previous version of the profile with SAP R/3 transaction RZ10. You can also try to manually correct the wrong parameters. Otherwise, restore the last version of the profiles from backup storage, such as TSM.

Use the following commands to check the profiles after a modification:

```
/sapmnt/<SID>/exe/sappfpar pf=/sapmnt/<SID>/profile/DEFAULT.PFL
check /sapmnt/<SID>/exe/sappfpar \
pf=/sapmnt/<SID>/profile/<SID>_<Instance>_<hostname> check
```

Further information can be gathered from the
/usr/sap/<SID>/<instance>/work/dev_w* log file.

Example 14-32 Output of sappfpar command

```
Shared memory resource requirements estimated
=====
Nr of shared memory descriptors required for
Extended Memory Management (unnamed mapped file)..: 1
Total Nr of shared segments required.....: 5
System-imposed number of shared memories..: 11
Shared memory segment size required min..: 54000000 ( 51.5 MB)
System-imposed maximum segment size.....: 2147483648 (2048.0 MB)
R/3-imposed maximum segment size.....: 2147483647 (2048.0 MB)

Swap space requirements estimated
=====
Shared memory.....: 271.2 MB
..in pool 10 16.1 MB, 304% used !!
..in pool 40 13.6 MB, 1579% used !!
..not in pool 0.4 MB
Processes.....: 32.2 MB
Extended Memory .....: 128.0 MB
-----
Total, minimum requirement.....: 431.5 MB
Process local heaps, worst case..: 762.9 MB
Total, worst case requirement....: 1194.4 MB

Errors detected.....: 2
Warnings detected.....: 0
```

- If **sappfpar** finishes with return code zero, check the following process trace files for further information:
 - /usr/sap/<SID>/<instance>/work/dev_w*
 - /usr/sap/<SID>/<instance>/work/dev_ms
 - /usr/sap/<SID>/<instance>/work/dev_disp

14.4.5 Cleanup shared memory [showipc and cleanipc]

This task handles orphaned shared memory elements. Semaphores or shared memory segments should not remain in existence after the SAP R/3 instance has been shut down.

Attention: Never use the **cleanipc** command on an active instance!

The steps in the following list should be executed sequentially:

- The existence of orphaned elements prevents the restart of an instance. In that case, these elements must be removed manually by executing the following commands:

```
/sapmnt/<SID>/exe/showipc <instance_number>  
/sapmnt/<SID>/exe/cleanipc <instance_number>
```

- If the **showipc** command shows remaining elements for the AIX user ora<sid> and <sid>adm after the **cleanipc** command was run, use the operating system command **ipcrm** to remove these parts:

```
/usr/sbin/ipcs -ms | /usr/bin/grep -E 'db2<sid>|ora<sid>|<sid>adm'  
/usr/sbin/ipcrm -m <id>  
/usr/sbin/ipcrm -s <id>
```

Attention: Remaining database related shared memory segments and semaphores must be removed manually by using the **ipcrm** command.

14.4.6 SAP R/3 connection to database [R3trans]

The SAP R/3 tool **R3trans** checks whether the database is running and the SAP R/3 instance can connect to the database.

R3trans (DB2)

Use the following list to check the environment and shared libraries for the SAP R/3 work processes:

- The **R3trans** command connects to the database and should finish with return code 0. A higher value or further messages indicates that an error has occurred. In this case, read the **R3trans** log file, as shown in Example 14-33. The syntax of the relevant command is:

```
/usr/bin/su - <sid>adm  
/sapmnt/<SID>/exe/R3trans -d
```

Further information can be gathered from the /home/<sid>adm/trans.log log files.

Example 14-33 Output of the R3trans command

```
This is R3trans version 6.05 (release 46C - 25.04.00 - 13:20:00).  
2EETW169 no connect possible: "DBMS = DB6    --- DB2DBDFT = 'SLZ'"  
R3trans finished (0012).
```

- The **R3trans** tool connects to the database and tries to create a new DB2 agent process. If this creation fails, **R3trans** also fails. The log file trans.log contains detailed information and, usually, a reason for failing (see

Example 14-34). You can also refer to the available online help for database error codes (see “Online help” on page 471 for further reference).

Example 14-34 Log file trans.log

```

4 ETW000 R3trans version 6.05 (release 46C - 25.04.00 - 13:20:00).
4 ETW000 =====
4 ETW000 date&time   : 01.08.2001 - 14:32:43
...
4 ETW000 [developertrace,0] *** ERROR in DB6Connect[dbdb6.c, 1291]
4 ETW000 [developertrace,0] &+ 0|      |      |      DB6Connect( SQLConnect
SQLSTATE=57030: [IBM][CLI Driver] SQL1226N  The maximum number of coordinating
agents are already started.  SQLSTATE
4 ETW000 [developertrace,0] &+ 0|      |      |      =57030
4 ETW000 [developertrace,0] *** ERROR in DB6Connect[dbdb6.c, 1291]
4 EETW169 no connect possible: "DBMS = DB6      --- DB2DBDFT = 'SLZ'"

```

- Increase the maximum number of agent processes by adjusting the database manager parameter:

```

/usr/bin/su - db2<sid>
/db2/<SID>/sqllib/bin/db2 update db manager cfg using maxagents <nn>

```

- Another possible problem is an insufficient number of maximum processes on AIX. This value should be set according to the SAP R/3 installation manual to at least 500 using the following commands to list and adjust the maxuproc parameter:

```

/usr/sbin/lssattr -El sys0 -a maxuproc
/usr/sbin/chdev -l sys0 -a maxuproc=500

```

R3trans (Oracle)

Use the following list to check the environment and shared libraries for the SAP R/3 work processes:

- The **R3trans** command connects to the database and should finish with return code 0. A higher value or further messages indicates that an error has occurred. In this case, read the R3trans log file as shown in Example 14-35. The syntax of the relevant command is:

```

/usr/bin/su - <sid>adm
/sapmnt/<SID>/exe/R3trans -d

```

Further information can be gathered from the /home/<sid>adm/trans.log log file.

Example 14-35 Output of the R3trans command

```

This is R3trans version 6.05 (release 45B - 08.04.99 - 13:23:00).
2EETW169 no connect possible: "DBMS = ORACLE --- dbs_ora_tnsname = 'ZMB'"
R3trans finished (0012).

```

- The **R3trans** tool connects to the database and tries to create a new shadow process. If the connect to the listener process or the creation of a new shadow process fails, **R3trans** fails also. The log file `trans.log` contains detailed information and, usually, the reason for the failure (see Example 14-36 for details). For database error codes, online help is available (see “Online help (Oracle)” on page 472 for further information).

Example 14-36 Log file `trans.log`

```

4 ETW000 R3trans version 5.34 (release 31H - 01.08.97 - 16:30:12).
4 ETW000 =====
4 ETW000
4 ETW000 control file: <no ctrlfile>
4 ETW000 date&time   : 01.08.2001 - 13:55:00
Wed Aug  1 13:55:11 2001
***LOG BY2=>  sql error 20 performing CON [dblnk   0488]
***LOG BY0=>  [dblnk   0488]
2EETW169 no connect possible: "DBMS = ORACLE --- dbs_ora_tnsname = 'MUC'"

```

- If necessary, increase the maximum number of shadow processes by adjusting the parameter in the file `init<SID>.ora` as follows:

```

/usr/bin/vi /oracle/<SID>/<version>/dbs/init<SID>.ora
processes = <number_of_SAP_work_process + 10>
sessions  = <processes> * 1.2

```

- Another possible problem is an insufficient number of maximum processes on AIX. The minimum value, according to the SAP R/3 installation manual is 500. Under certain circumstances (refer to Section 11.4.3, “Asynchronous disk I/O” on page 338), specify a higher value. The following commands list and adjust the `maxuproc` parameter:

```

/usr/sbin/lssattr -El sys0 -a maxuproc
/usr/sbin/chdev -l sys0 -a maxuproc=500

```

14.4.7 Online help

An online help facility is especially useful in error situations, because it delivers error reasons, and sometimes tasks, to solve the situation.

Online help (DB2)

The DB2 diag log file, as well as the work process log files often only contain short messages and codes. The following list shows how you can get more information:

- Codes with a 5-digit message number are called DB2 message codes or SQL state. The DB2 message code 57030, for example, contains the message text shown in Example 14-37 on page 472. The following command displays

the text for this DB2 message number. It is important to include any leading zeros. The syntax of the relevant command is:

```
/usr/bin/su - db2<sid>
/db2/<SID>/sqllib/bin/db2 ? <message_number>
/db2/SLZ/sqllib/bin/db2 ? 57030
```

Example 14-37 Output of db2 command (Message number)

SQLSTATE 57030: Connection to application server would exceed the installation-defined limit.

- ▶ The DB2 SQL error codes of the failing statement are very useful, because the online help gives a very detailed explanation what happened and some recommendations to solve the problem. Example 14-38 shows, for instance, the help text for the SQL error code -289. It is important to use a 4-digit format and to include any leading zeroes. The syntax of the relevant command is:

```
/usr/bin/su - db2<sid>
/db2/<SID>/sqllib/bin/db2 ? SQL<sql_error>
/db2/SLZ/sqllib/bin/db2 ? SQL0289
```

Example 14-38 Output of db2 command (SQL error code)

SQL0289N Unable to allocate new pages in table space
" <tablespace-name>".

Explanation: One of the following conditions is true:

...

2. All the containers assigned to this DMS table space are full.

This is the likely cause of the error.

3. The table space object table for this DMS table space is full.

...

Details can be found in the system error log and/or the database manager error log.

User Response: Perform the action corresponding to the cause of the error:

...

2. add new container(s) to the DMS table space and try the operation again, after the rebalancer has made the new pages available

...

sqlcode: -289

sqlstate: 57011

Online help (Oracle)

The Oracle alert log file, as well as the work process log files, often only contain short messages with a certain message number. The Oracle error 0020, for example, contains the message text shown in Example 14-39 on page 473.

The following command displays the text for these Oracle message numbers:

```
/usr/bin/su - ora<sid>  
/oracle/<SID>/<version>/bin/oerr ora <errcode>
```

Example 14-39 Output of oerr command

```
00020, 00000, "maximum number of processes (%s) exceeded"  
// *Cause: All process state objects are in use  
// *Action: Increase maximum processes - init.ora parameter "processes"
```

14.4.8 Problems with brarchive and brbackup

The **brarchive** and **brbackup** tools are used to store the database and their log files in the backup system, which is at least a tape drive. Our solution models use TSM as backup system. The error situations below refer to this environment. The following list contains steps to locate the reason for the error situation:

- If **brbackup** or **brarchive** have been killed without cleaning up their lock files after that, all subsequent **brbackup/brarchive** operations terminate with return code 3. Check the current **brarchive** log file (see Example 14-40). The syntax of the relevant commands is:

```
/usr/bin/pg $( ls -tr /oracle/<SID>/saparch/a* | /usr/bin/tail -1 )  
/usr/bin/pg $( ls -tr /oracle/<SID>/sapbackup/b* | /usr/bin/tail -1 )
```

Further information can be gathered from the following log files:

- /oracle/<SID>/saparch/a*
- /oracle/<SID>/sapbackup/b*

Example 14-40 Log file of brarchive

```
BR020E BRARCHIVE already running or was killed.  
BR021I Please delete the file /oracle/MUC/saparch/.lock.bra if BRARCHIVE was  
killed  
BR022E Setting BRARCHIVE lock failed.  
BR007I End of offline redo log processing: adfwapj cps 2001-07-31 17.52.3  
BR005I BRARCHIVE terminated with errors.
```

- In the case described above, remove the lock files manually and run **brarchive** or **brbackup** again. It is always a good idea to check whether these processes exist or not before removing any lock files:

```
/usr/bin/ps -ef | /usr/bin/grep -E 'brbackup|brarchive'  
/usr/bin/rm /oracle/<SID>/saparch/.lock.bra  
/usr/bin/rm /oracle/<SID>/sapbackup/.lock.brb  
/usr/bin/rm /db2/<SID>/saparch/.lock.bra
```

- If the archive file system saparch is full, the **brarchive** process terminates with return code 5, because it cannot write its own log file. As a result, no log

file is created, and a detailed output of the error messages is only visible for the caller program. In this case, you only have the chance to see the error messages if you execute **brarchive** or **brbackup** from the command line.

- ▶ If the reason for the termination of **brarchive** command was a full file system, there are two possibilities. Either move some files out of the file system or set a temporary environment variable to redirect the log file. Then run **brarchive** again. You may add the variable SAPARCH to the following file:

```
/usr/bin/vi /home/<sid>adm/.cshrc
setenv SAPARCH $HOME/tmp
```

- ▶ Schedule a regular archiving of offline redo logs, as defined in your backup and recovery concept, to prevent similar errors in the future.
- ▶ The log files of **brarchive** and **brbackup** contain some useful information, if there are any existing problems regarding the TSM server or backint configuration, as shown in Example 14-41. The **brbackup** tool is not applicable in conjunction with DB2, because these backups go directly into TSM:

```
/usr/bin/pg $( ls -tr /oracle/<SID>/saparch/a* | /usr/bin/tail -1 )
/usr/bin/pg $( ls -tr /oracle/<SID>/sapbackup/b* | /usr/bin/tail -1 )
```

Example 14-41 Log file of brarchive command

```
BKI0008I: Number of bytes to save: '218.620 MB'.
BKI0026I: Time: 16:31:20 Object: 1 of 11 in process: /oracle/MUC/saparch/MU
1_2120.dbf Size: 20.000 MB. MNGM-Class: DEFAULT, ADSM server: TERRA_DB, Red
ANS1329S (RC29) Server out of data storage space
BKI1125W: The object '/oracle/MUC/saparch/MUCarch1_2120.dbf (1.Copy)' will
tried [2].
BKI0026I: Time: 16:31:39 Object: 1 of 11 in process: /oracle/MUC/saparch/MUCar
ANS1329S (RC29) Server out of data storage space
BKI1125W: The object '/oracle/MUC/saparch/MUCarch1_2120.dbf (1.Copy)' will
tried [3].
BKI0026I: Time: 16:31:58 Object: 1 of 11 in process: /oracle/MUC/saparch/MU
1_2120.dbf Size: 20.000 MB. MNGM-Class: DEFAULT, ADSM server: TERRA_DB, Red
ANS1329S (RC29) Server out of data storage space
BR233E Error in backup utility while saving file /oracle/MUC/saparch/MUCarch1_
```

14.5 AIX commands

This section contains AIX commands (in alphabetical order) that are useful for locating faulty components.

Attention: The description given for each command is only valid in the context of the relevant flow chart. It does not cover *all* possible situations. The descriptions are tailored to the situations that are outlined in the flow charts (see Figure 14-2 on page 441 and Figure 14-3 on page 442).

14.5.1 Connectivity check [ping]

If there is a problem with the network, execute the **ping** command to an appropriate TCP/IP address. The steps below should be executed as long as a probable error reason is found:

- If response packets are received, at least the physical network layer works correctly and there is no routing problem. If you try to determine SAP R/3 and database related problems, the TCP/IP address <addr> is the address of an adapter in the front-end network. In case of a failing backup task <addr>, refer to an adapter in the backup network. Execute the following command:

```
/usr/bin/ping <addr>
```

- A further check is to transfer a huge amount of data to check all network layers, including performance critical settings and parameters. For this step, it is recommended to have appropriate reference values in place. The symbol <if> means the AIX name for the interface with the TCP/IP address <addr>. The **netstat** command shows the network throughput in packages per second, as shown in Example 14-42.

Execute the following command to issue the data transfer:

```
/usr/bin/rsh <addr> dd if=/dev/kmem 0<&- | /usr/bin/dd of=/dev/null &
```

Execute the following command to display the transfer statistics:

```
/usr/bin/rsh <addr> '/usr/bin/netstat -I <if> 1'
```

Example 14-42 Output of the netstat command

input	(en2)	output	input	(Total)	output
packets	errs	packets errs colls	packets	errs	packets errs colls
7512	0	2679 0 0	7517	0	2680 0 0
7580	0	2695 0 0	7616	0	2723 0 0
7762	0	2759 0 0	7839	0	2834 0 0
7615	0	2716 0 0	7618	0	2717 0 0
7640	0	2721 0 0	7645	0	2722 0 0
7611	0	2710 0 0	7620	0	2714 0 0
7723	0	2746 0 0	7732	0	2747 0 0
7539	0	2683 0 0	7643	0	2785 0 0
7526	0	2682 0 0	7535	0	2683 0 0

14.5.2 Login to a host

Before executing any command, a **telnet** logon to the appropriate host is necessary. The following steps are possible:

- Pay attention within clustered environments. If you want to log in to a certain service use the address on the front-end network to which the service is bound; otherwise, you may connect to the wrong node.

The abbreviations for target hosts that are used in the flow charts translate to the following:

- DB_node** Database host
- CI_node** Host where the SAP R/3 central instance is installed
- AP_node** Application server

- Every host in the models that are described in this book has an interface to the control network (See Chapter 6, “Network” on page 149). Try the IP address of this interface to log in for administrative tasks on a certain node. If a connect to a certain service via the front-end network fails, try to connect via the control network. Within a clustered environment, check all possible notes in the cluster for the service, before you draw any conclusions.

14.5.3 Performance problems [vmstat]

Sometimes a system administrator hears complaints about poor response times of the system. A system is usually limited either by I/O throughput, memory capacity, or CPU performance. Chapter 11, “Performance” on page 305 contains detailed information on performance tuning. Use the following steps to determine possible reasons for poor response times:

- Check the average paging rate per minute. This value can be measured with the **vmstat** command. The columns **pi** and **po** show the current paging, as can be seen in Example 14-43. If there is a high paging rate for more than five minutes, there is a problem. In that case, check the AIX parameters described in the next step and adjust them if necessary. You can find further information in Section 11.5.1, “The vmstat command” on page 345. The syntax of the relevant command is:

```
/usr/sbin/vmstat 10
```

Example 14-43 Output of the vmstat command (memory constraint)

kthr		memory			page				faults				cpu			
r	b	avm	fre	re	pi	po	fr	sr	cy	in	sy	cs	us	sy	id	wa
1	5	3788039	245	0	70	93	1638	34272	0	2131	9205	3921	2	11	47	40
0	4	3788039	262	0	57	119	434	8105	0	1930	5238	1766	0	4	75	21
0	6	3788039	244	0	58	101	964	25034	0	2167	13396	4728	2	2	60	37

0	10	3788039	24	0	83	59	1683	28003	0	2174	8696	4454	3	2	51	44
2	10	3788044	247	0	52	180	1486	35867	0	2139	6941	3770	5	10	35	49
2	6	3788044	243	0	38	190	994	20806	0	2152	6140	2753	10	11	43	36
1	6	3788044	152	0	26	224	985	9582	0	2083	7601	2099	9	6	50	35
1	5	3788044	217	0	17	242	743	14459	0	2258	6824	2990	5	2	65	27

- The amount of memory which can be used for file caching is important for a database server. If this value is too high, the database shared memory segments interfere with the file cache in a counterproductive fashion. The setting below limits the file cache to 10 percent of total memory. You can find further information for the `vmtune` command in Section 11.5.2, “The vmtune command” on page 347. The syntax of the relevant command is:

```
/usr/samples/kernel/vmtune -p 5
/usr/samples/kernel/vmtune -P 10
```

- Another point to consider are the I/O operations. One sequential read access can get 16 (default 8) blocks from disks. You can find further about this topic in “Recommendations for an SAP R/3 system” on page 343. The syntax of the relevant command is:

```
/usr/samples/kernel/vmtune -r 2
/usr/samples/kernel/vmtune -R 16
```

- The CPU utilization is calculated as the sum of user and system time. This value can be measured with the `vmstat` command. The columns `us` and `sy` show the current CPU utilization, as shown in Example 14-44. If this value is above 95 percent for more than five minutes, the CPU may be the limiting component. The syntax of the relevant command is:

```
/usr/sbin/vmstat 10
```

Example 14-44 Output of the `vmstat` command (CPU constraint)

kthr		memory			page				faults				cpu			
r	b	avm	fre	re	pi	po	fr	sr	cy	in	sy	cs	us	sy	id	wa
8	1	256367	7136	0	0	0	0	0	0	833	18180	625	32	66	2	0
9	1	256367	7136	0	0	0	0	0	0	837	18088	622	33	67	0	0
8	1	256368	7125	0	0	0	0	0	0	850	18480	615	29	71	0	0
8	1	256368	7125	0	0	0	0	0	0	840	19743	489	24	76	0	0
7	1	256368	7125	0	0	0	0	0	0	830	20029	458	25	75	0	0
8	1	256368	7125	0	0	0	0	0	0	835	19976	457	25	75	0	0
9	1	256827	6458	0	0	0	0	0	0	856	18536	795	30	70	0	0
9	1	257218	5957	0	0	0	0	0	0	917	17427	966	32	68	0	0
11	1	258734	4198	0	0	0	0	0	0	906	15476	1181	41	59	0	0
10	1	258887	3907	0	0	0	0	0	0	893	14810	917	44	56	0	0

- SAP R/3 uses semaphores to avoid interfering tasks on application instances. These semaphores are set in conjunction with predefined operations from

any transaction, for example, to update a financial record or write on file on the operating system level. If a semaphore remains active for more than a minute, the reason for that should be discovered by an SAP R/3 specialist.

- ▶ The database request time should have the right proportion to the average response time. If this ratio is lower than 1:2, there may be a problem with the database performance. Repeat the steps in this section on the database server.
- ▶ If neither of these tuning hints or a better organization of the daily SAP R/3 workload solves the problem, assign the problem situation to a Certified SAP R/3 Basis Consultant or contact SAP for an equivalent service, such as SAP Early Watch. The last chance is the procurement of new servers.

Related publications

The publications listed in this section are considered particularly suitable for a more detailed discussion of the topics covered in this redbook.

IBM Redbooks

For information on ordering these publications, see “How to get IBM Redbooks” on page 483.

- ▶ *AIX Version 4.3 Differences Guide*, SG24-2014
- ▶ *AIX 5L Differences Guide*, SG24-5765
- ▶ *AIX 5L Performance Tools Handbook*, SG24-6039
- ▶ *AIX 5L Workload Manager (WLM)*, SG24-5977
- ▶ *Backing Up DB2 Using Tivoli Storage Manager*, SG24-6247
- ▶ *Database Performance on AIX in DB2 UDB and Oracle Environments*, SG24-5511
- ▶ *Designing an IBM Storage Area Network*, SG24-5758
- ▶ *Getting Started with Tivoli Storage Manager: Implementation Guide*, SG24-5416
- ▶ *HACMP/ES Customization Examples*, SG24-4498
- ▶ *IBM SAN Survival Guide*, SG24-6143
- ▶ *Implementing ADSM Server-to-Server Configurations*, SG24-5244
- ▶ *Inside the RS/6000 SP*, SG24-5145
- ▶ *Introduction to Storage Area Network, SAN*, SG24-5470
- ▶ *IP Network Design Guide*, SG24-2580
- ▶ *Migrating to HACMP/ES*, SG24-5526
- ▶ *NIM: From A to Z in AIX 4.3*, SG24-5524
- ▶ *Planning and Implementing an IBM SAN*, SG24-6116
- ▶ *PSSP Version 3 Survival Guide*, SG24-5344
- ▶ *R/3 Data Management Techniques Using Tivoli Storage Manager*, SG24-5743
- ▶ *RS/6000 ATM Cookbook*, SG24-5525

- ▶ *RS/6000 SP Software Maintenance*, SG24-5160
- ▶ *RS/6000 SP System Performance Tuning Update*, SG24-5340
- ▶ *RS/6000 SP and Clustered @server pSeries Systems Handbook*, SG24-5596
- ▶ *RS/6000 Systems Handbook 2000 Edition*, SG24-5120
- ▶ *RS/6000 SP Cluster: The Path to Universal Clustering*, SG24-5374
- ▶ *Tivoli Storage Management Concepts*, SG24-4877
- ▶ *Understanding SSA Subsystems in Your Environment*, SG24-5750
- ▶ *Understanding and Using the SP Switch*, SG24-5161
- ▶ *Using ADSM to Back Up Databases*, SG24-4335
- ▶ *Workload Management: SP and Other RS/6000 Servers*, SG24-5522

Other resources

These publications are also relevant as further information sources:

- ▶ *AIX V4.3 Network Installation Management Guide and Reference*, SC23-4113
- ▶ *SCSI Tape Drive, Medium Changer, Library Device Drivers User's Guide*, GC35-0154
- ▶ *RS/6000 Adapters, Devices, and Cable Information for Multiple Bus Systems*, SA38-0516
- ▶ *PSSP: Installation and Migration Guide*, GA22-7347
- ▶ *PSSP: Diagnosis Guide*, GA22-7350
- ▶ *Tivoli Storage Manager for AIX Administrator's Guide Version 4 Release 2*, GC35-0403
- ▶ *Tivoli Storage Manager SAN Tape Library Sharing*, REDP0024
- ▶ "Configuring and Tuning IBM Storage Systems In an Oracle Environment", IBM/Oracle International Competency Center, Version 1.0, May 25, 1999
- ▶ *IBM @server pSeries Clustered Computing*, White Paper, June 2001
- ▶ *IBM SAP Marketing White Paper: 64 Bit SAP R/3 on the RS/6000*, ISICC, Version 1.3, April 2001
- ▶ *Network Load for Release 4.6*, SAP Document, Version 2.5, January 2000
- ▶ Costales, et al, *sendmail*, O'Reilly and Associates, Inc., 1997, ISBN 1565922220
- ▶ *Oracle8 Backup and Recovery Handbook* (comes with product)

- ▶ *Performance Management Guide, AIX 5L Version 5.1, Second Edition*, April 2001 (part of AIX 5L online documentation)
- ▶ *R/3 Heterogeneous System Copy, Release 4.6D*, Material number 51010925 (English) (comes with product)
- ▶ *R/3 Homogeneous System Copy, Release 4.6D*, Material number 51010924 (English) (comes with product)
- ▶ *RS/6000 SP: SP Switch Performance*, August 1999, Version 3 (IBM White Paper)
- ▶ *SAP Technical Infrastructure - Network Integration of SAP Servers*, SAP Document, June 2001
- ▶ *SAP R/3 in Switchover Environments*, SAP Document, September 1999
- ▶ *Orb, AIX Mirror Write Consistency with Oracle Databases*, found at:
<http://dscrs6k.aix.dfw.ibm.com>
- ▶ *Backup & Restore Concepts for mySAP.com System Landscapes*, found at:
<http://service.sap.com/atg>
- ▶ *Database Layout for SAP Installations with DB2 UDB for Unix and Windows*, found at:
<http://service.sap.com>
- ▶ *Database Layout for R/3, Installations under ORACLE*, 50038473, found at:
<http://service.sap.com>
- ▶ *Braden, Disk Sizing, Data Layout and Tuning for AIX - Presentations*, found at:
<http://dscrs6k.aix.dfw.ibm.com/>
- ▶ *IBM eServer Clustered Computing White Paper*, found at:
<http://www-1.ibm.com/servers/de/eserver/pseries/library/specsheets>
- ▶ *IBM Magstar Tape Drives -- AIX High Availability SAN Failover for 3590*, found at:
http://www.storage.ibm.com/hardsoft/tape/3590/prod_data/magstarwp.pdf
- ▶ *Adams III, IBM Subsystem Device Driver / Data Path Optimizer on an ESS*, found at:
<ftp://ftp.software.ibm.com/storage/subsystem/tools>
- ▶ *SAP R3 in Switchover Environments*, Document 50020596, found at:
<http://service.sap.com/systemmanagement>

- ▶ Diether, *SAP R/3 and HACMP Setup and Implementation*, found on the IBM intranet at:
<ftp://sicc980.isicc.de.ibm.com/perm/hacmp>
- ▶ *The IBM @server pSeries 680: Reliability, Availability, Serviceability*, found at:
http://www.ibm.com/servers/eserver/pseries/hardware/whitepapers/p680_reliability.html

Referenced Web sites

These Web sites are also relevant as further information sources:

- ▶ <http://service.sap.com/> (requires registration)
- ▶ <http://service.sap.com/instguides> (requires registration)
- ▶ <http://service.sap.com/network> (requires registration)
- ▶ <http://service.sap.com/notes> (requires registration)
- ▶ <http://service.sap.com/osdbmigration> (requires registration)
- ▶ <http://service.sap.com/platforms> (requires registration)
- ▶ <http://service.sap.com/rrr> (requires registration)
- ▶ <http://service.sap.com/systemmanagement> (requires registration)
- ▶ <http://www.ibm.com/servers/aix/products/aixos/specs/index.html>
- ▶ <http://www.ibm.com/servers/eserver/pseries>
- ▶ http://www.ibm.com/servers/eserver/pseries/library/hardware_docs/options.html
- ▶ <http://www.mesa.nl/pub/monitor>
- ▶ <http://www.sap.com/benchmark>
- ▶ <http://www.storage.ibm.com/hardsoft/products/7133/7133-spec.htm>
- ▶ <http://www.storage.ibm.com/hardsoft/products/ess/ess.htm>
- ▶ <http://www.storage.ibm.com/hardsoft/products/expplus/expplus-spec.htm>
- ▶ <http://www.storage.ibm.com/hardsoft/products/ssa/>
- ▶ <http://www.storage.ibm.com/hardsoft/tape/3494/index.html>
- ▶ <http://www.storage.ibm.com/hardsoft/tape/3584/index.html>
- ▶ http://www.storage.ibm.com/hardsoft/tape/3590/prod_data/magstarwp.pdf
- ▶ <http://www.storage.ibm.com/snetwork/nas/>

- ▶ <http://www.storage.ibm.com/storage>
- ▶ <http://www-1.ibm.com/servers/de/eserver/pseries/library/specsheets>
- ▶ http://www-1.ibm.com/servers/eserver/pseries/hardware/system_perf.html
- ▶ http://www-1.ibm.com/servers/eserver/pseries/library/wp_systems.html
- ▶ <http://www.ibm.com/servers/aix>
- ▶ <http://aixpdslib.seas.ucla.edu/aixpdslib.html>
- ▶ <http://www.bull.de/pub/>

How to get IBM Redbooks

Search for additional Redbooks or redpieces, view, download, or order hardcopy from the Redbooks Web site:

ibm.com/redbooks

Also download additional materials (code samples or diskette/CD-ROM images) from this Redbooks site.

Redpieces are Redbooks in progress; not all Redbooks become redpieces and sometimes just a few chapters will be published this way. The intent is to get the information out much quicker than the formal publishing process allows.

IBM Redbooks collections

Redbooks are also available on CD-ROMs. Click the CD-ROMs button on the Redbooks Web site for information about all the CD-ROMs offered, as well as updates and formats.

Special notices

References in this publication to IBM products, programs or services do not imply that IBM intends to make these available in all countries in which IBM operates. Any reference to an IBM product, program, or service is not intended to state or imply that only IBM's product, program, or service may be used. Any functionally equivalent program that does not infringe any of IBM's intellectual property rights may be used instead of the IBM product, program or service.

Information in this book was developed in conjunction with use of the equipment specified, and is limited in application to those specific hardware and software products and levels.

IBM may have patents or pending patent applications covering subject matter in this document. The furnishing of this document does not give you any license to these patents. You can send license inquiries, in writing, to the IBM Director of Licensing, IBM Corporation, North Castle Drive, Armonk, NY 10504-1785.

Licensees of this program who wish to have information about it for the purpose of enabling: (i) the exchange of information between independently created programs and other programs (including this one) and (ii) the mutual use of the information which has been exchanged, should contact IBM Corporation, Dept. 600A, Mail Drop 1329, Somers, NY 10589 USA.

Such information may be available, subject to appropriate terms and conditions, including in some cases, payment of a fee.

The information contained in this document has not been submitted to any formal IBM test and is distributed AS IS. 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 the customer's operational environment. While each item may have been reviewed by IBM for accuracy in a specific situation, 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.

Any pointers in this publication to external Web sites are provided for convenience only and do not in any manner serve as an endorsement of these Web sites.

The following terms are trademarks of other companies:

Tivoli, Manage. Anything. Anywhere., The Power To Manage., Anything. Anywhere., TME, NetView, Cross-Site, Tivoli Ready, Tivoli Certified, Planet Tivoli, and Tivoli Enterprise are trademarks or registered trademarks of Tivoli Systems Inc., an IBM company, in the United States, other countries, or both. In Denmark, Tivoli is a trademark licensed from Kjøbenhavns Sommer - Tivoli A/S.

C-bus is a trademark of Corollary, Inc. in the United States and/or other countries.

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

Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States and/or other countries.

PC Direct is a trademark of Ziff Communications Company in the United States and/or other countries and is used by IBM Corporation under license.

ActionMedia, LANdesk, MMX, Pentium and ProShare are trademarks of Intel Corporation in the United States and/or other countries.

UNIX is a registered trademark in the United States and other countries licensed exclusively through The Open Group.

SET, SET Secure Electronic Transaction, and the SET Logo are trademarks owned by SET Secure Electronic Transaction LLC.

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

Abbreviations and acronyms

ABAP/4	Advanced Business Application Programming 4th Generation	ESS	Enterprise Storage Server
AIX	Advanced Interactive Executive	EWA	SAP EarlyWatch Alert
AP-Server	Application Server	FC	Feature Code
APAR	Authorized Program Analysis Report	FC	Fibre Channel
API	Application Programming Interface	FC-AL	Fibre Channel Arbitrated Loop
ARP	Address Resolution Protocol	FDDI	Fibre Distributed Data Interface
ATM	Asynchronous Transfer Mode	FI	Financial accounting
CCMS	SAP Computing Center Management System	FICON	Fiber Connection
CI	Central Instance	GUI	Graphical User Interface
CRM	Customer Relationship Management	HACMP	High Availability Cluster Multi-Processing
CTS	SAP Correction and Transport System	HACMP/ES	High Availability Cluster Multi-Processing Enhanced Scalability
CWS	Control Workstation	HTML	Hypertext Markup Language
C-SPOC	Cluster Single Point Of Control	HW	Hardware
DB	Database	I/O	Input/Output
DBA	Database Administration	IBM	International Business Machines Corporation
DB2 UDB	DB2 Universal Database	IDES	International Demo and Education System
DNS	Domain Name Services	IMG	SAP Implementation Guide for R/3 Customizing
DSS	Decision Support System	IOS	Internet Operating System
ECC	Error Checking and Correction	IP	Internet Protocol
EDI	Electronic Data Interchange	ISICC	IBM SAP International Competence Center
ERP	Enterprise Resource Planning	ISL	Inter Switch Link
ESCON	Enterprise Systems Connection	IT	Information Technology
		ITSO	International Technical Support Organization
		JBOD	Just a bunch of disk
		JFS	Journaled File System

km	Kilometer	RFC	Remote Function Call
LCD	Liquid Crystal Display	RFC	Request For Comments
LED	Light Emitting Diode	rpm	revolutions per minute
LPP	Licensed Program Products	RSCT	Reliable Scalable Cluster Technology
LSC	Low Speed Connection	SAA	System Administration Assistant
LV	Logical Volume	SAN	Storage Area Network
LVM	Logical Volume Manager	SAP	Systeme Anwendungen und Produkte
m	Meter	SAPS	SAP Application Benchmark Performance Standard
MIB	Management Information Base	SCSI	Small Computer System Interconnect
MM	Material Management	SCM	Supply Chain Management
MTU	Maximum Transfer Unit	SD	Sales and Distribution
MWC	Mirror Write Consistency	SDD	Subsystem Device Driver
NAS	Network Attached Storage	SDR	System Data Repository
NFS	Network File System	SID	SAP System Identification
NIM	Network Installation Manager	SLA	Service Level Agreement
NTP	Network Time Protocol	SMB	Server Message Block
NVRAM	Non Volatile Random Access Memory	SMIT	System Management Interface Tool
ODM	Object Data Manager	SMP	Symmetric Multiprocessor
OLTP	Online Transaction Processing	SNMP	Simple Network Management Protocol
OSS	Online Service System	SP	Scalable Power
PSSP	Parallel System Support Programs	SPoF	Single Point of Failure
PC	Personal Computer	SPOT	Shared Product Object Tree
PCI	Peripheral Component Interconnect	SSCR	SAP Software Change Registration
PP	Physical Partition	SSA	Serial Storage Architecture
PTF	Program Temporary Fix	STP	Spanning Tree Protocol
PVC	Permanent Virtual Circuit	SVC	Switched Virtual Circuit
RAID	Redundant Array of Independent Disk	TCO	Total Cost of Ownership
RAS	Reliability, Availability and Serviceability	TDP	Tivoli Data Protection
RC	Return Code	TMS	Transport Management System
RDBMS	Relational Database Management System		

<i>TSM</i>	Tivoli Storage Manager
<i>TX</i>	Transaction
<i>UNI</i>	User Network Interface
<i>UTP</i>	Unshielded Twisted Pair
<i>VLAN</i>	Virtual Local Area Network
<i>VMM</i>	Virtual Memory Manager
<i>WCOLL</i>	Working Collection
<i>WSM</i>	Web SMIT

Archived

Index

Symbols

{\$nopage}Interface Specific Network Options (IS-NO), see ISNO 180

Numerics

10/100 Mbps Ethernet PCI Adapter 55
2032-064 144
2104 Expandable Storage Plus 47
2108-G07 142
2109 148
32-bit 307
3494 54
3590 54
4-Port 10/100 Base-TX Ethernet 56
6227 146
6228 146
64-bit 67, 307
7133 Serial Disk System 48

A

ABAP dumps 425
Access characteristics 110
Access communication 161
Access network 70, 161
Active tablespaces 112
Adapter failure 231
Address space 307
Admin 73
Administrate LPP sources 283
Administration domains 19
Administration server 73
Administrative schedules 206
Administrator server 281
Advanced Business Application Programming (ABAP/4) 39
aioservers 338, 340
AIX
 Documentation 75
AIX buffer cache 121
AIX ES Extended Memory 325
AIX ES shared memory 325
AIX features 329

AIX file system buffer cache 121
AIX I/O processing 122
AIX installation 283
AIX kernel 307
AIX manual pages 282
AIX Workload Manager (WLM) 59
AL08 443
Analyzing disk access 114
Application
 Service level agreement 26
Application monitoring with HACMP/ES 227
Application server 39, 67
 Distribution 84
Application server instance 9
Application server instances 39
Architecture 7, 8
Archive 187, 188
Archived log files 191
Archiver stuck 79
ARP cache 294
Asset management
 Service level agreement 27
Asynchronous disk I/O 338
Asynchronous I/O 298
Asynchronous I/O failed 451
ate command 281
ATM
 Bandwidth 159
ATM connections 142
atmstat command 178
Audience of this redbook 2
autoconfig 339
Automatic takeover 225
Automation 19
automount attribute 286
Availability 12
 Basic model 77
 Definition 14
Availability considerations 69

B

backfm utility 193
Background processing 123, 302

- BACKINT 64
- backint 193
- backint command 384
- Backup 187
- Backup and recovery 185, 429
 - Concept 12
 - Service level agreement 26
 - Testing 220
- Backup and restore methods 196
- backup command 193
- backup db command 195
- Backup devices 204
- Backup network 71, 165
- Backup subsystem 71
- Backups 188
- Balance I/O 98
- Bandwidth requirements 151, 230
- Bandwidth SCSI technology 102
- Basic model 2, 65
- Batch (B) 40
- Batch processes 151, 153, 361
- Batch processing 10, 68
- Batch queues
 - Service level agreement 29
- Big volume group 126
- Blocking disk I/O 340
- Boot IP label 223, 233
- Boot logical volume 125
- bootlist command 126
- bosboot command 126
- brarchive command 72, 193, 220
 - Problems 473
- brbackup command 72, 220, 265
 - Problems 473
- brdb6brt command 379
- Bright idea 3
- Brocade SilkWorm 142
- brrestore command 72
- Buffer tuning 317
- Business logic 39
- Business process continuity 88
- Bypass-card 129

C

- Cabling ESS 129
- Cache mirror write consistency 124
- Cascading resource groups 234
- Catalyst Family 57

- CCMS 401
- Central instance 9, 66
 - Minimal configuration 66
- Central Instance (CI) 41
- Central mailbox 281
- Central system 9
- Change management
 - Service level agreement 27
- chdev command 177, 345
- Checkpoint not complete 454
- ChecksumOffload 179
- ChecksumTCP 179
- Chipkill memory 43
- chvg command 126, 129
- Cisco 57
- Classic modem 256
- Classic SP 44
- Clean up jobs 300
- cleanipc command 468
- Client copy 303
- Client schedules 205
- Client segments 332
- Cluster nodes 224
- Cluster software 225
- Cluster solution 80
- Cluster state 227
- Cluster topology 229
- Clustered Enterprise Servers 44
- Collocation 199
- Command line prompt 290
- Commands
 - ate 281
 - atmstat 178
 - backint 384
 - backup 193
 - backup db 195
 - bootlist 126
 - bosboot 126
 - brarchive 72, 193, 473
 - brbackup 72, 265, 473
 - brdb6brt 379
 - brrestore 72
 - chdev 177, 345
 - chvg 126, 129
 - cleanipc 468
 - cu 281
 - dscdb6up 390
 - dsh 288
 - dsmformat 212

- entstat 178
- errpt 358, 409
- exportfs 289
- extendvg 126
- fddistat 178
- filemon 349
- importvg 389
- iostat 348
- kinit 414
- klist 414
- lgtst 467
- lsattr 176, 341
- lscons 286
- lsdev 176, 341
- lslv 125
- lsnrctl 465
- mail 411
- memlimits 466
- migratepv 126
- mirrorvg 125
- mkdev 341
- mkitab 287
- mksysb 192, 283
- monitor 355
- netstat 289, 351, 412
- no 183
- ping 289, 475
- piodigest 293
- R3trans 469
- rcp 293
- rdist 287, 293
- replacepv 126
- restore 193
- restvg 196
- rexec 287
- rlogin 287
- rsh 287
- sapdba 265, 384
- saplicense 373
- sappfpar 313, 467
- savevg 196
- sendmail 411
- showipc 468
- svrmgrl 386, 391, 450
- swcons 286
- telnet 476
- tokstat 178
- topas 326, 353
- trcoff 349

- trcon 349
- trcstop 349
- truncate 392
- varyonvg 129
- vmstat 346, 476
- vmtune 121, 336, 347
- Common system landscape 10
- Compatibility
 - Storage subsystems 46
- Computational pages 335
- Concurrent maintenance 42
- Concurrent microcode update 130
- Concurrent users 152
- Configuration fi 291
- Configuration information 19
- Configuration snapshots 20
- Consistency check 273
- Consistent naming 18
- Console redirect 286
- Constraints 16
- Container size 119
- Control network 71, 166
- Control Workstation 280
- Control Workstation (CWS) 43, 184, 210
- Control workstation (CWS) 414, 433
- Copy group 201
- Copy pools 200
- Core dumps 20
- Cost structure 24
- cu command 281
- Customer Relationship Management (CRM) 42
- Customer service center
 - Service level agreement 27

D

- Data consistency
 - Ensuring 267
- Data file size 119
- Data striping 108
- Data transfer times 97
- Database
 - Documentation 75
- Database buffers 121
- Database consistency 264
- Database copy 368, 378
- Database errors 419
- Database free space 416
- Database growth

- Service level agreement 29
- Database layer 109
- Database layout 111
- Database layout DB2 119
- Database layout Oracle 120
- Database level actions 390
- Database performance data 113
- Database processes
 - DB2 452
 - Oracle 453
- Database reconnect 243
- Database redo logs 110
- Database server 9, 38, 66
- Database system
 - Monitoring 416
- Database system check 418
- Databases 62
- Daylight saving time 302
- DB administrative tasks 421
- DB02 374, 402, 416, 419
- DB13 400, 417
- DB2 Command Line Processor 379
- DB2 Control Center 63
- DB2 database layout 119
- DB2 recovery history files 193
- DB2 Universal Database 62
- db2diag.log 449
- db2uext.err 449
- DBR3CP.R3S 376
- DBshadow 269
- DBSNP 392
- DDLOG 392
- Decision Support Systems (DSS) 123
- Deferred paging space allocation 334
- Define Resource Groups 236
- Departments 22
- Design white papers 17
- Destage 109
- Development system 402
- df command 289
- Dialog (D) 40
- Dialog processes 151
- Direct disk access 120
- Directors 142
- Disaster safety 72
- Disaster-tolerant 2, 72
- Disaster-tolerant model 86, 252
- Disaster-tolerant system 14
- Disk adapter cache 108

- Disk cost efficiency 100
- Disk I/O time 97
- Disk interconnection 101
- Disk layout 98, 109
 - Example 113
- Disk majority (Quorum) 133
- Disk mechanic 97
- Disk pools 200
- Disk price performance ratio 101
- Disk random operations 97
- Disk sequential operations 98
- Disk sizing 98
- Disk storage 46, 69, 95
- Disk technical specifications 97
- Disk technologies 51
- Disk wearout 98
- Dispatcher 40, 327
- Distributing disk workload 114
- Documentation 20
- Documentation server 282
- Domain Name Service (DNS) 387
- Downtime 12, 14
- Drawer SSA 128
- dscdb6up command 390
- dsh command 288
- dsh script 288
- DSM_CONFIG 198
- DSM_DIR 198
- dsmformat command 212
- Dynamic Kernel 58

E

- Early paging space allocation 334
- Easy administration interface 279
- Efficiency 13
- Efficient disk layout 98
- Electronic Data Interchange 9
- Enqueue (E) 40
- Enqueue reconnect 243
- Enterprise Fabric Connectivity (EFC) 144
- Enterprise Fibre Channel Director 144
- Enterprise Resource Planning (ERP) 21, 42
- Enterprise servers 42
- Enterprise Storage Server 106
- entstat command 178
- Environment definitions 18
- Environmental constraints 17
- Error Checking and Correction (ECC) 42

- Error logging 59
- Error situation 14, 439
- errpt command 358, 409
- ESS
 - Bay 108
 - Bay adapter 108
 - Cabling 129
 - Concurrent microcode update 130
 - Configuration
 - Basic 129
 - Disaster-tolerant 134
 - Fault-tolerant 132
 - SAN 136
 - Data striping 108
 - Disk adapter cache 108
 - Eight-pack 108
 - Fast write cache 108
 - Fiber optic cables 134
 - FlashCopy 109, 118
 - Full stripe destage 109
 - Host bay adapter 108
 - Logical Volume sizing 117
 - Overview 107
 - Partition pattern 117
 - Price advantage 101
 - Rank 108
 - Read performance 108
 - Sample partitioning 118
 - Subsystem Device Driver (SDD) 130
 - Write preempt 108
- EtherChannel 164
- Ethernet
 - Bandwidth 159
 - Switches 174
- Event 227
- Event script 227
- Exclude file 196
- Executor 31
- Expandable Storage Plus 47
- exportfs command 289
- Extended Remote Copy 50
- Extended Shared Memory (ESM) 311
- extendednetstats 352
- Extender, Fibre Optical 103
- extendvg command 126
- EXTSHM 311, 315

F

- Fall back system 265
- Fast write cache 108
- Fault tolerance 42, 222
- Fault-tolerant 2, 72
- Fault-tolerant model 77
- Fault-tolerant system 13
- fddistat command 178
- Fiber Connection (FICON) 142
- Fiber optic cables (ESS) 134
- Fiber optic converters 255
- Fiber optics 102
- Fibre Channel
 - Arbitrated Loop 53
 - Bandwidth 102
 - Host Bay Adapters 147
 - Switch 142
 - Switched Fabric 103
- Fibre Channel (FC) 50, 102, 139, 269
- Fibre Channel technology 52
- Fibre optic cables 141
- Fibre Optical Extender 103
- File access characteristics 110
- File pages 335
- File synchronization 291
- File system buffer cache 121
- File system cache 121
- File system full 450, 459
- File systems 410
- File types 110
- filemon command 349
- FlashCopy 50, 109, 118, 272, 371
- flow_ctrl 178
- Flowcharts 439
- Forced synchronization 124
- Forced varyon 135
- Forced volume groups varyon 135
- Free space 317
- Free space problem 455
 - DB2 456
 - Oracle 457
- Front-end communication 161
- Front-end network 70, 161, 230
 - Bandwidth 163
- Front-end traffic 230
- FTAB 312
- Full stripe destage 109

G

- Gateway (G) 40
- Gigabit Ethernet 164, 179, 230
- Gigabit Ethernet - SX PCI Adapter 56
- Global configuration files 75
- Global power supply 77
- Graceful takeover 227
- Growth characteristics 115
- GUI buffers 312
- Guidelines 17
- Guidelines for tablespaces 118

H

- HACMP 80, 226, 236
 - Adapter failure 231
 - Adapters 231
 - Application start and stop 241
 - Automatic operation 249
 - Boot IP label 223, 233
 - Cascading resource groups 234
 - Cluster diagram 248
 - Cluster topology 229
 - config-too-long 247
 - Configure Adapters 236
 - Database reconnect 243
 - Define Application Servers 236
 - Disk change 258
 - Enqueue reconnect 243
 - File system change 258
 - Hanging NFS mounts 238
 - High water mark 245
 - I/O pacing 245
 - Implementation 235
 - Install patches 256
 - Low water mark 245
 - Maintenance 256
 - Major device number 240
 - Monitoring 415
 - NFS considerations 238
 - NFS mounts 239
 - Node synchronization 244
 - Priority, Resource group 234
 - Resource group priority 234
 - Resource groups 231
 - Resources 234
 - Rotating resource groups 234
 - SAP GUI clients 247
 - SAP R/3 license 246
 - SAP R/3 profiles 237
 - saplicense 247
 - SAPLOCALHOST 237
 - Serial link 231
 - Service address 231
 - Service IP label 223, 233
 - Standby address 231
 - Standby IP label 223, 233
 - Standby subnet 223
 - Starting SAP R/3 241
 - State-transition diagram 248
 - Stopping SAP R/3 243
 - Symbolic link 239
 - syncd 245
 - Topology 229
 - Visible clinfo 248
 - Visible clstrmgr 248
- HACMP classic 227
- HACMP/ES 227
- Hard disaster 86
- Hardware address takeover 227
- Hardware separation 273
- hd5 125
- Head seek time 97
- Heap 319
- Heartbeat 230
- Heartbeat detection rate 246
- Heartbeats 226
- Heterogeneous system copy 368
- High availability 13
- High Availability Cluster Multi Processing (HACMP) 61, 221
- High performance storage 108
- High Voltage Differential (HVD) 53
- High-performance Parallel Interface (HiPPI) 141
- History file 290
- Hit ratio 317
- Homogeneous system copy 368
- Host adapter connectors 141
- Host Adapters 55
- Host spool access method 303
- Hot spare disk 126
- Hot swappable disks 43
- Hot-plug PCI slots 43
- Housekeeping 20, 300

I

- I/O balancing 98

- I/O pacing 245, 343
- I/O processing overview 122
- I/O throughput gain 123
- I/O time 97
- IBM ^pSeries 42
- IBM 10/100/1000 Base-T Ethernet PCI adapter 56
- IBM 3584 UltraScalable Tape Library 53
- IBM Enterprise Storage Server 142
- IBM Enterprise Storage Server (ESS) 50
- IBM Magstar 3590 Tape Subsystem 142
- IBM Magstar MP 3570 Tape Subsystem 142
- IBM Magstar technology 54
- IBM Network Attached Storage 200 49
- IBM Network Attached Storage 300 49
- IBM PCI 4-channel Ultra3 48
- IBM RS/6000 42
- IBM SAN Data Gateway 142
- IBM SAN Fibre Channel Switch 142
- IBM SAN Switch 2109 148
- IBM SAP International Competence Center (ISICC) 45
- IBM SSA technology 48
- IBM StorWatch 51
- IBM Ultrastar disk drives 47
- IBM Ultrium LTO technology 53
- ICNV 374
- Implementation guideline 17
- Implementation of HACMP for SAP R/3 228
- importvg command 389
- Informix 62
- init 184
- inittab 184
- Inode management 122
- Installation manuals 17
- Instance buffers 68, 312
- Integration tests 37
- Inter Switch Link 83
- Inter Switch Link (ISL)
 - Supported distance 160
- Interactive users 68
- Interconnection technology 101
- Interface availability
 - Service level agreement 30
- Internal architecture 36
- Internet Operating System (IOS) 57
- Interoperability
 - Storage subsystems 47
- Invalid loop configuration 129
- iostat command 348
- IP aliases 230
- IP label naming concept 277
- IP label takeover 227
- IP labels 277
- IP labels and subnets 223
- IP names and resolution 289
- IREC 312
- ISNO 180
- IT Infrastructure 11, 22

J

- JFS 120
- Jfslog 127
- Journalized file system cache 120
- Journalized file systems 120
- Journalized file systems log 127
- Jumbo frames 175
- jumbo_frames 178

K

- Kerberos 388
- kernel processes (KPROC) 338
- kinit command 414
- klist command 414
- Korn shell scripts 225
- kprocprio 339

L

- LAN Switching 57
- Landscape-wide operation 32
- LANG 290
- Late paging space allocation 334
- Latency requirements 151
- Latency, rotational 97
- Layout criteria 111
- Least Recently Used (LRU) 312
- les 291
- lgtst command 467
- Libelle 269
- Library sharing 213
- Licensed Program Product (LPP) 74
- Licensed Program Products (LPP) 280
- Linear Tape Open (LTO) 53
- Listener process 465
- Local Area Network (LAN) 55
- Local name resolution 289
- Lock entries 425

- Log file system 285
- Log files 20, 187, 434
- log files 187
- log_archive 449
- log_dir 449
- Logical database error 263
- Logical layer 109
- Logical Volume Manager 59, 109, 120
- Logical Volume Manager (LVM) 191, 270
- Logical volume names 127, 278
- Logical Volume naming concept 278
- Logical Volume sizing 117
- Logical volumes (LV) 411
- Login prompt 291
- Logon groups 467
- Low Speed Connection (LSC) 151
- Low Voltage Differential (LVD) 53
- LPP source administration 283
- LPP sources 283
- lrud kernel process 333
- lsattr command 176, 341
- lscons command 286
- lsdev command 176, 341
- lslv command 125
- lsnrctl command 465
- LTO Ultrium 148

M

- Magstar 3494 Tape Library 54
- Magstar 3590 54
- mail command 411
- Main memory buffering 300
- Maintainability 13
 - Storage subsystems 46
- Maintenance 14
- Maintenance of clusters 256
- Maintenance window 14
- Major device number 240
- Manageability
 - Basic model 76
 - Storage subsystems 47
- Management class 202
- Manual and automatic takeover 224
- Manual pages 282
- Manual takeover 224
- MAX_SIZE_MB 326
- maxfree 337
- Maximum Transfer Unit (MTU) 159

- maxperm 336
- maxpgahead 337
- maxpout 344
- maxreqs 339, 340
- maxservers 338
- maxuproc 340
- McDATA ED-6064 144
- McDATA ES-3016 143
- McDATA ES-3032 143
- media_speed 178
- memlimits command 466
- Memory free list 333
- Memory history buffer 333
- Memory pages 307
- Menu buffer 313
- Message (Msg) 40
- Middleware 60
- migratepv command 126
- minfree 337
- Minimal central instance 66
- Minimum downtime 189
- minperm 336
- minpgahead 337
- minpout 344
- minservers 338
- Mirror Write Consistency (MWC) 123
- Mirror write consistency cache 124
- mirrorvg command 125
- Missing disk 129, 135
- Missing indices 419
- mkdev command 341
- mkitab command 287
- mksysb command 74, 192, 209, 283
- Mode Conditioning Patch Cord 104
- MONI 392
- monitor command 355
- Mountgroup 285
- MTU 180
- Multiple SAP R/3 system 295
- MWC 123
- MWC cache 124

N

- Nagle algorithm 298
- Naming concept 277
- Naming convention 276
- Naming conventions 18, 277
- NAS technology 52

- netstat command 289, 351, 412
- Network
 - Service level agreement 26
- Network adapters
 - Configuration 175
- Network Attached Storage (NAS) 49, 139
- Network bandwidth 152
- Network component connectors 141
- Network configuration
 - Changes for shadow database 269
- Network File System (NFS) 139
- Network infrastructure 150
- Network Installation Manager (NIM) 59, 209, 283
- Network layout
 - Basic configuration 167
 - Disaster-tolerant configuration 172
 - Fault-tolerant configuration 169
- Network options
 - Recommended values 183
 - System-wide 184
- Network routes 412
- Network Time Protocol (NTP) 284
- New Dimension products 3
- NFS and user IDs 279
- NFS cross mounts 234
- NFS mount hangs 289
- no command 183
- Node synchronization 244
- Non Volatile Random Access Memory (NVRAM) 107
- Non-ambiguous names 18
- Non-blocking disk I/O 340
- Non-run-time requirements 13
- NSORDER 388
- NSORDER variable 290
- Number of CPUs 267
- Number of users
 - Service level agreement 28
- Number of work processes 327
- Number range buffering 299
- numperm 336

O

- Obviousness 20
- ODM update 293
- Offline backups 14
- Online help
 - DB2 471

- Oracle 472
- Online Transaction Processing 10
- Online Transaction Processing (OLTP) 123
- Online workload 123
- Operating system 58
 - Backup images 218
 - Image 281
- Operating system requirements 17, 125
- Operations daily checks 433
- Optical Extender 103
- Optimizer statistics 417
- Oracle 62
- Oracle control file 299
- Oracle database layout 120
- Oracle recommendations 298
- Organizational aspects 21
- Organizational constraints 16
- OS07 443
- OSMON 392
- OSS1 400
- Output distribution
 - Service level agreement 26
- Outsourcing 24
- Overview 2
- Overview I/O processing 122

P

- Page fault 333
- Page frame table 333
- Page replacement 333
- Page stealer 333
- Paging 360
- Paging space 110, 296, 334
- PAHI 392
- Parallel System Support Program (PSSP) 60
- Parity calculation 106
- Partition level striping 116
- Partition pattern 117
- PCI Multimode Fibre ATM Adapter 56
- PCI Unshielded Twisted Pair Adapter 57
- Peer-to-Peer Remote Copy 50
- Performance 12
 - Basic model 76
 - Storage subsystems 47
- performance and workload 426
- Performance classifications 44
- Performance critical storage objects 110
- Performance data 113

- Performance optimized data handling 101
- Performance penalty 100
- Persistent errors 409
- Persistent segments 332
- Personal constraints 16
- Perspectives 280
- Physical database errors 264
- Physical layer 109
- Physical Partition level striping 116
- ping command 289, 475
- piodigest command 293
- Pitfall ahead 3
- Placeholders 368
- Planned downtime 14
- Planned downtimes
 - Service level agreement 29
- Platforms 42
- Point-in-time recovery 264
- Policy domain 201
- Pools 314
- Poor man's striping 116
- Possible platforms 42
- Power outages 14
- Power3 High Node 43
- Power3 Thin Node 43
- Power3 Wide Node 43
- Predictive failure analysis 42
- Presentation buffer 312
- Price advantage 101
- Principle of this redbook 3
- Principles of administration 276
- Principles of the SP system 280
- Print management
 - Service level agreement 29
- Printer backend 294
- Printer queues 412
- Printing
 - Service level agreement 26
- Printing time out 294
- Private Memory 319
- Private segments 307
- Problem determination 437
- Problem management
 - Service level agreement 27
- Problem resolution
 - Service level agreement 30
- Processor modes 331
- Production system 11, 37
- Program buffer 312

- Program Temporary Fixes (PTF) 74
- Program temporary fixes (PTF) 18
- Project life cycle 2
- Protocol layers 140
- PSALLOC 325
- PSALLOC environment variable 334
- PSAPBTABD 115
- PSAPROLL 112
- PSAPTEMP 112

Q

- qdaemon 293
- Quality assurance system 11, 37
- Quorum 129, 131
- Quorum buster 255
- Quorum buster disk 133
- Quorum checking 133
- Quorum setting 126

R

- R3INST 371
- R3load 373
- R3SETUP 370, 382
- R3trans command 469
- Rack servers 42
- RAID 99
- RAID 0 to RAID 10 99
- RAID 5 272
 - Data strip 105
 - Data stripe 105
 - Parity calculation 106
 - Parity checksum 105
- RAID 5 penalty 100
- Random online workload 123
- Random operations 97
- RAS 42
- Raw logical volumes (LV) 120
- rcp command 293
- RDBMS cache 121
- RDDIMPDP 459
- rdist command 287, 293
- Read performance 108
- Recipient 31
- Reclamation 199
- Recovery 187, 188
- Recovery scenario 267
- Redbooks Web site 483
 - Contact us 24

- Redo logs 262
 - Archiving 265
 - Copying 263
 - Space requirement 266
- Redundancy 12, 13, 42
 - Storage subsystems 47
- Redundant array of independent disks (RAID) 99
- Redundant Inter Switch Links 160
- Relational Database Management System (RDBMS) 38, 121
- Reliability 12
 - Basic model 76
- Remote Function Calls 9
- Remote library 72
- Reorganizing
 - DB2 tables 420
 - Oracle tables 421
- Repage fault 333
- replacev command 126
- Report RDDIMPDP 459
- Report RTCCTOOL 302
- Repository buffer 312
- Requestor 31
- Requirements 12, 13
 - Runtime 12
- Resource group 226
- Resource groups 231
- Resources 234
- Response Time
 - Service level agreement 28
- Response time 12, 152
- Restore 187, 188
- restore command 193
- Restore time
 - Service level agreement 30
- restvg command 196
- Reuse delay 199
- rexec command 287
- rfc1323 181
- rlogin command 287
- Roles relationship 22
- roll_extension 325
- Rootvg mirroring 125
- Rotating resource groups 234
- Rotational latency 97
- Routine maintenance 14
- RS/6000 42
- RS/6000 SP 43
- RS-232 cabling 255
- rsh command 287
- RSMEMORY 320
- Rules for interaction 31
- Run queue 330, 361
- Runtime
 - Service level agreement 28
- Run-time requirements 12
- rx_checksum 178
- RZ04 397
- RZ10 397
- RZ12 397
- RZ20 401, 428
- RZ21 401

S

- SA38 401
- Safety 13
- Sample partitioning 118
- SAP APO 3
- SAP Business Warehouse 3
- SAP CRM 3
- SAP DB 62
- SAP DBA 63
- SAP GUI 9
- SAP Kernel CD 373
- SAP notes database 17
- SAP profiles and logs 467
- SAP R/3
 - Administration principles 276
 - Application layer 8
 - Application server 39
 - Application server instances 39
 - Architecture 7, 8
 - Availability criteria 29
 - Basic model 2
 - Batch processing 10
 - Building blocks 35
 - Business logic 39
 - Capacity criteria 28
 - Checks 427
 - Cluster solution 80
 - Database backup 191, 192
 - Database layer 8
 - Database server 38
 - Design 2
 - Development system 36
 - Disaster-tolerant model 2, 86
 - Extended Memory 318

- Fault-tolerant model 2, 77
- Housekeeping 422
- Implementation 2
- Infrastructure 1
- Infrastructure model 222, 260, 306
- infrastructure model 366
- Instance buffers 312
- Locks 40
- Maintenance 3
- Modules 10
- Network layout 167
- Network layouts 149
- Networks 70
- Operation 2
- Operational guidelines 30
- Operations 21
- Performance criteria 28
- Presentation layer 8
- Processes 40, 460
- Requirements 2
- Roles 21, 22
- Roll area 319
- Service level agreement 27
- Software architecture 36
- Stacked layers 11
- Start scripts 18
- Stop script 18
- System copy 365
- System landscape 10, 367
- Three-tier architecture 8
- Tuning 320
- SAP R/3 client copies 303, 369
- SAP R/3 database performance data 113
- SAP R/3 installation 294
- SAP R/3 Library 75
- SAP R/3 license 246
- SAP R/3 operation 301
- SAP R/3 performance data 113
- SAP R/3 profiles 301
- SAP R/3 reorganization 300
- SAP R/3 system log 423
- SAP R/3 Transport system 369
- SAP Service Marketplace 372
- SAP Service marketplace 17
- SAP standards 17, 276
- SAPconnect 400
- sapdba command 265, 384
- SAPDBA logs 193
- sapdba_role.sql 392
- saplicense command 247, 373
- SAPLOCALHOST 237
- SAPLOGON.INI 388
- SAPNet 85, 388, 400
- saposcol 315
- sappfpar command 313, 467
- saprouter 85
- SAPS 45, 85
- SAPUSER table 392
- savevg command 196
- sb_max 180
- Scalability 13
 - Basic model 76
 - Storage subsystems 46
- Scalable platform 38
- SCC4 399
- SCC5 401
- Scheduling 205
- SCOT 400
- Script naming concept 278
- Scripts names 278
- SCSI bus arbitration 102
- SCSI bus contention 102
- SCSI implementation 102
- SCSI Parallel Interface 102
- SCSI RAID Adapter 48
- SCSI solution 51
- SCSI-3 standard 102
- SDBAD 392
- SDBAH 392
- SDCC 401
- SE03 375, 399
- SE06 394
- SE11 419
- SE14 374, 419
- SE16 395
- SE37 320
- Security 13
 - Storage subsystems 47
- Seek time 97
- sendmail command 411
- Sequential background processing 123
- Sequential operations 98
- Sequential-Access Read Ahead 337
- Serial bus protocol 103
- Serial Disk System 48
- Serial storage architecture 51
- Serial Storage Architecture (SSA) 48, 102
- Server AP 78

- Server communication 70, 162
- Server Consolidation 295
- Server DB/CI 78
- Server journal 301
- Server Message Block (SMB) 139
- Server network 162, 230
- Server selection 45
- Server systems
 - Service level agreement 25
- Server traffic 230
- Service address 231
- Service hours
 - Service level agreement 29
- Service IP address 223
- Service IP label 223, 233
- Service IP subnet 223
- Service level agreement 16, 24
 - Criteria 25
- Service network 224
- Service processor 43
- Service provider model 24
- SGEN 401
- Shadow database 86, 90, 262
 - Implementation 268
 - Implications 266
- Shared Common System Area (SCSA) 315
- Shared Product Object Tree (SPOT) 74
- Shared segments 307, 309
- Short nametab 312
- showipc command 468
- SICK 402
- Side effects 20
- Single disk throughput 98
- Single point of control 280
- Single points of failure (SPoF) 12, 41, 77, 128, 231
- Single system operation 33
- Sizing disk 98
- SM02 401
- SM12 425
- SM13 374, 397, 424
- SM21 402, 423, 443
- SM35 374, 398
- SM36 423
- SM37 375, 395, 424
- SM50 447
- SM51 402, 423, 444
- SM54 399
- SM55 399
- SM58 398
- SM59 400
- SM61 398
- SM63 397
- SM65 398
- SM69 399
- Small Computer System Interconnect (SCSI) 47
- Small shadow database 269
- SMLG 397
- Soft disaster 86
- Software Change Registration (SSCR) 404
- Software components 58
- Software development
 - Common methodology 36
 - Life cycle 36
- Software distribution 73
- Software installation 59
- Source system 368
 - Actions 379
- SP concepts 280
- SP daemons 413
- SP log files 414
- SP Switch 44
 - Bandwidth 160
- SP Switch2 44
- SP system 43, 413
 - Nodes 43
- SP12 396, 425
- SPAD 396, 402
- Spanning Tree Protocol 246
- SP-attached servers 44
- Special purpose instances 69
- Spool (S) 40
- SPOT 284
- Spreadsheet 114
- SPRO 404
- SSA
 - Attenuation 104
 - Bandwidth 102
 - Bypass-card 129
 - Configuration
 - Basic 128
 - Disaster-tolerant 133
 - Fault-tolerant 131
 - Drawer 128
 - Fibre Optical Extender 103
 - Implementation 103
 - Loop principle 103
 - Mode Conditioning Patch Cord 104
 - Optical Extender 103

- Path light loss 104
- Ring topology 103
- SSA Optical Extenders 133
- SSAA 428
- ST02 322, 426, 445
- ST03 398, 426, 445
- ST04 419
- ST06 326, 357
- ST22 425
- Stacked layers 11
- Standard processes 23
- Standby address 231
- Standby IP label 223, 233
- Standby IP subnet 223
- Standby subnet 223
- startsap 463
- Status of recovery 269
- STMS 394, 395, 448
- Storage
 - Service level agreement 25
- Storage Area Network (SAN)
 - Considerations 146
- Storage Area Networks (SAN) 137
- Storage management 63
- Storage pools 199
- Storage subsystem
 - SPoF 79
- Storage subsystems 46
- strict_maxperm 336
- Stripe destage 109
- Structural changes 268
- SU01 400
- Subsystem Device Driver (SDD) 130, 147
- Supply Chain Management (SCM) 42
- svrmgrl command 386, 391, 450
- Swapping 317
- swcons command 286
- switch network 70
- Switches
 - Configuration 174
 - SPoF 79
- Symmetric multiprocessor (SMP) 330
- Synchronization 279
- Synchronization points 190
- Synchronizing 279
- System buffers 426
- system call (SVC) 308
- System configuration
 - Documentation 74

- System console log 412
- System copy 367
- System Data Repository (SDR) 60
- System documentation 282
- System landscape 10, 36, 367
 - Additional 37
 - Minimal configuration 37
- System level action 387
- System management 408
- System Management Interface Tool (SMIT) 59
- System management techniques 12
- System memory 426
- System outages 13, 260
- System programmers 2

T

- Table buffers 312
- Tablespace activity 112
- Tablespace administration 119
- Tablespace fragmentation 119
- Tablespace growth characteristics 115
- Tablespace PSAPBTABD 115
- Tablespace PSAPROLL 112
- Tablespace PSAPTEMP 112
- Tablespaces 416
- Takeover 82, 223, 279
- Tape backups 52
- Tape library 72
- Tape management 212
- Tape pools 200
- Target mode SSA 255
- Target system 368
 - Actions 380
- TCP nodelay 298
- TCP/IP options 175
- TCP/IP over ATM 159
- tcp_mssdflt 181
- tcp_nodelay 181
- tcp_pmtu_discover 180
- tcp_recvspace 181
- tcp_sendspace 181
- Technical constraints 16
- Technical specifications 97
- telnet command 476
- TemSe consistency 425
- terminal server 282
- Test systems 273
- thewall 180

- Think time 152
- Thread 330
- Three-tier client/server 8
- Throughput 12
- Time master 280
- Time synchronization 284
- Time zones 279
- Tivoli Data Protection (TDP) 186, 382
- Tivoli Data Protection for R/3 64
- Tivoli Storage Manager (TSM) 63, 71, 186, 382
 - Concept 197
- TLOCK 392
- tokstat command 178
- Tool lsof 287
- Tool monitor 287
- Tool samba 287
- topas command 326, 353
- Topology 229
- Topology constraints 17
- Total costs of ownership (TCO) 17
- Tower servers 42
- TPFET 392
- TPFHT 392
- Traffic front end 230
- Traffic server 230
- Transaction
 - DB6COCKPIT 113
 - ST04 113
- Transactions 443
- Transfer time 97
- Transport directory 464
- Transport management system
 - Service level agreement 30
- Transport Management System (TMS) 36, 191, 279, 448
 - Files 194
- Transport tool 395
- trcoff command 349
- trcon command 349
- trcstop command 349
- Troubleshooting procedure 437
- Troubleshooting tasks 448
- truncate command 392
- TSLE4 392
- TSM
 - Documentation 75
- TSM administrative schedules 431
- TSM client schedules 429
- TSM Disaster Recovery Manager 200

- TSM scheduler daemons 433
- TSM scheduling 205
- TSM server activity log 432
- TSM server configuration
 - Documentation 75
- TSM storage pools 199
- TSM virtual node names 197
- TTAB 312
- Tuning 320
- Two-tier implementation 66
- TX DB6COCKPIT 113

U

- udp_pmtu_discover 180
- udp_recvspace 181
- udp_sendspace 181
- Unification 18, 276
- Unplanned downtime 14
- Update (U) 40
- Update records 424
- Update statistics 417
- Update V3 300
- Uppercase filenames 297
- use_isno 180
- User connections 327
- User context 318
- User Exits 268
- User ID management 280
- User limits 296

V

- varyonvg command 129
- Vaulting 72, 211
- Version control 203
- Version history 403
- Virtual address space 307
- Virtual Memory (VM) 306, 332
- Virtual Memory Manager (VMM) 307, 332
- vmstat command 346, 476
- vmtune command 121, 336, 347
- Volume group 277
- Volume group size 126
- Volume groups naming concept 277
- VP01 396

W

- Wait queue 330

WCOLL 288
Web server 282
Wide Area Network (WAN) 265
Working collection 288
Working segments 332
Workload 113
Workload analysis 426
Workload Manager (WLM) 59, 295
Write preempt 108

X

xload 371

Z

ztta 325



Redbooks

A Holistic Approach to a Reliable Infrastructure for SAP R/3 on AIX



A Holistic Approach to a Reliable Infrastructure for SAP R/3 on AIX

Design, set up, and manage a highly available IBM @server pSeries environment

Implement a disaster proven backup and recovery solution

Plan optimized network and disk layouts

SAP R/3 has become one of the most widespread Enterprise Resource Planning software products in the world. It is the application that is at the core of every company's business. With the incorporation of Web-initiated e-business transactions in SAP R/3, it now has to be available literally all the time.

This IBM Redbook shows a holistic approach for a reliable and highly available SAP R/3 infrastructure based on IBM @server pSeries. It is a collection of knowledge gathered from a team of SAP R/3 basis consultants that have gained their experience over many years by working in numerous projects.

This redbook provides a high-level overview of the SAP R/3 architecture and contains the high availability design criteria that are essential for IT managers and solution architects. It includes tips and tricks for database copies, performance optimization, and troubleshooting. This redbook also features a wealth of practical information and hints, which are needed during implementation, for IT specialists, SAP R/3 administrators, and system administrators.

This redbook is an invaluable source of information for all professionals that maintain a reliable SAP R/3 environment.

INTERNATIONAL TECHNICAL SUPPORT ORGANIZATION

BUILDING TECHNICAL INFORMATION BASED ON PRACTICAL EXPERIENCE

IBM Redbooks are developed by the IBM International Technical Support Organization. Experts from IBM, Customers and Partners from around the world create timely technical information based on realistic scenarios. Specific recommendations are provided to help you implement IT solutions more effectively in your environment.

For more information:
ibm.com/redbooks