Consolidating Multiple SAP Systems on One IBM eserver™ pSeries

- Implementing complex SAP system landscapes
- Concepts and experiences with LPAR on AIX 5L
- Manage workloads for multiple SAP systems

Olaf Siebenhaar
Carol Davis
Cathrin Klonowski
Matthias Koechl
Walter Orb
Norbert Pistoor

Redpaper

ibm.com/redbooks

International Technical Support Organization

Consolidating Multiple SAP Systems on One IBM eServer pSeries

December 2002

Note: Before using this information and the product it supports, read the information in “Notices” on page vii.

First Edition (December 2002)

This edition applies to the AIX 5L operating system and SAP Version 4.6.

Contents

Notices vii
Trademarks viii

Preface ix
The team that wrote this Redpaper ix
Become a published author xi
Comments welcome. xi

Part 1. SAP on AIX installation concepts 1

Chapter 1. Technology overview 3

Chapter 2. Situation and scope 5
2.1 Degree of flexibility and complexity. 6
2.1.1 Permanent relocations 7
2.1.2 Temporary relocations 7
2.1.3 Higher availability 7
2.2 Design principles. 8
2.2.1 Uniformity 8
2.2.2 The SAP system as a unit. 9
2.2.3 Disjoint units 9
2.2.4 Reusability and standardization 9

Chapter 3. Solution 11
3.1 Operating system 12
3.1.1 AIX parameter. 12
3.1.2 Numeric user and group ID 12
3.2 IP names and host names. 14
3.3 Disk storage 18
3.3.1 Volume groups 18
3.3.2 File systems 19
3.3.3 File system layout 22
3.3.4 Oracle file systems 25
3.3.5 Symbolic links 26
3.3.6 SAP file systems and symbolic links. 28
3.4 Oracle database 30
3.4.1 Oracle Installer 31
3.4.2 Kernel extensions 31
3.4.3 Oracle listener. 31

3.5 SAP instance . . . 32
3.5.1 Instance number . . . 33
3.5.2 Transport directory . . . 34
3.5.3 Performance collector . . . 36
3.5.4 SAP license key . . . 37
3.6 Backup and recovery . . . 37
3.6.1 Backup objects . . . 37
3.6.2 Backup domains . . . 39
3.7 AIX Workload Manager . . . 41
3.7.1 Positioning of AIX Workload Manager . . . 42
3.7.2 Description of AIX Workload Manager . . . 43
3.8 Logical partitions (LPARs) . . . 47
3.8.1 Sizing considerations . . . 47
3.8.2 LPAR overhead . . . 48
3.8.3 Cross LPAR impact . . . 48
3.8.4 Architectural considerations . . . 48
3.8.5 Summary of results for a 32-way p690 . . . 49
3.8.6 Configuration alternatives . . . 50
3.8.7 Single partition . . . 50
3.8.8 Traditional LPARs . . . 51
3.8.9 Affinity LPARs . . . 52
3.8.10 Balanced memory . . . 53
3.8.11 mySAP applications . . . 53
3.8.12 Dynamic LPAR and AIX 5L Version 5.2 . . . 53
3.8.13 Impact on databases . . . 54
3.8.14 Monitoring dynamic LPARs . . . 54
3.9 Differences between p690 LPARs and stand-alone nodes . . . 55

Chapter 4. Conclusion of Part 1 . . . 57

Part 2. Regulating system resources with AIX 5L Workload Manager . . . 59

Chapter 5. Setup and scenarios . . . 61

Chapter 6. Scenario 1: Equally important systems . . . 67
6.1 Test 1: Workload distributed evenly . . . 68
6.1.1 Test 1 summary . . . 71
6.2 Test 2: Workload distributed unevenly . . . 72
6.3 Test 3: Stress test with equal shares . . . 74
6.3.1 Test 3 summary . . . 77
6.4 Test 4: Stress test with unequal shares . . . 77
6.4.1 More users in the high load system . . . 78
6.4.2 More users in the low load system . . . 80
6.4.3 Test 4 summary . . . 82

Chapter 7. Scenario 2: One system more important than others 83
7.1 Test 5: Workload distributed evenly 84
7.1.1 Test 5 summary 87
7.2 Test 6: workload distributed unevenly 88

Chapter 8. Scenario 3: One system less important than others 93
8.1 Test 7: Workload distributed evenly 94
8.2 Test 8: Workload distributed unevenly 97
8.2.1 Test 8 summary 99

Chapter 9. Conclusions and recommendations 101

Appendix A. Implementation and configuration details 103
Syntax description 103
Sample Oracle files 104
Oracle TNS names file 104
Oracle listener configuration file 104
Sample IBM Tivoli Storage Manager files 105
Sample IBM Tivoli Storage Manager client option file. 105
Sample IBM Tivoli Storage Manager include/exclude files 106
Workload Manager 106
Processor binding with WLM. 106
WLM setup 109

Appendix B. Products 111

Related publications 113
IBM Redbooks 113
Other resources 113
How to get IBM Redbooks 114
IBM Redbooks collections. 114

Notices

This information was developed for products and services offered in the U.S.A.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described 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:
IBM Director of Licensing, IBM Corporation, North Castle Drive Armonk, NY 10504-1785 U.S.A.

The following paragraph does not apply to the United Kingdom or any other country where such provisions are inconsistent with local law: INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some states do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this IBM product and use of those Web sites is at your own risk.

IBM may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to the names and addresses used by an actual business enterprise is entirely coincidental.

COPYRIGHT LICENSE:
This information contains sample application programs in source language, which illustrates programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to IBM, for the purposes of developing, using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. IBM, therefore, cannot guarantee or imply reliability, serviceability, or function of these programs. You may copy, modify, and distribute these sample programs in any form without payment to IBM for the purposes of developing, using, marketing, or distributing application programs conforming to IBM's application programming interfaces.

Trademarks

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

AIX 5L™
AIX®
Balance®
DB2®
Enterprise Storage Server™
IBM eServer™
IBM®
Perform™
pSeries™
Redbooks (logo)™
Redbooks™
RS/6000®
Tivoli®

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

Lotus®
Word Pro®

The following terms are trademarks of other companies:

ActionMedia, LANDesk, MMX, Pentium and ProShare are trademarks of Intel Corporation in the United States, other countries, or both.

Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

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

C-bus is a trademark of Corollary, Inc. in the United States, other countries, or both.

UNIX is a registered trademark of The Open Group in the United States and other countries.

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.

Preface

With the introduction of ever more powerful UNIX servers, such as the IBM @server pSeries 690 models, server consolidation is back on the agenda in many computer centers. As a widely used business application, SAP R/3 is often on top of the list when applications are identified that are currently running on several smaller servers, but could probably be consolidated onto a single, more powerful server.

This IBM Redpaper covers two major trends in the IT business: the technological improvements of server platforms, providing increasingly powerful and flexible servers, such as the IBM @server pSeries 690, and the close integration of business processes along the value chain, resulting in multifunctional application portfolios, such as mySAP.com.

Today, server consolidation is a must for many IT sites. Minimized total cost of ownership (TCO) and complexity, with the maximum amount of flexibility, is a crucial goal of nearly all customers. This Redpaper describes how you can exploit the technical features of the IBM @server pSeries platform in order to accomplish these requirements. It is intended to help IT architects and specialists in designing, implementing, and using mySAP.com consolidation scenarios. It includes the newest key information about AIX 5L, logical partitioning (LPAR), and AIX Workload Manager (WLM). The concepts presented in this paper are field-tested best practices.

This IBM Redpaper brings together discussions of the various aspects that have to be considered when an attempt is made to consolidate multiple SAP R/3 systems on a single IBM @server pSeries server. Not all of the ideas presented here are entirely new. Some have been published in other places. But all of them have been field tested and enriched with the latest experiences, which the authors are delighted to share.

The team that wrote this Redpaper

This Redpaper was produced by a team of specialists from around the world.

Olaf Siebenhaar is a Certified Consulting IT Specialist in Global Service in IBM Germany. He has worked at IBM for 10 years and has nine years of experience in the AIX and SAP fields. His areas of expertise include HACMP and SAP, as well as IBM Tivoli Storage Manager. He holds a degree in Computer Science from the Technical University Dresden, Germany.

Carol Davis is a pSeries and SAP performance specialist working for the IBM SAP International Competence Center in Walldorf. She has 12 years experience with RS/6000 and seven years experience in SAP basis, infrastucture, and benchmarking. Her primary focus is technical enablement: paving the way for new platform functionality in the SAP environment and ensuring the quality and performance of new SAP functionality on the pSeries platform. She holds a BSc in Computer Science from Lafayette, Louisiana.

Cathrin Klonowski is an IT Specialist from IBM Munich, Germany. She works in the Pre-Sales Technical Support Team in the Web Server Sales division of EMEA Central Region. Her main responsibility is AIX and WLM technical support to customers. She holds a Master's Degree in Mathematics and Physics from York University, UK.

Matthias Koechl joined IBM D&R Boeblingen in 1984 as a development engineer. Then, he worked for IBM industry and pSeries sales as a sales representative before joining the ISICC in 1999. He now holds a position as Senior Technical Marketer for the pSeries brand. His responsibility is enablement and alignment of technical development between SAP and pSeries. He holds a diploma in Electrical Engineering.

Walter Orb is a technical consultant working at the International SAP/IBM Competence Center in Walldorf, Germany. He has more than nine years of experience with SAP on AIX, with a major focus on system performance, benchmarks, and large-scale system tests. Walter holds a Master's Degree in Mathematics and Economics from University of Mainz, Germany.

Norbert Pistoor is an Advisory IT Specialist with IBM @server pSeries Pre-Sales Technical Support in Munich, Germany. He has more than 12 years of experience with RS/6000 and pSeries, including seven years in benchmarking and performance analysis. He holds a Ph.D. in Physics from University of Mainz, Germany.

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

We would like to thank Tobias Müller and Basil Moshous from IBM Global Service BMS/ERP, who contributed significant enhancements during the many reviews of this document. We are also very grateful to our colleagues from the IBM SAP International Competence Center (ISICC) in Walldorf, who supported us with much useful information and by setting up the test environment used in the WLM project described in Part 2 of this paper. Many thanks also to Diana Gfrörer from IBM Software Group for her valuable support during the publishing process.

Become a published author

Join us for a two- to six-week residency program! Help write an IBM Redbook dealing with specific products or solutions, while getting hands-on experience with leading-edge technologies. You'll team with IBM technical professionals, Business Partners and/or customers.

Your efforts will help increase product acceptance and customer satisfaction. As a bonus, you'll develop a network of contacts in IBM development labs, and increase your productivity and marketability.

Find out more about the residency program, browse the residency index, and apply online at:

ibm.com/redbooks/residencies.html

Comments welcome

Your comments are important to us!

We want our papers to be as helpful as possible. Send us your comments about this Redpaper 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:

  IBM Corporation, International Technical Support Organization
  Dept. JN9B Building 003 Internal Zip 2834
  11400 Burnet Road
  Austin, Texas 78758-3493

Part 1

SAP on AIX installation concepts

1

Technology overview

The evolution of processor and storage technologies have a great impact on the architecture of IT infrastructures. They were subject to change since the first version of SAP R/3 initiated an unappeasable demand for highly available performance and storage capacity. This was the most significant challenge for the infrastructure in the past and will also be in the future.

During the first half of the 1990s, one single SAP system per node was suitable. Moreover, most productive systems needed additionally associated nodes, so-called application servers. Increasing performance and reliability by simply replicating application server nodes led to complex environments that often resulted in poor systems management. The reason for these complex constructions was the limited computing power of a single node. This limitation was softened during the second half of the 1990s.

Big symmetric multiprocessor (SMP) nodes with higher clock rates and increased memory provided the possibility to install more than one system on a node. This new situation had some side effects regarding system operations. For example, release planning processes had to pay attention to different database or SAP versions, or both, to avoid unresolvable conflicts. Furthermore, no functionality existed that prevented negative, interfering effects.

In 2000, Workload Manager for AIX (WLM) was announced. In addition, multiple SAP system installations became more and more popular because of the permanently increasing number of SAP systems at customer sites. The general availability of this functionality to separate the workloads of dedicated systems eliminated the last obstacle for consolidating several systems in one node. Unfortunately, as of today, not all main performance monitoring functions inside SAP are adjusted for such an environment.

Some customers expanded the usage of their SAP systems and consequently model more business processes. This often caused an increased number of SAP systems used and a stronger demand on flexibility. The life cycle of these systems differed extremely. Renaming, removal, or deletion became more and more common system administration tasks. The nature of infrastructure and system installation can either encourage or thwart this.

In 2001, new IBM @server pSeries hardware technology with logical partitioning was generally available. Logical partitioning creates the possibility to define logical partitions (LPARs) that are adapted to customer needs regarding the number of processors, assigned memory, and I/O adapters. From an SAP system standpoint, there is no difference from a collection of stand-alone nodes. The minimum requirements for one LPAR limit the maximum number of LPARs in one hardware system. Therefore, multiple system installations become necessary again if more SAP systems have to be installed than there are LPARs available. This is a very common situation at present.

The transition of storage technology from SCSI to Fibre Channel demanded remarkable expenses in cost and time. The old SCSI technology was a point-to-point connection. Any reassignment of disk storage capacity was not possible without complicated physical reconstructions. Beginning with the usage of intelligent storage subsystems, such as IBM Enterprise Storage Server, the need for this reconstruction has gone away. This was replaced by a much easier reconfiguration task. The new Fibre Channel-based storage area network (SAN) delivers the same functionality with nearly unlimited distances between storage components and nodes compared to the SCSI distance limitation of approximately 24 meters. The feature of easy reconfigurable storage devices in conjunction with the possibility of multiple installations of SAP systems on big SMP nodes fulfills a new class of customer requirements, those desiring relocatable SAP systems on different levels of automation.

2

Situation and scope

A fundamental requirement to minimize the effort for system relocations is an appropriate hardware infrastructure running adequately installed systems. This IBM Redpaper provides you with field-tested hints and tips, as well as an overview of the most common obstacles and how to avoid them.

2.1 Degree of flexibility and complexity

During their life cycle, SAP systems are installed and removed at least once. Since single SAP system environments have lost their distinctiveness, some additional operations became common, for example, renaming and relocating SAP systems. Every installed system has a certain degree of flexibility that determines the efforts of that operation to a large extent. Various details and many related topics according to the infrastructure for SAP systems are discussed in *A Holistic Approach to a Reliable Infrastructure for SAP R/3 on AIX*, SG24-5050.

One or more systems can be installed on a node or LPAR. Basically, all of these systems in any configuration can be installed relocatable, but some principles have to be taken into consideration to ensure full relocatability. The installation of more than one system on a node undeniably increases the complexity, but well thought out concepts and their rigid application can limit the handling effort. The following degrees of complexity can be defined:

- One system per node or LPAR
- Several systems on one node or LPAR
- Several relocatable systems on one node or LPAR

One SAP system per LPAR or node is a possible solution for a small number of systems within a small environment. In this scenario, every system is installed on dedicated nodes or logical partitions, consisting at least of one processor, memory, disk capacity, and network interfaces. This scenario is valid as long as the number of systems does not exceed the number of available nodes. Otherwise, more than one system has to be installed on one node. The limitation of single systems per LPAR within an IBM @server pSeries 690 environment is the number of available processors and I/O adapters. A large number of SAP systems per IBM @server pSeries 690 hardware unit each installed in its separate LPAR require AIX adapter sharing because of the otherwise unreasonably high number of adapters per p690. The maximum number of processors within one p690 system is another unbreakable limit as long as every LPAR needs at least one dedicated assigned processor. However, sharing modes of adapters and processors on a finer granularity are not yet available.

Several systems installed on a node have a strong impact on availability, maintainability, and performance. A sufficient maintainability is only achievable by an installation based on well thought out concepts. To prevent disturbing influences, SAP does not support the installation of productive and non-productive SAP systems on the same node. Without effective resource management, theoretically, every system might monopolize the complete processor and memory capacity. An active component, the Workload Manager AIX 5L, can prevent such situations.

In Part 2 of this paper, we evaluate some exemplary Workload Manger (WLM) configurations. Currently, SAP systems tolerate WLM, but from an SAP Computing Center Management System (CCMS) point of view, WLM is ignored. A stronger integration in SAP systems is planned for future SAP releases. WLM is imperatively necessary to assure a quality of service for several systems in a node. The resulting overall performance for simultaneous operation of several systems is, therefore, a matter of sizing.

2.1.1 Permanent relocations

Any SAP system can be permanently relocated. This is also known as a host name change within the R/3 system copy guide. A reason for a relocation might be a strategic decision for manual workload balancing to respond to altered long-term performance values. The activities that are necessary for permanent relocation of a system are a subset of the activities for changing the SAP system identifier. The most significant result of this task is the renamed instance identifier. This causes discontinued or lost data for all information related to the instance name, for example, performance data, operation modes, and inbound RFC connections. The advantage of this field-tested method is full support by SAP.

2.1.2 Temporary relocations

A system can be temporarily relocated to a more powerful node for special tasks, such as database loads or reorganizations. On partitionable hardware, such as IBM @server pSeries 690, a similar result can be achieved without any relocations by allocating additional processors or memory resources, or both, to a logical partition. Both methods have the same goal of providing sufficient resources to the system, but they have different basic approaches. A system relocation moves a system, whereas a resource allocation reassigns the resources. The reassignment of resources is almost transparent for an SAP system except for some profile modifications that might become necessary, whereas a relocation impacts much more, such as a moved IP address or modified instance names.

2.1.3 Higher availability

Relocatability also offers new opportunities for system maintenance. For example, it is possible to temporarily relocate all systems of a node to another node. Subsequently, the operating system of the original node can be maintained without generating any interfering effects. This technique allows nearly zero down time during maintenance activities on a node or the network connections. Furthermore, relocatability of a system is the fundamental prerequisite for high availability solutions. These solutions can be triggered manually, as well as automatically, but the kind of initiator has no influence on the SAP system installation prerequisites. These are discussed in detail in *A Holistic Approach to a Reliable Infrastructure for SAP R/3 on AIX*, SG24-5050.

In this solution, all systems on each node or LPAR are installed ready for manual relocation to a node or LPAR in the other hardware system. Both systems should have enough performance capacity to temporarily deliver sufficient performance for all SAP systems on both hardware units. Furthermore, it can be disaster tolerant if both hardware systems are placed into adequately separated locations.

2.2 Design principles

During the last 10 years, the IBM @server pSeries hardware technology has been significantly enhanced. All components were improved for better performance and more flexibility, for example, faster processor units, larger memories, and storage components connected by fibers. An infrastructure built of such modern components enables new designs. The design principles in this document try to exploit these new possibilities.

All field experiences described in this document are gathered during the design and implementation of a complex infrastructure using IBM @server pSeries 690 Regatta at customer sites. The operating systems AIX 5L runs together with High-Availability Cluster Multiprocessing/Enhanced Scalability (HACMP/ES). The applications used SAP R/3, SAP APO, and SAP BW on Oracle Version 8.

2.2.1 Uniformity

A good comparability of nodes gives the system administrator the possibility to cross check between nodes. This is invaluable during troubleshooting and error fixing. Of course, no node can be absolutely identical to another, because every node needs a unique identity, such as host names and network addresses. Related to all possible differences, only very few are absolutely necessary. It is up to the system administrator to keep all nodes in sync as well as possible.

Every node within an environment that is eligible for an SAP system must comply to all requirements according to the appropriate *SAP Installation Guides*. All locally existing software versions have to be taken into consideration. Therefore, prospective release planning might be useful.

2.2.2 The SAP system as a unit

A fundamental design principle is the approach to interpreting an SAP system and the associated database as an indivisible unit. Every existing or planned SAP system is interpreted as such a unit that needs certain resources. This unit is named and unambiguously identifiable by the SAP system identifier (SID). Minimal prerequisites for an active SAP system of any kind are always the following hardware resources:

- Memory
- Processor
- Disk storage capacity
- Network connection

A correctly installed and customized operating system is also necessary. All components used by an SAP system and its associated database are either system specific or unspecific.

2.2.3 Disjoint units

In an environment with multiple systems on a node, every system might have its own copies of computing resources. If a copy of a resource is always associated to only one dedicated system, the resource is automatically a candidate to be system specific. It is not system specific if a copy of a resource is dedicated to more than one system in any constellation. A copy of a system-specific resource can never be dedicated to more than one system. In contrast, a system-unspecific resource is never dependent on a single system or dedicated to any system. Only one copy per node is possible. That also applies to installations of several systems on a node. These rules ensure the existence of disjoint units that can be maintained without any disturbing influences.

2.2.4 Reusability and standardization

Another design target of all concepts and solutions within this document is the best possible reduction of all administration efforts. This can be achieved by putting the following simple designs principles into practice:

- Best possible level of standardization of all systems
- Best possible level of synchronization between all systems
- Rigid orientation on reusability

A thorough implementation of these principles on all infrastructure components and levels enables an overall low administration effort. Every infrastructure is subject to change during its life cycle. It also has to be flexible and extendable to sustain a low administration effort for reconstructions. Furthermore, an appropriate scalability of the infrastructure is a fundamental prerequisite to preserve the flexibility over its lifetime.

A reference scenario for benchmarking a certain infrastructure might be the evaluation of the necessary administration effort for certain tasks. Installing and removing a system can be extremely time consuming when it shares the node with other systems. The relocation of a system is not only a simple addition of both efforts, rather it requires a certain degree of standardization across all involved nodes.

3

Solution

A stable and scalable infrastructure for SAP can be built by taking a few fundamental design principles into consideration. These principles are outlined in the following sections and used for all concepts and solutions within this document to achieve a best possible reduction of all administration efforts.

3.1 Operating system

The number of installed SAP systems is constantly increasing in most customer environments. Systems are newly installed, renamed, or sometimes relocated. Permanent relocation means uninstalling on a node and transferring the whole system to another node. Basically, a relocation is always possible, but might require a higher administration effort.

3.1.1 AIX parameter

Every node within an environment that is eligible for an SAP system must at least comply to all the requirements stated in the *SAP Installation Guides*. These can be different from node to node because of different SAP releases. To allow for planned system relocations in the future and to respect the design principle of uniformity, all system-specific differences between nodes should be avoided.

Some well-known components should be uniform across all nodes, for example, AIX software including program temporary fixes (PTFs) and their maintenance level. Furthermore, some designated system settings must be consistent, for example, the maximum number of processes and file size for every user. These settings are stored in the /etc/security/limits file that can be synchronized across nodes using the AIX command **rdist**. The operating system has to be enabled for 64 bit if one of the databases is a 64-bit database. The Portable Stream Environment must be activated on all designated SAP nodes.

The architecture of SAP uses relatively large shared and private memory areas. See *Configuration of R/3 on hosts with much RAM*, SAP Online Service System (OSS) Note 146528 for more information. The total capacity of these areas might not fit into the physically available memory. In this case, a memory over-commitment causes heavy paging activities at operating system level. See *AIX 5L Differences Guide Version 5.1 Edition*, SG24-5765 for more information. The size of the paging space should be, as a rule of thumb, three to four times the memory capacity to provide sufficient swap space. All locally installed systems will be affected if the system runs out of paging space. Further details are discussed in *Configuration of R/3 on hosts with much RAM*, SAP OSS Note 146528.

3.1.2 Numeric user and group ID

An SAP system is very seldom an isolated system. File transfers between systems are daily business. If these transfers are done through AIX utilities, such as Network File System (NFS) or the **rcp** command, the AIX file permissions come into effect. These permissions are linked to the numeric user and group identity. Nasty side effects occur if these numeric identities are not consistent across all involved nodes. Fortunately, it is relatively easy to avoid such situations. For example, all AIX users and groups can always be maintained on a dedicated node. Then, certain files are distributed across all other nodes to synchronize the user and group identities. The numeric identifiers should be chosen according to a coordinated convention because of the importance for future extensions. The following proposal is documented if such a convention does not exist for the planned environment. The numeric identifier of an AIX user, a password, and settings are stored in the AIX configuration files /etc/passwd, /etc/security/passwd, /etc/security/limits, and /etc/security/user.

A concept for numeric user identifiers considering all aspects previously discussed is important for a consistent and flexible installation. Such a concept is outlined in Example 3-1.

The instance number is an important parameter to form a valid numeric user identifier. This numeric value is defined during the installation of every SAP instance and has to be chosen according to the *SAP Installation Guides*. Only values between 00 and 97 are valid. All instances on a node need unequivocal numbers. Distinct numbers within a certain collection of nodes are useful to avoid conflicts during future relocations. For more details, see 3.5, “SAP instance” on page 32.

Example 3-1 Naming concept: User identifier

```
instance_number ::=  00 .. 97   Number of SAP instance [consecutive
                                numbering within a node/LPAR category]

domain_base ::=      2500       Special user
                     3000       Basenumber for SAP system with
                                instance number 00 in domain A
                     4000       Basenumber for SAP system with
                                instance number 00 in domain B
                     5000       Basenumber for SAP system with
                                instance number 00 in domain C

user_number ::=      0          sidadm
                     1          orasid
                     3          FTP user
                     5          Systems management user

uid ::=                         domain_base + ( instance_number * 10 ) +
                                user_number
```

User identifiers derived from this concept have the following values:

- 2503: FTP user
- 3000: SAP administrative user (sidadm) of instance 0 in domain A
- 5051: Oracle user (orasid) of instance 05 in domain C

The numeric identifier of an AIX group and its settings are stored in the files /etc/group and /etc/security/group. A concept for numeric group identifiers considering all aspects previously discussed is important for a consistent and flexible installation. Such a concept is outlined in Example 3-2.

Example 3-2 Naming concept: Group identifier

```
group_number ::=      1      bin

                      2      sys

                      210    SAP administrative group (sapsys)

                      211    Oracle administrative group (oper)

                      212    Oracle database group (dba)
```

A group identifier derived from this concept is the value 212 for the group named dba. This group, for example, is the default group for all Oracle administrative users on AIX.

3.2 IP names and host names

An SAP system has some predefined requirements for the environment it is installed on. These requirements are described in the appropriate *SAP Installation Guide*. A general rule is that each SAP system needs an IP label name that is resolvable as the host name and a unique port number for its services.

If a node contains several systems, either one dedicated system or all locally installed systems can be relocated at the same time. However, an SAP system, or more precisely an SAP instance, requires a fixed IP address that it is bound to. At the moment, it is requested that this address is the resolvable IP address of the host name, as discussed in *SAP R/3 in switchover environments*, SAP AG. An exception exists only for high availability solutions without identity takeover, such as HACMP. An SAP system within such a solution can, under certain circumstances, temporarily violate this rule.

This rule might soften in the future to allow a more flexible handling of SAP systems. For example, it could be conceivable to relocate dedicated single systems between several nodes. In this solution, every system would be assigned its own virtual IP address, and these virtual IP addresses are relocated according to the system. They are implemented as IP aliases that are associated to the appropriate network adapter. Besides the capability of the node to provide all resources for an acceptable performance, some implementation details limit the number of systems on a node in that scenario. The number of IP aliases per network adapter in AIX 5L can be increased at least up to eight. However, the maximum number of IP addresses per network is limited by the network mask of the accessed network that is derived from the existing network design.

A temporary relocation of all SAP systems on a node at the same time is technically possible and fully supported by SAP. In this scenario, all systems and the associated IP address are relocated together. Because this does not include a node identity takeover, certain SAP profile parameters are necessary, such as the instance profile parameter SAPLOCALHOST. From an SAP system standpoint, there is no difference between traditional, automatically triggered high availability solutions and manually initiated temporary relocations. These implementations are completely identical.

If all nodes belong to the same IP access network, basically all systems can be relocated to any node seen from this standpoint. Of course, further restrictions exist for storage accessibility or aspects of maintainability.

The basic principle for handling IP addresses during system relocations is shown in Figure 3-1 on page 16 and Figure 3-2 on page 17. Figure 3-1 on page 16 shows the base IP configuration of a relocatable system. Two systems are installed on every node and represented as rectangles in the figure. All SAP systems on a node are bound to an IP address that is identically named as the host name. These relations are expressed by named arrows crossing ellipses that represent the network adapters. Furthermore, in every node, an additional network adapter is installed to ensure network accessibility in all situations, for example, during maintenance activities. This special topic is discussed in more detail in *A Holistic Approach to a Reliable Infrastructure for SAP R/3 on AIX*, SG24-5050.

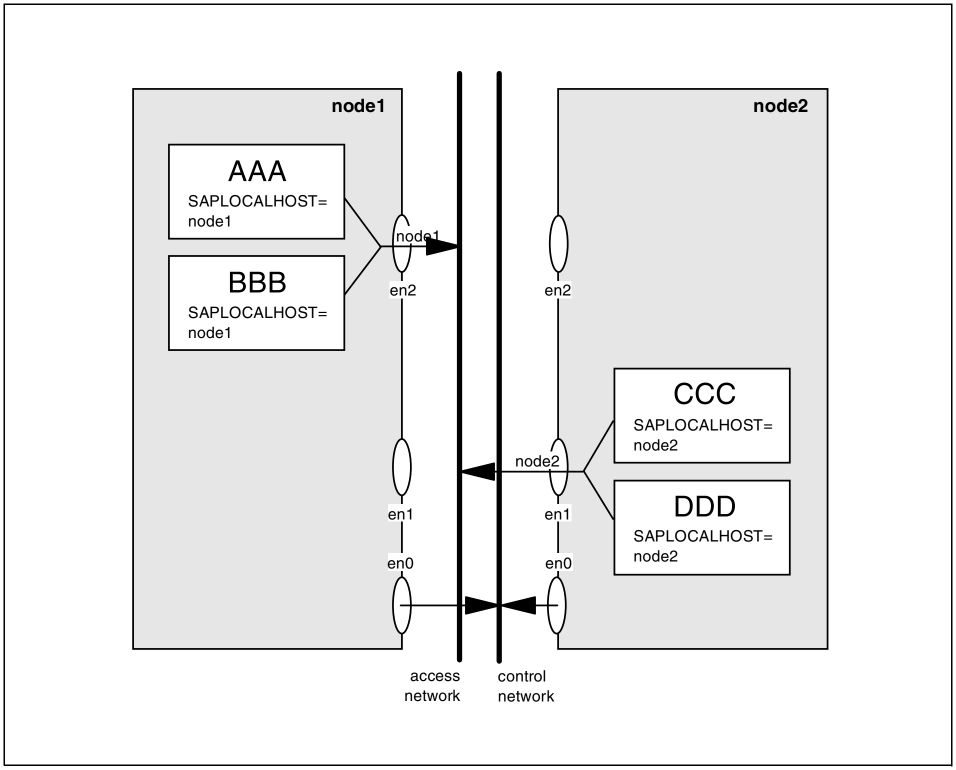

Figure 3-1 IP configuration (normal case)

Figure 3-2 on page 17 shows the situation after a relocation of all systems from node1 to node2. The associated IP label node1 was transferred together with the systems AAA and BBB. The simplest case for transferring an IP label is to reconfigure an additional adapter with this value, which is used in the example. It is also possible to establish the IP label as an IP alias on an adapter that is already in use to connect the access network. Some high availability solutions, such as HACMP Version 4.4, support these scenarios only in conjunction with special extensions. Its implementation restrictions should be minded.

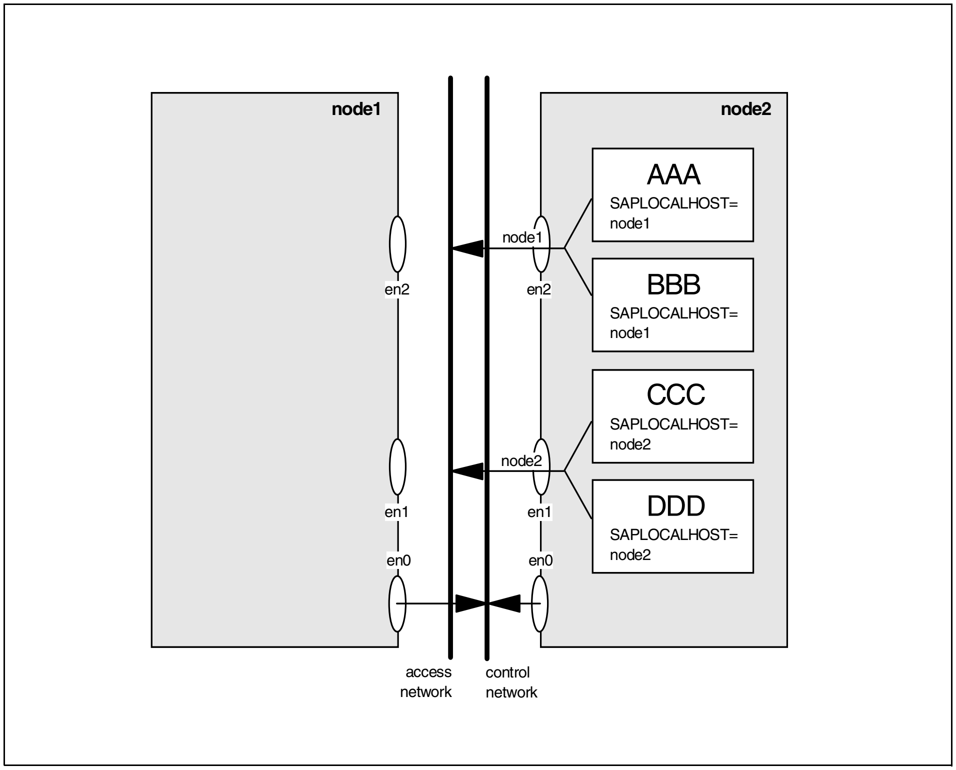

Figure 3-2 IP configuration (after relocation)

An IP address is always associated to the Media Access Control (MAC) address of the appropriate network adapter. Because an adapter supports only one MAC address at the same time, every relocated IP address needs a dedicated network adapter. Otherwise, only an IP address relocation without MAC address relocation is possible. This causes the loss of all connections if the ARP table entries in all components connected to the network segment have not been flushed. This behavior is not critical as long as the nodes are connected to a routed network, because ARP tables on routers are flushed regularly. However, this behavior can be critical for clients connected over switched networks.

3.3 Disk storage

The directory structure for installations of complex software products, such as SAP, is always predetermined. Standard *SAP Installation Guides* recommend the mapping of this structure to dedicated file systems. These recommendations are based particularly on features and limitations of older hardware components, for example, single stand-alone SCSI disks. Modern disk subsystems are working with write caches and very large read caches and have much better performance characteristics.

Usually, all disks are protected against data loss with RAID1 or RAID5. This protection is only sufficient in case of hardware failures. Disaster tolerant environments require at least two physically isolated storage locations. The internal AIX Logical Volume Manager (LVM) or special software products of the storage subsystem, such as Peer-to-Peer Remote Copy (PPRC), can be used to synchronize the disks over these locations.

Generally, the physical implementation should have no impact on the volume group and file system layout. Performance considerations are nonsensical for single disks within storage subsystems, because their performance is mostly determined by available read and write caches, as well as the utilization of internal crossbar switches and bus systems. The smallest design entities are disk groups, or so-called ranks.

3.3.1 Volume groups

A suitable mapping of the directory structure to file systems, logical volumes, and volume groups is, besides the correct implementation of the storage area network (SAN), very important to ensure the relocatability of a system. All files belonging to a system should be separated from files associated to the local node. This design principle allows for future, unproblematic system relocations.

A concrete implementation based on this principle contains one root volume group and several system-specific volume groups. The root volume group accommodates only system-independent files, as listed in Table 3-3 on page 33. There are some system-specific files that are stored only temporarily, for example, offline redo logs or dumps during database reorganizations. It is a special case if they are stored in a common file system that is installed in the root volume group. This is an exemption, but in this case, it does not prevent successful system relocations, because all of these files are temporary.

A naming concept for volume groups, considering all aspects previously discussed, is important for a consistent and flexible installation. Such a concept is outlined in Example 3-3 on page 19.

Example 3-3 Naming concept: Volume groups

```
vgname ::=    <ident> '.' <type> 'vg' [ '.' <number> ]

ident ::=     { <node_id> | <SID> }

node_id ::=   <unambiguous identifier of the node or LPAR>

SID ::=       <ID of the system>  SAP system identifier
                                  [capital letter]

type ::=      sapr3             R/3 system
              sapbw             BW system
              saplc             liveCache
              sdb               Shadow database
              arch              Offline redo logs
              reorg             Reorganization dump

number ::=    01..99            Consecutive numbering of volume groups
                                for a system
```

A volume group name derived from this concept is, for example, AAA.sapr3vg.01. This name describes the first volume group of the SAP R/3 system AAA. A volume group containing a common file system to store offline redo logs on node NODE17 might be named NODE17.archvg. For more details about common file systems, see 3.3.4, “Oracle file systems” on page 25.

3.3.2 File systems

Basically, two design alternatives exist for file systems. Either all files are spread over some relatively small file systems or a few large file systems. Both design alternatives have their own special advantages and disadvantages. All installations have to obey the predefined directory structure. All directory paths must be accessible, but the original directory structure can be preserved through the installation of links.

Splitting the storage capacity into small file systems according to the *SAP Installation Guides* prevents disturbing space competing effects. On the other hand, this layout implies a relatively high waste of capacity, because every file system needs sufficient spare capacity. The possibility that a file system accidentally runs out of space exists for every file system. Many existing file systems increase this value noticeably for the complete SAP system.

For storage capacity implemented on modern cached and striped disk subsystems, a few large file systems are better than many small ones in regard to administration efforts and flexibility. This layout based on large file systems avoids these disadvantages, but disobeys some rules of the *SAP Installation Guides*. Relatively large spare capacities decrease the likelihood of space contention. It also has the advantage of optimal resource utilization and flexibility, because this spare capacity can be used by all systems. Annoying space transferring tasks from one file system to another one are no longer necessary.

On the other hand, large file systems also have certain space limitations. These limitations are either caused by physically available disk space or by the internal file system structure, for example, the bit length of internal pointers. The journaled file system (JFS) parameters shown in Table 3-1 define the internal structure and influence the maximum file system size. They must be chosen during the creation of a file system and cannot be modified later.

The new Journaled File System 2 (JFS2) is available only on AIX 5L, but has a maximum size of 5 GB and no internal structural limits. For more information, see *AIX Logical Volume Manager from A to Z: Introduction and Concepts*, SG24-5432. Standard journaled file systems on AIX 5L still have the following size limitations.

Table 3-1 Journaled file system parameter on AIX for JFS

Maximum size (GB)	NBPI	Minimum AG size	Fragment size
8	512	8	512, 1024, 2048, 4096
16	1024	8	512, 1024, 2048, 4096
32	2048	8	512, 1024, 2048, 4096
64	4096	8	512, 1024, 2048, 4096
128	8192	8	512, 1024, 2048, 4096
256	16384	8	512, 1024, 2048, 4096
512	32768	16	1024, 2048, 4096
1024	65536	32	2048, 4096
1024	131072	64	4096

As can be seen from Table 3-1, the maximum JFS size on AIX is 1 terabyte. However, it is not a good idea to exhaust this maximum because of time consuming file system checks that might be necessary in case of a serious system crash. To ensure a reasonable run time for file system checks, it is better to split the storage capacity into some medium-sized file systems. This way, every file system can be checked independently and in parallel, which decreases the total run time to a fraction of the undivided run time. A few medium-sized file systems, perhaps 250 GB, are better than one very large file system.

The effectively usable number of disks is dependent on the kind of mirroring used, eventual existing HACMP clusters, and the volume group type. Table 3-2 shows the derived maximum number of disks per volume group. Every HACMP cluster requires an extra quorum disk that cannot be used to store application data, and therefore, the number of usable disks is decreased by one. This special disk ensures that only one copy of a file system can be active at a time.

Table 3-2 Maximum number of effectively usable disks per volume group

Volume group type	Native	LMV mirroring	LVM mirroring and HACMP
Standard volume group	32	16	14
Big volume group	128	64	62

All file systems are installed into underlying logical volumes. Every logical volume is created within a volume group. All logical volumes, according to an SAP system and the database, should be grouped as a unit to decrease the administration effort. Creating all logical volumes of every system within a dedicated volume group fulfills this recommendation.

A naming concept for logical volumes, considering all aspects previously discussed, is important for a consistent and flexible installation. Such a concept is outlined in Example 3-4 on page 22.

Example 3-4 Naming concept: Logical volumes

```
lvname ::=    <ident> '.'  <type>  [ '.'  <number > ]

ident ::=     { <SID> | <node_id> }

SID ::=       <ID of system>     SAP system identifier
                                 [capital letter]

node_id ::=   <unambiguous identifier of the node or LPAR>

type ::=      lvdata             Tablespace data files
              lvoracle           Database executables
              lvglobal           SAP executable, transport data, profiles
              jfslog             Log of journaled file systems
              lvarch             Oracle offline redo logs
              lvtrans            SAP transport files
              lvquorum           Place holder on HACMP cluster quorum
                                 disks

number:       01..99             Consecutive numbering of logical volumes
```

A logical volume name derived from this concept is, for example, AAA.lvdata.01. This name describes the first logical volume containing tablespace data files of the SAP R/3 system AAA. A logical volume containing a common file system to store offline redo logs on node NODE17 might be named NODE17.lvarch. For more information about common file systems, see 3.3.4, “Oracle file systems” on page 25.

3.3.3 File system layout

A file system layout that meets all the previously discussed criteria is shown in Figure 3-3 on page 24 and Figure 3-4 on page 25. The skeletons of these figures is the standard directory structure of SAP and Oracle database for two systems on one node that are named *AAA* and *BBB*. The layout can be expanded by repeating all parts related to system AAA. Figure 3-3 on page 24 shows all the directories belonging to a database and Figure 3-4 on page 25 shows all the directories belonging to the SAP R/3 or BW instances. A rectangle describes the logical volume and the mount point of the associated file system. Graphically overlapping rectangles are equivalent to hierarchically over-mounted file systems.

Most Oracle database-related files are stored in subdirectories of /oracle. There are only a few small files located in /etc, which do not matter. The /oracle directory is a file system containing no system-specific permanent data, but rather system-independent software, such as Oracle client software and Oracle stage area. The file system is installed into the logical volume <node>.lvoracle that is located in the root volume group. Other file systems within the root volume group contain only common AIX files, for example, AIX executables and libraries. Instance- and database-specific files are located in instance specific file systems that are created on instance-specific volume groups.

The logical volume AAA.global belongs to the volume group AAA.sapr3vg.01 and contains the file system /global.AAA for SAP binaries and configurations, as well as client transport system files. Some small file systems, according to the *SAP Installation Guide*, have been combined to these bigger ones. The original directory structure is preserved through the installation of links. These are shown in Figure 3-4 on page 25. The /global.AAA/extdata directory provides system-specific storage capacity for intra-system file exchanges.

The logical volume AAA.lvoracle belongs to the volume group AAA.sapr3vg.01 with the file system /oracle/AAA installed containing Oracle binaries and configuration files. A hierarchically mounted file system /oracle/AAA/sapdata1 is used to store the database tablespace files. It is installed in the logical volume AAA.lvdata.01 and also belongs to the volume group AAA.sapr3vg.01.

One sapdata file system is sufficient for relatively small SAP systems. In that case, data and index tablespace files are mixed into one directory. However, at least two file systems are necessary, mapped on different disks if it is planned to separate index and data files. For very large systems, additional sapdata file systems can be added in pairs as long as there is enough space available within the volume group. Whenever the maximum number of disks per volume group has been reached, a new volume group can be created. See Table 3-2 on page 21.

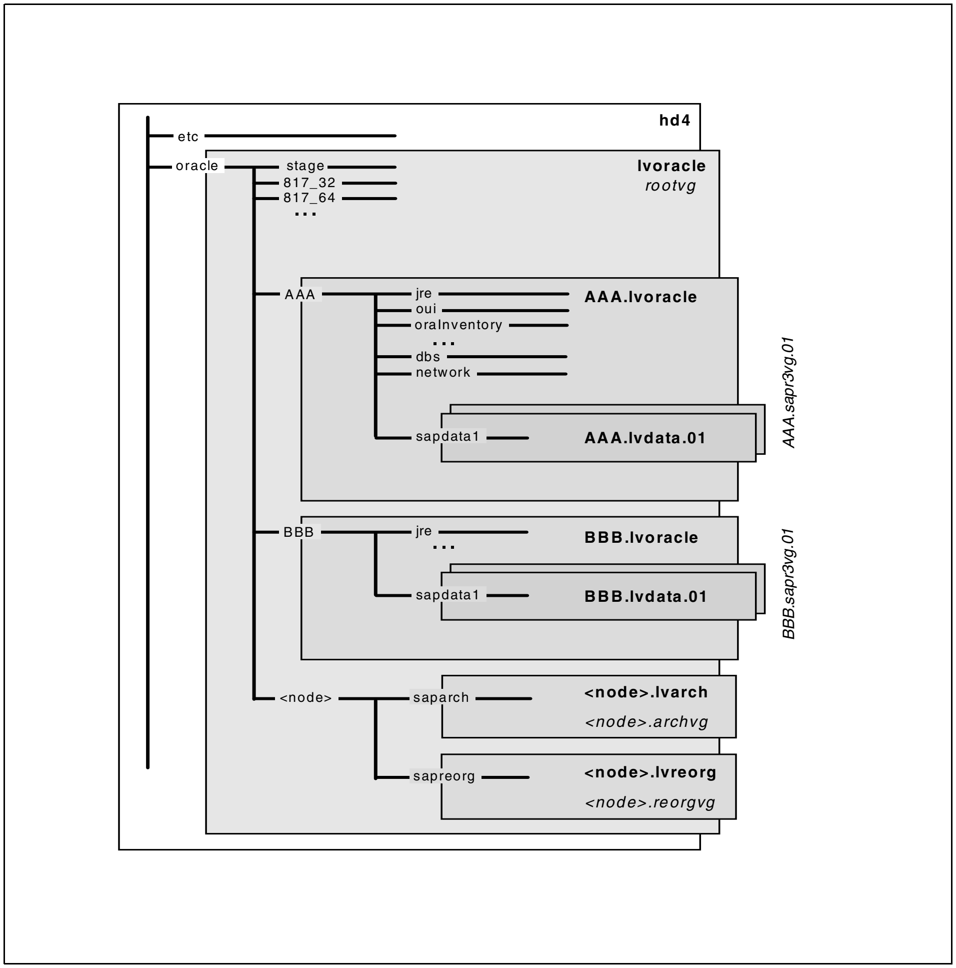

Figure 3-3 Volume groups and file systems for Oracle databases

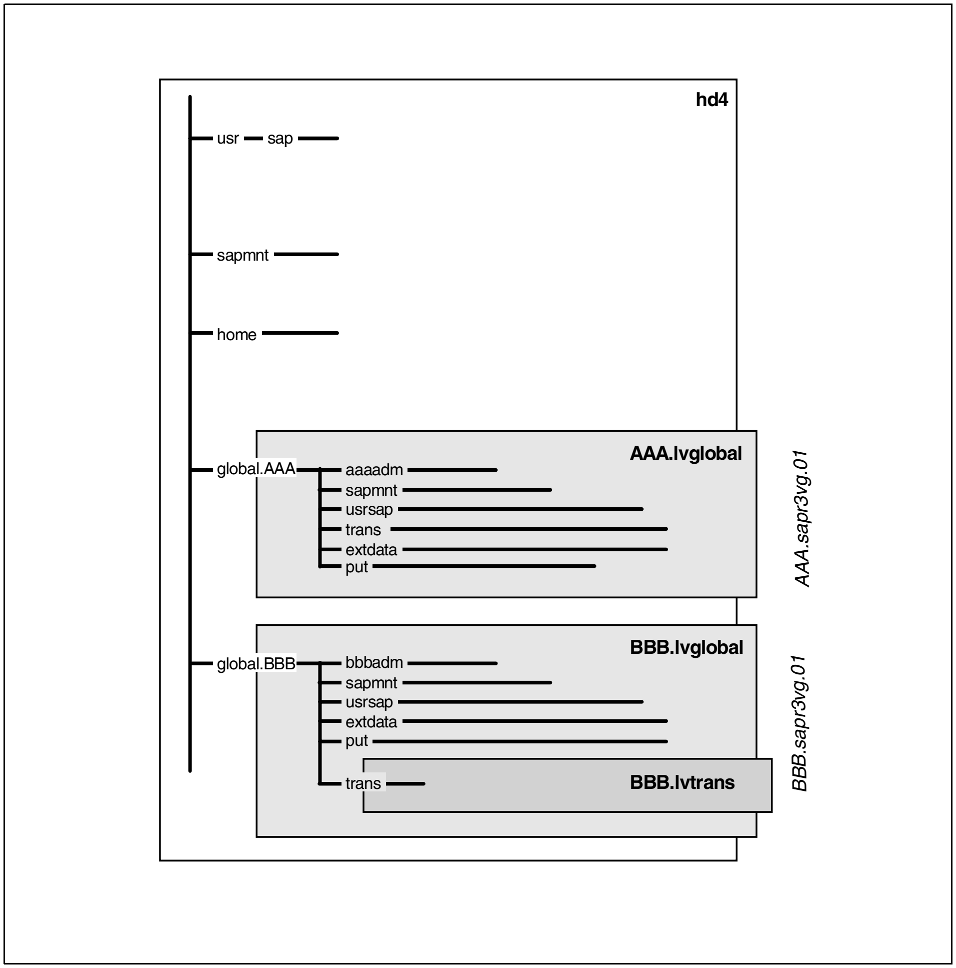

Figure 3-4 Volume groups and file systems for SAP instances

3.3.4 Oracle file systems

A vital part of an SAP system is always the database. According to the standard *SAP Installation Guide*, it is possible to install several Oracle databases on a node. The following concept enhances the file system structure for a more effective system administration as previously outlined. Some directories are replaced with links to new locations to hide these modifications. Therefore, all files can always be accessed over their unmodified directory path. The following concept is based on Oracle, but can be adapted easily for other widespread databases, such as DB2 UDB.

The /oracle/SID/saparch directory is a space-critical component. If any database is not able to save the offline redo logs, an archiver stuck will occur. Having enough available space avoids this situation. However, it might be a waste of storage capacity to assign enough spare capacity to every system, because it is difficult to determine the specific amount of sufficient free space. Changing database workloads can result in very different numbers of offline redo logs. An environment without a common archive log space works well as long as the file system is perfectly monitored and maintained twenty-four hours a day. If any system runs out of space, a manual intervention is required. However, the usage of a common file system with a large amount of free space for offline redo logs dramatically reduces the likelihood of a space shortage. This solution also ensures stable system operation in an environment with less perfect system monitoring.

A modification of the oracle parameter log_archive_dest in the init<SID>.ora file to the new value /oracle/<SID>/saparch/global redirects all archive redo logs to the common file system. A link is not a sufficient solution, because normally, a copy of the database control file is located in the /oracle/SID/saparch/cntrl directory. When using a link, this system-specific control file would be mistakenly stored in the common file system for temporary objects.

Every Oracle database, like some other databases, writes all information required for recovery activities into special files, the so-called online redo logs. They exist in two different directory paths. Because of its importance for recoverability and performance, it is highly recommended by Oracle and SAP to store both copies in different file systems using different disks. However, a disk storage subsystem with a write cache facility providing RAID5 level disks is a suitable alternative for recoverability and performance. Therefore, in our example, no dedicated file systems are necessary for online redo logs.

Some files within an Oracle database environment are database independent, for example, the Oracle stage area and the Oracle client software. They are stored in the original path of the system independent file system /oracle.

3.3.5 Symbolic links

The Oracle standard installation assumes that there are dedicated file systems for offline redo logs and dumps during reorganizations. Current databases are often bigger than 100 GB. Any reorganization of such a database can require a lot of space to store temporary export/import files. The /oracle/SID/sapreorg directory of all systems was relocated to a dedicated file system to improve the space utilization. Because parallel reorganizations of two or more systems are seldom, every system can normally use approximately the whole space capacity.

An issue of the file system concept is the limitation of the number of file systems. Therefore, there is normally one file system for tablespace files. Only for larger databases are more file systems necessary. Missing file systems were replaced with symbolic links to hide this modification from Oracle binaries. To ensure relocatability of the system, no link can contain any system-specific part, for example, link sapdata2 at the directory path /oracle/SID points to sapdata1 and not to /oracle/AAA/sapdata1.

The link concept that is derived from the file system concept previously discussed is shown in Figure 3-5 on page 28. The skeleton of the figure is the standard directory structure for two Oracle databases on one node that are named AAA and BBB. The layout can be expanded by repeating all parts related to system AAA.

An arrow describes a link that points to the directory where the files are physically stored. Arrows with a dotted line represent variable links. They are used if a directory path does not contain any system-specific part to distinguish between systems in a multiple system environment. The corresponding directory is replaced by a link to solve this situation. Unfortunately, this solution requires an intervention to adjust the links before certain tasks, such as Oracle upgrades, can be started. Moreover, it prevents the running of such tasks concurrently.

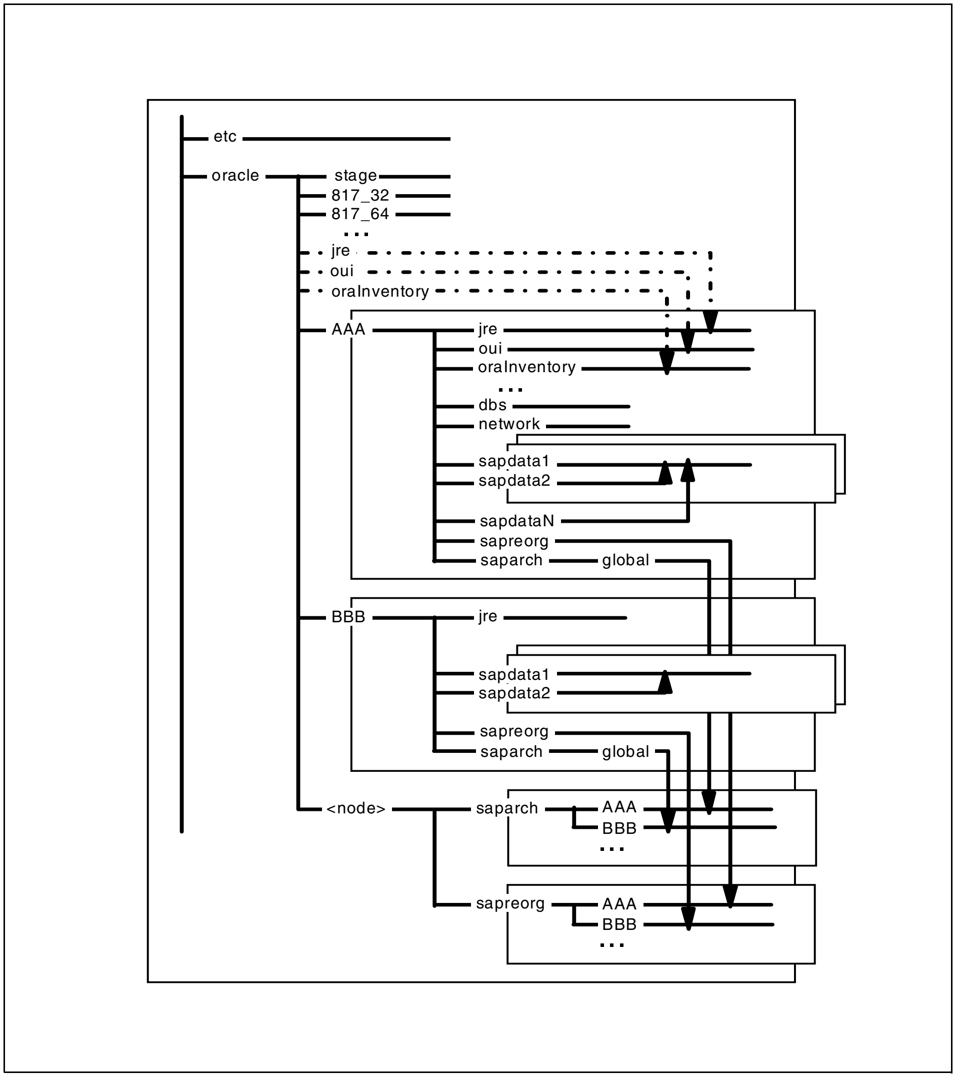

Figure 3-5 Links for an Oracle database installation

3.3.6 SAP file systems and symbolic links

Fortunately, it is almost possible to install several SAP systems on one node according to the standard *SAP Installation Guide*. The concept previously outlined enhances the file systems structure for a more effective system administration. The link concept that is derived from the file system concept is shown in Figure 3-6 on page 30. The skeleton of the figure is the standard directory structure of SAP for two systems on one node that are named AAA and

BBB. The layout and can be expanded by repeating all parts regarding system AAA.

An arrow describes a link that points to a directory where the files are physically stored. Arrows with a dotted line represent variable links. They are used if a directory path does not contain any system-specific parts to distinguish between systems in a multiple system environment. The corresponding directory is replaced by a link to solve this situation. Unfortunately, this solution requires an intervention to adjust the link before certain tasks, such as an SAP upgrade, can be started. Moreover, it prevents the running of such tasks concurrently.

The original directories are replaced with links to new locations to hide these modifications. All files can be accessed over their unmodified path. The following modifications are the most significant on the standard file system layout:

- /home/sidadm moved to /global.SID/sidadm and replaced with a link.
- /usr/sap/SID moved to /global.SID/usrsap and replaced with a link.
- /sapmnt/SID moved to /global.SID/sapmnt and replaced with a link.

If a directory path contains any system-specific part, for example, the system ID, a multiple systems installation is possible. Unfortunately, there are a few directories without such a system-specific part. The most important example is the directory /usr/sap/put that is used during SAP upgrades only. A suitable solution to solve this conflict is to establish an variable link from /usr/sap/put to /global.SID/put that has to be adjusted before every SAP system upgrade. This complies with the rule of unmodified paths for all SAP binaries. However, it has the side effect that concurrent SAP upgrades on a node will no longer be possible.

Another directory without any system-specific part is /usr/sap/tmp. It is only used by the program saposcol and can remain as a directory within the root volume group.

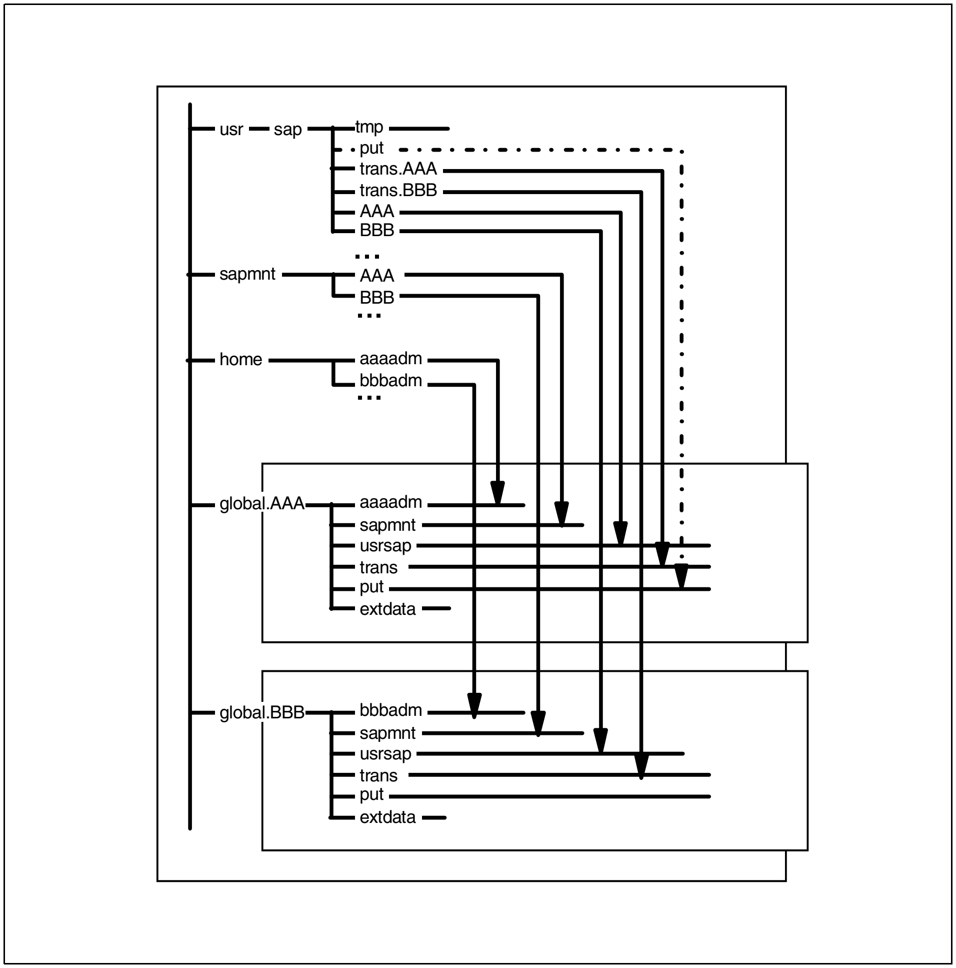

Figure 3-6 Links for an SAP instance

3.4 Oracle database

An installation of more than one database per node does not necessarily follow standard installation procedures. Often, they require some adaptations to the installation process.

3.4.1 Oracle Installer

Normally, an Oracle database is installed using Oracle Installer. This program (runInstaller) saves information about installation states and configurations in files in a path without any system-specific part. For more information, see *Creating a new or 2nd Oracle SID with runInstaller*, SAP OSS Note 350251. This prevents concurrent Oracle installations or upgrades of different databases running on the same node. This affects the /oracle/inventory, /oracle/jre, and /oracle/oui directories. They are replaced with symbolic links that point to directories according to the database to be installed. These links have to be adjusted for every installation and reinstallation before Oracle Installer can be started. A similar solution modifies the /etc/oraInventory file defining the installation path. For more information, see *Creating a new or 2nd Oracle SID with runInstaller*, SAP OSS Note 350251.

The /etc/oratab file contains a list of all existing databases and is used by Oracle Installer. To ensure correct function during database updates within an relocatable system environment, this file must contain the identifiers of all designated Oracle databases. The file can be maintained with a normal text editor for adding missing entries to prevent problems after system relocations between different nodes.

3.4.2 Kernel extensions

Up to Oracle Version 8.x, every Oracle database requires a loaded Oracle-specific kernel extension. The binary is delivered by the Oracle Corporation. Only one Oracle kernel extension, /etc/pw-syscall, can be loaded and active at the same time. However, this does not cause any problems according to Oracle support, because the kernel extension is downward compatible to Oracle 7.x. To ensure the relocatability, the kernel extension must be installed and loaded on every designated target node. Beginning with Oracle Version 9, a kernel extension is no longer necessary.

3.4.3 Oracle listener

There are two architectural choices for installing Oracle listeners. Either one common listener exists for all databases, or every database has a dedicated listener process. The second alternative is recommended because the existence of only one listener holds the possibility of interfering workloads and interfering maintenance tasks. Relocatable systems always require dedicated Oracle listener processes.

The installation of dedicated Oracle listeners for all databases on a node requires a dedicated port number for every database listener service. If relocatable databases are planned, these numbers have to be unique on all designated nodes. The listener configuration listener.ora and names file tnsnames.ora for every system are configured with the correlating system identifier and port number. An example is shown in “Sample Oracle files” on page 104. Customer-specific parts have to be individually adjusted and are highlighted. The files are usually located in /oracle/<SID>/<db_version>/network/admin.

The instance number is an important parameter to form a valid port number. This numeric value is defined during the installation of every SAP instance and has to be chosen according to the *SAP Installation Guide*. For more details, see 3.5, “SAP instance” on page 32. A concept for port numbers considering all aspects previously discussed is important for a consistent and flexible installation. Such a concept is outlined in Example 3-5.

Example 3-5 Naming concept: Port number

```
port_number ::=        1700 + instance_number

instance_number ::=    00 .. 97  Number of SAP instance
                                 [consecutive numbering within a node/LPAR
                                 domain]
```

For example, port number 1710 belongs to the Oracle listener of a system with the SAP instance number 10.

3.5 SAP instance

An installation of more than one SAP system per node does not necessarily follow standard installation procedures. This often requires some adaptations of the installation process. However, minimal prerequisites for an active SAP system of any kind are always some hardware resources, for example, memory, processor, and storage capacity. All components used by an SAP system and its associated database are either system specific or unspecific.

Table 3-3 Components of SAP systems

Type	Component	Resource
Global	AIX	► Sufficient paging space ► lpp sources ► Customized AIX environment ► Enabled Portable Stream Environment ► IP label with its name identical to the host name
	Oracle	► Oracle kernel extension ► Oracle stage area
	SAP	► SAP hardware key ► SAP performance collector
System specific	Oracle	► Disk space for database files ► Oracle binaries ► Customized environment of the Oracle user ► Customized Oracle profile
	SAP	► SAP binaries ► SAP license key ► Customized environment of the SAP user ► Customized SAP profile

3.5.1 Instance number

An SAP instance is unambiguously identified by a unique combination of a system identifier (three characters) and system number (two digits). For more information, see *Several systems/instances on one UNIX computer*, SAP OSS Note 21960. The system identifier is normally specified according to a separate naming convention and can be chosen without any architectural limitations. In contrast, the system number must be at least unique within the node or LPAR where the system is installed. If you plan to relocate a system in the future, all instance numbers on all eligible nodes must comply to the requirement of uniqueness.

The instance number is an important parameter to identify a specific instance on a node. This numeric value is defined during the installation of every SAP instance and has to be chosen according to the *SAP Installation Guide*. Only values between 00 and 97 are valid. All instances on a node need unequivocal numbers. Distinct numbers within a certain collection of nodes are useful to avoid conflicts during future relocations. Careful planning is essential to avoid any unnecessary administration efforts for renaming a system. One conceivable approach is to assign all nodes or LPARs to a domain for productive and non-productive systems. A mixed operation of these two system categories on a node is not allowed by SAP. Therefore, it is very unlikely that a system is relocated to a node into a foreign domain.

Example 3-6 Naming concept: Instance number

```
instance_number ::=    00 .. 97   System number of SAP instance
                                  [consecutive numbering within a
                                  node/LPAR domain]
```

For example, the instance number 10 belongs to the productive SAP system within the domain for productive systems. No other productive system is installed using this instance number, whereas a development system with instance number 10 might exist.

3.5.2 Transport directory

The transport system is a fundamental piece of every SAP system. Some SAP transactions and special programs belong to the transport system. They use an AIX directory for data exchange between systems that are grouped together as a transport domain. The directory is unique within every transport domain and mostly mounted over Network File Systems (NFS). If several systems are installed on a node, they are often members of different transport domains.

The standard path of the transport directory is /usr/sap/trans. This path is unusable in an environment with several systems on a node, because the path is not system specific. However, the transport directory can be customized using the SAP transaction STMS. The new path includes at least its own system identifier. The mapping between the transport directory from the systems point of view to the physical storage location ensures NFS mounts or static links. Network File Systems are used for mapping between nodes. Links are used for the same assignment on local nodes.

The use of a foreign system identifier, for example, the identifier of the transport domain controller, is not recommended. If, for example, two relocatable systems are installed on two nodes and belong to the same transport domain, they are using the same local transport directory. These equally named transport directories might conflict during the unmount if one of these systems is relocated back to the original node.

A concept for naming the local transport directory considering all aspects previously discussed is important for a consistent and flexible installation. Such a concept is outlined in Example 3-7 on page 35.

Example 3-7 Naming concept: Transport directory

```
SID ::=               <ID of system>     SAP system identifier
                                         [capital letter]

local_trans_dir ::= '/usr/sap/trans.' <SID>
```

A transport directory name derived from this concept is, for example, /usr/sap/trans.AAA. This name describes the local transport directory of the SAP system AAA. It is mounted using NFS or directly linked to the path /sapmnt.DDD/trans, where DDD is the name of the transport domain controller.

Figure 3-7 on page 36 shows an implementation of several transport domains in a node. In this example, two systems on different nodes belong to domain AAA and another two belong to domain BBB. The transport data for domain AAA is physically stored on node1. System AAA accesses the data over the directory path /usr/sap/trans.AAA that is linked to /sapmnt.AAA/trans. On node2, system BBB uses a remote Network File System mount to access files on the transport domain controller AAA.

The standard transport directory is not usable in a multiple system environment. Alternative directories have to be defined on every SAP system. That requires some customization of values within the SAP transaction STMS, as well as the default profile DEFAULT.PFL with the transaction RZ10. The values of the following parameters have to be modified:

- `DIR_TRANS=/usr/sap/trans.SID`
- `DIR_EPS_ROOT=/usr/sap/trans.SID/EPS`

The modification of the default profile follows the usual process to customize any profile parameters and requires a restart of the SAP system. Adjustments with transaction STMS take effect immediately. The menu path is as follows:

overview

 system

  transporttool

   -> TRANSDIR /usr/sap/trans.SID

   -> EPS_ROOT /usr/sap/trans.SID

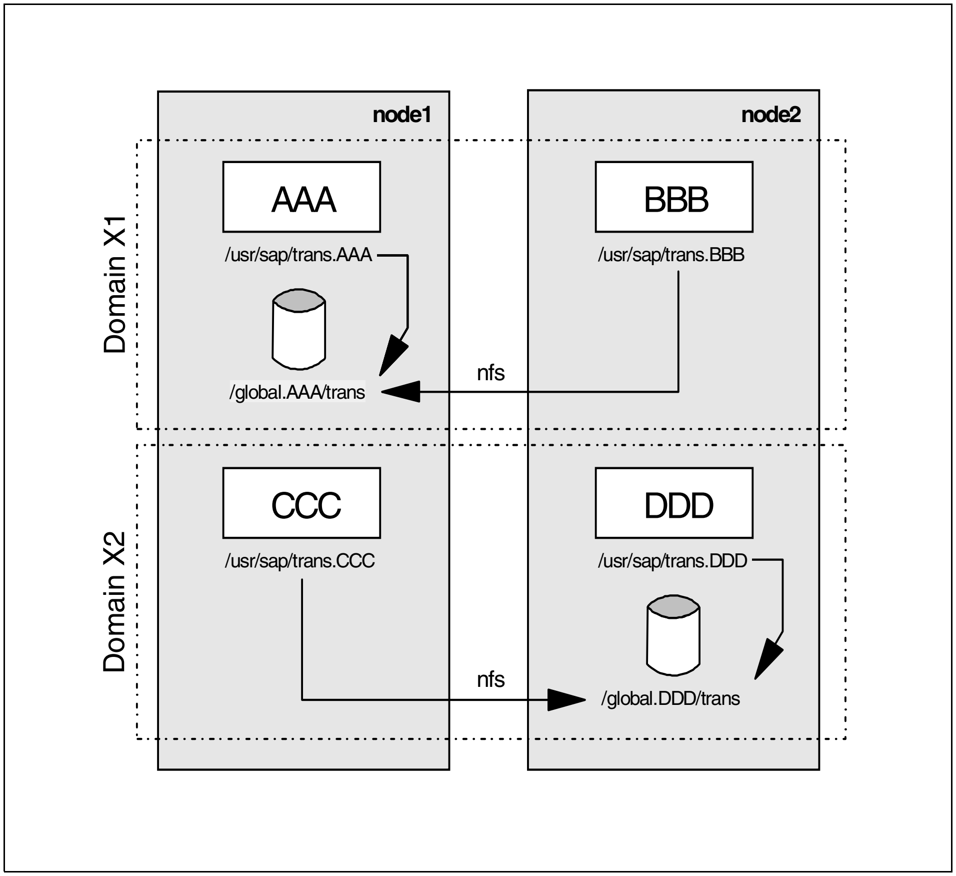

Figure 3-7 Transport directory

3.5.3 Performance collector

Although more than one system is running on a node, only one performance collector process should be active. If the collector is already running, further startup attempts fail. However, the program instance that is started first is accessible by all other systems. The SAP Computing Control Management Center (CCMS) does not recognize the multisystem environment, and all systems see the same global values.

Starting the performance collector from the system that started first after reboot has one big disadvantage. The file system /global.SID is locked again. Moreover, the instance has been stopped. This prevents unmounts and possible relocations. An alternative solution is to install a system-independent performance collector, for example, under /etc/sap/saposcol. The collector can be started from the inittab during the boot process:

```
mkdir /usr/sap/tmp
/etc/sap/saposcol
mkitab –i strload saposcol:2:once:su – sidadm –c
/etc/sap/saposcol –l
```

3.5.4 SAP license key

The SAP license key is absolutely necessary to start the system. The key is dependent on the three character system identifier and a hardware key. This license key is generated on the base of the IBM @server pSeries machine ID number and the system identifier. If an SAP system is relocated, an appropriate SAP license key is required. All keys are stored in the SAP database and are validated during every system startup.

For high availability solutions, it is possible to request more than one license key at the same time. All other relocations require newly applied license keys. No new license key is required after relocating within one IBM @server pSeries model 690, because the machine ID number is identical for all logical partitions.

3.6 Backup and recovery

Stable backup and recovery procedures are very important prerequisites for a reliable system. Especially complex installations, such as several SAP systems on a node, need a carefully designed and implemented solution. An operation belonging to a system must have no impact on files belonging to another system. Sensitive operations here include renaming, relocating, or removing a system. Therefore, the concept of backup objects was introduced.

3.6.1 Backup objects

A backup object is a well-defined collection of files or directories that are stored together in the backup system. Such collections are perhaps all the database and SAP binaries of a system or all tablespace files of a single database. The collection itself has to be distinct to prevent accidental overwriting of existing files belonging to foreign, perhaps running, systems. If IBM Tivoli Storage Manager is used to process all backups, every backup object is associated with a dedicated virtual node. In case of data loss, for example, due to hardware failure, the entire virtual node can be restored as an entity. Always restoring complete backup objects containing the latest file versions is easy and always produces a consistent environment.

An alternative solution is to store all files belonging to a node, including system-specific ones, in one common backup object. This is much more fault prone in case of a selective restore of a single system, because all files or directory trees have to be selected manually.

At least one backup object exists for every node or LPAR. This object is named AIX. It contains the operating system binaries, configurations, and other files within the root volume group rootvg. The backup versions within IBM Tivoli Storage Manager are not usable to process a complete reinstall from an uninstalled or destroyed node. An installable image of the root volume group is necessary for this task and can be created by the operating system command **mksysb**. This command saves, as a standard behavior, the file system structure of the root volume group, all boot information, and all files within the volume group. For a successful reinstall, only the AIX executables and libraries are necessary. However, these concepts also propose to store some system-independent files into the root volume group. These files will be saved in the virtual node AIX. These files are not required for the first reinstallation step, and therefore, it is unnecessary to save them in the bootable system image. To exclude certain files or directory trees, a special file, /etc/exclude.rootvg, can be customized. It contains at least a line /oracle/stage/ to limit the boot image to the absolute necessary size.

For every installed SAP system, there are two additional backup objects named EXE and DB. The object EXE contains all database and SAP binaries, as well as their configuration files. The backup copies of all tablespace files, control files, and online and offline redo logs belong to the virtual node DB.

A naming concept for virtual nodes considering all aspects previously discussed is important for a consistent and flexible installation. Such a concept is outlined in Example 3-8 on page 39.

Example 3-8 Naming concept: Virtual IBM Tivoli Storage Manager nodes

```
tsm_node ::=  <node_id> [ <SID> [ <suffix> ] ]

node_id ::=   <unambiguous identifier of the node or LPAR>

SID ::=       <ID of system>     SAP system indentifier, for backup objects
                                 EXE and DB only
                                 [capital letter]

suffix ::=    DB                 Database tablespace files and redo logs
```

An IBM Tivoli Storage Manager virtual node name derived from this concept is, for example, NODE17AAA_DB. This name describes a virtual node storing the tablespace data files of the SAP system AAA. A virtual node containing all SAP and Oracle executables belonging to this system (AAA) is named NODE17AAA. All AIX executables are associated with the virtual node named NODE17. The name of this node is identical to the host name. This allows the use of automatically generated IBM Tivoli Storage Manager passwords and is a prerequisite for this feature due to limitations in the IBM Tivoli Storage Manager client for AIX implementations.

3.6.2 Backup domains

Normally, a special definition file, the include/exclude file, is used to exclude files or to map existing files to dedicated virtual nodes and management classes. This constitutes a remarkable administration effort, because every system requires its own specially customized include/exclude file. A solution using the domain feature of IBM Tivoli Storage Manager with less administration effort is described in this chapter.

Introducing domain definitions for every IBM Tivoli Storage Manager client has a strong simplifying effect in conjunction with generalized include/exclude lists. Only the system-specific IBM Tivoli Storage Manager client configuration files still contain system-specific information. The include/exclude list is very simple and describes only generic rules. Therefore, the include/exclude list can be maintained on a dedicated master node and distributed to all other nodes. An example for such a file and a suitable IBM Tivoli Storage Manager client configuration is shown in “Sample IBM Tivoli Storage Manager client option file” on page 105 and “Sample IBM Tivoli Storage Manager include/exclude files” on page 106. Customer-specific parts have to individually adjusted and are highlighted. The path of the configuration file is, for example, /tsm/conf/client. It is important to save this file to ensure the ability of a complete disaster recovery.

The suggested path is contained within the root volume group for which an installable image is regularly created.

According to the design principle *disjoint units* every file is associated to exactly one domain, and therefore, a domain should never overlap other domains. A domain concept that meets all criteria previously discussed is shown in Figure 3-8 on page 41. The skeleton of the figure is the standard directory structure of SAP and Oracle database for two systems on a node. They are named AAA and BBB. This layout can be expanded through repeating all part belonging to system AAA. The top of the figure shows all the directories of a database, and the lower part shows all the directories belonging to the SAP instances. A rectangle describes a domain and the path name to the directory. Graphically overlapping rectangles are equivalent to distinct domains.

All instance- and database-specific files are located in four instance-specific directory paths: /oracle/<SID>, /global.<SID>, /oracle/<node>/saparch/<SID>, and /oracle/<node>/sapreorg/<SID>. These directories are interpreted as one domain that is defined according to the virtual node AAAEXE. System-specific interface data is stored in the directory /global.<SID>/exdata, complying with these concepts. These files belong to the IBM Tivoli Storage Manager domain AAAEXE. Database tablespace files and online redo logs are excluded from this domain. They are associated with an extra domain and virtual node AAADB. This virtual node is used by database backup utilities, for example, IBM Tivoli Data Protection for SAP R/3, as a storage location for Oracle data files and archived redo logs.

The root volume group contains only common files, for example, AIX executables and libraries, as well as Oracle client software. This volume group, excluding all other domains, is associated with the virtual node AIX.

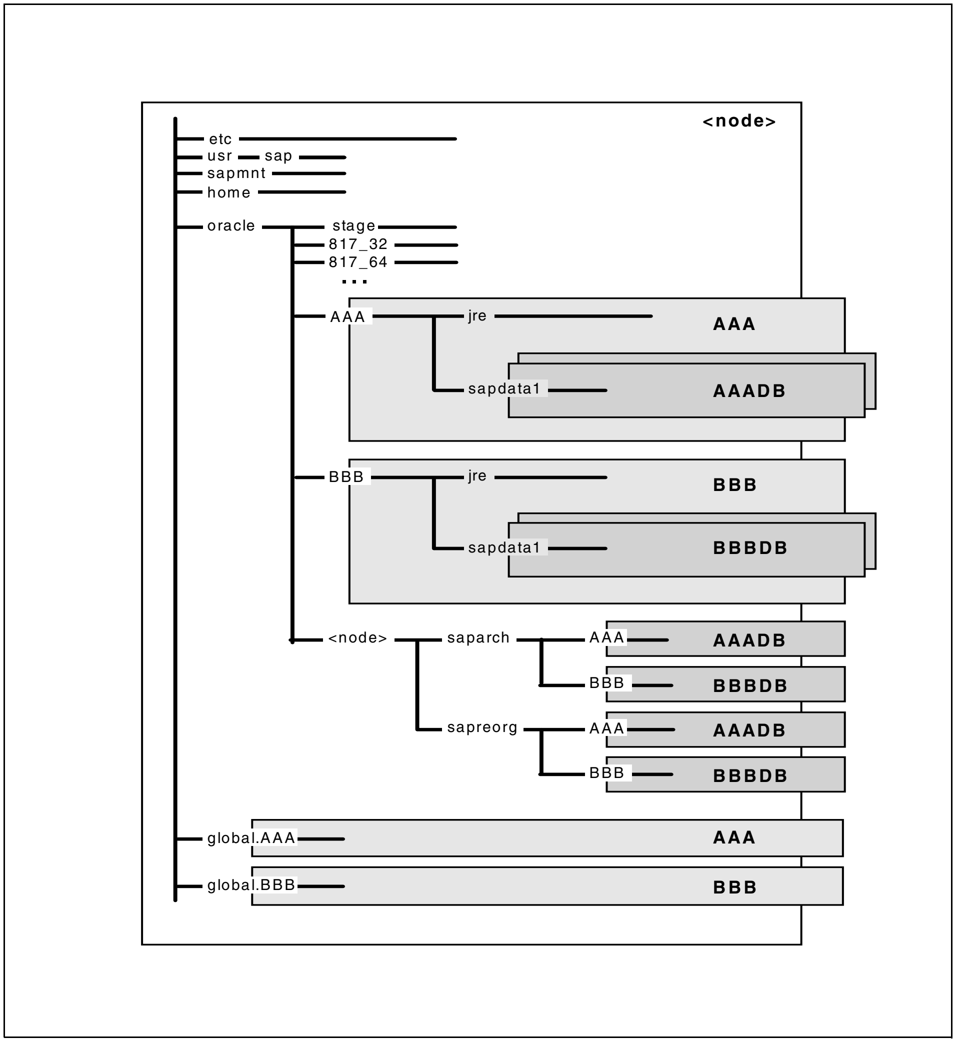

Figure 3-8 Backup domains

3.7 AIX Workload Manager

Part 2 of this document contains a dedicated section about WLM usage in SAP environments based on an evaluation project. Therefore, this section just provides a brief positioning and introduction of Workload Manger (WLM) mechanisms.

3.7.1 Positioning of AIX Workload Manager

After the introduction of LPARs, the question may rise: Why do we need WLM?

WLM is another option for partitioning resources, but without operating system (OS) isolation. WLM can run on any pSeries servers that are not LPAR enabled, as well as within LPARs themselves, in order to prioritize specific application tasks. Although SAP has been restrictive in supporting multiple R/3 systems within a single OS image for a long time, they have changed their position. *Several systems/instances on one UNIX computer*, SAP OSS Note 21960, states that users can consolidate mySAP.com applications, as follows:

- Several instances of the same system on one server

  Multiple instances of a single SAP system can run concurrently on the same server to exploit available hardware resources. For example, such a setup can be used to separate application workload on an instance level, to set up dedicated RFC instances, or to provide more than one spool work process prior to SAP Release 4. Several instances always place higher demands on resources than a single instance, because selected shared memory areas (for example, Program buffers and Table buffers) are duplicated. If SAP memory constraints caused by the 32-bit SAP kernel are the issue, the preferred solution is to move to a 64-bit environment rather than splitting the SAP system into multiple instances.

- Several central systems on a server

  Several central SAP systems (that is, a database and central instance together) can run concurrently on a single host or partition.

  This applies to multiple systems implementations independent of WLM usage. WLM can help to effectively dedicate resources to preferred applications. Per SAP the following restrictions apply:

  - Because of possible negative effects on production system stability, a combination of test, consolidation and productive systems on the same host is not recommended.
  - You should avoid combining 32-bit and 64-bit databases on the same server. If a mix is required, compatibility has to be verified with the database vendor.
  - Several SAP systems on a single server affect each other in terms of stability and performance. However, it is difficult to determine exactly how and when they impact each other.

  - As far as the technical implementation is concerned, simultaneous operation of several systems or instances on a single host is mainly a matter of sizing. This means that the hardware resources required must be planned accurately and adapted to the operating system, database, and SAP parameters. These parameters depend on the individual requirements of the particular user and their original setup.

- Tune application performance

  Another feature beyond the flexible resource management facilities of WLM is its capability to bind processes to specific processor groups. This function can be used to gain some performance improvements in highly tuned environments, for example, benchmarks. Memory access latency is critical for an SAP application server instance. To improve the throughput of a single application server instance, WLM can be used to bind the work processes of an application server instance to a group of processors. This has the effect that work processes (especially under high load) will not float between processors improving the data affinity in a second- or third-level cache.

  As a trade-off, this tuning restricts dynamics and flexibility across application instances. If one application server is very busy, and another is not, the busy application server cannot utilize free resources outside its processor group. A single SAP system using multiple application server instances will usually be using load balancing for online users, and batch jobs are normally distributed over available batch capable instances. This restrictive binding is only likely to be a problem where workload cannot be balanced across the available servers, or when separate systems are consolidated on one server. The performance improvement is typically below 5%. Therefore, WLM processor binding in productive environments is not recommended because of restrictions to flexibility.

3.7.2 Description of AIX Workload Manager

Workload Manager is a no-cost AIX tool included with the AIX base, introducing more control over how CPU and real memory resources are allocated to processes by means of resource prioritization. AIX 5L introduces additional manageable resources, that is I/O facilities, such as disk and LAN I/O. WLM can be used to prevent different classes of jobs from interfering with each other and to allocate resources based on the requirements of different groups of users.

WLM is rolled out in several stages improving functionality over time. The first stage shipped with AIX 4.3.3 delivered the basic ability to allocate CPU and physical memory resources to classes of jobs and allows processes to be automatically assigned to classes based on user, group, or application. Early tests have been performed by the IBM SAP International Competence Center (ISICC) at the Montpellier test site showing the feasibility of WLM in a mySAP.com application environment. This has been documented in several white papers, available through the ISICC. During this early phase, however, WLM showed some functional deficiencies and reliability issues that limited the practical use of the solution. Therefore, we recommend using WLM as shipped with AIX 5L Version 5.1 and later.

In this second stage, WLM allows a hierarchy of classes to be specified, allocation of I/O bandwidth to classes of jobs, processes to be assigned to classes by the characteristics of a process, manual assignment of processes to classes, and application tagging.

The main use of WLM is intended for large systems with many processors, which are often used for server consolidation. Another use of WLM is to provide isolation between jobs with very different system requirements. This can prevent effective starvation of workloads with certain behaviors (for example, interactive or low CPU usage jobs) from workloads with other behaviors (for example, batch jobs). This usage has two different issues that WLM must address:

- Adaptive goals for amount of resources available to different workloads
- Maximum and minimum boundaries on the amount of resources that a workload may exhaust (maximum) and that which it is guaranteed (minimum)

The WLM principle is in some contrast with logical partitions (LPARs) and physical partitions (PPARs), with which it should not be confused: LPAR and PPAR divide a physical server into several independent domains, each running its own operating system instance. In contrast to this, WLM does *not* partition hardware and operating system, but instead is a means of allocating resources, such as CPU and memory, to specific processes (programs and applications), or users, or groups of users, or both. WLM does not guarantee the processes exclusive or dedicated use of the resources, but provides resource access according to priority. Partitioned domains provide operating system and application isolation, such that incidents within one domain should not affect the others. This cannot be achieved with the single operating system based WLM.

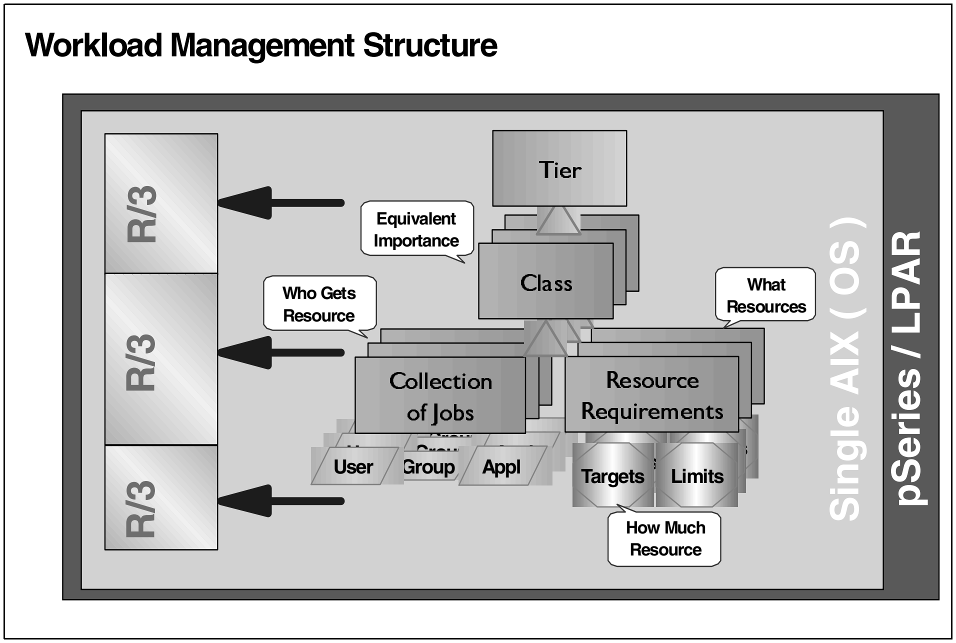

Figure 3-9 Workload management structure

Figure 3-9 shows the components of WLM used to control several independent SAP R/3 systems within a single AIX image. There is no difference, whether it is implemented on a discrete server or on an LPAR. The following introduces terms related to WLM resource management:

Class A class is a collection of processes (and their associated threads) that have a single set of resource limitation values and target shares applied to them. When the term class is used, this includes both subclasses and superclasses.

Superclass A superclass is a class with subclasses associated with it. No processes can belong to the superclass without also belonging to a subclass. A superclass has a set of class assignment rules that determines which processes will be assigned to the superclass. A superclass also has a set of resource limitation values and resource target shares that determine the amount of resources that can be used by processes that belong to the superclass. These resources are divided among the subclasses based on the resources limitation values and resource target shares of the subclasses.

Subclass A subclass is a class associated with exactly one superclass. Every process in the subclass is also a member of the superclass. Subclasses inherit attributes from their superclass and only have access to resources that are available to the superclass. A subclass has a set of class assignment rules that determine which of the processes assigned to the superclass will belong also to the subclass. A subclass also has a set or resource limitation values and resource target shares that determine the resources that can be used by processes in the subclass. These resource limitation values and resource target shares indicate how much of the resources available to the superclass (the target for the superclass) can be used by processes in the subclass.

Class assignment rule A class assignment rule indicates what set of attribute values of a process (or a system state) will result in a process being assigned to a particular class (superclass or subclass within a superclass).

Process attribute value A process attribute value is the value that a process has for some attribute of the process. (The process attributes can include user ID, group ID, application path name, and so on.)

Resource limitation values Resource limitation values are a set of values that Workload Manager should attempt to maintain for a set of resource utilization values. (These are the ranges for Workload Manager to maintain.) These limits are completely independent of the resource limits specified with setrlimit().

Resource target share Resource target shares are the shares of a resource that should be available to a class (subclass or superclass). These shares are used with other class shares at the same level (subclass or superclass) and tier to determine the desired distribution of the resources between classes at that level and tier.

Resource utilization value A resource utilization value is the amount of a resource that a process or set of processes is currently using in a system. (Whether it is one process or a set of processes is determined by the scope of process resource collection.)

Process class properties The process class properties are the set of properties that are given to a process based on the classes (subclass and superclass) it was assigned to. (This includes resource set assignments.)

Class authorizations The class authorizations are a set of rules that indicate which users and groups are allowed to perform operations on a class or processes and threads in a class. (This includes the authorization to manually assign processes to a class, or to create subclasses of a superclass.)

Class tier The tier value for a class is the position of the class in the hierarchy of resource limitation desirability for all classes. The resource limits (including the resource targets) for all classes in a tier will be satisfied before any resource is provided to lower tier classes. Tiers are provided at both the superclass and subclass levels. Resources are provided to superclasses based on their tiers. Within a superclass, resources are given to subclasses based on their tier values within the superclass. Therefore, the superclass tier is the major differentiator in resource distribution, with the subclass tier providing an additional smaller differentiator with the tier of superclasses.

3.8 Logical partitions (LPARs)

With AIX 5L Version 5.1 and the Regatta-type product line, IBM @server pSeries introduced LPAR capabilities. The following is intended to provide an SAP performance and sizing focused overview of LPAR usage.

This content is based on measurements performed by ISICC Walldorf using a 16-way and 32-way 1.3 GHz p690 systems and project feedback collected so far. Therefore, it cannot cover all flavors of production customer environments, but is intended as a generic guideline.

3.8.1 Sizing considerations

The ISICC system capacity tables show capacity figures reflecting the physical (CPU) layout of servers. This includes the LPAR-capable members of the Regatta family (p630, p650, p670, and p690).

3.8.2 LPAR overhead

The numbers in the pSeries sizing table do not show specific figures for certain LPAR sizes, although LPAR usage can introduce some impact on the achievable SAPS numbers. SAPS is a relative performance value created by SAP that is based on well-defined benchmark procedures. Thus far, our experiences have shown the following:

- Operating Regatta in LPAR mode introduces a slight performance decrease of 3% in terms of overall system capacity compared to the SMP mode.
- This overhead is introduced by switching from SMP mode to LPAR mode and is independent of number and size of LPARs.

3.8.3 Cross LPAR impact

When partitioning a server, there are still some components that are shared across the physical system, for example, the backplane. This can introduce contention when LPARs are heavily loaded. Performance evaluations using the SAP Sales and Distribution (SD) benchmark environment showed:

- LPARs do not affect each others in terms of throughput when running at the SAP recommended CPU utilization of 65%.
- Running multiple LPARs close to 100% CPU utilization showed a throughput degradation of about 5% in each heavily loaded partition.

The ISICC system capacity tables show values for non-partitioned servers in symmetric multiprocessing (SMP) mode. When you plan to use partitioned systems, reduce the SAPS values in the table by 3% for LPAR overhead. For LPARs with a smaller number of processors than published in the capacity table, use a linear dependency between the smallest SAPS value published and number of processors for the planned LPAR. One processor LPARs are technically feasible, but usually not recommended for production systems in order to guarantee better and more consistent response times.

3.8.4 Architectural considerations

Although there can be a minor cross LPAR impact, the use of LPARs can help to avoid potential application scalability issues on larger SMPs caused by resource contention. The ISICC has performed a series of tests using different LPAR sizes and setup modes.

3.8.5 Summary of results for a 32-way p690

Figure 3-10 compares the throughput of various LPAR configurations with a reference measurement using the whole system in SMP mode. System throughput was determined using the SAP SD benchmark.

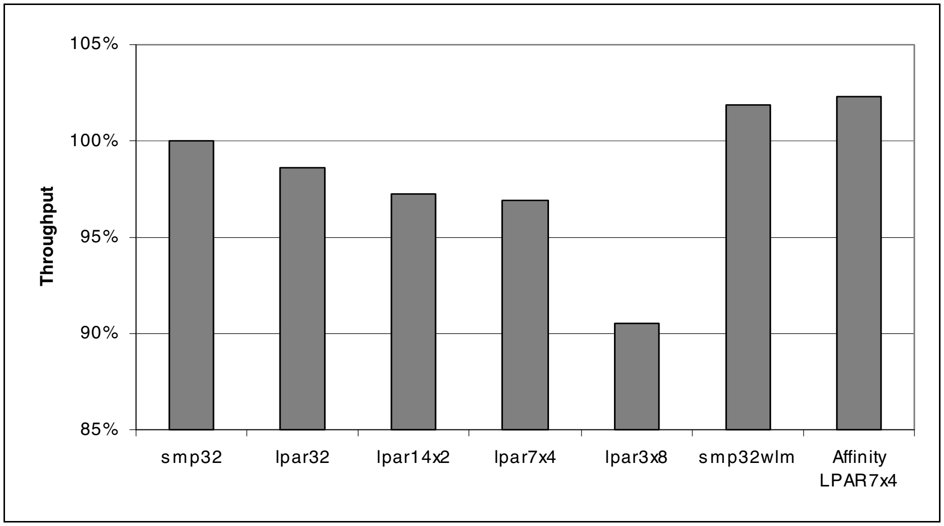

Figure 3-10 Relative throughput of LPARs

The best results were achieved using affinity LPARs, so there is some benefit to be gained by partitioning the machine. The difference between the affinity LPARs and the SMP with WLM, however, was negligible. Affinity partitioning carves up a Regatta into symmetrical LPARs of four or eight processors. The processors and memory of a LPAR being aligned at multi-chip module (MCM) boundaries shortens memory access path length at the cost of system flexibility. The processor and memory assignments are fixed, but I/O is still user selectable. Obviously, the true “logical” partitioning mode of the p690 demands some overhead in terms of memory access compared to architectures that are limited to physical boundaries. Physical partitions (PPARs) provide hardware-enforced CPU to memory affinity. WLM introduces processor affinity by OS means with similar effect.

Considering a balance between flexibility and performance, we recommend implementing 4-way LPARs for 3-tier OLTP environments. Other applications, such as APO and liveCache, typically benefit from coexisting on a single server, and we do not recommend separating the individual components (application server, database, liveCache) across LPARs.

3.8.6 Configuration alternatives

There are a number of alternative ways in which the resources of a single Regatta system can be distributed. Each of these has its purpose and its advantages and disadvantages. Which meets the requirements of an individual customer situation depends on the priority different aspects have in the given customer environment. This section documents the pros and cons of the various alternatives.

3.8.7 Single partition

This alternative uses a single partition with or without Workload Manager in SMP or LPAR mode.

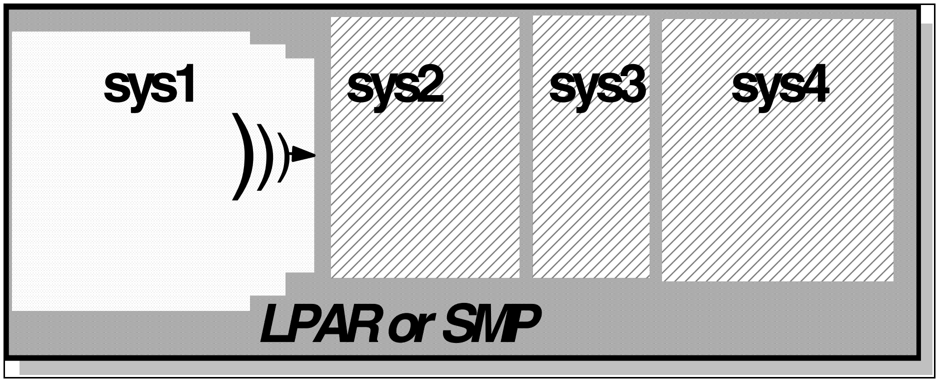

Figure 3-11 Single partition mode

Plus

This configuration provides the most flexible resource distribution. WLM can be used to enforce some level of guaranteed resource level for each component, while also allowing components to take advantage of idle resources should these be available. WLM resource groups can also be changed on the fly to be more or less restrictive according to customer needs.

Minus

The negative side of this flexibility includes an increased complexity in configuration (WLM) and monitoring. Monitoring is particularly an issue if unrelated SAP systems are consolidated on a single server, because SAP Computing Center and Management System (CCMS) cannot currently see beyond a single system boundary. Systems running together on a single OS will have less protection from each other in the case of a catastrophic error caused by any one system. This could also lead to maintenance problems because of potential software prerequisite conflicts of different application components.

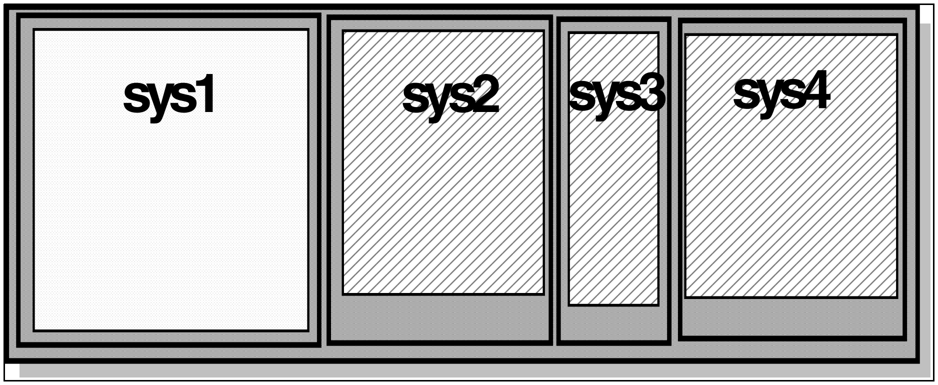

Figure 3-12 Traditional LPARs

3.8.8 Traditional LPARs

This method uses traditional LPARs.

Plus

Traditional LPARs allow a flexible distribution of resources within LPAR boundaries. Each LPAR can be configured according to the specific needs of the occupant application. There is no predetermined memory size or limitation nor limit on the number of processes beyond the minimum requirements. The LPARs provide a protection boundary between the systems. Test and development systems can exist on the same server in separate LPARs. The operating system level maintenance effects only the specific LPAR allowing testing of new operating system releases or fixes, or both.

Minus

Idle resources beyond the partition boundary cannot be utilized. The partition must be allocated resources according to its peak requirement, and these allocations are basically static. Dynamic LPARs (DLPARs), as introduced with AIX 5L Version 5.2, can help to shift resources according to expected load profiles.

3.8.9 Affinity LPARs

This method of configuring LPARs carves up a Regatta into symmetrical LPARs of four or eight processors. The processors and memory of a LPAR are aligned at MCM boundaries. This allows for a simplified setup on the Hardware Management Console (HMC), because partitions are created automatically as 8-way or 4-way partitions. The processor and memory assignments are fixed, but I/O is still user selectable. It is not possible to run a mix of affinity partitions and traditional partitions concurrently. The type of the first partition started after power-on determines whether the system will run with affinity or traditional partitions.

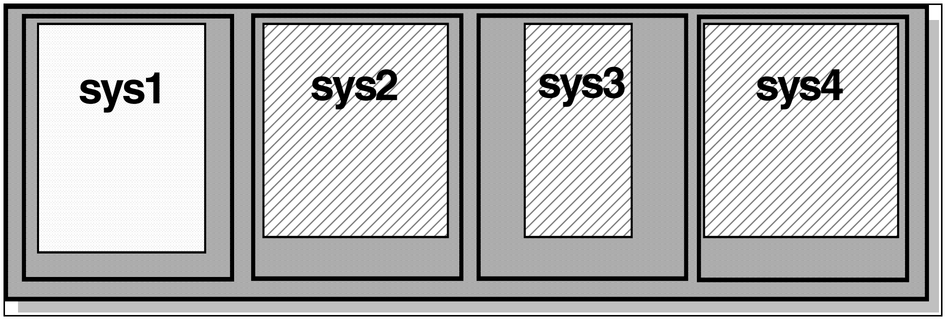

Figure 3-13 Affinity LPARs

Plus

Affinity LPARs provide the safety aspects of the traditional LPARs, as previously documented, and have proven to be the variant with the best overall performance.

Minus

Affinity LPARs provide the least flexible resource distribution because they are fixed configurations that divide the machine into equal portions on MCM boundaries. In this configuration, it is not possible to give one partition an extremely large memory, for example, or assign a different number of processors to one LPAR and then to another. In fact, it is not currently possible to define LPARs of different sizes even if the division would be on MCM boundaries, such as a mixture of 8-way and 4-way LPARs. This operation mode does not allow dynamic reconfiguration of resources.

3.8.10 Balanced memory

The memory configuration has an impact on overall system throughput. This is most noticeable when running under high load. To achieve optimum, balanced, and predictable performance, the system should be configured using all available memory slots.

3.8.11 mySAP applications

Production systems should be separated from non-production (test, development, consolidation, integration, and so on) systems and reside in separate LPARs. The same is true for online transaction processing (OLTP) versus online analytical processing (OLAP) (BW, APO) systems (see *Several systems/instances on one UNIX computer*, SAP OSS Note 21960, for more details).

3.8.12 Dynamic LPAR and AIX 5L Version 5.2

AIX 5L Version 5.2 introduces dynamic LPAR capabilities, meaning LPAR resources can be reallocated without rebooting affected partitions and restarting applications running on these.

An SAP kernel patch is required to ignore Dynamic Reconfiguration (DR) signals. Otherwise, the default action of SAP to an unknown signal would be to shut down the instance. The required patch will most likely already be part of the AIX 5L Version 5.2 enabled SAP kernel once it is released. If not, there is a possibility of setting profile parameters to ignore the relevant signals. This is documented in *Work processes terminate when reconfiguring LPARs*, SAP OSS Note 569569.

SAP does not actively support DLPAR. That means SAP will continue to run after a configuration change to an LPAR, but will not react automatically to the changed environment. SAP does not allow either starting or stopping SAP work processes dynamically or changing memory configuration parameters on a running system (currently, this is not planned for future releases either).

Most of the SAP instance memory is allocated to individual user contexts, which when using the AIX alternative memory management, are outside of the statically allocated SAP memory buffers. Therefore, adding memory to an LPAR with a running instance might help to sustain a peak in user load without changes to a running SAP instance. However, if the additional user load also causes other static shared SAP memory areas to fill up, the overall instance performance would be hindered, and the SAP instance would have to be stopped and restarted with a different memory parameterization. The instance would also have to be restarted with a different work process configuration if the additional load causes an over usage of the currently allocated work processes.

Alternatively, you could start a second SAP instance instead of having to restart a running instance. Note that you could basically achieve the same effect by starting the additional instance on the other LPAR without having to move resources first.

SAP applications do not prevent the removal of resources. They do not pin memory or bind processes to specific processors, so the removal of CPUs or memory from an LPAR should work without problems. Again, the SAP kernel will not be able to react to the changed environment, and taking too much memory away could easily lead to severe paging and massive performance problems on the running instance. A production SAP LPAR can be protected from the unwanted removal of resources using appropriate LPAR configuration settings on the Hardware Management Console. DR event scripts can be implemented that would decline the removal of resources announced by a DR request.

3.8.13 Impact on databases

DB2 and Oracle will support the use of dynamic LPARs improving DLPAR integration over time. Initially, there will not be many automatic adjustments to the change of system resources. It will be the responsibility of an administrator to provide dynamic reconfiguration scripts using the features of the DR application framework.

For example, Oracle 9i Release 2 dynamically detects changes in the number of available processors within the LPAR and adjusts the CPU_COUNT parameter. CPU_COUNT affects certain Oracle behaviors, such as determining the degree of parallelism for parallel query. Oracle does not detect changes in the amount of physical memory allocated to the LPAR. It does, however, support the ability to dynamically change the size of most of its memory areas, such as the size of the database buffer cache.

3.8.14 Monitoring dynamic LPARs

Currently, there is no integration between OS and SAP facilities tracking dynamic reconfiguration events. SAP monitoring tools (CCMS) do not keep histories of the available hardware resources. A basis administrator could be wondering about sudden changes in CPU utilization without being able to check in CCMS that these changes were caused by a DLPAR reconfiguration. For the same reason, the validity of SAP Early Watch sessions and IBM Insight analysis might be jeopardized when a clear relationship between available resources and system load cannot be established for a certain point in time. This will always be the case whenever DR events take place during the monitoring period.

3.9 Differences between p690 LPARs and stand-alone nodes

An IBM @server pSeries 690 in LPAR mode can be initially interpreted as a collection of pSeries stand-alone nodes. The similarity of both hardware technologies overcomes the existing differences. One of the most important similarities is the possibility to use the same adapters and operating system. The process of hardware resource assignments, such as processor, memory, or I/O adapter, to dedicated LPARs opens a new dimension of flexibility. An IBM @server pSeries 690 hardware system can be interpreted as a pool of hardware components of which independent units can be composed. Each unit largely resembles a stand-alone node regarding hardware and software.

The IBM @server pSeries 690 has some special and very useful standard features, for example, distinguished remote manageability and reliability, availability, and serviceability (RAS) support, including call home capability. It unveils new dimensions of daily system handling tasks, so far only known to IBM RS/6000 SP. Common operation tasks, such as an AIX installation or power on/off of any LPAR, can be undertaken without any local presence.

From a system operation and maintenance perspective, the option for static configurable processor or memory assignments, or both, is a powerful feature. It allows you to construct units of processing power that are fully adapted to business requirements. Besides the total number of processors and memory capacity, the number of available I/O adapters is another limitation. Possibilities of resource assignments through simple configuration changes enable easy adoptions on changed business needs in the course of time, for example, long term resource balancing. In contrast, stand-alone nodes always require partially complicated physical disassembling tasks to redistribute components, such as memory, between different nodes.

AIX 5L is required for installations of IBM @server pSeries 690. However, the compatibility of AIX 5L for 32-bit applications and the easily realizable adaptations for 64-bit applications reduce the effective difference between a p690 and stand-alone nodes, but there are still some constraints for certain software products, such as Oracle or HACMP. No software that is only available for a lower AIX release than 5L can be installed on this hardware. The use of any software that is not currently released for AIX 5L needs stand-alone pSeries nodes.

Supported AIX, DB, and SAP release combinations can be found at the SAP Service Marketplace (requires registration) using the alias "platforms" at:

http://service.sap.com/platforms

An infrastructure based on IBM @server pSeries 690 using logical partitioning have currently the following technicalities from an SAP systems perspective:

- Installation of several systems on LPARs have the same complexity as on stand-alone nodes.
- There is no difference according to the installation guidelines between LPARs and a bunch of large SMP nodes.
- Workload Manager is tolerated, but not integrated by SAP CCMS.
- Workload Manager is highly recommended from a technical perspective.
- All logical partitions use the same SAP hardware license key.

The total flexibility of resource assignments exists only for separate logical partitions within one single physical system. Unfortunately, certain maintenance activities on the hardware system, for example, microcode upgrades, require a power off or reboot of all partitions. This causes a noticeable interruption for all systems. Within an environment based on stand-alone nodes, such interruptions are limited on single nodes and just a matter of high availability. But an installation, as discussed in *A Holistic Approach to a Reliable Infrastructure for SAP R/3 on AIX*, SG24-5050, containing at least two IBM @server pSeries can minimize such disturbing maintenance interruptions.

4

Conclusion of Part 1

The evolution of hardware platforms running any UNIX derivatives continues at high speed. During the last years, AIX has demonstrated its excellent stability and flexibility within many customer projects. Today, more than 7,000 customers use the pSeries as their SAP platform of choice. The combination of AIX 5L and the new IBM @server pSeries hardware technology provides a convincing foundation for a leading-edge SAP infrastructure.

Customers often expand the use of their SAP systems and model more business processes. This can lead to an increased number of SAP systems and a stronger demand on flexibility. The life cycles of these systems differ considerably. Renaming, removal, or deletion become more and more common system administration tasks. However, the kind of infrastructure and system installation can support or hinder this. Some field tested aspects of a flexible infrastructure are outlined in Part 1.

An integrated mix of generally available components, such as IBM @server pSeries 690, AIX 5L, Workload Manager, HACMP, and Tivoli Storage Manager, installed according to field-tested concepts covers most situations and can meet customer demands. It is a proven, flexible, and stable platform to master upcoming future requirements.

In addition to the expected performance enhancements, some additional improvements for IBM @server pSeries and AIX 5L are desired, for example:

- Sub-processor allocation
- Adapter sharing between LPARs
- Inter-LPAR communication without LAN adapters
- DLPAR and WLM integration in SAP CCMS

WLM, LPAR, and DLPAR provide a flexible framework to support changing SAP resource requirements, as well as evolving customer business needs. Although the growing number of SAP systems in some environments generates an increasing degree of complexity, advanced architectural features, thorough planning, and consequent use of best practices can ease the handling of such complexity.

A solution is always customer specific, and the best result is the most suitable solution according to the customer's business requirements. Every infrastructure should be derived from these business needs. The resulting topology might contain systems that are relocatable between different hardware systems for high availability solutions and advanced system maintenance tasks.

The combination of pSeries hardware and software features, together with best practices provided by IBM Services, result in a flexible and proven infrastructure for SAP applications, meeting today's rapidly changing business demands.

Part 2

Regulating system resources with AIX 5L Workload Manager

5

Setup and scenarios

When consolidating multiple applications or application instances onto a single server, one of the topics to deal with is the problem of resource contention. Because different applications running on one large server are now competing for system resources, how can we ensure that each application gets the resources it needs in order to obtain the desired throughput and response time? Without Workload Manager (WLM), the increased load of one SAP R/3 system will directly affect the behavior of all other systems. There is no way to limit one system from consuming resources required by others. With Workload Manager, however, it is possible to balance requests of competing workloads and to control system resources. WLM is a configurable component of the operating system provided at no additional charge that was introduced first in AIX 4.3 and further enhanced in AIX 5L. By classifying processes, the system administrator can allocate resources between applications without having to partition the system.

The project setup consists of an IBM @server pSeries 680 Model S85 (six CPUs and 16 GB memory) as the System Under Test (SUT) with four SAP R/3 systems installed on SSA disks and an RS/6000 SP High Node as the load generator. Details about software releases, support packages, fixes, and PTFs are described in Appendix B, “Products” on page 111.

For the workload, we used the transactions defined in the SAP Standard Benchmark Tool for Sales and Distribution (SD). This load generator simulates a certain amount of users in a given run for a specific SAP system. The workload generated is thus well-defined by the number of users simulated. This can be varied independently for each SAP R/3 system for which we measured the average response time for the transactions in seconds (sec) and the total throughput in terms of dialog steps per second (DS/sec). For each of the test scenarios, we selected different WLM configurations according to desired resource regulations. We then varied the workload for the SAP R/3 systems. To demonstrate the effects of different WLM configurations, we compared the performance of the systems in terms of average response times and throughput.

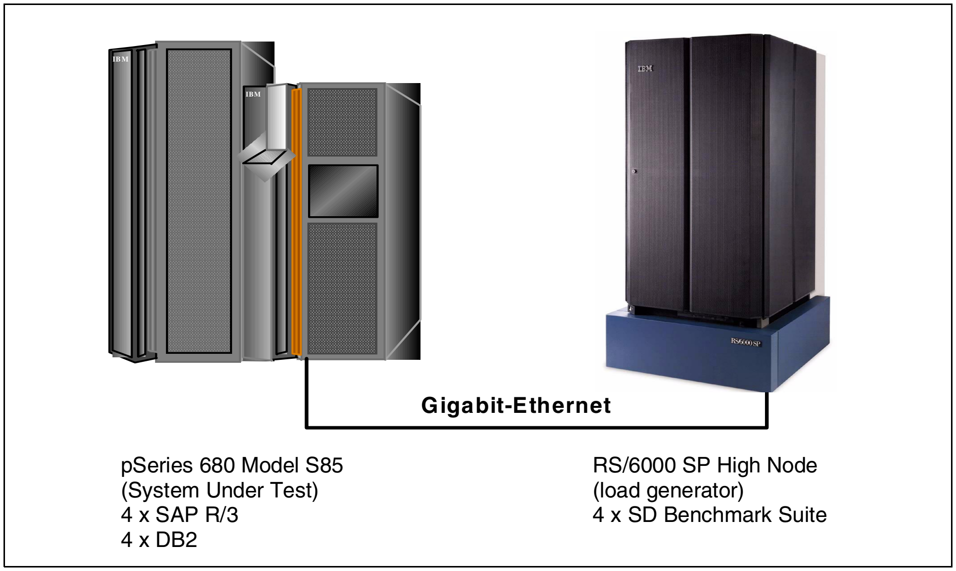

Figure 5-1 Test setup

We considered the following base scenarios:

- Scenario 1
  Tests 1-4: All four SAP R/3 systems are equally important. No system should be allowed to get more than its fair share of the CPU time available.
- Scenario 2
  Tests 5-6: One of the four SAP R/3 systems is more important than the others. This preferred system should get all the CPU resources needed. The others should only get the CPU time that is not used by the preferred system.
- Scenario 3
  Tests 7-8: One of the four SAP R/3 systems is less important than the others and should only get the CPU time not needed by the other three preferred systems.

For our WLM configurations, we created four classes, one for each SAP system: sd1cl, sd2cl, sd3cl, and sd4cl. The classes System and Default are created automatically and have not been modified for the purpose of these tests. The assignment rules of our WLM setup are shown in "WLM setup" on page 109. Even though WLM can control CPU, memory, and disk I/O bandwidth, this first part of the project is concerned with CPU resource allocation only.

Workload Manager can run in passive and active mode, which helps to define WLM configuration strategies. In passive mode, Workload Manager classifies new and existing processes and gathers statistics about their resource usage, but does not regulate this usage. In this mode, the processes compete for resources exactly as they would if WLM was off.

One of the recommendations for the use of WLM is to adjust shares and tiers first to get closer to the required resource allocation goals and to use limits only if absolutely necessary. Therefore, we set up our test scenarios using configurations with different shares and tiers only, leaving limits unchanged. For Scenario 1, where all four systems are equally important, the WLM configurations differ only in their distribution of shares between the four systems. For Scenarios 2 and 3, we wanted to prioritize workloads and have thus varied tier configurations with equal shares. Details of these configurations are explained in the individual test descriptions and are shown in Table 5-1, which summarizes the tier configurations for the three scenarios in symbolic form. The classes are not represented by their proper class name, but instead by numbers 1 to 4.

Table 5-1 Scenario description in terms of tiers and classes

	Tiers	Classes
Scenario 1	0 1 2	- 1 2 3 4 -
Scenario 2	0 1 2	- 1 2 3 4
Scenario 3	0 1 2	- 2 3 4 1

For each of the three scenarios, we conducted various tests that can be further classified as base and stress tests. With the base tests, we explored the general behavior of the systems. With the stress tests, we added some more users to one of the systems and compared the results to the ones obtained for the base tests. Table 5-2 on page 64 gives an overview of these tests.

Table 5-2 Overview of test scenarios

	Base tests	Stress tests
Scenario 1 - 1 2 3 4 -	Equal shares Equal users (Test 1a)	More users (Test 3)
	Unequal shares Equal users (Test 1b)	
	Unequal shares Unequal users (Test 2)	More users on high load (Test 4a) More users on low load (Test 4b)
Scenario 2 - 1 2 3 4	Equal shares Equal users (Test 5)	More users on high priority (Test 6a)
		More users on low priority (Test 6b)
Scenario 3 - 2 3 4 1	Equal shares Equal users (Test 7)	More users on high priority (Test 8a)
		More users on low priority (Test 8b)

In Scenario 1: Equally important systems, all systems had an equal number of shares and users (Test 1a). We then varied the number of shares for one system (Test 1b), and finally, we added more users to one of the systems and adjusted the shares in such a way that all four systems achieved equal average response times (Test 2). Based on Test 1, we then added more users to one of the systems while leaving the shares equal (Test 3). Based on Test 2, we first added more users to the system with high load that already had a higher workload than the others but also more shares (Test 4a), and then we added more users to one of the systems with low load (Test 4b).

In Scenario 2: One system more important than others, all four systems had equal amounts of shares. For the base test, we gave all four systems the same workload, which means equal numbers of users, but we assigned one system class to a higher tier (Test 5). We stress tested the systems by adding more users either to the system in tier 1 (Test 6a) or to one of the systems in tier 2 (Test 6b).

In Scenario 3: One system less important than others, we left the amount of shares unchanged as in Scenario 2. We again started the base test where all four systems have the same workload, but this time, one system is in a less prioritized tier (Test 7). We put stress on the systems by adding more users first to one of the systems with high priority (Test 8a), and then to the system with low priority (Test 8b).

In addition, whenever we talk about equal share distributions in our setup, we gave each system 80 shares. However, the results would have been exactly the same for other amounts of shares as long as the proportions are kept equal. The relative number, not the absolute number, of shares is important. This means, we could have used a share distribution of 1/1/1/1 as well for the same results. We chose a distribution of 80/80/80/80 for demonstration purposes. This makes it easier to clearly show the effects of changing WLM configurations.

6

Scenario 1: Equally important systems

For this scenario, we performed the following tests:

- Test 1: Workload distributed evenly. All four systems have the same amount of users.
- Test 2: Workload distributed unevenly. One system has more users than the others.
- Test 3: Stress test with equal shares. Based on Test 1.
- Test 4: Stress test with unequal shares. Based on Test 2.

6.1 Test 1: Workload distributed evenly

In the first test, each SAP system had the same number of users. We compared the results for WLM in passive and active mode. The WLM mode p in the tables stands for passive mode. First, we used CPU share configurations where all systems have equal shares (80/80/80/80) and then where one of the systems has more CPU shares than the others (160/80/80/80 and 240/80/80/80). We looked at the effects of an increase in CPU shares on response times and throughput for all systems. We tried to find the "optimal" workload where we have high throughput but still acceptable average response times. We started with all four SAP systems having 50 users each, increasing the number of users up to 90 each, and obtained the following results.

Table 6-1 Test 1: 4 x 50 users (p = passive)

Users	Shares	Average response time (sec)				Throughput (DS/sec)			
		WLM p	WLM active			WLM p	WLM active		
			x=80	x=160	x=240		x=80	x=160	x=240
50	x	0.26	0.26	0.25	0.25	4.88	4.88	4.89	4.89
50	80	0.26	0.26	0.27	0.27	4.88	4.88	4.88	4.88
50	80	0.26	0.26	0.27	0.27	4.88	4.88	4.88	4.88
50	80	0.27	0.26	0.27	0.27	4.88	4.88	4.88	4.88
					Total	19.52	19.52	19.52	19.53

Table 6-2 Test 1: 4 x 60 users (p = passive)

Users	Shares	Average response time (sec)				Throughput (DS/sec)			
		WLM p	WLM active			WLM p	WLM active		
			x=80	x=160	x=240		x=80	x=160	x=240
60	x	0.27	0.28	0.25	0.25	5.85	5.85	5.86	5.87
60	80	0.27	0.28	0.29	0.29	5.85	5.84	5.84	5.85
60	80	0.27	0.28	0.29	0.29	5.85	5.84	5.84	5.84
60	80	0.27	0.28	0.29	0.29	5.85	5.84	5.84	5.84
					Total	23.41	23.37	23.38	23.40

Results: For the 4x50 users test, the total throughput of the systems is unchanged by different share configurations, that is, 19.5 DS/sec. There is little influence on the response times when we increase the number of shares for one system. If we give one system three times as many shares, it will respond 7% faster (0.25 versus 0.27 sec). Response times of the other three systems are almost unchanged.

We then increased the number of users to 60 for each system and obtained similar results as for the 4x50 users test. The reason is, that overall, there is still hardly any competition for CPU resources, but occasionally, there are peaks, and the system resources are fully used. In this case, the system with more shares will get more resources and has, therefore, on average, an improved response time of 14% (0.25 versus 0.29 sec). Changes in throughput between passive mode and different WLM configurations are not significant. The total throughput is 23.4 DS/sec.

Table 6-3 Test 1: 4 x 70 users (p = passive)

Users	Shares	Average response time (sec)				Throughput (DS/sec)			
		WLM p	WLM active			WLM p	WLM active		
			x=80	x=160	x=240		x=80	x=160	x=240
70	x	0.31	0.31	0.26	0.26	6.80	6.80	6.83	6.84
70	80	0.31	0.30	0.32	0.31	6.80	6.80	6.80	6.80
70	80	0.31	0.31	0.32	0.31	6.80	6.80	6.80	6.80
70	80	0.31	0.31	0.32	0.32	6.80	6.80	6.80	6.80
					Total	27.20	27.21	27.22	27.23

Table 6-4 Test 1: 4 x 80 users (p = passive)

Users	Shares	Average response time (sec)				Throughput (DS/sec)			
		WLM p	WLM active			WLM p	WLM active		
			x=80	x=160	x=240		x=80	x=160	x=240
80	x	0.37	0.37	0.27	0.27	7.73	7.73	7.80	7.80
80	80	0.37	0.37	0.39	0.39	7.73	7.73	7.71	7.71
80	80	0.37	0.37	0.39	0.39	7.73	7.72	7.71	7.71
80	80	0.37	0.37	0.39	0.39	7.73	7.72	7.70	7.72
					Total	30.91	30.90	30.93	30.94

Table 6-5 Test 1: 4 x 90 users (p = passive)

Users	Shares	Average response time (sec)				Throughput (DS/sec)			
		WLM p	WLM active			WLM p	WLM active		
			x=80	x=160	x=240		x=80	x=160	x=240
90	x	0.71	0.60	0.29	0.29	8.41	8.50	8.76	8.76
90	80	0.72	0.59	0.65	0.66	8.41	8.50	8.47	8.46
90	80	0.73	0.60	0.65	0.66	8.40	8.49	8.46	8.45
90	80	0.71	0.60	0.66	0.67	8.41	8.50	8.46	8.45
					Total	33.63	34.00	34.14	34.12

Results: The tendency of getting an improved average response time for the system with more shares is also confirmed if we change the setup to 4x70 users. The improvement in response time is 16% (0.26 versus 0.31 sec). The total throughput is 27.2 DS/sec with a very small, but not significant, increase in throughput for the system with more shares.

If we further increase the number of users per system to 80, that is 320 users in total, the average response time of the system with three times as many shares compared to the other systems improves by 31% (0.27 versus 0.39 sec). This result indicates that CPU resources are well used. There is resource competition for the whole duration of the run. Total throughput is 30.9 DS/sec. There is again a small difference in throughput between systems with different shares, but it is still very small (about 1%).

Changing the number of users to 90 each shows a big increase in average response times for all four systems due to fierce CPU resource competition. As expected, the system with more shares gets most resources. Its average response time is reduced by 56% (0.29 versus 0.66 sec) compared to the other systems. However, the resources available are not sufficient to maintain “acceptable” response times for all systems. The total throughput is now about 34.1 DS/sec. The difference in throughput between systems with different shares increases as well. It is now 3.5% higher for the system with three times as many shares.

6.1.1 Test 1 summary

The effects of using WLM configurations for equal numbers of users for each system, and therefore equal workload but different amounts of shares, are summarized and displayed in Figure 6-1. We compare the average response times for WLM in passive mode and for WLM active with a share configuration of 240/80/80/80 for the systems with 80 shares each and the system with 240 shares. The diamonds represent the total throughput for the different sets of users.

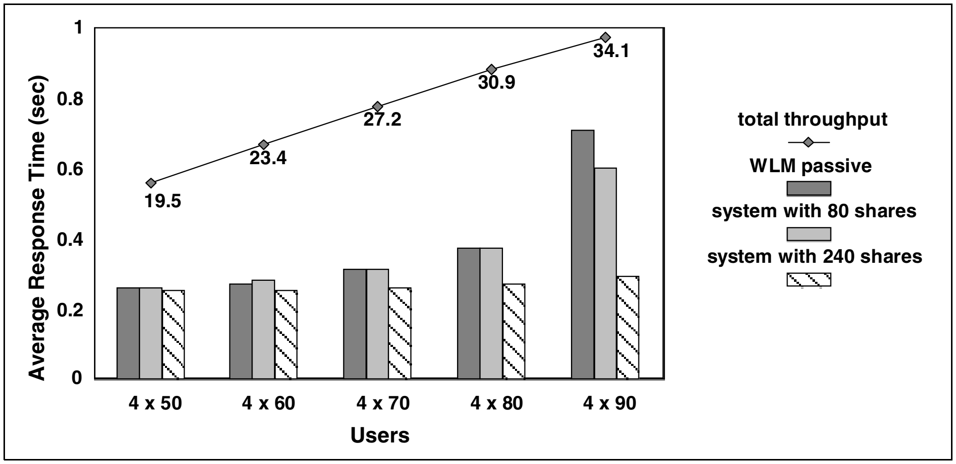

Figure 6-1 Test 1 summary

Test 1 produced the following results:

- The system with a higher share distribution has improved average response times compared to the other systems.
- The difference in average response times between systems with unequal shares is bigger the more workload there is. That means, if we give one system three times as many shares, for 4x50 users, we have an improved response of 7%; for 4x90 users, it is 58%.
- For a given number of users, there is no significant change in total throughput if we increase the shares for one system.
- The throughput increases the more users there are on the systems. As long as there are enough CPU resources available, the throughput per user is constant, that is, 351 DS/h/user. If there is too much resource competition, this throughput per user can not be maintained. It decreases to 340 DS/h/user for the 4x90 users setup.
- We get “optimal” workload if we have 80 users per system. Therefore, we based the following tests on this setup.

Table 6-6 Test 1 summary: Average response time

	Average response time (sec)				
Users	**4x50**	**4x60**	**4x70**	**4x80**	**4x90**
80 shares	0.27	0.29	0.31	0.39	0.66
240 shares	0.25	0.25	0.26	0.27	0.29
Advantage	7%	13%	16%	31%	58%

Table 6-7 Test 1 summary: Throughput

	Throughput				
Users	**4x50**	**4x60**	**4x70**	**4x80**	**4x90**
Total (DS/sec)	19.5	23.4	27.2	30.9	34.1
Per user (DS/h)	351	351	351	348	340

6.2 Test 2: Workload distributed unevenly

We gave one system 120 users, and the other three systems still had 80 users each. We tried to find a distribution of CPU shares that results in equal average response times for all four systems. We compared the results with WLM in passive mode and with three different distributions of CPU shares: 80/80/80/80, 120/80/80/80, and 140/80/80/80.

Table 6-8 Test 2: 1 x 120 users, 3 x 80 users (p = passive)

Users	Shares	Average response time (sec)				Throughput (DS/sec)			
		WLM p	WLM active			WLM p	WLM active		
			x=80	x=120	x=140		x=80	x=120	x=140
120	x	0.85	1.09	0.57	0.34	11.07	10.83	11.37	11.62
80	80	0.50	0.35	0.55	0.65	7.62	7.74	7.59	7.52
80	80	0.50	0.36	0.55	0.64	7.62	7.73	7.59	7.53
80	80	0.50	0.36	0.55	0.6	7.62	7.72	7.59	7.53
Total						33.94	34.03	34.14	34.18

Results: Without WLM, the system with the high load has average response times that are 70% higher than for the systems with the low load (0.85 versus 0.50 sec). If each system has the same amount of CPU shares, the average response time improves for the systems with low load (0.36 versus 0.50 sec), but get worse for the system with the high load (1.09 versus 0.85 sec). We obtained the best results if the CPU shares are distributed in proportion to the number of users with average response times that are nearly the same for all systems (0.55 to 0.57 sec). If the CPU shares for the system with the high load are increased further, this system has a clear advantage over the systems with the low load. A graphical summary of the results is displayed in Figure 6-2.

The total system throughput is the same for all configurations (approximately 34 DS/sec). With WLM in passive mode, the system with high load has 1.45 times the throughput of the system with low load. This ratio declines to only 1.40 for equal CPU shares. We again obtained the best results when the shares are in proportion to the number of users. We have a ratio of 1.50 that is exactly the same as the ratio of the users (120/80). With 140 CPU shares, this throughput ratio further increases to 1.54, as shown in Table 6-9 on page 74.

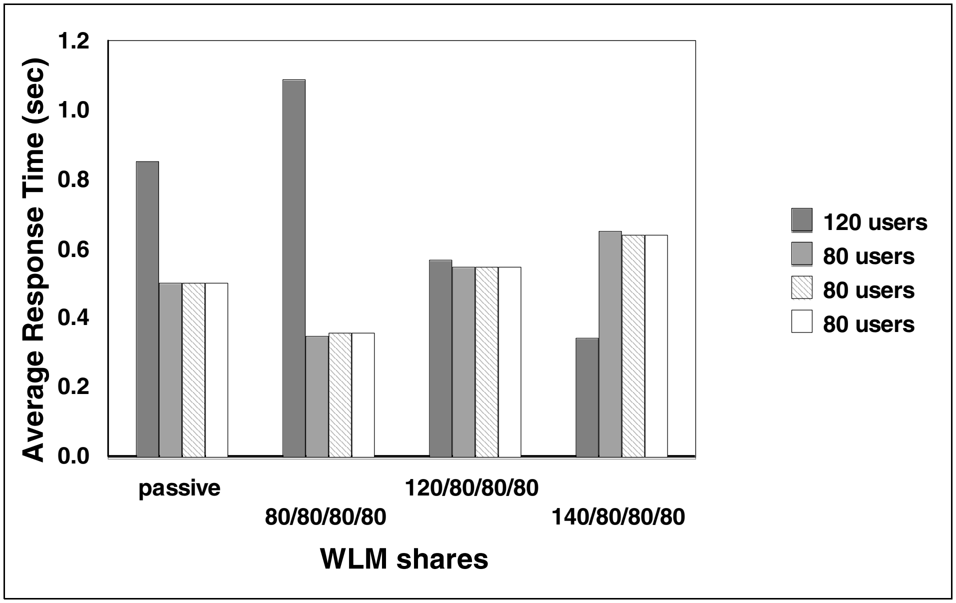

Figure 6-2 Test 2 summary

Table 6-9 Test 2: Throughput ratio between systems with high and low loads

Users	Shares	Throughput (DS/sec)			
		WLM passive	WLM active		
			x=80	x=120	x=140
120	x	11.07	10.83	11.37	11.62
80	80	7.62	7.74	7.59	7.52
Throughput ratio		1.45	1.40	1.50	1.54

6.3 Test 3: Stress test with equal shares

We left the number of users for three systems unchanged at 80. We increased the number of users for one system in the first step to 120, and then to 140. All four systems had the same amount of CPU shares, that is 80/80/80/80. With this setup, we wanted to answer the following questions:

- How are the average response times and throughput of all four systems affected if we increase the workload of one of the systems?
- Can we protect individual systems from the negative influence of each other?

Table 6-10 Test 3: 1 x 120 users, 3 x 80 users, 80/80/80/80 shares

Users	Average response time (sec)		Users	Average response time (sec)	Users	Average response time (sec)
	WLM passive	WLM active		WLM active		WLM active
120	0.85	1.09	80	0.37	90	0.60
80	0.50	0.35	80	0.37	90	0.59
80	0.50	0.36	80	0.37	90	0.60
80	0.50	0.36	80	0.37	90	0.60

Table 6-11 Test 3: 1 x 120 users, 3 x 80 users, 80/80/80/80 shares

Users	Throughput (DS/sec)		Users	Throughput (DS/sec)	Users	Throughput (DS/sec)
	WLM passive	**WLM active**		**WLM active**		**WLM active**
120	11.07	10.83	80	7.73	90	8.50
80	7.62	7.74	80	7.73	90	8.50
80	7.62	7.73	80	7.73	90	8.49
80	7.62	7.72	80	7.73	90	8.50
Total	33.94	34.03	Total	30.90	Total	34.00

Results: The findings for comparing active and passive modes for a system with increased workload are described in 6.2, “Test 2: Workload distributed unevenly” on page 72. We now compared the setup of 1x120 users and 3x80 users in active mode (equal shares) with the base setup of 4x80 users. We found that increasing the workload of one system from 80 to 120 users means that the average response time for this system deteriorates from 0.37 to 1.09 sec. This is an increase of 195%. The average response time for the other systems is almost unchanged (0.36 versus 0.37 sec). Therefore, if one system has a suddenly increased workload, but all systems have equal shares, only this system will suffer in response time. The other three systems are protected from the negative influence of the system with the high load.

Considering the total number of users (360), we also compared these results with a setup where all four systems have 90 users each. In this case, we have an average response time for each system with 90 users of 0.60 sec. Therefore, for the systems with low load, the average response time is better if the number of users is not equally distributed between all four systems (0.60 versus 0.36 sec). The opposite is true for the system with high load (0.60 versus 1.09 sec).

The total throughput for all systems with 80 users each is 30.9, compared to 34.0 DS/sec if we have 40 more users on the system. The throughput for that distribution of users is the same as for the 4x90 case, as we expected. The throughput of the systems with low load is not affected by the increase in workload for one system. For the system with high load, it increased from 7.73 to 10.83 DS/sec, but the dialog steps per user were unchanged.

We then increased the number of users for one system even more to 140, which equals a total number of users of 380, and compared the results with a setup where we distributed the users equally, that is 4x95 users. We encountered the previously described effects in a similar manner, as shown in Figure 6-3 on page 77.

Table 6-12 Test 3: 1 x 140 users, 3 x 80 users, 80/80/80/80 shares

Users	Average response time (sec)		Users	Average response time (sec)	Users	Average response time (sec)
	WLM passive	**WLM active**		**WLM active**		**WLM active**
140	1.83	1.93	80	0.37	95	0.90
80	0.60	0.38	80	0.37	95	0.90
80	0.60	0.38	80	0.37	95	0.87
80	0.60	0.37	80	0.37	95	0.85

Table 6-13 Test 3: 1 x 124 users, 3 x 80 users, 80/80/80/80 shares

Users	Throughput (DS/sec)		Users	Throughput (DS/sec)	Users	Throughput (DS/sec)
	WLM passive	**WLM active**		**WLM active**		**WLM active**
140	11.85	11.76	80	7.73	95	8.74
80	7.55	7.72	80	7.73	95	8.73
80	7.55	7.72	80	7.73	95	8.75
80	7.55	7.72	80	7.73	95	8.78
Total	34.50	34.92	Total	30.90	Total	34.99

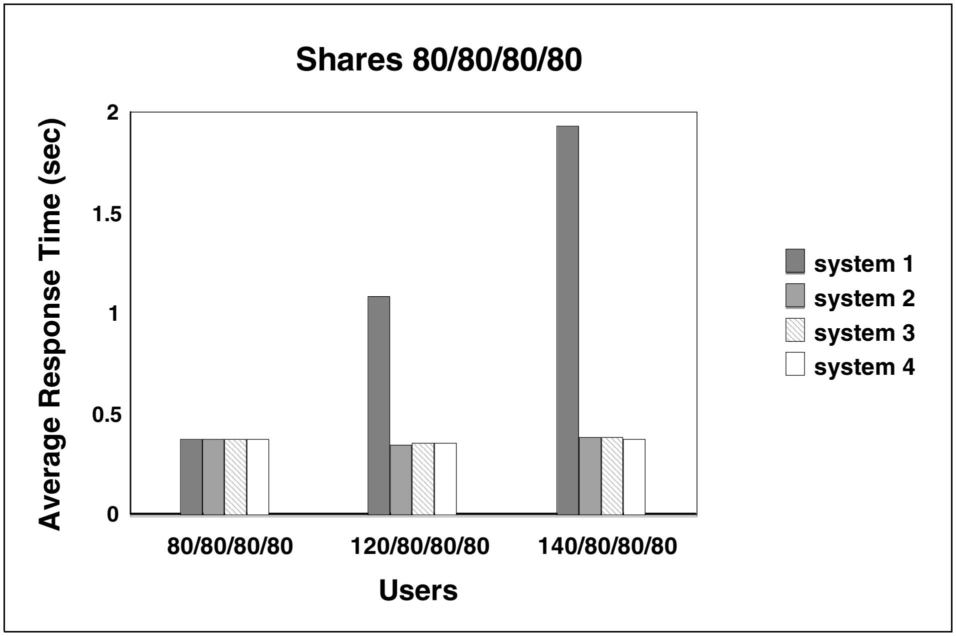

Figure 6-3 Test 3 summary

6.3.1 Test 3 summary

We can protect individual systems from the negative influence of each other. Therefore, if we increase the workload for one system, but all four systems have an equal amount of shares, only the system with increased load suffers considerably in response time. The average response times for the systems with low loads do not suffer and remain unchanged.

The average throughput per user is not affected by the increase of workload for one system. It does not significantly change for any system.

6.4 Test 4: Stress test with unequal shares

We used the "good working" setup of 6.2, "Test 2: Workload distributed unevenly" on page 72 and distributed the amount of shares in proportion to the number of users (users 120/80/80/80 and shares 120/80/80/80). We now had two possibilities to stress test the systems: either by putting more users in the system with high load or in the system with low load.

6.4.1 More users in the high load system

First, the system with the high load got additional users, 140 and 160, while the share distribution remained at 120/80/80/80.

Table 6-14 Test 4: 1 x 140 users, 3 x 80 users

Shares	Users	Average response time (sec)		Users	Average response time (sec)
		WLM passive	WLM active		WLM active
120	140	1.83	1.32	120	0.57
80	80	0.60	0.68	80	0.55
80	80	0.60	0.68	80	0.55
80	80	0.60	0.67	80	0.55

Table 6-15 Test 4: 1 x 140 users, 3 x 80 users

Shares	Users	Throughput (DS/sec)		Users	Throughput (DS/sec)
		WLM passive	WLM active		WLM active
120	140	11.85	12.39	120	11.37
80	80	7.55	7.50	80	7.59
80	80	7.55	7.50	80	7.59
80	80	7.55	7.50	80	7.59
	Total	34.50	34.89	Total	34.14

Results: Comparing the results for passive and active mode for the system with high load, we found that in active mode, due to a higher share distribution, the average response time improves by 28% (1.83 versus 1.32 sec). The systems with low load suffered in response time by about 12% (0.60 versus 0.68 sec). If we now compare the results with the base setup of 7.2, “Test 6: workload distributed unevenly” on page 88, we find that the average response times of all four systems suffer if we increase the number of users of the system with the high load to 140. This system suffered by 132% (1.32 versus 0.57 sec) and the average response times of the systems with low load suffered by 24% (0.55 versus 0.68 sec).

The throughput per user of the system with the high load is slightly better in active than in passive mode. The opposite is true for the systems with low load, but the difference is very small.

If we increase the number of users of the system with the high load even further to 160, we get similar results for average response times and throughput, as shown in the following tables and Figure 6-4 on page 81.

Table 6-16 Test 4: 1 x 160 users, 3 x 80 users

Shares	**Users**	**Average response time (sec)**		**Users**	**Average response time (sec)**
		WLM passive	**WLM active**		**WLM active**
120	160	2.48	2.11	120	0.57
80	80	0.65	0.70	80	0.55
80	80	0.65	0.69	80	0.55
80	80	0.65	0.71	80	0.55

Table 6-17 Test 4: 1 x 160 users, 3 x 80 users

Shares	**Users**	**Throughput (DS/sec)**		**Users**	**Throughput (DS/sec)**
		WLM passive	**WLM active**		**WLM active**
120	160	12.83	13.23	120	11.37
80	80	7.52	7.48	80	7.59
80	80	7.52	7.49	80	7.59
80	80	7.52	7.48	80	7.59
	Total	35.39	35.67	Total	34.14

6.4.2 More users in the low load system

We again used the setup of 6.2, “Test 2: Workload distributed unevenly” on page 72, (users 120/80/80/80 and shares 120/80/80/80) as a basis to compare and distributed the amount of shares in proportion to the number of users. In addition, we increased the number of users of one of the low load systems from 80 to 120, but left the share distribution for this system at 80. With this test, we were interested in the change of response time and throughput for the following:

- The system with the high load having 120 shares
- The system with the high load having 80 shares
- The influence on the other two systems with the low load

Table 6-18 Test 4: 2 x 120 users, 2 x 80 users

Shares	Users	Average response time (sec)		Users	Average response time (sec)
		WLM passive	**WLM active**		**WLM active**
120	120	2.19	0.72	120	0.57
80	80	0.77	0.74	80	0.55
80	120	2.15	2.61	80	0.55
80	80	0.77	0.71	80	0.55

Table 6-19 Test 4: 2 x 120 users, 2 x 80 users

Shares	Users	Throughput (DS/sec)		Users	Throughput (DS/sec)
		WLM passive	**WLM active**		**WLM active**
120	120	9.86	11.21	120	11.37
80	80	7.43	7.46	80	7.59
80	120	9.89	9.53	80	7.59
80	80	7.44	7.49	80	7.59
	Total	34.61	35.69	Total	34.14

Results: When comparing passive and active modes, we see that the system with high load with 120 shares improved significantly in average response time (2.19 versus 0.72 sec). WLM in active mode ensured an average response time that is the same as for the two systems with low load, which responded almost equally as in passive mode. The second system with high load with a share distribution of 80 suffered significantly in response time in active mode (2.61 versus 2.15 sec). However, if we compare the results with the base setup of 1x120 users and 3x80 users having the same share distribution, we see that the average response time decreases for all four systems. The difference is smallest for the two systems with low load and the system with high load with more shares. The system with increased workload, but unchanged share distribution of 80, suffered the most (0.55 versus 2.61 sec).

The throughput per user was lower for both systems with high load in passive WLM mode. Activating WLM increased the throughput for the system with high load with more shares. Compared to the base setup, the throughput per user decreased for all four systems. The system with high load and 80 shares again suffered the most.

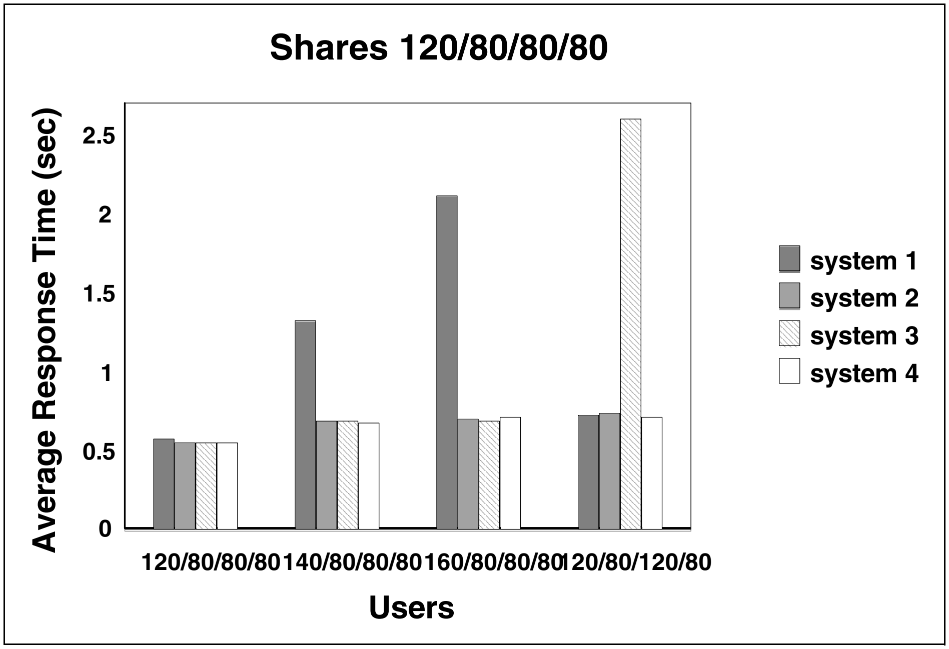

Figure 6-4 Test 4 summary

6.4.3 Test 4 summary

If we increase the workload of the system with the high load (140 and 160 users with 120 shares), the average response time for this system deteriorates significantly compared to having 120 users. The systems with low load are affected as well, but not as badly. Comparing the average response time of the system with high load with passive mode, it benefits from having a high share distribution in WLM active mode. The total throughput per user stays almost the same for all setups.

Increasing the workload of one of the systems with low load means that this system suffers significantly in throughput and average response times compared to the base setup and to passive mode. The two systems with low load have almost the same response times in passive and active mode, but lower response compared to the base setup. The second system with high load and increased share distribution achieves active mode response times equal to the two systems with the low load. This is a deterioration if we compare the results to the base setup, but a big improvement if we compare it to passive mode.

7

Scenario 2: One system more important than others

In Scenario 1: Equally important systems (Tests 1-4), we demonstrate the effects of different WLM share distributions on the four SAP R/3 systems. For all tests in Scenario 2 and 3 (Tests 5-8), we kept the shares for each system constant at 80 and examined the effects of changing tier distributions. In Scenario 2, we defined one system to be more important than the other three. The system should get preferred treatment over the other systems independent of its workload and should only give the resources left on the server to the less important systems. We implemented this behavior by assigning the WLM class of the important system to tier 1 and classes of the other three systems to tier 2. In general, there are 10 available tiers from 0 to 9 with tier 0 being the most important and tier 9 being the least important one.

7.1 Test 5: Workload distributed evenly

We increased the number of users per system from 4x50 to 4x90 users in a similar manner as we did in 6.1, "Test 1: Workload distributed evenly" on page 68, and compared the results of WLM in passive and active mode with a tier configuration of 1/2/2/2. We then increased the tier separation for the last two setups to 1/6/6/6, where we have a lot of resource competition, to see if a bigger tier separation has greater impact on our results, as suggested by *AIX 5L Workload Manager (WLM)*, SG24-5977. We obtained the following results.

Table 7-1 Test 5: 4 x 50 users

Users	Tier	Average response time (sec)		Throughput (DS/sec)	
		WLM passive	WLM active	WLM passive	WLM active
50	1	0.26	0.25	4.88	4.88
50	2	0.26	0.26	4.88	4.88
50	2	0.26	0.27	4.88	4.88
50	2	0.27	0.26	4.88	4.88
			Total	19.52	19.52

Table 7-2 Test 5: 4 x 60 users

Users	Tier	Average response time (sec)		Throughput (DS/sec)	
		WLM passive	WLM active	WLM passive	WLM active
60	1	0.27	0.25	5.85	5.86
60	2	0.27	0.28	5.85	5.84
60	2	0.27	0.28	5.85	5.84
60	2	0.27	0.28	5.85	5.85
			Total	23.41	23.40

Table 7-3 Test 5: 4 x 70 users

Users	Tier	Average response time (sec)		Throughput (DS/sec)	
		WLM passive	WLM active	WLM passive	WLM active
70	1	0.31	0.26	6.80	6.83
70	2	0.31	0.32	6.80	6.79
70	2	0.31	0.32	6.80	6.79
70	2	0.31	0.32	6.80	6.79
			Total	27.20	27.21

Table 7-4 Test 5: 4 x 80 users

Tier	Users	Average response time (sec)			Throughput (DS/sec)		
		WLM passive	WLM active		WLM passive	WLM active	
			x=2	x=6		x=2	x=6
1	80	0.37	0.26	0.27	7.73	7.81	7.80
x	80	0.37	0.40	0.39	7.73	7.70	7.72
x	80	0.37	0.39	0.39	7.73	7.70	7.71
x	80	0.37	0.40	0.39	7.73	7.71	7.71
				Total	30.91	30.91	30.93

Table 7-5 Test 5: 4 x 90 users

Tier	Users	Average response time (sec)			Throughput (DS/sec)		
		WLM passive	WLM active		WLM passive	WLM active	
			x=2	x=6		x=2	x=6
1	90	0.71	0.28	0.28	8.41	8.77	8.77
x	90	0.72	0.67	0.68	8.41	8.44	8.44
x	90	0.73	0.68	0.66	8.40	8.44	8.45
x	90	0.71	0.69	0.67	8.41	8.42	8.44
				Total	33.63	34.07	34.10

Results: Generally, the preferred system of tier 1 has the best average response times compared to the other systems in tier 2. The effect is very small for the 4x50 users setup, because there are enough resources on the server. However, the more we increase the total number of users, and therefore the competition for resources, the bigger the difference is between average response times of the preferred system and the other three systems. For example, for the 4x60 users setup, we have an improved average response time of 11% (0.25 versus 0.28 sec). Looking at the 4x80 users setup, we have an improvement of 31% (0.27 versus 0.39 sec). These results clearly show that the system in tier 1 gets most resources and the other systems get only the resources left on the server.

Comparing the passive and active modes, we found that up to the 4x80 users setup there are slightly better response times for the less important systems in passive mode. For the 4x60 users setup, it is 0.27 versus 0.28 sec (4%), and for the 4x80 users setup, it is 0.37 versus 0.40 sec (7%). The average response times for the preferred system are better if WLM is in active mode. This is what we expected and what we wanted to achieve. Because it is in a better tier, it gets resource priority. Using the same examples, the results for the 4x60 users setup are 0.27 versus 0.25 sec (7%), and for the 4x80 users setup, 0.37 versus 0.26 sec (30%). For a total of 360 users, that is 4x90 users, the difference in average response time of passive and active WLM modes for the preferred system is 61%. The average response times of all four systems are better in active mode than in passive mode, not just for the preferred system. For the 4x90 users setup, we have to bear in mind that there is permanent overcommitment of CPU resources, so we do not obtain acceptable results anymore.

The total throughput is the same in active and passive mode for up to 4x80 users. It is slightly better in active mode for the 4x90 users setup. The throughput of the preferred system in all setups is higher than for the other systems. Looking at the 4x60 users setup, for example, the preferred system has 5.86 versus 5.84 DS/sec. For 4x80 users setup, we have 7.81 versus 7.70 DS/sec. This is a difference in throughput of only about 1%.

According to the *AIX 5L Workload Manager (WLM)*, SG24-5977, more emphasis can be put on resource allocation if tier separation is increased. However, comparing our test results of the average response times and throughput with an increase of tier separation from 1/2/2/2 to 1/6/6/6, we found no evidence for such an effect for the software release we used (bos.mp 5.1.0.15).

7.1.1 Test 5 summary

Putting one system in a prioritized tier ensures that this preferred system gets all necessary resources. It has, therefore, the same average response time for each number of users (approximately 0.26 sec), as shown in Figure 7-1.

The preferred system (tier 1) has better average response times in WLM active mode than in passive mode and compared to the other systems.

All other systems suffer in response time. The difference increases the more shortage of resources there is.

The total throughput of the systems is the same for WLM in active and passive mode.

There is a slight increase of throughput for the preferred system, and therefore, a small decrease in throughput for the other, less important systems comparing WLM active and passive modes.

Increasing the separation of tiers between systems has no effect.

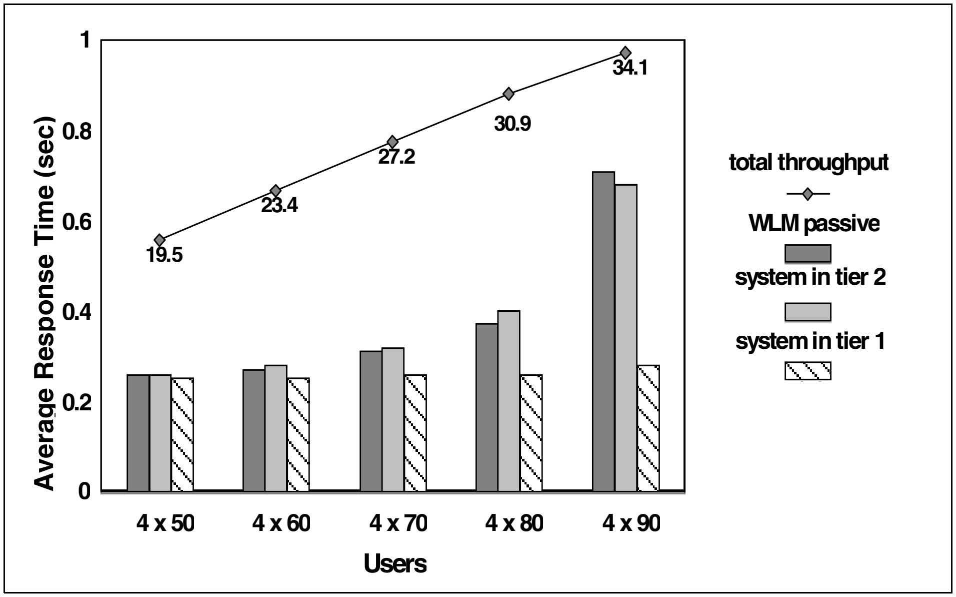

Figure 7-1 Test 5 summary

In 6.1, “Test 1: Workload distributed evenly” on page 68, we measured the performance for setups with 4x50 users up to 4x90 users, giving one system a higher share distribution. In 7.1, “Test 5: Workload distributed evenly” on page 84, we used the same user setups, but this time, we put one system in a prioritized tier. The results are almost identical, as can be seen by comparing the diagrams of Test 1 and Test 5 and the summary tables of both tests. That means, if we have four systems with equal numbers of users, and we want to prioritize one system to have best performance, we can either give it more shares than the others or set it in a higher prioritized tier. This worked fine for a setup where the systems have equal workload. Both configurations, however, reacted differently to an additional load, as we show in Tests 6 and 8.

Table 7-6 Test 5 summary: Average response time

	Average response time (sec)				
Users	**4x50**	**4x60**	**4x70**	**4x80**	**4x90**
Tier 2	0.26	0.28	0.32	0.39	0.67
Tier 1	0.25	0.25	0.26	0.27	0.28
Advantage	4%	11%	19%	31%	59%

Table 7-7 Test 5 summary: Throughput

	Throughput				
Users	**4x50**	**4x60**	**4x70**	**4x80**	**4x90**
Total (DS/sec)	19.5	23.4	27.2	30.9	34.1
Per user (DS/h)	351	351	351	348	340

7.2 Test 6: workload distributed unevenly

We start again with the base setup of 4x80 users, where the preferred system is in tier 1, and the others are in tier 2. We first increased the workload of the preferred system. Next, we increased the workload of one of the less important systems and determined what influence these changes had on average response time and throughput. The important system now has 120 users, while the other systems of tier 2 have 80 users each.

Table 7-8 Test 6: 1 x 120 users, 3 x 80 users

Tier	Users	Average response time (sec)		Users	Average response time (sec)
		WLM passive	WLM active		WLM active
1	120	0.85	0.29	80	0.26
2	80	0.50	0.72	80	0.40
2	80	0.50	0.69	80	0.39
2	80	0.50	0.71	80	0.40

Table 7-9 Test 6: 1 x 120 users, 3 x 80 users

Tier	Users	Throughput (DS/sec)		Users	Throughput (DS/sec)
		WLM passive	WLM active		WLM active
1	120	11.07	11.68	80	7.81
2	80	7.62	7.47	80	7.70
2	80	7.62	7.49	80	7.70
2	80	7.62	7.47	80	7.71
	Total	33.94	34.10	Total	30.91

Results: If we increase the workload of the preferred system, we have no significant change in average response time for this system. All other systems suffer in average response and throughput compared to passive mode and compared to the base setup of 4x80 users. Therefore, putting one system in a prioritized tier ensures that it will get CPU resource priority.

We then increased the workload of one of the less important systems, which are all in tier 2. The important system of tier 1 has 80 users. We obtained the following results.

Table 7-10 Test 6: 1 x 120 users, 3 x 80 users

Tier	Users	Average response time (sec)		Users	Average response time (sec)
		WLM passive	WLM active		WLM active
1	80	0.50	0.28	80	0.26
2	120	0.85	1.67	80	0.40
2	80	0.50	0.35	80	0.39
2	80	0.50	0.38	80	0.40

Table 7-11 Test 6: 1 x 120 users, 3 x 80 users

Tier	Users	Throughput (DS/sec)		Users	Throughput (DS/sec)
		WLM passive	WLM active		WLM active
1	80	7.62	7.79	80	7.81
2	120	11.07	10.34	80	7.70
2	80	7.62	7.71	80	7.70
2	80	7.62	7.72	80	7.71
	Total	33.9	33.6	Total	30.9

Results: If we increase the workload for one of the less important systems, there is no significant change in average response times and throughput for the preferred system. It still got all the CPU resources needed, and therefore, we had the same results for this system as for the 4x80 users base setup. The non-preferred system with increased workload (120 users) suffered significantly in response time (0.40 versus 1.67 sec). The other two less important systems of tier 2 with 80 users each had only a very small improvement in response time and throughput (0.40 versus 0.38 sec). That means, in summary, increasing the workload of one of the less important systems has no negative influence on all other systems other than on its own performance. Its response time and throughput per user suffer significantly.

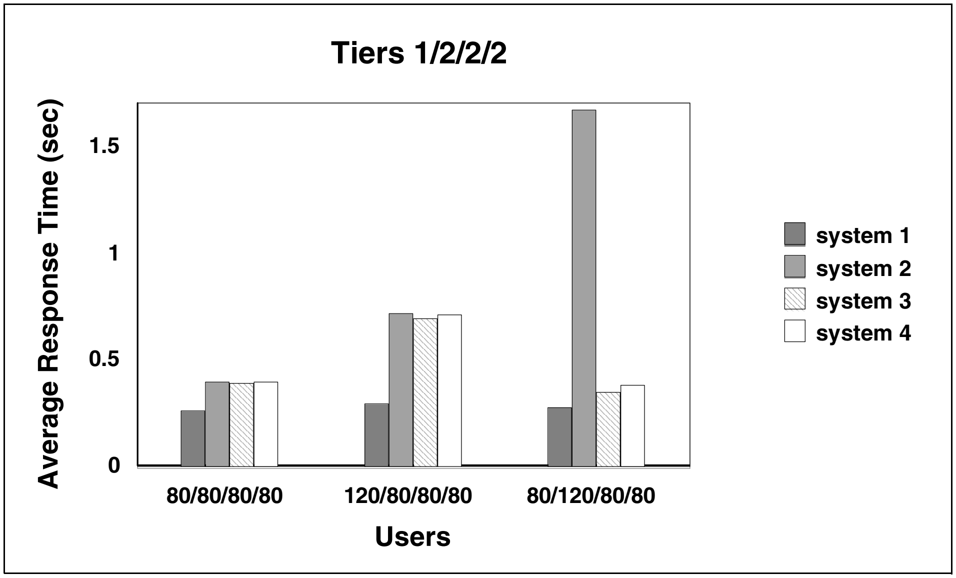

Figure 7-2 Test 6 summary

8

Scenario 3: One system less important than others

Another possibility for WLM to control CPU resources is to have one system that is not important and should only run when the other three more important systems do not need the resources. This setup can be desirable so that background jobs do not interfere with regular working hours. For this setup, we put the less important system in tier 2 and the other three systems in tier 1 and repeated the runs of 7.1, “Test 5: Workload distributed evenly” on page 84 and 7.2, “Test 6: workload distributed unevenly” on page 88.

8.1 Test 7: Workload distributed evenly

We ran tests as in 7.1, “Test 5: Workload distributed evenly” on page 84, starting with 4x50 users and going up to 4x90 users with a tier distribution of 2/1/1/1 and 6/1/1/1 for the last two setups. We obtained the following results.

Table 8-1 Test 7: 4 x 50 users

Tier	Users	Average response time (sec)		Throughput (DS/sec)	
		WLM passive	**WLM active**	**WLM passive**	**WLM active**
2	50	0.26	0.28	4.88	4.88
1	50	0.26	0.25	4.88	4.89
1	50	0.26	0.25	4.88	4.88
1	50	0.27	0.25	4.88	4.89
			Total	19.52	19.53

Table 8-2 Test 7: 4 x 60 users

Tier	Users	Average response time (sec)		Throughput (DS/sec)	
		WLM passive	**WLM active**	**WLM passive**	**WLM active**
2	60	0.27	0.31	5.85	5.83
1	60	0.27	0.26	5.85	5.85
1	60	0.27	0.26	5.85	5.86
1	60	0.27	0.26	5.85	5.85
			Total	23.41	23.39

Table 8-3 Test 7: 4 x 70 users

Tier	Users	Average response time (sec)		Throughput (DS/sec)	
		WLM passive	WLM active	WLM passive	WLM active
2	70	0.31	0.41	6.80	6.74
1	70	0.31	0.28	6.80	6.82
1	70	0.31	0.28	6.80	6.82
1	70	0.31	0.28	6.80	6.82
			Total	27.20	27.19

Table 8-4 Test 7: 4 x 80 users

Tier	Users	Average response time (sec)			Throughput (DS/sec)		
		WLM passive	WLM active		WLM passive	WLM active	
			x=2	x=6		x=2	x=6
x	80	0.37	0.56	0.57	7.73	7.57	7.58
1	80	0.37	0.31	0.31	7.73	7.77	7.77
1	80	0.37	0.31	0.31	7.73	7.77	7.76
1	80	0.37	0.31	0.31	7.73	7.77	7.77
				Total	30.91	30.80	30.88

Table 8-5 Test 7: 4 x 90 users

Tier	Users	Average response time (sec)			Throughput (DS/sec)		
		WLM passive	WLM active		WLM passive	WLM active	
			x=2	x=6		x=2	x=6
x	90	0.71	1.35	1.29	8.41	7.94	7.99
1	90	0.72	0.36	0.37	8.41	8.69	8.69
1	90	0.73	0.36	0.37	8.40	8.70	8.70
1	90	0.71	0.37	0.37	8.41	8.70	8.69
				Total	33.63	34.07	34.06

Results: This is the opposite case of Test 5, but works in a similar way. The preferred systems had improved average response times compared to the less important system and compared to WLM in passive mode. For the 4x80 users setup, for example, we obtained 0.56 versus 0.31 sec comparing the less important and the preferred systems. That is a difference of 45%. Looking now at passive and active modes, we had for the preferred systems an average response of 0.37 versus 0.31 sec, which means they respond 16% faster in active mode. Looking at the throughput, we found that the total throughput is the same for WLM in active and passive modes. The throughput of the preferred system was slightly better in active mode and also better than the less important system. Comparing the throughput for the preferred systems to the less important system, we obtained 7.77versus 7.57 DS/sec, a difference of 3%.

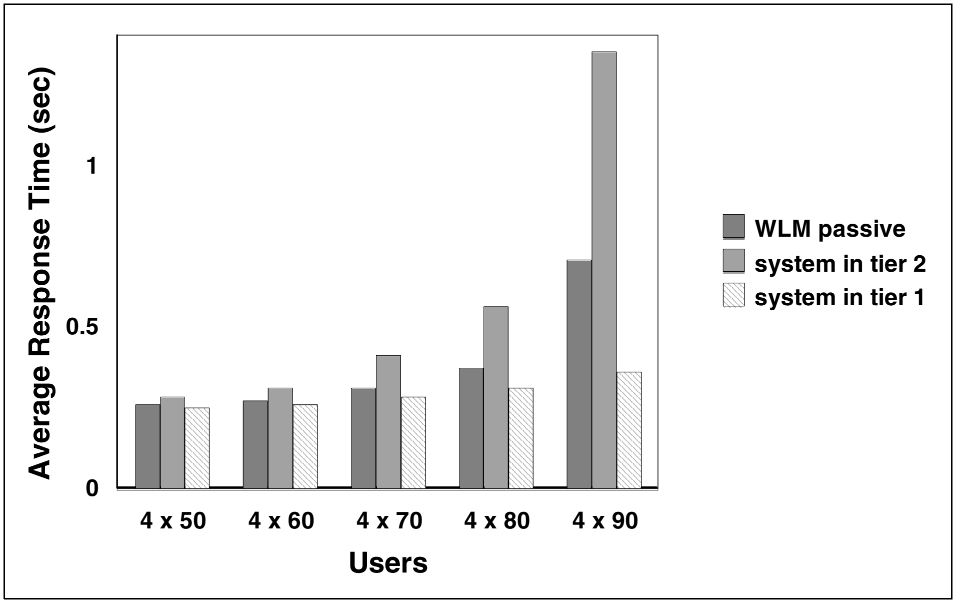

Figure 8-1 Test 7 summary

With a change of tier distribution from 2/1/1/1 to 6/1/1/1, we again had no change in results, which confirms our findings of 7.1, “Test 5: Workload distributed evenly” on page 84. For the software release we used, a greater tier separation has no impact on the results.

8.2 Test 8: Workload distributed unevenly

We ran the same tests as in 7.1, "Test 5: Workload distributed evenly" on page 84, but this time with a tier distribution of 2/1/1/1. Therefore, one system is in a lower prioritized tier 2 and should get only resources that are not needed by the other three systems. From our base setup of 4x80 users, we increased the workload of the less important system assigned to tier 2.

Table 8-6 Test 8: 1 x 120 users, 3 x 80 users

Tier	Users	Average response time (sec)		Users	Average response time (sec)
		WLM passive	**WLM active**		**WLM active**
2	120	0.85	1.22	80	0.56
1	80	0.50	0.34	80	0.31
1	80	0.50	0.34	80	0.31
1	80	0.50	0.34	80	0.31

Table 8-7 Test 8: 1 x 120 users, 3 x 80 users

Tier	Users	Throughput (DS/sec)		Users	Throughput (DS/sec)
		WLM passive	**WLM active**		**WLM active**
2	120	11.07	10.7	80	7.57
1	80	7.62	7.75	80	7.77
1	80	7.62	7.75	80	7.77
1	80	7.62	7.75	80	7.77
	Total	33.94	33.9	Total	30.88

Results: The average response time of the less important system deteriorates from 0.56 sec to 1.22 sec. The other three systems respond almost in the same way as for the 4x80 users setup (0.31 versus 0.34 sec). This shows that increasing the workload of the less important system does not significantly influence the important ones. They still get the resources required as before. We see a similar behavior for the throughput.

We then increased the workload of one of the important systems in tier 1. The less important system of tier 2 has 80 users.

Table 8-8 Test 8: 1 x 120 users, 3 x 80 users

Tier	Users	Average response time (sec)		Users	Average response time (sec)
		WLM passive	WLM active		WLM active
2	80	0.50	1.34	80	0.56
1	120	0.85	0.49	80	0.31
1	80	0.50	0.30	80	0.31
1	80	0.50	0.30	80	0.31

Table 8-9 Test 8: 1 x 120 users, 3 x 80 users

Tier	Users	Throughput (DS/sec)		Users	Throughput (DS/sec)
		WLM passive	WLM active		WLM active
2	80	7.62	7.06	80	7.57
1	120	11.07	11.5	80	7.77
1	80	7.62	7.77	80	7.77
1	80	7.62	7.77	80	7.77
	Total	33.94	34.0	Total	30.9

Results: The system with the increased workload has decreased average response times compared to the base setup of 4x80 users. We have seen in 7.2, “Test 6: workload distributed unevenly” on page 88 that we can achieve an average response time of 0.29 sec for a system with 120 users in tier 1. But for our system, we only had a response of 0.49 sec. Therefore, this system did not get all the resources it needed even though it is in a prioritized tier, instead it competes with the other two systems of the same tier that all have the same amount of shares. These systems, however, did not suffer at all by the increase of the total workload. The less important system had decreased response times and throughput compared to the base setup as expected because it only got the remaining resources. Compared to passive mode, only the less important system of tier 2 suffered in performance. The important systems all had improved results in WLM active mode.

8.2.1 Test 8 summary

If we increase the workload of the less important system (tier 2), it suffers significantly in response time and throughput. The other three important systems (tier 1) are hardly affected by the increase of total workload.

If we increase the workload of one of the important systems (tier 1), this system suffers in response time and throughput, and the other two preferred systems are not affected. The less important system gets only the leftover resources and is, therefore, badly influenced. Comparing active and passive modes, the three important systems benefit in performance in active mode. On the other hand, the less important system suffers, as we expected to see. These results are displayed in Figure 8-2.

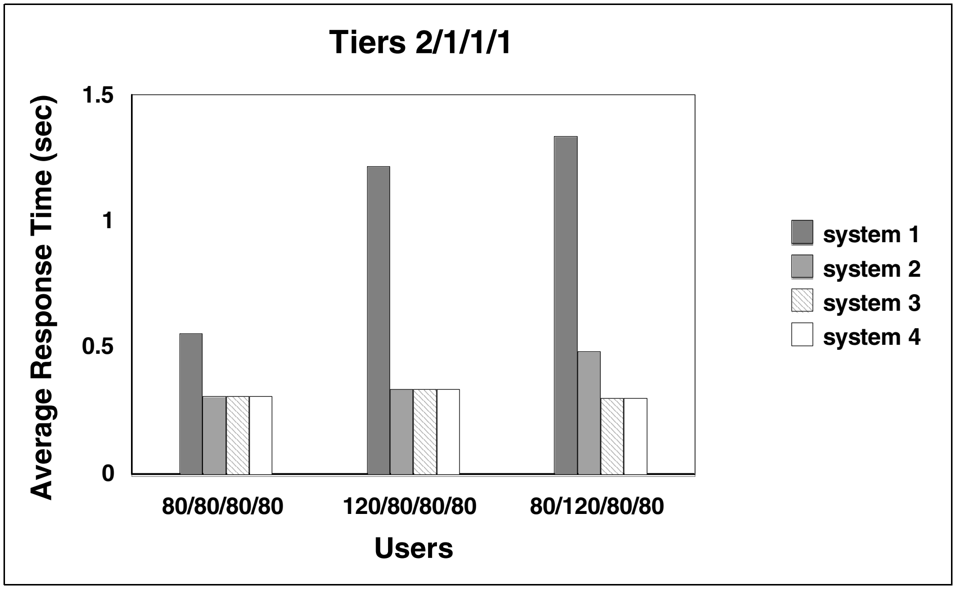

Figure 8-2 Test 8 summary

9

Conclusions and recommendations

With this project, we were able to confirm that with AIX 5L Workload Manager it is possible to regulate CPU resources used by different SAP R/3 systems that are consolidated on one IBM @server pSeries node. This can be done by first defining the classes for the different systems and then by allocating either different amounts of shares or by assigning these classes to different WLM tiers as described in this paper.

A generous allocation of shares to a single system, 6.1, “Test 1: Workload distributed evenly” on page 68, has the same effect on average response times and throughput as putting one system in a preferred tier as in 7.1, “Test 5: Workload distributed evenly” on page 84. This works fine if we have the same workload on each system. However, both configurations behave differently when putting an additional load on one of the systems.

If we assign one system to a preferred tier and increase the workload of either the preferred system or one of the others, the performance of the preferred system does not change. It still has priority access to the CPU resources as in 7.2, “Test 6: workload distributed unevenly” on page 88. If, however, all systems are in the same tier, adding more workload intensifies the competition for resources which results in lower performance of all systems as in 6.3, “Test 3: Stress test with equal shares” on page 74 and 6.4, “Test 4: Stress test with unequal shares” on page 77. This can be adjusted by distributing shares according to the workload as in 6.2, “Test 2: Workload distributed unevenly” on page 72.

If we have a single system in a tier of lower priority as in 8.1, “Test 7: Workload distributed evenly” on page 94, and we increase the load on this system, we do not have an impact on the systems with high priority as in 8.2, “Test 8: Workload distributed unevenly” on page 97 (Test 8a). If we, on the other hand, increase the load of one of the systems with high priority, this system has a decreased response. It competes with the other two systems of the same tier that also have the same amount of shares. As expected, the other two systems of high priority do not suffer in performance. The system with low priority gets only the leftover CPU resources, and therefore, has the most deterioration in performance 8.2, “Test 8: Workload distributed unevenly” on page 97 (Test 8b).

Before you start implementing different WLM configurations, however, it is very important that you understand the requirements of your workload and decide what you want to achieve using Workload Manager. First, monitor the system performance using WLM in passive mode. Then, implement different WLM configurations according to your resource allocation goals. Monitor the system performance again in order to decide which configuration works best in your environment. WLM configurations can be modified and updated while WLM is running and take immediate effect without having to stop and restart WLM.

To help you plan your WLM strategies, here are some guidelines and configuration steps you should follow. Generally, you want to separate the interactive jobs that typically consume very little CPU time but require quick response times from batch jobs that are very CPU- and memory- intensive. In a database environment, for example, you want to separate online transaction processing (OLTP) traffic from queries of data mining. It is recommended to use resource shares rather than limits. WLM sees shares as goals to achieve, which allows greater system flexibility than imposed limits. Even if the shares set up are not optimal, the system is still able to balance the load reasonably well, as our different test scenarios have shown. With hard limits, WLM can do little to prevent applications from being starved of resources and should, therefore, not be modified. To promote a ranking of jobs, use tiers. Bear in mind, however, that processes assigned to higher tiers do not compete for resources with processes of lower tiers. Make sure that there are enough resources available for processes in lower tiers.

A

Implementation and configuration details

The following sections contain a syntax description and sample files and examples.

Syntax description

`token ::=  <token1> <token2> <token3>`	Sequence of token1, token2, and token3
`token ::=  { <token1> | <token2> }`	Alternative token1 or token2
`token ::=  <token1> +  <token2>`	Numeric addition of token1 and token2
`token ::=  <token1> * 10`	Numeric multiplication of token1
`token ::=  [ <token1>]`	Optional token1
`token ::=  ‘XX’`	Constant with value XX
`token ::=  00 .. nn`	Numeric range from 00 to nn

Sample Oracle files

The following sections contain sample Oracle files.

Oracle TNS names file

File tnsnames.ora:

```
<SID>.WORLD=
  (DESCRIPTION =
     (SDU = 32768)
     (ADDRESS_LIST =
        (ADDRESS =
          (COMMUNITY = SAP.WORLD)
          (PROTOCOL = TCP)
          (HOST = <node>)
          (PORT = <portnumber>)
        )
     )
     (CONNECT_DATA =
        (SID = <SID>)
        (GLOBAL_NAME = <SID>.WORLD)
        )
  )
```

Oracle listener configuration file

File oracle listener.ora:

```
LISTENER =
  (ADDRESS_LIST =
     (ADDRESS=
        (PROTOCOL=IPC)
        (KEY= <SID>.WORLD)
     )
     (ADDRESS=
        (PROTOCOL=IPC)
        (KEY= <SID>)
     )
     (ADDRESS =
        (COMMUNITY = SAP.WORLD)
        (PROTOCOL = TCP)
        (HOST = <node>)
        (PORT = <portnumber>)
     )
  )

STARTUP_WAIT_TIME_LISTENER = 0
```

```
CONNECT_TIMEOUT_LISTENER = 10
TRACE_LEVEL_LISTENER = OFF
SID_LIST_LISTENER =
  (SID_LIST =
     (SID_DESC =
        (SDU = 32768)
        (SID_NAME = <SID>)
        (ORACLE_HOME = /oracle/<SID>/817_32)
     )
  )
```

Sample IBM Tivoli Storage Manager files

The following sections provide a sample IBM Tivoli Storage Manager client option file and sample Tivoli Storage Manager include/exclude files.

Sample IBM Tivoli Storage Manager client option file

File dsm.aix.opt:

```
servername       <tsmserver>
compressalways   yes
followsymbolic   yes
```

File dsm.<SID>_exe.opt:

```
servername       <tsmserver>
virtualnodename  <virtualnode>
compressalways   yes
domain           /global.<SID>
domain           /oracle/<SID>
domain           /oracle/<SID>/sapdata1
```

Sample IBM Tivoli Storage Manager include/exclude files

File include.aix:

```
INCLUDE.FILE            /.../*                AIX
* exclude database
EXCLUDE.FS              /oracle/???
EXCLUDE.FILE            /oracle/???/.../*
* exclude SAP
EXCLUDE.FS              /global.???
EXCLUDE.FILE            /global.???/.../*
EXCLUDE.DIR             /.../lost+found
EXCLUDE.FILE            /dev/.../*
EXCLUDE.FILE            /.../core
```

File include.exe:

```
INCLUDE.FILE            /.../*            EXEC
* exclude tablespaces, redo logs, and control files
EXCLUDE.FILE            /oracle/.../ctrl*.ctl
EXCLUDE.FILE            /oracle/.../*.dbf
EXCLUDE.FILE            /oracle/.../*.dmp
EXCLUDE.FILE            /oracle/.../*.data*
EXCLUDE.FILE            /usr/sap/.../PAGFIL*
EXCLUDE.FILE            /usr/sap/.../ROLLFL*
EXCLUDE.DIR             /.../lost+found
EXCLUDE.FILE            /.../core
```

Workload Manager

The following sections describe processor binding with WLM and WLM setup.

Processor binding with WLM

The following shows an example of how to achieve SAP instance binding to processor groups for multiple instances of a single SAP system. Different SAP systems with a distinct SAP system name run under different AIX user IDs and do not require application tagging, because WLM can differentiate the processes by user ID and group. Complete the following steps:

1. Define a group of processors as a resource set:

   ```
   lsrset -vn 'ISICC’
   T  Name          Owner   Group   Mode   CPU  Memory  Resources
   s  ISICC/cpugp1  root    system  rwr-r-  4       0   sys/cpu.0000
                                                        sys/cpu.0001
                                                        sys/cpu.0002
                                                        sys/cpu.0003
   ```

```
s  ISICC/cpugp2  root    system  rwr-r-  4       0   sys/cpu.0004
                                                     sys/cpu.0005
                                                     sys/cpu.0006
                                                     sys/cpu.0007
s  ISICC/cpugp3  root    system  rwr-r-  4       0   sys/cpu.0008
                                                     sys/cpu.0009
                                                     sys/cpu.0010
                                                     sys/cpu.0011
s  ISICC/cpugp4  root    system  rwr-r-  4       0   sys/cpu.0012
                                                     sys/cpu.0013
                                                     sys/cpu.0014
                                                     sys/cpu.0015
s  ISICC/cpugp5  root    system  rwr-r-  4       0   sys/cpu.0016
                                                     sys/cpu.0017
                                                     sys/cpu.0018
                                                     sys/cpu.0019
s  ISICC/cpugp6  root    system  rwr-r-  4       0   sys/cpu.0020
                                                     sys/cpu.0021
                                                     sys/cpu.0022
                                                     sys/cpu.0023
s  ISICC/cpugp7  root    system  rwr-r-  4       0   sys/cpu.0024
                                                     sys/cpu.0025
                                                     sys/cpu.0026
                                                     sys/cpu.0027
s  ISICC/cpugp8  root    system  rwr-r-  4       0   sys/cpu.0028
                                                     sys/cpu.0029
                 .                                   sys/cpu.0030
                                                     sys/cpu.0031
```

2. WLM classes

   Define a class for each SAP instance. The following shows the class definition stanza for application server instance 01:

```
aps01:
      description = "appl server 01"
      tier   = 2
      inheritance = "yes"
      authuser = "b01adm"
      authgroup = "sapsys"
      adminuser = "root"
      admingroup = "system"
      rset   = "ISICC/cpugp1"
```

3. WLM rules

   Create rules to assign a tag to each class:

```
Class     resvd   User    Group   Application  Type    Tag
aps00     -       -       -       -            -       aps00
aps08     -       -       -       -            -       aps08
aps07     -       -       -       -            -       aps07
aps06     -       -       -       -            -       aps06
aps05     -       -       -       -            -       aps05
aps04     -       -       -       -            -       aps04
aps03     -       -       -       -            -       aps03
aps02     -       -       -       -            -       aps02
aps01     -       -       -       -            -       aps01
```

4. Program settag

   Compile and link the program settag that is tagging all processes of a SAP instance:

```
#include <unistd.h>
#include <stdio.h>
#include <errno.h>
#include <sys/wlm.h>

/* Program for launching and tagging an application */

main (argc,argv)
   char **argv;
   int argc;
{
   int rc,flags;
   if (argc != 3) {
       usage(argv[0]);
       exit(1);
   }
      flags= WLM_VERSION|SWLMTAGINHERITFORK|SWLMTAGINHERITEXEC;
   if(wlm_initialize(WLM_VERSION)){
      perror("wlm_initialize");
      exit(1);
      }
   if(wlm_set_tag(argv[1],&flags)){
      perror("wlm_set_tag");
      exit(2);
      }
   if (execl("/usr/bin/sh", "sh", "-c", argv[2],0)){
      perror("execl"); printf("Problem launching app...\n");
      exit(3);
      }
   /* printf("Just launched %s and should be tagged as
              %s\n",argv[2],argv[1]); */
   exit(0);
}

usage(char *cp)
{
   printf("Usage: %s tag_name program_name \n",cp);
   printf("where: tag_name is the rule tag that program_name will inherit
\n");
}
```

5. Modifications on SAP scripts

   Modify the startsap script supplied by SAP to start an instance with the appropriate application tag. WLM will recognize the tag of a starting process and will classify the process automatically using the above rules file. WLM will then use the class definitions and assign the process to the appropriate resource group. For more details about application tags, see the IBM Redbook *AIX 5L Workload Manager (WLM)*, SG24-5977. The sample program for application tagging presented in this IBM Redbook was modified to allow starting an executable with command line arguments as required for the sapstart executable.

   The following shows the required changes to the startsap script in order to start the application server work processes with a WLM tag:

   ```
   MYNAME=$( basename $0 )
   MYTAG=aps${MYNAME##*_}
   settag $MYTAG "$SAPSTART pf=$PROFILE_DIR/$START_PROFILE >> $LOGFILE 2>&1"
   ```

WLM setup

The WLM assignment rules used in Part 2, “Regulating system resources with AIX 5L Workload Manager” on page 59 are as follows:

```
Class        User                      Group  Application  Type  Tag
System       root                      -      -            -     -
sd1cl        sd1adm, db2sd1            -      -            -     -
sd2cl        sd2adm, db2sd2            -      -            -     -
sd3cl        sd3adm, db2sd3            -      -            -     -
sd4cl        sd4adm, db2sd4            -      -            -     -
Default      -                         -      -            -     -
```

Attributes with a hyphen (-) are not specified. sd1adm is the SAP administration user of system1, and db2sd1 is the installation user.

B

Products

All concepts use IBM @server pSeries hardware. The following software product versions were the base for this document:

- SAP R/3 4.6
- SAP Business Warehouse (BW) 3.0
- SAP Advanced Planner and Optimizer (APO)
- AIX 4.x, AIX 5L
- Oracle 8.x
- DB2/6000
- HACMP 4.4

To implement the test setups described in Part 2, “Regulating system resources with AIX 5L Workload Manager” on page 59, we used the following software packages:

- AIX 5L Version 5.1, bos.mp 5.1.0.15
- DB2/6000, Release 6.1.0.31, PTF U473545
- SAP Release 4.6C, Hot Package Level 15 for ABA, Support Package
- R/3 Support Package, HR Support Package, SAP Kernel Release 4.6D with Patch Level 797

Related publications

The publications listed in this section are considered particularly suitable for a more detailed discussion of the topics covered in this Redpaper.

IBM Redbooks

For information on ordering these publications, see “How to get IBM Redbooks” on page 114.

- *A Holistic Approach to a Reliable Infrastructure for SAP R/3 on AIX*, SG24-5050
- *AIX 5L Differences Guide Version 5.1 Edition*, SG24-5765
- *AIX 5L Workload Manager (WLM)*, SG24-5977
- *AIX Logical Volume Manager from A to Z: Introduction and Concepts*, SG24-5432

Other resources

These publications are also relevant as further information sources. You must be registered as an SAP Service Marketplace user to access the following resources. The registration requires an SAP installation or customer number. To register, go to:

`http://service.sap.com`

- *Configuration of R/3 on hosts with much RAM*, SAP Online Service System (OSS) Note 146528
- *Creating a new or 2nd Oracle SID with runInstaller*, SAP OSS Note 350251
- *Network Integration of SAP Servers*, SAP AG
- *SAP R/3 in switchover environments*, SAP AG
- *Several systems/instances on one UNIX computer*, SAP OSS Note 21960
- *Work processes terminate when reconfiguring LPARs*, SAP OSS Note 569569

How to get IBM Redbooks

You can order hardcopy Redbooks, as well as view, download, or search for Redbooks at the following Web site:

ibm.com/redbooks

You can also download additional materials (code samples or diskette/CD-ROM images) from that site.

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.

IBM Redbooks

Consolidating Multiple SAP Systems on One IBM @server pSeries

Consolidating Multiple SAP Systems on One IBM eserver pSeries

Redpaper

Implementing complex SAP system landscapes

Concepts and experiences with LPAR on AIX 5L

Manage workloads for multiple SAP systems

With the introduction of ever more powerful UNIX servers, such as the IBM eserver pSeries 690 models, server consolidation is back on the agenda in many computer centers. As a widely used business application, SAP R/3 is often on top of the list when applications are identified that are currently running on several smaller servers, but could probably be consolidated onto a single, more powerful server. This IBM Redpaper covers two major trends in the IT business: the technological improvements of server platforms, providing increasingly powerful and flexible servers, such as the IBM eserver pSeries 690, and the close integration of business processes along the value chain, resulting in multifunctional application portfolios, such as mySAP.com.

Today, server consolidation is a must for many IT sites. Minimized total cost of ownership (TCO) and complexity, with the maximum amount of flexibility, is a crucial goal of nearly all customers. This Redpaper describes how you can exploit the technical features of the IBM eserver pSeries platform in order to accomplish these requirements. It is intended to help IT architects and specialists in designing, implementing, and using mySAP.com consolidation scenarios. It includes the newest key information about AIX 5L, logical partitioning (LPAR), and AIX Workload Manager (WLM). The concepts presented in this paper are field-tested best practices.

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