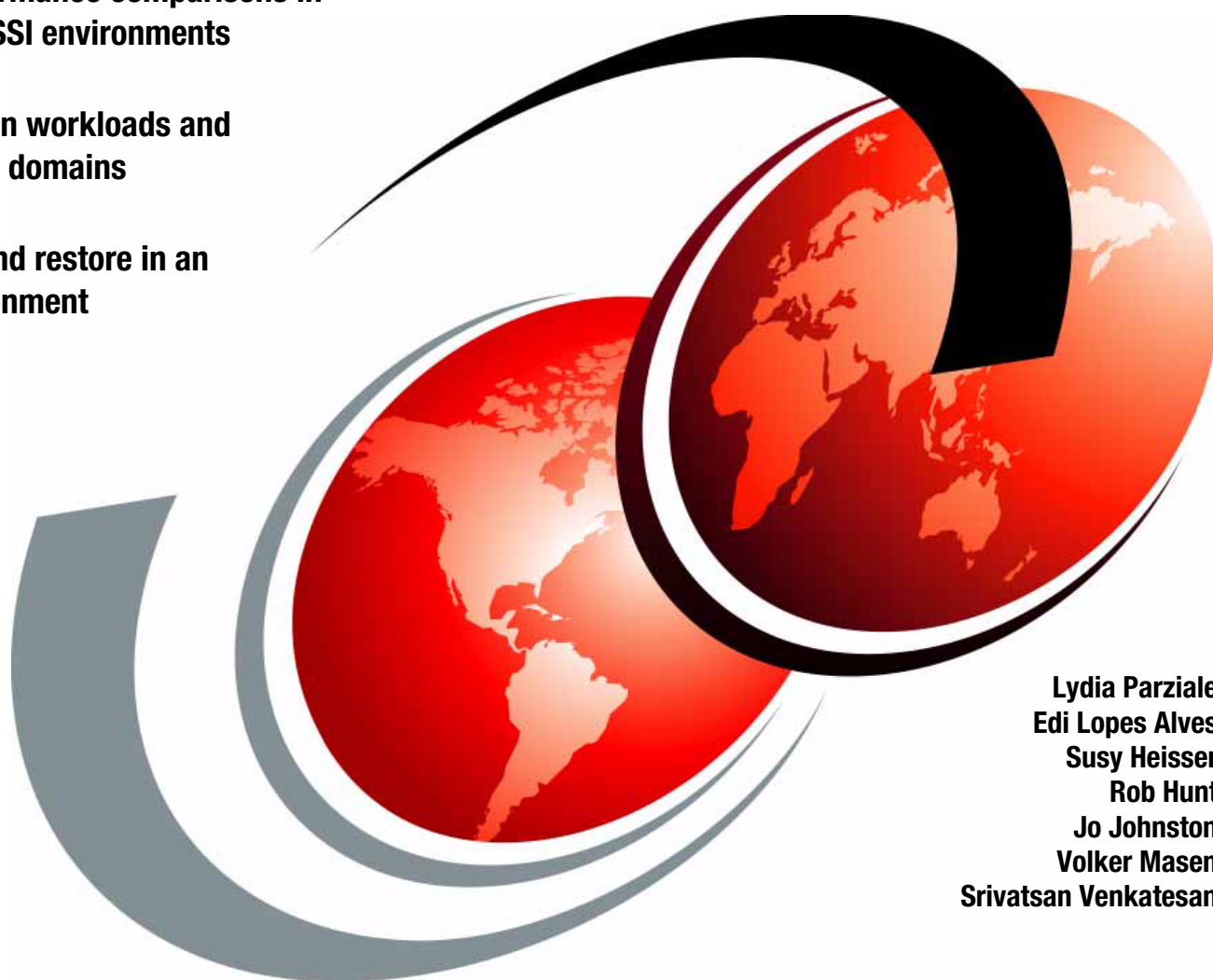


# Using z/VM v 6.2 Single System Image (SSI) and Live Guest Relocation (LGR)

LGR performance comparisons in different SSI environments

Application workloads and relocation domains

Backup and restore in an SSI environment



Lydia Parziale  
Edi Lopes Alves  
Susy Heisser  
Rob Hunt  
Jo Johnston  
Volker Masen  
Srivatsan Venkatesan

# Redbooks





International Technical Support Organization

**Using z/VM v 6.2 Single System Image (SSI)  
and Live Guest Relocation (LGR)**

August 2012

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

**First Edition (August 2012)**

This edition applies to Version 6, Release 2, of z/VM and IBM Backup and Restore Manager for z/VM Version 1, Release 2.

© Copyright International Business Machines Corporation 2012. All rights reserved.

Note to U.S. Government Users Restricted Rights -- Use, duplication or disclosure restricted by GSA ADP Schedule Contract with IBM Corp.

# Contents

<b>Notices</b> .....	vii
Trademarks .....	viii
<b>Preface</b> .....	ix
The team who wrote this book .....	ix
Now you can become a published author, too! .....	xi
Comments welcome. ....	xi
Stay connected to IBM Redbooks .....	xi
<b>Chapter 1. Overview of SSI and LGR</b> .....	1
1.1 Single system image (SSI) .....	2
1.1.1 SSI commands .....	3
1.1.2 SSI mapping of real devices using equivalency identifiers (EQIDs) .....	4
1.1.3 SSI coordinated MAC addresses .....	4
1.1.4 SSI use of ISFC links .....	4
1.1.5 SSI members .....	5
1.2 Changes for SSI in the USER Directory .....	5
1.2.1 Single-configuration virtual machine definition .....	6
1.2.2 Multi-configuration virtual machine definition .....	6
1.2.3 DIRMAINT commands .....	7
1.3 Live guest relocation (LGR) .....	7
<b>Chapter 2. Lab environment</b> .....	9
2.1 Overview of our four member cluster .....	10
2.2 Overview of our two member cluster. ....	12
<b>Chapter 3. Applications setup</b> .....	15
3.1 SAP and DB2 on System z .....	16
3.2 IBM Trade Performance Benchmark. ....	17
3.3 Workload on non SCSI disks .....	18
3.4 Workload with DB2 logs on SCSI disks .....	19
3.5 Workload with DB2 logs and WebSphere Application Server on SCSI disks .....	21
<b>Chapter 4. Using SCSI for file systems</b> .....	23
4.1 SCSI lab setup .....	24
4.2 Special setup in an SSI cluster .....	24
4.3 Hardware prerequisites .....	26
4.4 Understanding the WWPNs on the z Hardware .....	26
4.4.1 WWPNs used in our setup .....	26
4.4.2 Reading the WWPN over the HMC. ....	27
4.4.3 Reading the WWPN from z/VM. ....	27
4.4.4 NPIV .....	28
4.5 Defining the volume in the storage controller .....	29
4.6 Defining the FCP channels to z/VM .....	31
4.6.1 Defining FCP channels dynamically .....	31
4.6.2 Defining FCP channels permanently. ....	32
4.7 Defining the SCSI file system in Linux .....	33
4.7.1 Setting up multipath .....	37
4.7.2 Define the volume group. ....	38
4.7.3 Creating a file system .....	39

<b>Chapter 5. Relocation domains</b> . . . . .	41
5.1 Review of relocation domain concepts . . . . .	42
5.2 Assigning relocation domains to a virtual guest . . . . .	42
5.3 Using the default relocation domain named SSI . . . . .	44
5.3.1 All four members active in the cluster . . . . .	44
5.3.2 Only members of a System z z196 are active in the cluster . . . . .	45
5.4 Using relocation domains to reflect the architecture level . . . . .	47
5.5 Relocation to a different architecture domain . . . . .	49
5.6 Using relocation domains with special hardware or software features . . . . .	51
5.7 Using relocation domains for different business purposes . . . . .	52
5.8 Upgrading your environment . . . . .	55
<b>Chapter 6. Performance topics</b> . . . . .	59
6.1 Install and set up the IBM Performance Toolkit for VM in an SSI environment . . . . .	60
6.1.1 Activate the IBM Performance Toolkit web interface . . . . .	60
6.1.2 Activate Linux guest monitoring by IBM Performance Toolkit . . . . .	61
6.1.3 Monitoring multiple z/VM members from a central monitor machine . . . . .	62
6.2 Monitoring SSI-relevant data in z/VM 6.2 . . . . .	65
6.2.1 Activate monitoring of SSI data . . . . .	65
6.2.2 New performance data screens in z/VM 6.2 supporting the SSI function . . . . .	66
6.3 Introduction to performance aspects of LGR . . . . .	69
6.3.1 Monitor records for relocation information . . . . .	70
6.4 Sources of additional information . . . . .	72
<b>Chapter 7. Benchmarks for relocating Linux on System z guests using LGR</b> . . . . .	73
7.1 Relocation benchmark dependent on relocation options . . . . .	74
7.2 Relocation benchmark dependent on the number of CTCs . . . . .	76
7.3 Two Linux guests in a four cluster SSI system dependent on different LPARs and processors . . . . .	78
7.4 Two Linux guests in two and four cluster systems dependent on SCSI and non-SCSI . . . . .	81
<b>Chapter 8. IBM Backup and Restore Manager for z/VM</b> . . . . .	83
8.1 Overview of the IBM Backup and Restore Manager for z/VM . . . . .	84
8.2 Installation of IBM Backup and Restore Manager for z/VM . . . . .	85
8.2.1 Prerequisite: Create a Shared File System server and file pool . . . . .	85
8.2.2 Prerequisite: Install REXX library . . . . .	88
8.2.3 Userids used for Backup and Restore Manager for z/VM . . . . .	88
8.2.4 Set up service machines . . . . .	92
8.2.5 Final installation steps . . . . .	92
8.3 Set up the configuration file . . . . .	93
8.4 Back up and restore a single configuration user (a USER directory entry) . . . . .	93
8.4.1 How to back up a single configuration user (a USER directory entry) . . . . .	93
8.4.2 How to restore a data for a single configuration user (a USER directory entry) . . . . .	95
8.4.3 Known issues and workaround . . . . .	96
8.5 Back up and restore a multi-configuration user (an IDENTITY directory entry) . . . . .	97
8.5.1 How to back up a multi-configuration user (an IDENTITY directory entry) . . . . .	97
8.5.2 How to restore an IDENTITY . . . . .	97
8.6 Backup and restore commands . . . . .	98
8.6.1 BKR VOL . . . . .	98
8.6.2 BKR JOB . . . . .	98
8.6.3 BKR USER . . . . .	99
8.6.4 BKR LIST . . . . .	99
<b>Appendix A. Hints and tips</b> . . . . .	101

SCSI connection .....	102
SSI and IBM VM Backup and Restore Manager .....	102
LGR and performance .....	102
<b>Related publications</b> .....	103
IBM Redbooks publications .....	103
Other publications .....	103
Online resources .....	103
Help from IBM .....	104
<b>Index</b> .....	105



# 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 websites are provided for convenience only and do not in any manner serve as an endorsement of those websites. The materials at those websites are not part of the materials for this IBM product and use of those websites 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 illustrate 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.

# Trademarks

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corporation in the United States, other countries, or both. These and other IBM trademarked terms are marked on their first occurrence in this information with the appropriate symbol (® or ™), indicating US registered or common law trademarks owned by IBM at the time this information was published. Such trademarks may also be registered or common law trademarks in other countries. A current list of IBM trademarks is available on the Web at <http://www.ibm.com/legal/copytrade.shtml>

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

AIX®	OMEGAMON®	Tivoli®
CICS®	RACF®	WebSphere®
DB2®	Rational®	z/OS®
DirMaint™	Redbooks®	z/VM®
DS8000®	Redpapers™	z10™
FICON®	Redbooks (logo)  ®	zEnterprise®
IBM®	RMF™	zSeries®
IMS™	System z10®	
MVS™	System z®	

The following terms are trademarks of other companies:

Linux is a trademark of Linus Torvalds in the United States, other countries, or both.

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

# Preface

In this IBM® Redbooks® publication, we expand upon the concepts and experiences described in *An introduction to z/VM Single System Image (SSI) and Live Guest Relocation (LGR)*, SG24-8006. An overview of that book is provided in Chapter 1, “Overview of SSI and LGR” on page 1.

In writing this book, we re-used the same lab environment used in the first book, but expanded it to include IBM DB2® v10 on Linux on System z®, two IBM WebSphere® Application Server environments, and added a WebSphere application, used for performance benchmarking, which provided a workload that allowed us to observe the performance of the WebSphere Application Server during relocation of the z/VM® 6.2 member that was hosting the application server.

Additionally, this book examines the use of small computer system interface (SCSI) disks in the z/VM v6.2 environment and the results of using single system images (SSI) and live guest relocation (LGR) in this type of environment.

In the previous book, a detailed explanation of relocation domains was provided. In this book, we expand that discussion and provide use cases of relocation domains in different situations.

Finally, because the ability to back up and restore your data is of paramount importance, we have provided a discussion about how to use one tool, the IBM Backup and Restore Manager for z/VM, which can be used in the new z/VM6.2 environment. We provide a brief overview of the tool and describe the changes in the installation process as a result of using single system image clusters. We also demonstrate how to set up the configuration file, and how to back up and restore both a user and an identity.

This publication is intended for IT architects who will be responsible for designing the system and IT specialists who will have to build the system.

## The team who wrote this book

This book was produced by a team of specialists from around the world working at the International Technical Support Organization, Poughkeepsie Center.

**Lydia Parziale** is a Project Leader for the ITSO team in Poughkeepsie, New York, with domestic and international experience in technology management including software development, project leadership, and strategic planning. Her areas of expertise include business development and database management technologies. Lydia is a certified PMP and an IBM Certified IT Specialist with an MBA in Technology Management. She has been employed by IBM for over 25 years in various technology areas.

**Edi Lopes Alves** is an IT Specialist with the IBM System z Strategic Outsourcing Delivery team in São Paulo, Brazil. She has over 20 years of experience as a VM system programmer and IBM Content Manager for solutions in the finance area. Edi is a certified z/Series Specialist with a Masters degree in e-business at ESPM in Sao Paulo. She currently supports IBM z/VM and Linux in IBM Global Accounts (IGA) and her area of expertise is System z, z/VM and Linux on System z. Edi has co-authored two Redbooks about Linux system performance and tuning and z/VM basics.

**Susy Heisser** is an IT Specialist in the IBM Development Lab in Böblingen, Germany. She is a member of the infrastructure team of the Firmware Design Support Department. Her areas of expertise include z Hardware, z/VM and z/Linux. Susy joined IBM over 25 years ago working for VSE Software Development in Böblingen. After her assignment to Networking Test in Raleigh, NC, she joined the Hardware Competence Center. For the past nine years she has been responsible for the hardware and software environment for simulation and firmware tools on System z.

**Rob Hunt** is an Accredited IT Specialist with the IBM System z Strategic Outsourcing Team in the United Kingdom. Rob has over 24 years of experience with IBM in storage and systems management, supporting MVS™, z/VM, and z/OS® environments. He has provided IBM mainframe storage and VM support in the government, financial, and insurance sectors. Rob now works in the IBM Systems z server team as a z Series systems programmer and mainframe network team leader, supporting z Series and z/VM based hardware and software for external commercial and IBM internal mainframe accounts.

**Jo Johnston** is a Certified IT Specialist and Chartered Engineer who works in the IBM System z Strategic Outsourcing Team in the United Kingdom. She has worked on IBM mainframe systems as a systems programmer supporting z/VM, z/OS, MVS, CICS®, DB2, and IMS™ for more than 30 years. She joined IBM in 2001 working in Strategic Outsourcing Commercial Platform Support, where she provided day-to-day z/OS, CICS, DB2, and IMS support for customer systems that had been outsourced to IBM. Jo then moved to the IBM System z Technical Sales Team, specializing in WebSphere Business Integration products on System z, with specific responsibility for WebSphere Application Server, WebSphere Process Server, and WebSphere Service Registry and Repository. She now works in the System z Database team, supporting not only DB2 and Adabas on z/OS, but also WebSphere Application Server on z/OS and Tivoli® Storage Productivity Center for Replication on z/OS and z/VM. Jo has co-authored two IBM Redpapers™ documents about WebSphere Process Server on System z and a Redbooks publication about z/VM 6.2 SSI and LGR.

**Volker Masen** is an IT specialist in the IBM Development Lab in Böblingen, Germany. He started his career 20 years ago in the System z environment, supporting the library and build management environment around ISPF/SCLM. After spending several years supporting development environments for the IBM Rational® brand, he moved back into the System z environment several years ago as a system programmer for z/VM in Böblingen and a specialist for other virtualization environments (VMware, KVM). Volker co-authored the Redbooks publication about z/VM 6.2 SSI and LGR.

**Srivatsan Venkatesan** is an IT Specialist in the Systems and Technology Group in IBM USA. He has enthusiastically gained one year of experience in the z/VM and Linux on System z field. He holds a degree in Computer Science from the University of South Carolina. His areas of expertise include Linux and middleware on System z.

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

Roy P. Costa, Robert Haimowitz, David Bennin  
International Technical Support Organization, Poughkeepsie Center

Bill Bitner, Mark Lorenc, John Franciscovich, Emily Kate Hugenbruch, and Xenia Tkatschow  
IBM USA

Tracy Dean  
IBM USA

Oliver Petrik  
IBM Germany

Thanks to the authors of the first book, *An introduction to z/VM Single System Image (SSI) and Live Guest Relocation (LGR)*, SG24-8006:

Anthony Bongiorno, Howard Charter, Jo Johnston, Volker Masen, Clovis Pereira, Sreehari Somasundaran, Srivatsan Venkatesan

## Now you can become a published author, too!

Here's an opportunity to spotlight your skills, grow your career, and become a published author—all at the same time! Join an ITSO residency project and help write a book in your area of expertise, while honing your experience using leading-edge technologies. Your efforts will help to increase product acceptance and customer satisfaction, as you expand your network of technical contacts and relationships. Residencies run from two to six weeks in length, and you can participate either in person or as a remote resident working from your home base.

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

[ibm.com/redbooks/residencies.html](http://ibm.com/redbooks/residencies.html)

## Comments welcome

Your comments are important to us!

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

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

[ibm.com/redbooks](http://ibm.com/redbooks)

- Send your comments in an email to:

[redbooks@us.ibm.com](mailto:redbooks@us.ibm.com)

- Mail your comments to:

IBM Corporation, International Technical Support Organization  
Dept. HYTD Mail Station P099  
2455 South Road  
Poughkeepsie, NY 12601-5400

## Stay connected to IBM Redbooks

- Find us on Facebook:

<http://www.facebook.com/IBMRedbooks>

- Follow us on Twitter:

<http://twitter.com/ibmredbooks>

- Look for us on LinkedIn:

<http://www.linkedin.com/groups?home=&gid=2130806>

- ▶ Explore new Redbooks publications, residencies, and workshops with the IBM Redbooks weekly newsletter:  
<https://www.redbooks.ibm.com/Redbooks.nsf/subscribe?OpenForm>
- ▶ Stay current on recent Redbooks publications with RSS Feeds:  
<http://www.redbooks.ibm.com/rss.html>



## Overview of SSI and LGR

This chapter contains an overview of the IBM Redbooks publication, *An introduction to z/VM Single System Image (SSI) and Live Guest Relocation (LGR)*, SG24-8006, where a description was provided on how to set up the z/VM clusters that were used in this book.

In that book, single system image (SSI) architecture and live guest relocation (LGR) are introduced and explained. An overview of multi-system virtualization with z/VM single system image (VMSSI) features and operations was provided, as well as a description of the difference between a z/VM SSI cluster and a stand-alone non-SSI z/VM system.

We described how to install, define, and set up an SSI cluster with four z/VM members. Also we demonstrated how to convert a non-SSI z/VM system to be a member of an SSI cluster. In our two member cluster, we included setting up IBM RACF®. We explained how the new VMSSI feature interacted with TCP/IP, Dirmaint, Programmable Operator, the IBM Performance Toolkit, RACF, and RSCS.

We provided an introduction to live guest relocation (LGR) and described some of the major attributes of LGR, as well as some of the factors that affected relocation. We identified the supported configurations for relocation and discussed requirements for memory and paging during LGR. We tested the relocation of various Linux guests between members of the SSI cluster and described which factors stopped a guest from being eligible for relocation.

We described the business benefits of SSI and LGR and the fact that it is not a high availability solution, but that it can be used for scheduled maintenance, software upgrades, and load balancing.

In this chapter, we provide an overview of single system image and live guest relocation.

## 1.1 Single system image (SSI)

The z/VM single system image feature (VMSSI) is an optional priced feature that is new with z/VM version 6.2.

A z/VM single system image (SSI) cluster is a multi-system environment on which the z/VM systems can be managed as a single resource pool and guests can be moved from one system to another while they are running. Each SSI member is a z/VM logical partition (LPAR) connected via channel-to-channel (CTC) connections.

A z/VM SSI cluster consists of up to four z/VM systems in an Inter-System Facility for Communications (ISFC) collection. Each z/VM system is a member of the SSI cluster. The cluster is self-managed by the z/VM control program (CP) using ISFC messages that flow across channel-to-channel devices between the members. All members can access shared DASD volumes, the same Ethernet LAN segments, and the same storage area networks (SANs). Figure 1-1 shows a four-member SSI cluster.

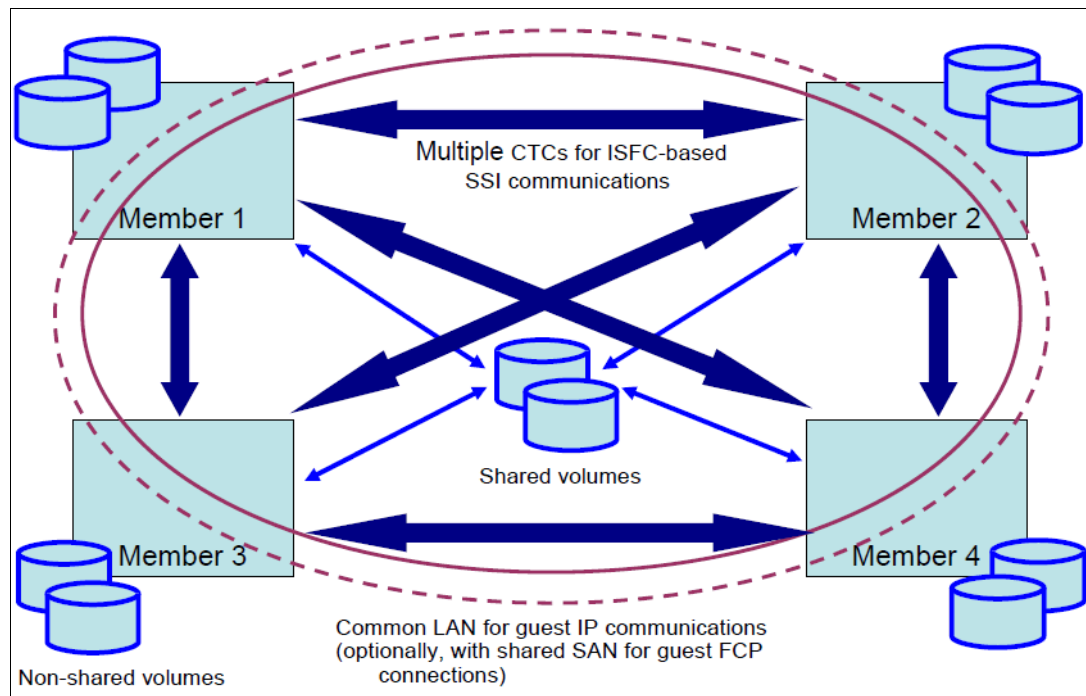


Figure 1-1 An SSI cluster with four members.

The z/VM single system image feature (VMSSI) enhances z/VM systems management, communications, disk management, device mapping, virtual machine definition management, installation, and service functions to enable up to four z/VM systems to share and coordinate resources within an SSI cluster.

In the SSI cluster, the members can be in various states. The overall mode of the cluster is dependent on the states of the individual members. The state of each member and the cluster mode determine the degree of communication, resource sharing, and data sharing among the members.

The different states a cluster can have are shown in Table 1-1 on page 3.

Table 1-1 SSI cluster modes of operation

SSI cluster modes	Description
STABLE	SSI cluster is fully operational
INFLUX	Joining or Leaving is in progress <ul style="list-style-type: none"> <li>• Cross-system functions are temporarily suspended</li> <li>• Negotiations for shared resources are deferred</li> <li>• Existing accesses are unaffected</li> </ul>
SAFE	<ul style="list-style-type: none"> <li>• A remote member's state cannot be determined</li> <li>• Any member is in Suspended state</li> <li>• Attempts to access shared resources fail</li> <li>• Existing accesses are unaffected</li> </ul>

For further details, refer to the manual *z/VM CP Planning and Administration* SC24-6178 and also to the IBM Redbooks publication, *An introduction to z/VM Single System Image (SSI) and Live Guest Relocation (LGR)*, SG24-8006, chapters 2, 4 and 5.

### 1.1.1 SSI commands

A number of new commands introduced with z/VM v6.2 allow you to query the state of your cluster. Table 1-2 provides a summary of these commands. For more information refer to the manual *z/VM V6R2.0 CP Commands and Utilities Reference*, SC24-6175.

Table 1-2 CP commands for SSI

Command	Function
SET SSI	Add or change existing entries in the SSI member list.
QUERY SSI	Displays the single system image (SSI) name, member status, and connectivity status.
AT -sysname- CMD	Use the AT command to remotely issue commands on active member systems in an SSI cluster.
QUERY -rdev- ID	Displays the device and control unit information from the sense ID data for a specified device address if they are known. It also displays the device equivalency ID (EQID) if one exists for the device.
SET -rdev- <NoEQID I EQID eqid> TYPE -type-	EQid eqid assigns the device equivalency ID (EQID) to the RDEV. The eqid is a string of 1-8 alphanumeric characters. Note that for CTCA, FCP, HiperSocket, and OSA devices, this EQID must be unique or be shared only by other devices of the same type. NOEQid removes a previously assigned EQID from this RDEV and reverts back to a system-generated EQID. If no EQID was previously assigned by a user, no action takes place.
QUERY user AT ALL	Command has been updated to display users logged on other cluster members.
VMRELOCATE	Moves an eligible, running z/VM virtual machine transparently from one z/VM system to another within an SSI cluster, and monitors and cancels virtual machine relocations that are already in progress.
DEFINE RELODOMAIN	Define or update an SSI relocation domain.
QUERY RELODOMAIN	List the members of one or more relocation domains.
SET VMRELOCATE	Dynamically control the relocation domain for a user.

### 1.1.2 SSI mapping of real devices using equivalency identifiers (EQIDs)

Real device (RDEV) mapping provides a means of identifying a device either by a CP-generated EQID or by a customer-generated EQID. This mapping is used to ensure virtual machines relocated by live guest relocation continue to use the same or equivalent devices following a relocation.

An administrator-assigned EQID must be unique or be shared only by other devices of the same type and with the same access rights.

Use the command **query eqid** to display the device EQIDs for a specific RDEV and to display the RDEVs associated with a specific EQID.

### 1.1.3 SSI coordinated MAC addresses

The assignment of MAC addresses to network interface cards (NICs) is coordinated across the SSI cluster so that even if a Linux on System z guest relocates across the cluster, they are accessible without any disruption of operations. SSI does not allow any member of the SSI cluster to have a MAC address that is already in use by another Linux guest within the cluster.

The assignment of MAC addresses extends the z/VM Ethernet virtual switch (VSWITCH) logic to coordinate its automatic MAC address assignment with all active members of an SSI cluster. Each system within the SSI cluster must have connectivity to the same physical and virtual LAN segment.

This requires the user to physically configure a global VSWITCH across the single system image. Spanning a VSWITCH across all members of an SSI cluster allows live guest relocation to migrate a virtual machine's network to any system within the cluster.

A global VSWITCH provides identical network connectivity across all active members within a single system image cluster. This is achieved by defining a VSWITCH with the same name across each z/VM image within the cluster. Each defined VSWITCH must also have one or more physical Open Systems Adapter (OSA) ports connected to the same physical LAN segment. Real OSA ports provide the connectivity necessary to access the virtual guest ports on each z/VM image.

#### Summary

- ▶ Assignment of MAC addresses by the control program (CP) is coordinated across an SSI cluster.
  - This ensures that new MAC addresses are not being used by any other member.
  - Guest relocation moves a MAC address to another member.
- ▶ Each member of a cluster should have identical network connectivity.
  - Virtual switches with same name defined on each member.
  - Same (named) virtual switches on different members should have physical OSA ports connected to the same physical LAN segment, assured by EQID assignments.

### 1.1.4 SSI use of ISFC links

Each member in the SSI cluster must have a direct ISFC connection to every other member in the SSI cluster. In other words, SSI traffic from one member to another never flows through an intermediate member. Each ISFC connection from one member to another is called an *ISFC logical link*, or simply a *logical link*. A logical link is composed of 1 - 16 channel-to-channel

(CTC) devices. Faster CTC speeds increase throughput and result in shorter relocations. There is always exactly one ISFC logical link between two members.

The ISFC SSI infrastructure provides tools that can be used for cross-system communication. This enhances the ISFC subsystem to improve the transport mechanism and provide convenient interfaces for exploitation by other subsystems with the CP nucleus.

ISCF links can be added and removed dynamically, depending on the throughput you need in your cluster, using the **activate** or **deactivate islink** commands (see Example 1-1). The **activate islink** command identifies a communication link to ISFC.

---

*Example 1-1 ISLINK commands*

---

```
activate islink 4050 node ITS0SSI5
deactivate islink 4050 node ITS0SSI5
```

---

For maximum throughput, when you are setting up your network, follow the guidelines for planning your network in an SSI cluster. The guidelines are in chapter 2 of the IBM Redbooks publication, *An introduction to z/VM Single System Image (SSI) and Live Guest Relocation (LGR)*, SG24-8006.

### 1.1.5 SSI members

Each instance of z/VM could be a member of a single system image cluster. System management of z/VM is simplified when z/VM instances are members of an SSI cluster because they can be serviced and administered as one system. You can also coordinate the joining and leaving of the cluster members and sharing of the cluster resources.

The SSI modes and member states are described fully in the manual *z/VM CP Planning and Administration version 6 release 2*, SC24-6178.

The information about the state of each member of the SSI cluster is held in the persistent data record (PDR), which is located on the shared common disk. This record contains a heartbeat mechanism, which ensures that a stalled or stopped member can be detected.

The following are events that can cause changes in the SSI member state, mode, or both:

- ▶ IPL
  - PDR initialization
  - Initial state and mode set to Down and Safe (local member only)
- ▶ Start and completion of join processing
- ▶ Changes in connectivity between any members
- ▶ Failure of a member to update its heartbeat
- ▶ State change notification from another member
- ▶ Shutdown or abend
- ▶ Set SSI down command

## 1.2 Changes for SSI in the USER Directory

In this section we describe new definitions in the z/VM directory for guests with single configuration and multiple configurations.

## 1.2.1 Single-configuration virtual machine definition

A single-configuration virtual machine definition consists of a user entry and any included profile entry. Only one virtual machine instance can be created from a single-configuration virtual machine definition. For example, you can specify a USER1 single-configuration virtual machine and log on to a z/VM system as USER1. In an SSI cluster, the virtual machine can be logged on to only one SSI member at a time. Your Linux guests are always defined as single users.

## 1.2.2 Multi-configuration virtual machine definition

A multi-configuration virtual machine definition consists of an identity entry, any included profile entry, and all associated subconfiguration entries. In an SSI-enabled source directory, this virtual machine definition allows multiple virtual machine instances to be defined, which enables the userid to be logged on concurrently to multiple members of the SSI cluster. Each of these virtual machine instances can have a different configuration from the others.

For example, you can define a MAINT multi-configuration virtual machine and concurrently log on to all the members of an SSI cluster as MAINT.

Figure 1-2 shows the definitions for a traditional USER called single-configuration virtual machine and shows the new type of user called IDENTITY or multi-configuration virtual machine.

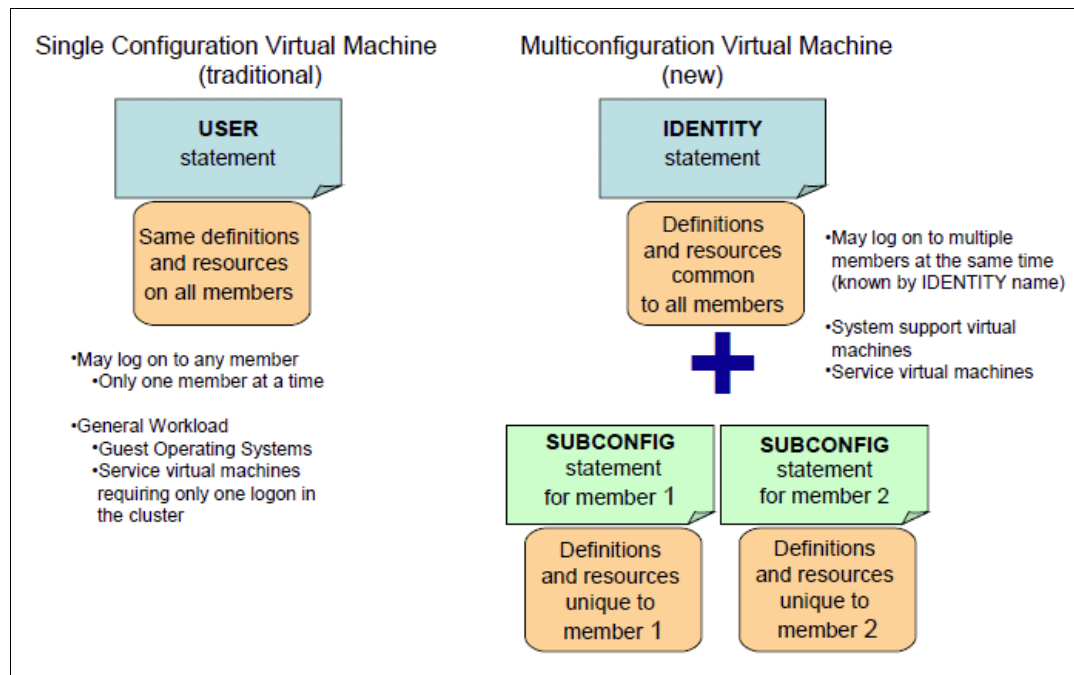


Figure 1-2 Single- versus multi-configuration definition

**Further explanation:** **IDENTITY** is used for system support RACF, RSCS, TCP/IP, PerfKit, PVM, MAINT, and so on. **USER** is for workload in general (for example, Linux guests and general CMS users).

### 1.2.3 DIRMAINT commands

IBM Directory Maintenance (DirMaint™) for z/VM is a CMS application that helps manage an installation's z/VM directory.

Table 1-3 contains a list of commands to the DIRMAINT machine that are new for SSI and IDENTITY.

Table 1-3 DIRMAINT commands for SSI

Command	Function
DIRM SSI	SSI operand prepares a source directory to be used on a nod in an SSI cluster.
DIMR UNDOSSI	The UNDOSSI operand rolls back the BUILD statement changes done by the SSI operand and removes the SSI operand from the DIRECTORY statement.
DIRM VMRELOCATE	The VMRELOCATE operand queries, updates, or deletes the relocation capability associated with a user or profile entry.
DIRM ADD subconfig BUILD ON member IN identity	The ADD operand has been updated for cloning the SUBCONFIG entries. This is described in more detail in <i>An introduction to z/VM Single System Image (SSI) and Live Guest Relocation (LGR)</i> , SG24-8006, section B.2.1 “Adding Identities.”

## 1.3 Live guest relocation (LGR)

With the IBM z/VM single system image, a running Linux on System z virtual machine can be relocated from one member system to any other, a process known as live guest relocation (LGR). LGR occurs without disruption to the business. It provides application continuity across planned z/VM and hardware outages and flexible workload balancing that allows work to be moved to available system resources.

**Operating systems supported:** Linux on System z is currently the only guest environment supported for relocation.

There are several reasons why you might need to relocate a running virtual server, such as:

- ▶ Maintenance of hardware or software
- ▶ Fixing performance problems
- ▶ Workload rebalancing

Relocating virtual servers can be useful for load balancing and for moving workload off of a physical server or member system that requires maintenance. After maintenance is applied to a member, guests can be relocated back to that member, thereby allowing you to maintain z/VM as well as keeping your Linux on System z virtual servers available.

Other approaches, such as Tivoli System Automation (TSA) in conjunction with application clustering techniques, offer availability, addressing unplanned outages as well. However, these require the customer to invest in substantial setup and customization, for example, preparing scripts to orchestrate TSA recovery actions. Moreover, due to the availability characteristics of System z and z/VM, LGR allows applications to remain available over planned outages with less impact to the application and less customer setup required.

In general, a guest can be relocatable when:

- ▶ It is a Linux on System z guest.
- ▶ It has enough resources available on the target system, such as memory, cpu, and so on.
- ▶ It has the same networking definition, for example VLAN, VSWITCH.
- ▶ It is disconnected and accessible when the guest is being relocated.
- ▶ It has the same EQIDs defined for the devices shared by the SSI members for relocation.

### **Before you relocate a guest**

There are some requirements for the SSI members regarding resources such as disks, memory, and networkings.

LGR is more fully described in chapter 3, “Live guest relocation (LGR) overview” in the IBM Redbooks publication, *An introduction to z/VM Single System Image (SSI) and Live Guest Relocation (LGR)*, SG24-8006.



## Lab environment

We set up two z/VM SSI clusters. One of them was a four member SSI cluster, the other was a two member SSI cluster. In this chapter, we provide a brief overview of both clusters. For a more detailed description of the clusters, see the IBM Redbooks publication, *An introduction to z/VM Single System Image (SSI) and Live Guest Relocation (LGR)*, SG24-8006.

## 2.1 Overview of our four member cluster

The four member SSI cluster is named ITSOSSIA. It consists of members ITSOSI1, ITSOSI2, ITSOSI3, and ITSOSI4, as shown in Figure 2-1.

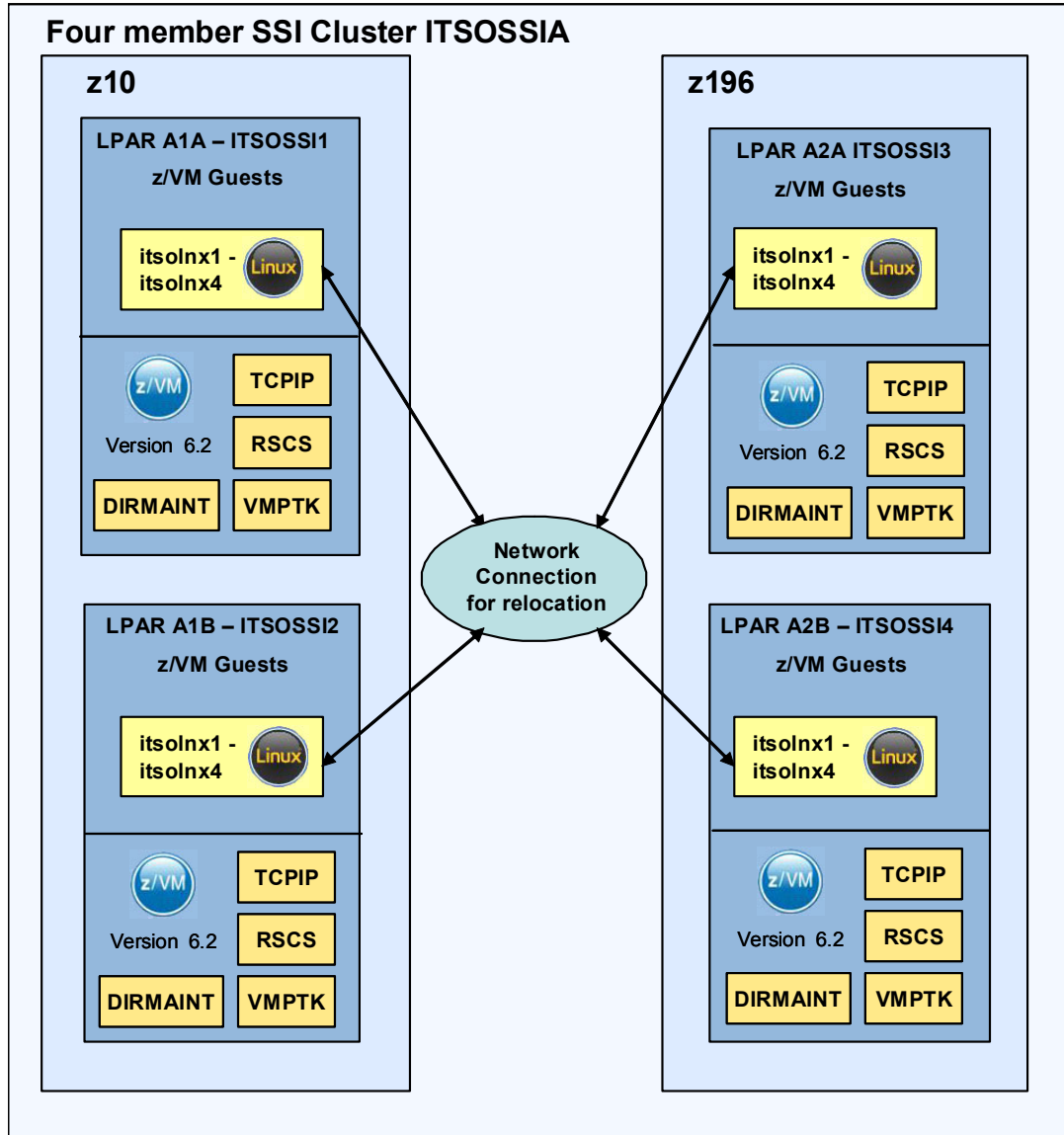


Figure 2-1 Our four member cluster

### Hardware definition

We used an IBM System z z10™ Enterprise Class machine with two LPARs, A1A and A1B, which hosted the two z/VM members ITSOSI1 and ITSOSI2. An IBM zEnterprise® 196 machine provided two more LPARs, A2A and A2B, which hosted the two z/VM members ITSOSI3 and ITSOSI4.

## Software

The following software products were customized for use:

- ▶ TCP/IP
- ▶ Remote Spooling Communication Subsystem (RSCS)
- ▶ Directory Manager (DIRMAINT)
- ▶ IBM Performance Toolkit for z/VM (VMPTK)

## Network configuration

The network configuration of our lab environment was as follows:

- ▶ Two CTCs between each LPAR using inter-system facility for communications (ISFC).
- ▶ One CTC for an RSCS connection between the four LPARs.
- ▶ One CTC from each LPAR to an external ITSO RSCS system.
- ▶ Each LPAR has a set of OSA cards connected to the same network.

## TCP/IP configuration

We provide the most important TCP/IP information for each of the z/VM members in Table 2-1.

*Table 2-1 z/VM TCP/IP setup for our four member cluster*

Hostname	ITSOSSI1	ITSOSSI2	ITSOSSI3	ITSOSSI4
Domain Name	itso.ibm.com	itso.ibm.com	itso.ibm.com	itso.ibm.com
IP address	9.12.4.232	9.12.4.233	9.12.4.234	9.12.4.235
Memory	8 GB	8 GB	8 GB	8 GB

## Linux on System z guests

Table 2-2 lists information about the z/VM guests that are candidates for relocation.

*Table 2-2 Linux guest information about our four member cluster*

Hostname	ITSOLNX1	ITSOLNX2	ITSOLNX3	ITSOLNX4
Domain Name	itso.ibm.com	itso.ibm.com	itso.ibm.com	itso.ibm.com
IP address	9.12.4.141	9.12.4.140	9.12.4.228	9.12.4.229
Memory	4 GB	4 GB	6 GB	6 GB
OS	RHEL 5	SLES 11	SLES 11	RHEL 5

## 2.2 Overview of our two member cluster

We named the two member SSI cluster ITSOSI5. It consists of members ITSOSI5 and ITSOSI6 as shown in Figure 2-2.

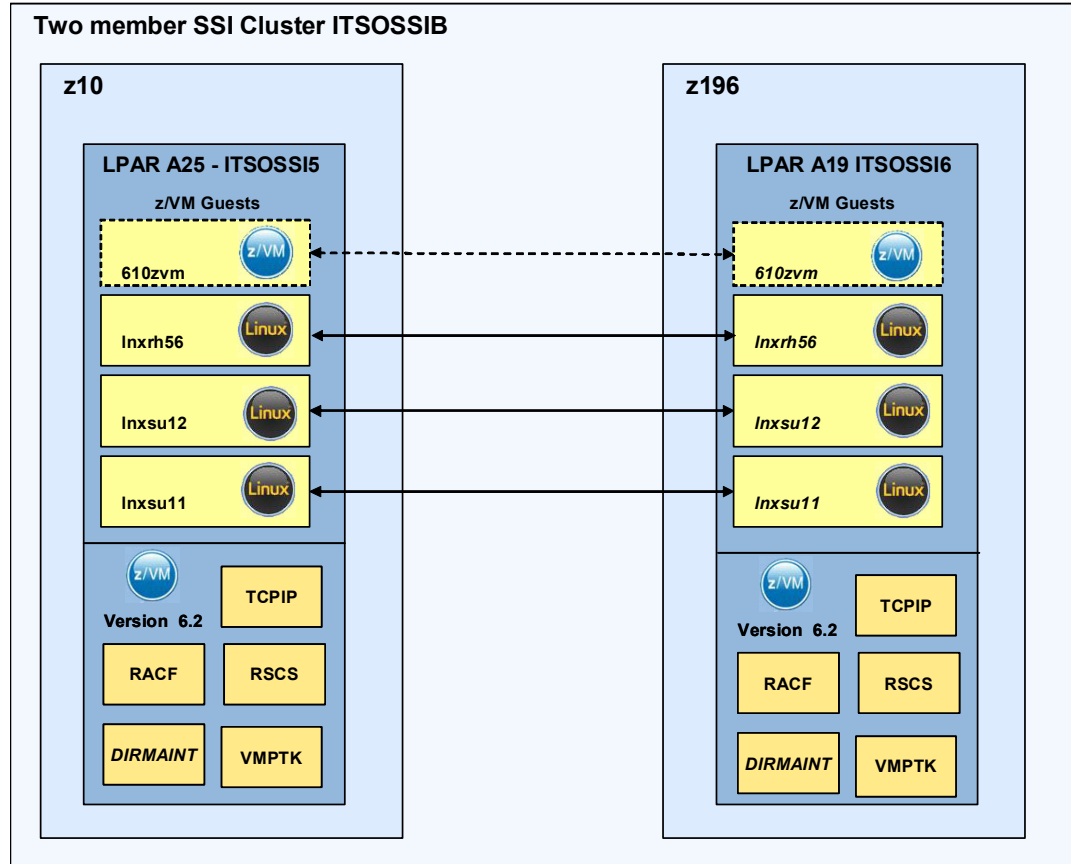


Figure 2-2 Our two member cluster

### Hardware definition

We used an IBM System z z10 Enterprise Class machine with an LPAR that we named A25, which hosted the z/VM member ITSOSI5. An IBM zEnterprise 196 had an LPAR named A19, which hosted the z/VM member ITSOSI6.

### Software

The following software was customized for use:

- ▶ TCP/IP
- ▶ Remote Spooling Communication Subsystem (RSCS)
- ▶ Directory Manager (DIRMAINT)
- ▶ Performance Toolkit for z/VM (VMPTK)
- ▶ IBM RACF Security Manager (RACF) - The RACF database is shared between both members.

## Network configuration

The network configuration of our lab environment was as follows:

- ▶ Two CTCs between each LPAR using inter-system facility for communications (ISFC).
- ▶ One CTC for an RSCS connection between the two LPARs.
- ▶ One CTC from each LPAR to an external ITSO RSCS system.
- ▶ Each LPAR has a set of OSA cards connected to the same network.

## TCP/IP configuration

The most important TCP/IP information for each z/VM member is shown in Table 2-3.

*Table 2-3 z/VM TCP/IP setup for our two member cluster*

Hostname	ITSOSSI5	ITSOSSI6
Domain Name	itso.ibm.com	itso.ibm.com
IP address	9.12.4.236	9.12.4.237
Memory	8 GB	8 GB

## Linux guests

Table 2-4 lists information about the z/VM guests that are candidates for relocation.

*Table 2-4 Linux guest information about our two member cluster*

Hostname	Inxsu11	Inxsu12	Inxrh56
Domain Name	itso.ibm.com	itso.ibm.com	itso.ibm.com
IP address	9.12.5.100	9.12.4.225	9.12.4.226
Memory	6 GB	6 GB	6 GB
OS	SLES 11	SLES 11	RHEL 56
Additional applications	SAP Application Server Central Instance	-	-

**Note:** The guests 610zvm and Inxsu12, shown in Figure 2-2 on page 12, were prepared for *An introduction to z/VM Single System Image (SSI) and Live Guest Relocation (LGR)*, SG24-8006 and are described in this chapter; however, we did not use them in the relocation tests for this book.





## Applications setup

This chapter describes the applications that we ran on the Linux on System z guests in our SSI clusters to produce a load on the system when we were relocating the Linux guests using LGR.

We installed a stocks and shares benchmarking application that runs on WebSphere Application Server and uses DB2 to store its data. We also describe the SAP and DB2 system that was used in the IBM Redbooks publication, *An introduction to z/VM Single System Image (SSI) and Live Guest Relocation (LGR)*, SG24-8006.

### 3.1 SAP and DB2 on System z

We have an SAP Application Server in a typical user environment on System z. This means, the IBM DB2 database server runs in a z/OS environment and the SAP Application Server Central Instance (CI) is located on a Linux on System z server. The SAP Application Server CI is running on z/VM system ITSOSI5 and the SAP Application Server Dialog Instance (DI) is running on an IBM AIX® system, as shown in Figure 3-1.

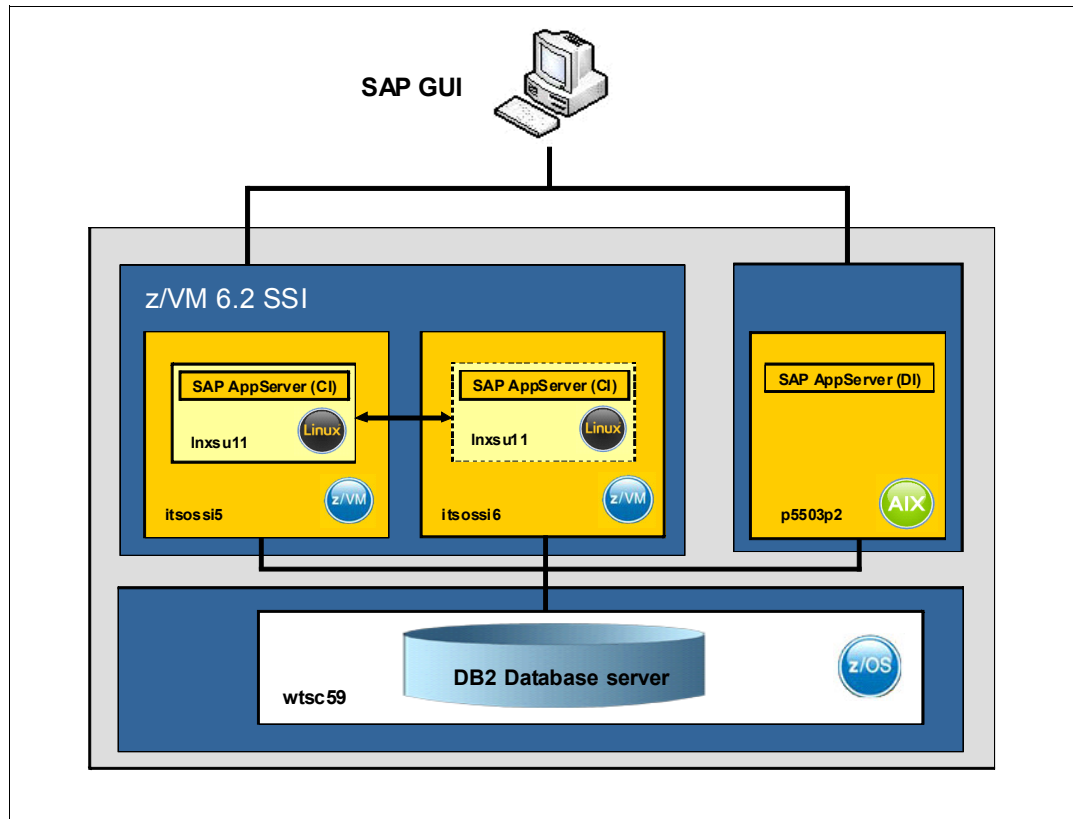


Figure 3-1 SAP solution on a System z with the relocation feature of z/VM 6.2

This system was described in detail in the IBM Redbooks publication, *An introduction to z/VM Single System Image (SSI) and Live Guest Relocation (LGR)*, SG24-8006. The Linux on System z guest running in our SSI cluster with two members is LNXSU11. Example 3-1 shows the z/VM directory entry for LNXSU11.

#### Example 3-1 Directory Entry from LNXSU11

```
USER LNXSU11 W76WD23E 6G 16G G
CPU 01
CPU 00
CPU 02
CPU 03
IPL 202
MACHINE ESA 4
OPTION CHPIDV ONE
CONSOLE 0009 3215 T
NICDEF C200 TYPE QDIO LAN SYSTEM VSWITCH1
NICDEF 6300 TYPE QDIO LAN SYSTEM VSW999 DEVICES 3
NICDEF 6303 TYPE QDIO LAN SYSTEM VSW199 DEVICES 3
```

```
NICDEF 7000 TYPE QDIO LAN SYSTEM VSW199 DEVICES 3
SPOOL 000C 3505 A
SPOOL 000D 3525 A
SPOOL 000E 1403 A
MDISK 0201 3390 1 10016 LX9980 MR
MDISK 0202 3390 1 10016 LX9981 MR
MDISK 0301 3390 1 10016 LX9982 MR
MDISK 0302 3390 1 10016 LX9983 MR
MDISK 0303 3390 1 10016 LX9984 MR
MDISK 0304 3390 1 10016 LX9985 MR
MDISK 0305 3390 1 10016 LX9986 MR
MDISK 0306 3390 1 10016 LX9987 MR
MDISK 0307 3390 1 10016 LX9988 MR
MDISK 0308 3390 1 10016 LX9989 MR
MDISK 0309 3390 1 10016 LX998A MR
```

---

We used LGR to relocate this guest between SSI cluster members ITSOSI5 and ITSOSI6.

## 3.2 IBM Trade Performance Benchmark

The IBM Trade Performance Benchmark sample, also known as the “Trade6” application, is a sample WebSphere end-to-end benchmark and performance sample application. This benchmark, designed and developed to cover the WebSphere programming model, provides a real world workload, driving WebSphere's implementation of J2EE 1.4 and web services, including key WebSphere performance components and features. Trade 6 simulates a stock trading application that allows you to buy and sell stock, check your portfolio, register as a new user, and so on. We used Trade 6 to generate workloads that we can analyze in terms of their impact on system performance. You can download the IBM Trade Performance Benchmark sample for WebSphere Application Server at no charge, after logging in, from the following website:

<https://www14.software.ibm.com/webapp/iwm/web/preLogin.do?source=trade6>

To run Trade6, you need the following:

- SUSE Linux Enterprise Server 8 SP 3, 2.4.21
- IBM DB2 UDB V8.2
- IBM WebSphere Application Server V6.0
- Rational Performance Tester V6.1

However, we installed it on the following software and it ran successfully:

- Red Hat Enterprise Linux Server release 5.6
- IBM DB2 UDB V10.1
- IBM WebSphere Application Server V7.0

We installed DB2 v10 following the standard installation instructions for Linux on System z. Before we started the installation, we disabled SELinux using the command:

```
echo 0 > /selinux/enforce
```

We enabled it again when the installation was completed using the command:

```
echo 1 > /selinux/enforce
```

We created the sample DB2 database to verify that DB2 had been successfully installed.

We installed the WebSphere Application Server following the standard installation instructions for Linux on System z. We selected a standalone application server and chose not to enable security or install the sample applications.

Trade6 can be used as a benchmark application to put a load on a Linux guest so that performance measurements can be taken. The Trade6 database was repopulated with base data before each test so that comparisons could be made between the tests that we ran.

### 3.3 Workload on non SCSI disks

The Linux on System z guest, LNXRH56, is running on Red Hat Enterprise Linux Server release 5.6. The setup of this guest is described in the IBM Redbooks publication *An introduction to z/VM Single System Image (SSI) and Live Guest Relocation (LGR)*, SG24-8006. We installed DB2 version 10.1 and WebSphere Application Server (WAS) version 7.0.0 with the Trade6 application running on the WebSphere Application Server, as shown in Figure 3-2.

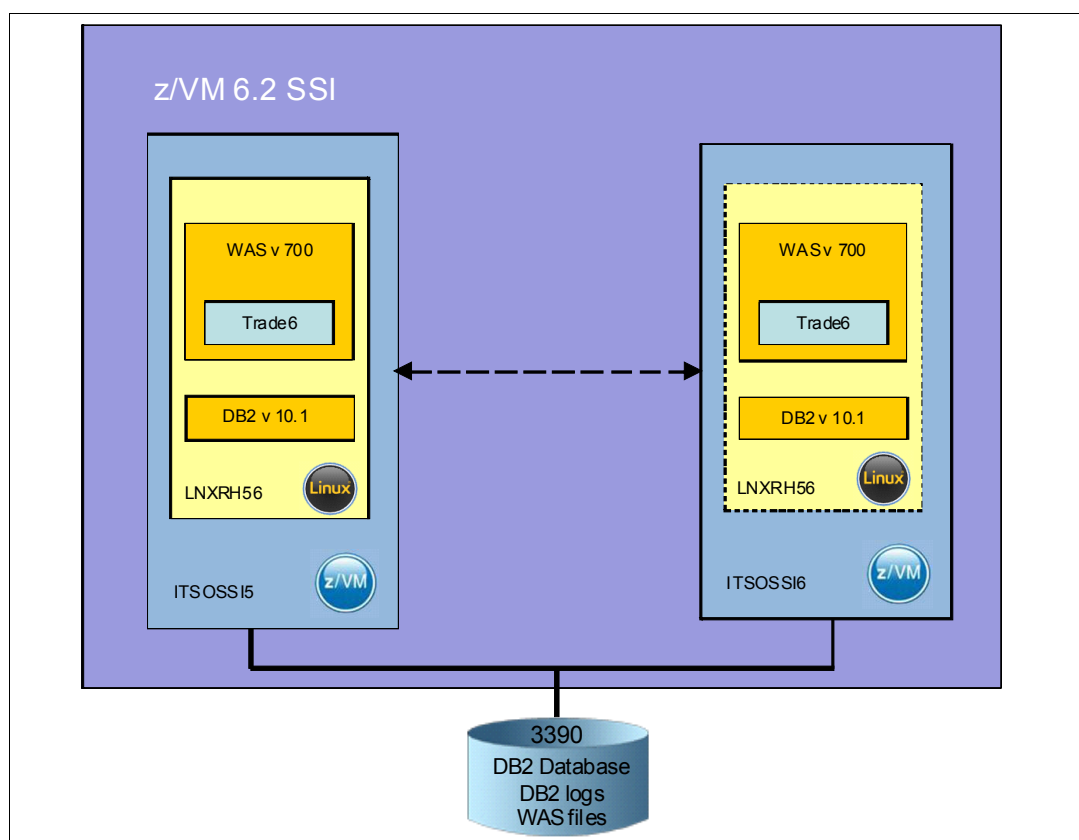


Figure 3-2 LNXRH56 z/VM guest with WebSphere Application Server, DB2, and Trade6

DB2 version 10.1 and WebSphere Application Server version 7.0.0 software were installed on minidisk 203 defined on a 3390 model 27 and attached to the z/VM guest LNXRH56. Example 3-2 on page 19 shows the directory entry for LNXRH56.

*Example 3-2 Directory entry for LNXRH56*

---

```
USER LNXRH56 HCHT57UI 6G 40G G
  INCLUDE LINDFLT
  IUCV ALLOW
  IUCV ANY
  MACH ESA 4
  OPTION CHPIDV ONE
  POSIXINFO UID 59
  NICDEF C200 TYPE QDIO LAN SYSTEM VSWITCH1
  MDISK 0201 3390 1 1000 LX9A29 MR
  MDISK 0202 3390 1001 9016 LX9A29 MR
  MDISK 0203 3390 1 15000 LX6030 MR
```

---

We used a workload generator to automatically run transactions on Trade6 and to put a load on the Linux guest when it was relocated using LGR. We relocated LNXRH56 between the z/VM systems ITSOSI5 and ITSOSI6, which are members of the SSI cluster ITSOSIB.

### 3.4 Workload with DB2 logs on SCSI disks

We used the Linux on System z guests, ITSOLNX3 and ITSOLNX4, that were set up for the IBM Redbooks publication, *An introduction to z/VM Single System Image (SSI) and Live Guest Relocation (LGR)*, SG24-8006. We installed DB2 version 10.1 on ITSOLNX4 and then installed WebSphere Application Server version 7.0.0 (WAS1) with the Trade6 application on ITSOLNX3, as shown in Figure 3-3. The dotted lines between the z/VM systems indicate the ability to relocate the Linux guests to any of the z/VM members in the SSI cluster. The DASD are accessible from any of the z/VM members.

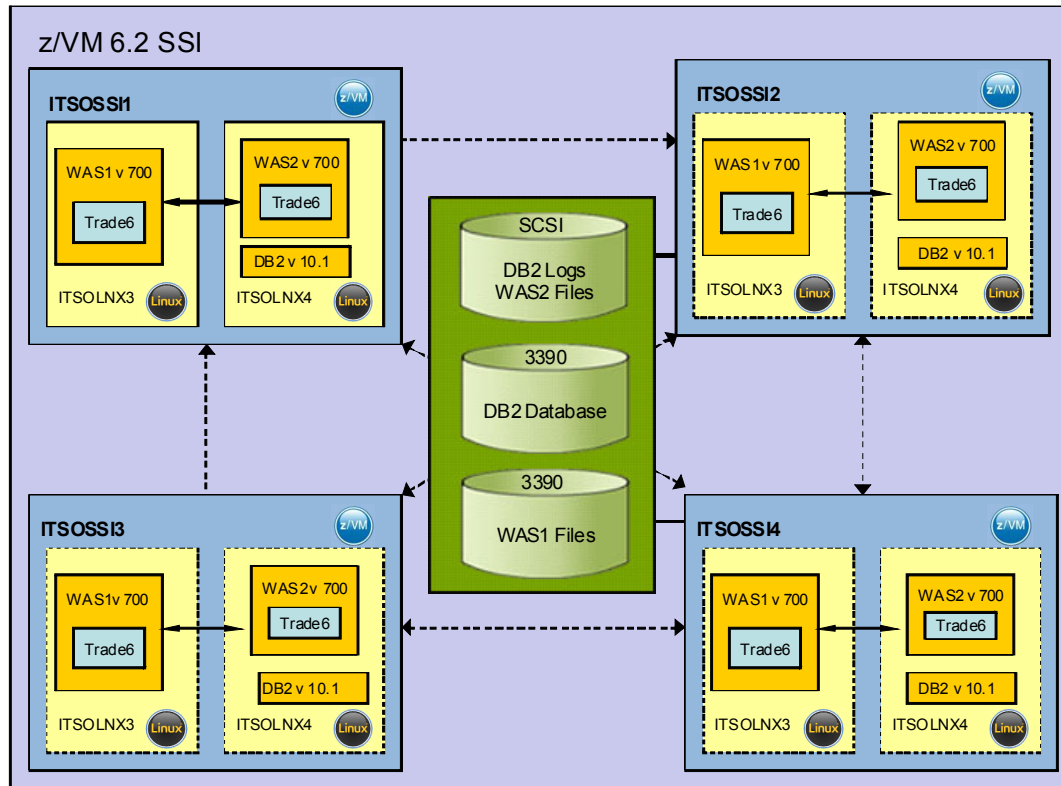


Figure 3-3 ITSOLNX3 and ITSOLNX4 Linux guests with WAS, DB2, and Trade6

We installed the DB2 Version 10.1 software on a minidisk defined on a 3390 model 27 attached to the Linux guest ITSOLNX4. We moved the DB2 logs on to the SCSI disks attached to ITSOLNX4 by using the command:

```
db2 update db config for trade6db using newlogpath DB2LOGS
```

DB2LOGS was the directory where the SCSI disks were mounted. In Chapter 4, “Using SCSI for file systems” on page 23 we describe how we set up the SCSI disks.

Example 3-3 shows the directory entry for ITSOLNX3.

*Example 3-3 Directory entry for ITSOLNX3*

---

```
USER ITSOLNX3 ITSOSI 6G 40G G
  INCLUDE LINDFLT
  IPL 202
  MACH ESA 2
  OPTION APPLMON LNKNOPAS
  MDISK 0201 3390 1 1000 LX9B25 MR
  MDISK 0202 3390 1001 9016 LX9B25 MR
```

---

Example 3-4 shows the directory entry for ITSOLNX4.

*Example 3-4 Directory entry for ITSOLNX4*

---

```
USER ITSOLNX4 ITSOSI 6G 32G G
  INCLUDE LINDFLT
  IPL 202
  MACH ESA 2
```

---

```

OPTION APPLMON LNKNOPAS
DEDICATE B800 B800 NOQIOASSIST
DEDICATE B900 B900 NOQIOASSIST
MDISK 0201 3390 0001 1000 LX9B26 MR
MDISK 0202 3390 1001 9016 LX9B26 MR
MDISK 0203 3390 3867 10000 LX6032

```

---

We installed the WebSphere Application Server version 7.0 software on a minidisk defined on a 3390 model 27 attached to the Linux guest ITSOLNX3 and defined a standalone application server. We installed the Trade6 application in this standalone application server on ITSOLNX3 and defined the database in DB2 on ITSOLNX4. We used a workload generator to put a load on the Trade6 application while the Linux guests were relocated to other members of the SSI cluster using LGR. We relocated ITSOLNX3 and ITSOLNX4 between ITSOSI1, ITSOSI2, ITSOSI3, and ITSOSI4, which are members of the SSI cluster ITSOSIA.

### 3.5 Workload with DB2 logs and WebSphere Application Server on SCSI disks

We installed another instance of the WebSphere Application Server (WAS2) on the ITSOLNX4 Linux guest, putting the software and WebSphere Application Server logs on the SCSI attached disks. See Figure 3-4 on page 21.

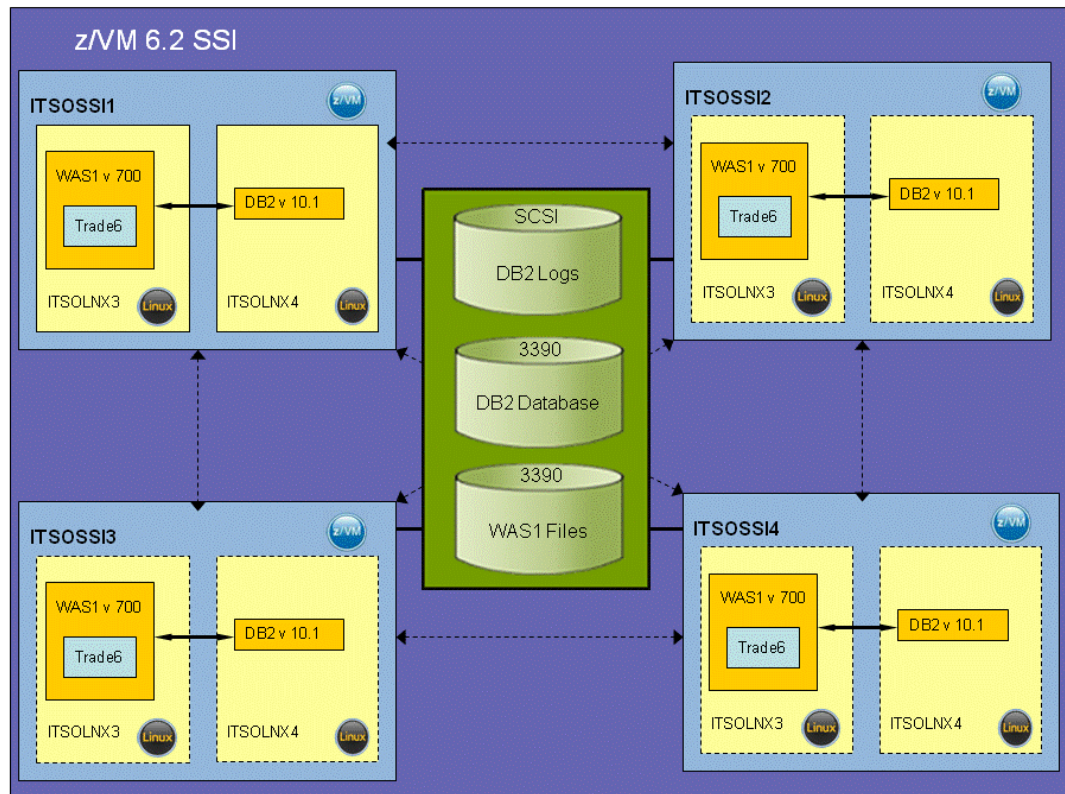


Figure 3-4 ITSOLNX3 and ITSOLNX4 Linux guests with DB2, Trade6, and WebSphere Application Server systems on both guests

The reason for installing WebSphere Application server on both ITSOLNX3 and ITSOLNX4 was to review the performance differences between relocating a Linux guest with WebSphere Application Server on a different guest from DB2 and relocating a Linux guest with WebSphere Application Server installed on SCSI disks.



## Using SCSI for file systems

In this chapter, we explain how to add a Small Computer System Interface (SCSI) file system to an existing Linux on System z system. We show the dependencies that the SCSI setup has in an SSI cluster and what must be defined to get LGR working. For this example, we assume that Linux is running in a z/VM v6.2 SSI cluster and that it is eligible for relocation.

## 4.1 SCSI lab setup

In our lab environment, the setup is a four member SSI cluster with two z/VM LPARs that are on a System z z10 machine and two z/VM LPARs that are on a System z z196 machine. We have two Fibre Channel Protocol (FCP) channels from each machine connected to the two switches. These are connected to the same DS8300. See Figure 4-1.

In this chapter, the following topics are discussed:

- ▶ SCSI example setup
- ▶ Special setup in an SSI cluster
- ▶ Hardware prerequisites
- ▶ Determining the dworldwide port name (WWPN) on System z hardware
- ▶ Defining a estorage area network (SAN) volume in the storage controller
- ▶ Defining the FCP channels to z/VM
- ▶ Defining the SCSI file system to the Linux on System z guest

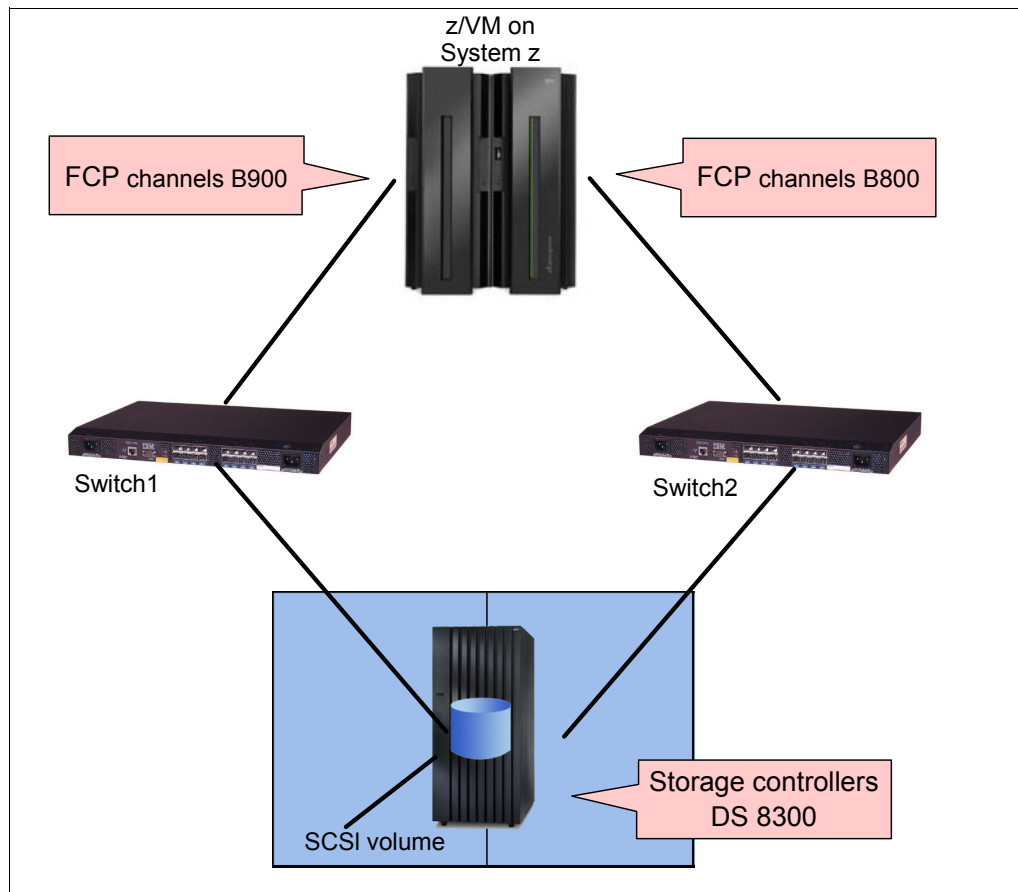


Figure 4-1 Basic lab environment

## 4.2 Special setup in an SSI cluster

In an SSI cluster, define your FCP connection from each member of your SSI cluster separately. For example, if each z/VM is installed in different LPARs or even on different

hardware, you will need to be able to reach your SAN (storage area network) volume in the DS8300 from each LPAR or machine. If you perform a live guest relocation of the Linux on System z guest that has mounted the file system and that corresponds to the SAN volume from one z/VM member to another, your application should not be interrupted.

In our example, the FCP channels are shown as B800 and B900, which are reachable in one LPAR via channel paths 7C and 7D, and in the other LPAR through channel paths 78 and 79. During the live guest relocation (using the **vmrelocate** command), the Channel Path Identifiers (CHPIDs) 7C and 7D in LPAR ITSOSI4 are detached, and on the LPAR ITSOSI1, CHPIDs 78 and 79 are attached to the Linux guest, without manual intervention. See Example 4-1 through Example 4-4.

---

*Example 4-1 FCP channels attached to ITSOLNX4*

---

```
q fcp
FCP B800 ATTACHED TO ITSOLNX4 B800 CHPID 7C
    WWPN C05076ECEA8014F0
FCP B900 ATTACHED TO ITSOLNX4 B900 CHPID 7D
    WWPN C05076ECEA801930
Ready; T=0.01/0.01 10:31:07
vmrelocate move itsolnx4 itsossi1
Relocation of ITSOLNX4 from ITSOSI4 to ITSOSI1 started
User ITSOLNX4 has been relocated from ITSOSI4 to ITSOSI1
Ready; T=0.01/0.01 10:32:49
```

---

After the Linux guest is relocated, the FCP channels on ITSOSI4 are detached. Example 4-2 shows that they are no longer active.

---

*Example 4-2 No FCP channels are active anymore*

---

```
q fcp
An active FCP was not found.
Ready; T=0.01/0.01 10:34:19
```

---

Before the relocation, the system ITSOSI1 has no active FCP channels.

---

*Example 4-3 No FCP channels to begin with*

---

```
q fcp
An active FCP was not found.
Ready; T=0.01/0.01 11:42:46
```

---

During the relocation, the FCP devices B800 and B900 are set online and attached to ITSOLNX4. They are accessed through different paths on system ITSOSI4. Example 4-4 shows that they now have different WWPNs.

---

*Example 4-4 FCP channels in ITSOSI1 after vmrelocate*

---

```
q fcp
FCP B800 ATTACHED TO ITSOLNX4 B900 CHPID 78
    WWPN C05076F77A001430
FCP B900 ATTACHED TO ITSOLNX4 B800 CHPID 79
    WWPN C05076F77A001540
Ready; T=0.01/0.01 11:43:00
```

---

## 4.3 Hardware prerequisites

To define a SCSI volume in a Linux system running under z/VM, you must implement certain hardware checks and definitions:

- ▶ FCP channels have hard-coded WWPNs in the FCP card, which are installed in the z hardware, and the WWPNs are different for each LPAR.
- ▶ The actual path from your z machine to the storage controller must be defined. This includes definitions in the IOCDs (Input/Output Configuration Data Set) of the machine and the switches.
- ▶ In your storage controller, you must define a volume, which is represented by a Logical Unit Number (LUN) and a WWPN. You must define a mapping for the volume WWPN to each of your FCP WWPNs, over which you want to reach the volume.

**Hint:** For redundancy reasons, it would be best to have two FCP channels over two different switches to each control unit of the DS8300. Thus if any hardware fails, there is an alternative path to reach your volume.

## 4.4 Understanding the WWPNs on the z Hardware

Because the WWPNs of the FCP adapters are hard coded, you must use the WWPNs that are supplied with the FCP card in the machine.

### 4.4.1 WWPNs used in our setup

In our four member cluster ITSOSSIA we use FCP devices B800 and B900. They are each defined in four LPARs: A1A, A1B, A2A, and A2B. B800 and B900 are connected to different switches, and each switch is connected to the DS8300. The WWPN is a unique number, which is the reason we end up with eight different WWPNs.

Example 4-5 shows the WWPNs from our environment.

*Example 4-5 z/VM cluster 1: P201 LPARs A1A, A1B, and P301 LPARs A2A, A2B (ITSOSSIA)*

---

Switch 1

P201 LPAR A1A CHPID 78 Device B800 WWPN c05076f77a001430  
P201 LPAR A1B CHPID 78 Device B800 WWPN c05076f77a0014b0  
P301 LPAR A2A CHPID 7C Device B800 WWPN c05076ecea801470  
**P301 LPAR A2B CHPID 7C Device B800 WWPN c05076ecea8014f0**

DS8300 S/N L3000 LUN ID 1000. WWPN 5005076305**00**C74C

Switch 2

P201 LPAR A1A CHPID 79 Device B900 WWPN c05076f77a001540  
P201 LPAR A1B CHPID 79 Device B900 WWPN c05076f77a0015c0  
P301 LPAR A2A CHPID 7D Device B900 WWPN c05076ecea8018b0  
**P301 LPAR A2B CHPID 7D Device B900 WWPN c05076ecea801930**

DS8300 S/N L3000 LUN ID 1000. WWPN 5005076305**08**C74C

---

LUN ID 1000 points to the same volume in the DS8300, but it has two different WWPNs, depending on which switch you connect to it. This means the host to LUN mapping in the DS8300 has to be made for both paths.

In Example 4-5, we show the setup on ITSOSI4, which is LPAR A2B, and we use CHPID 7C and 7D, shown in bold font in the example.

## 4.4.2 Reading the WWPN over the HMC

The identity of the WWPN of your FCP channel is obtained from the Hardware Management Console (HMC) on the Support Element of your System z hardware. Find it by clicking **CPC Configuration** → **FCP Configuration**.

The WWPN assigned to each device in each LPAR will be displayed. Figure 4-2 shows the data for FCP device B800. In LPAR A2A (ITSOSI4), it is assigned to CHPID 7C and has the WWPN c05076ecea8014f0. Here you can also verify that the N\_Port ID Virtualization (NPIV) mode is set.

Partition	CSS	IID	CHPID	SSID	Device Number	WWPN	NPIV Mode	Current Configured
A2B	02	0b	7c	00	b800	c05076ecea8014f0	On	Yes
A2B	02	0b	7c	00	b801	c05076ecea8014f4	On	Yes
A2B	02	0b	7c	00	b802	c05076ecea8014f8	On	Yes
A2B	02	0b	7c	00	b803	c05076ecea8014fc	On	Yes
A2B	02	0b	7c	00	b804	c05076ecea801500	On	Yes
A2B	02	0b	7c	00	b805	c05076ecea801504	On	Yes
A2B	02	0b	7c	00	b806	c05076ecea801508	On	Yes
A2B	02	0b	7c	00	b807	c05076ecea80150c	On	Yes
A2B	02	0b	7c	00	b808	c05076ecea801510	On	Yes
A2B	02	0b	7c	00	b809	c05076ecea801514	On	Yes
A2B	02	0b	7c	00	b80a	c05076ecea801518	On	Yes
A2B	02	0b	7c	00	b80b	c05076ecea80151c	On	Yes

Items found: 68 for LPAR A2B

Buttons: Apply, Transfer via FTP, Cancel, Help

Figure 4-2 Definition from HMC CHPID display

## 4.4.3 Reading the WWPN from z/VM

If you do not have access to the HMC, you can issue a **query fcp** command in z/VM. The FCP device must be attached to any user. You will see the WWPN on the display only if you have NPIV enabled on this channel. See Example 4-6.

*Example 4-6 Output from the query fcp command*

```
q fcp
FCP B800 ATTACHED TO ITSOLNX4 B800 CHPID 7C
    WWPN C05076ECEA8014F0
FCP B900 ATTACHED TO ITSOLNX4 B900 CHPID 7D
    WWPN C05076ECEA801930
```

#### 4.4.4 NPIV

NPIV(N-Port Id Virtualization) is required for security reasons. You can restrict the access to your SAN volumes, or to select LPARs or FCP channels.

Each virtual adapter has its own WWPN in the SAN, which is hard coded in the adapter card and cannot be changed by the customer.

SAN zoning is done in the switch. Only WWPNs that are in the same zone are able to communicate with each other.

LUN masking is done in the storage controller (for example, DS8300). You must create a mapping list defining which WWPNs are allowed to access your volume.

Without NPIV, only the first operating system that gets access to a disk volume can use it and you cannot control access to the disk volume.

Figure 4-3 provides a further explanation of why you should use NPIV.

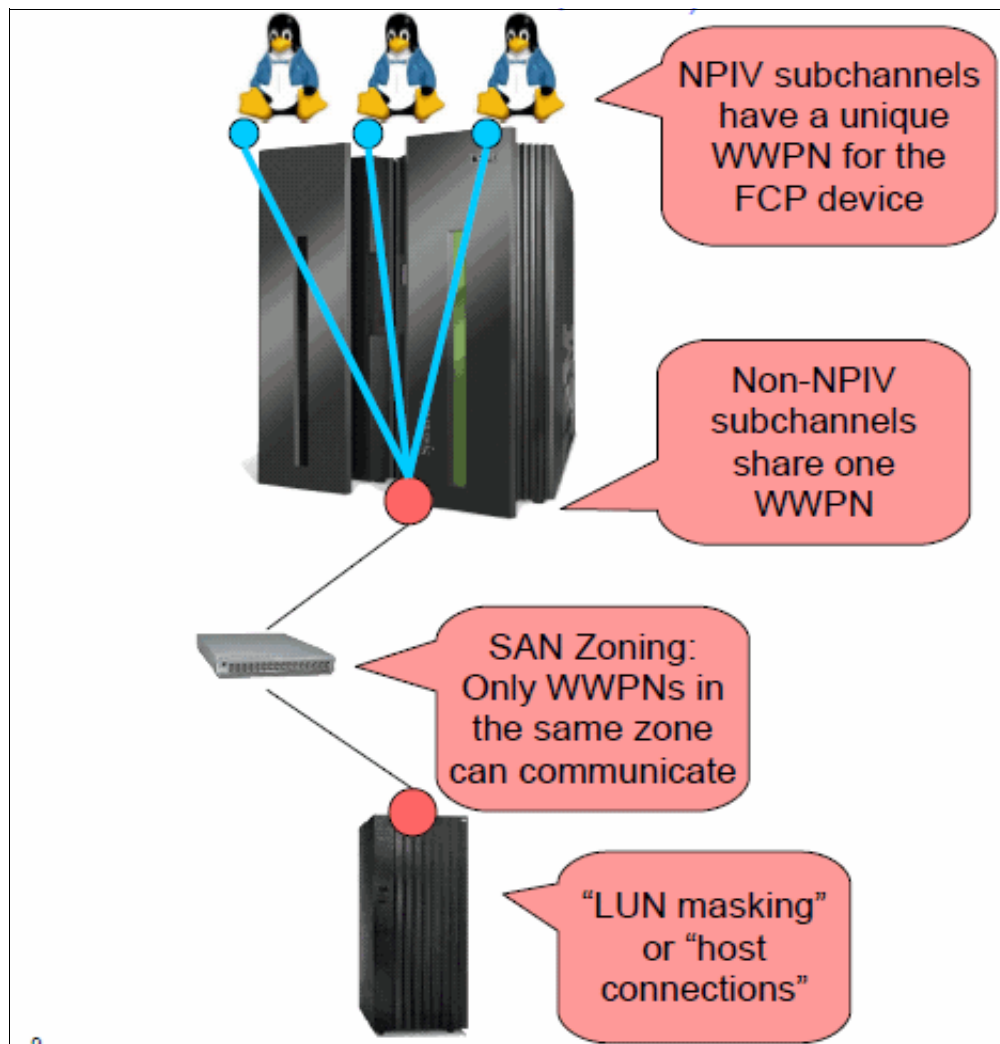


Figure 4-3 NPIV concept

## 4.5 Defining the volume in the storage controller

The definitions in the storage controller depend on the type of hardware that you use. In our environment, we used a DS8300. This controller has a GUI interface, where you can define volumes. This is shown in the following steps:

1. Define the host connections as shown in Figure 4-4.

Manage Host Connections			
Click on a host connection row in the table to view details about the connection. Select a host connection to perform an action.			
<div> </div> <div>Select action ▼</div>			
Select	Nickname	Status	Host Type
<input checked="" type="checkbox"/>	SSI4CHP7C	Normal	zLinux
<input type="checkbox"/>	SSI3CHP7C	Not logged-in	zLinux
<input type="checkbox"/>	SSI2CHP78	Not logged-in	zLinux
<input type="checkbox"/>	SSI1CHP78	Not logged-in	zLinux
<input type="checkbox"/>	SSI4CHP7D	Not logged-in	zLinux
<input type="checkbox"/>	SSI3CHP7D	Not logged-in	zLinux
Showing 1 - 6 of 16   Selected 1			

Figure 4-4 Defined host connections

2. Assign the host to a host port, which is the WWPN of the FCP channel. This is used to reach the Linux system on z/VM (Figure 4-5).

Host Ports defined			
Volume Groups accessed			
I/O Ports used			
The following host ports are defined in the selected host connection. This host connection has 1 host ports that are mapped to 1 volume groups through <Ar			
<div> </div> <div>Select action ▼</div>			
Select	Nickname	Status	WWPN
<input type="checkbox"/>	SSI4CHP7C	Normal	c05076ecea8014f0
Showing 1 - 1 of 1   Selected 0			

Figure 4-5 Connecting host ports

3. Define a volume group that corresponds to the host (Figure 4-6).

Open Systems Volume Groups									
Storage image: <span>IBMStoragePlex - &lt;75L3001&gt;</span> <span>Refresh</span> Last refresh: Tue Apr 10 13:57:32 EDT 2012 <span>Refresh for most current data</span>									
Volume Groups									
<div> </div> <div>Select action ▼</div>									
Select	Nickname	ID	Status	Addressing Method	Block size	# Volumes	Capacity	# Host Connections	
<input type="checkbox"/>	SVCCF8	0	Normal	Mask	512	16	320	8	
<input checked="" type="checkbox"/>	ZS2Z15A	1	Normal	Mask	512	1	10	8	
Showing 1 - 2 of 2   Selected 1									

Figure 4-6 Define the volume group

4. Define the size of the host and assign the host to a volume group (Figure 4-7).

**Volume Group Properties**

You can change the nickname of the volume group, or you can select volumes to add or remove from the volume group. The table displays all of the volumes that are assigned to the volume group.

Volume Group Nickname:  ID:

Addressing Method:  Block Size:

**Volumes Assigned**

Filter by LSS:

Select action:

Select	Nickname	ID	LUN ID	Storage Allocation	GiB
<input type="checkbox"/>	ZS2Z151000	1000		Standard	10.0

Showing 1 - 1 of 1 Selected 0

Figure 4-7 Volume group properties

5. Create an Open System Volume (Figure 4-8).

**Open Systems Volumes**

Last refresh: Tue Apr 10 13:57:32 EDT 2012 Refresh for most current data

[Back to Open Systems Volumes Main Page](#)

**Manage Volumes**

Select the filtering options to use for displaying volumes. The table is updated based on the filters that you select. To perform actions, select one or more volumes in the table.

Filter by:  Select LSS:

Select action:

Select	Nickname	ID	Status	Type	GiB	Storage Allocation
<input type="checkbox"/>	ZS2Z151000	1000	Normal	DS	10.0	Standard

Figure 4-8 Definitions of the SAN volume

6. You can identify the LUN of your volume using the command line interface of the DS8300. To get the LUN, issue the commands shown in Example 4-7, Example 4-8, and Example 4-9 from the Data Storage Control Line Interface (DSCLI):
  - a. List the extent pool. (Example 4-7).

Example 4-7 Output of `lsfbvol` command from the DSCLI

```
dscli> lsfbvol
Date/Time: April 11, 2012 1:37:24 PM EDT IBM DSCLI Version: 6.5.15.72 DS: IBM.2107-75L3001
Name      ID      accstate  datastate  configstate  deviceMTM  datatype  extpool  cap (2^30B)  cap (10^9B)  cap (blocks)
=====
ZS2Z151000 1000 Online   Normal    Normal      2107-900   FB 512    PO 10.0    -          20971520
SVCCF8_1001 1001 Online   Normal    Normal      2107-900   FB 512    PO 20.0    -          41943040
```

- b. List all defined volume groups (Example 4-8).

Example 4-8 Output of the `dscli lsvolgrp` command

```
dscli> lsvolgrp
Date/Time: April 11, 2012 1:39:34 PM EDT IBM DSCLI Version: 6.5.15.72 DS:
IBM.2107-75L3001
```

Name	ID	Type
SVCCF8	V0	SCSI Mask
<b>ZS2Z15A</b>	V1	SCSI Mask
All CKD	V10	FICON/ESCON All
All Fixed Block-512	V20	SCSI All
All Fixed Block-520	V30	OS400 All

c. Use the **showvolgrp - lunmap** command to get the LUN of your volume (Example 4-9).

*Example 4-9 Output of the dscli showvolgrp command*

---

```

dscli> showvolgrp -lunmap V01
Date/Time: April 9, 2012 12:41:51 PM EDT IBM DSCLI Version: 6.5.15.72 DS:
IBM.2107-75L3001
Name ZS2Z15A
ID V1
Type SCSI Mask
Vols 1000
=====LUN Mapping=====
vol lun
=====
1000 40104000

```

---

You will need the LUN number when you define your volume to the Linux guest, which is described in 4.7, “Defining the SCSI file system in Linux” on page 33.

You can find more information about LUNs in an IBM DS8000® at the following link:

<http://www.redbooks.ibm.com/abstracts/tips0598.html?Open>

## 4.6 Defining the FCP channels to z/VM

There are two ways to define the FCP channels for use in a z/VM environment for Linux servers: dynamic and permanent definitions.

After you set up and define your connections dynamically, you can attach them manually. When you have verified them, you should define them permanently. This applies to:

- ▶ FCP channels that must have an EQID that is unique for the cluster and Linux system, for relocation purposes
- ▶ FCP channels that must be dedicated to the Linux system

### 4.6.1 Defining FCP channels dynamically

You can issue the definition commands dynamically, as shown in Example 4-10.

*Example 4-10 EQID definition command*

---

```

set RDEV B800 EQID FCPEQID1 Type FCP
HCPZRP6722I Characteristics of device B800 were set as requested.
1 RDEV(s) specified; 1 RDEV(s) changed; 0 RDEV(s) created

set RDEV B900 EQID FCPEQID1 Type FCP

```

HCPZRP6722I Characteristics of device B900 were set as requested.  
1 RDEV(s) specified; 1 RDEV(s) changed; 0 RDEV(s) created

---

After the dynamic definition, you must manually attach the FCP channel to the Linux user, as shown in Example 4-11. First, check whether the path and the FCP channel are online. Then, attach the FCP channel to the Linux server.

*Example 4-11 ATTACH command*

---

**q chpid 78**

Path 78 online to devices B800 B801 B802 B803 B804 B805 B806 B807  
Path 78 online to devices B808 B809 B80A B80B B80C B80D B80E B80F  
Path 78 online to devices B810 B811 B812 B813 B814 B815 B816 B817  
Path 78 online to devices B818 B819 B81A B81B B81C B81D B81E B81F  
Path 78 online to devices B8FC B8FD  
Ready; T=0.01/0.01 10:58:01

**q chpid 79**

Path 79 online to devices B900 B901 B902 B903 B904 B905 B906 B907  
Path 79 online to devices B908 B909 B90A B90B B90C B90D B90E B90F  
Path 79 online to devices B910 B911 B912 B913 B914 B915 B916 B917  
Path 79 online to devices B918 B919 B91A B91B B91C B91D B91E B91F  
Path 79 online to devices B9FC B9FD  
Ready; T=0.01/0.01 10:58:24

**q b800 b900**

FCP B800 FREE , FCP B900 OFFLINE  
Ready; T=0.01/0.01 10:58:58

**att b800 itsolnx4**

FCP B800 ATTACHED TO ITSOLNX4 B800  
Ready; T=0.01/0.01 10:55:12

**att b900 itsolnx1**

FCP B900 ATTACHED TO ITSOLNX1 B900  
Ready; T=0.01/0.01 10:59:12

---

## 4.6.2 Defining FCP channels permanently

In this section, we describe the method to add the FCP definitions permanently to the z/VM environment. Do this using the following steps:

1. Add FCP definition statements to the z/VM PMAINT CF0 SYSTEM CONFIG file.

In our environment this is FCPEQID1, as shown in Example 4-12 on page 33. We used two FCP channels for redundancy reasons. The setup of your FCP channels and the corresponding EQIDs can vary depending on how many FCP channels you use for each machine, member of the cluster, and each Linux system you have defined. The setup of EQIDs is described in detail in chapter 2 of *An introduction to z/VM Single System Image (SSI) and Live Guest Relocation (LGR)*, SG24-8006.

#### Example 4-12 FCP statements in the SYSTEM CONFIG

```
/* ***** */
/* Set EQUID for the FCP channels used for SCSI */
/* ***** */

RDEV B800 EQID FCPEQID1 Type FCP
RDEV B900 EQID FCPEQID1 Type FCP
```

2. Add the FCP channels to the DIRMAINT directory entry of your Linux guest. We use DEDICATE statements because the relocation will take over this definition. Example 4-13

#### Example 4-13 DIRMAINT directory entry for Linux guest ITSOLNX4

```
USER ITSOLNX4 ITSOSI 6G 32G G
*
* RHEL 5.6
* 0201 = swap space
* 0202 = / root fs
*
INCLUDE LINDFLT
IPL 202
MACH ESA 2
* Option APPLMON is required for monitoring data collection
OPTION APPLMON LKNOPAS
DEDICATE B800 B800 NOQIOASSIST
DEDICATE B900 B900 NOQIOASSIST
MDISK 0201 3390 0001 1000 LX9B26 MR
MDISK 0202 3390 1001 9016 LX9B26 MR
MDISK 0203 3390 3867 10000 LX6032
*DVOPT LNK0 LOG1 RCM1 SMS0 NPW1 LNGAMENG PWC20111103 CRC>-
```

**Tip:** The DEDICATE statement in the Linux server directory is the same that is used to attach the device manually. The only difference between them is that DEDICATE is a permanent definition and ATTACH is valid only while the server is up. After an IPL of the z/VM system or a Linux guest logoff and logon, the ATTACH will be removed.

For more information about FCP channels in z/VM, see the IBM Redbooks publication *Fibre Channel Protocol for Linux and z/VM on IBM System z*, SG24-7266.

## 4.7 Defining the SCSI file system in Linux

Use the following steps to define the SCSI file system in Linux on System z:

1. Verify that the s390utils package is installed; this package contains the necessary tools to work with SCSI on Linux. Do this using the command in Example 4-14.

#### Example 4-14 Query command to verify that s390utils is installed

```
[root@itsolnx4 WAS]# rpm -qa | grep s390
s390utils-1.8.1-11.el5
```

2. Bring the FCP channels online using the **chccwdev** command. After the command is executed, a logical access to your FCP channel will be available in directory `/sys/bus/ccw/drivers/zfcp/` with the name of your FCP channel. The directory will contain

files that will be used to initialize the SCSI disk. We have two FCP channels so that multipathing can be enabled, as shown in Example 4-15.

*Example 4-15 Output of the chccwdev command to bring b900 FCP channel online*

---

```
[root@itsolnx4 ~]# chccwdev -e b800
Setting device 0.0.b800 online
Done
[root@itsolnx4 ~]# chccwdev -e b900
Setting device 0.0.b900 online
Done
```

---

3. Add the WWPN of the storage device to the FCP channel that was created. Use the **echo** command to write the WWPN to the port\_add file under the logical directory created in step 2. After adding the WWPN, another directory will be created to represent the WWPN. Each FCP channel will have a different WWPN so that there will be different paths to the same volume. In our example, the b800 FCP channel uses WWPN 0x500507630500c74c; b900 uses WWPN 0x500507630508c74c (Example 4-16).

*Example 4-16 Add the WWPN to the FCP channel*

---

```
[root@itsolnx4 0.0.b800]# pwd
/sys/bus/ccw/drivers/zfcp/0.0.b800
[root@itsolnx4 0.0.b800]# echo 0x500507630500c74c > port_add
[root@itsolnx4 0.0.b900]# pwd
/sys/bus/ccw/drivers/zfcp/0.0.b900
[root@itsolnx4 0.0.b900]# echo 0x500507630508c74c > port_add
```

---

4. Add the LUN to the WWPN. Use the **echo** command to write the LUN ID to the unit\_add file under the directory created in the previous step. An important point to note is that the LUN ID would stay constant in the setup because we want to reach the same volume, but through different paths (Example 4-17).

*Example 4-17 Add the LUN to the WWPN*

---

```
[root@itsolnx4 0x500507630500c74c]# pwd
/sys/bus/ccw/drivers/zfcp/0.0.b800/0x500507630500c74c
[root@itsolnx4 0x500507630500c74c]# echo 0x4010400000000000 > unit_add
[root@itsolnx4 0x500507630508c74c]# pwd
/sys/bus/ccw/drivers/zfcp/0.0.b900/0x500507630508c74c
[root@itsolnx4 0x500507630508c74c]# echo 0x4010400000000000 > unit_add
```

---

5. In order to have the SCSI disks available at each reboot, modify the /etc/zfcp.conf file and re-create the initial ram disk, as described here.

Each line in /etc/zfcp.conf represents a SCSI device. As shown in Example 4-18, the format is FCPID WWPN LUN.

*Example 4-18 /etc/zfcp.conf setup to have persistent SCSI devices*

---

```
[root@itsolnx4 scsivolgrp]# cat /etc/zfcp.conf
0.0.b800 0x500507630500c74c 0x4010400000000000
0.0.b900 0x500507630508c74c 0x4010400000000000
```

---

Re-create the initial ram disk by running the command shown in Example 4-19 as root and under the /boot directory.

*Example 4-19 Re-create the initial ram disk with the new configuration*

```
[root@itsolnx4 boot]# mkinitrd -v initrd-`uname -r` `uname -r`
Creating initramfs
Looking for deps of module ext3: jbd
Looking for deps of module jbd
Found root device dasdb1 for LABEL=/
Looking for driver for device dasdb1
Looking for deps of module ccw:t3990mE9dt3390dm0C: dasd_mod dasd_eckd_mod
Looking for deps of module dasd_mod
Looking for deps of module dasd_eckd_mod: dasd_mod
Looking for driver for device dasda1
Looking for deps of module ccw:t3990mE9dt3390dm0C: dasd_mod dasd_eckd_mod
Looking for deps of module ide-disk
Looking for deps of module dasd_fba_mod: dasd_mod
/sbin/scsi_id: option requires an argument -- s
/sbin/scsi_id: option requires an argument -- s
Looking for deps of module dm-mod
Looking for deps of module dm-mirror: dm-mod dm-log
Looking for deps of module dm-log: dm-mod
Looking for deps of module dm-zero: dm-mod
Looking for deps of module dm-snapshot: dm-mod
Looking for deps of module multipath
Looking for deps of module dm-multipath: scsi_mod scsi_dh dm-mod
Looking for deps of module scsi_mod
Looking for deps of module sd_mod: scsi_mod
Looking for deps of module scsi_dh: scsi_mod
Looking for deps of module dm-round-robin: scsi_mod scsi_dh dm-mod dm-multipath
Looking for deps of module dm-mem-cache
Looking for deps of module dm-region_hash: dm-mod dm-log
Looking for deps of module dm-message
Looking for deps of module dm-raid45: dm-message dm-mod dm-mem-cache dm-log
dm-region_hash
Using modules: /lib/modules/2.6.18-238.el5/kernel/fs/jbd/jbd.ko
/lib/modules/2.6.18-238.el5/kernel/fs/ext3/ext3.ko
/lib/modules/2.6.18-238.el5/kernel/drivers/s390/block/dasd_mod.ko
/lib/modules/2.6.18-238.el5/kernel/drivers/s390/block/dasd_eckd_mod.ko
/lib/modules/2.6.18-238.el5/kernel/drivers/s390/block/dasd_fba_mod.ko
...
/lib/modules/2.6.18-238.el5/kernel/drivers/md/dm-mem-cache.ko
/lib/modules/2.6.18-238.el5/kernel/drivers/md/dm-region_hash.ko
/lib/modules/2.6.18-238.el5/kernel/drivers/md/dm-message.ko
/lib/modules/2.6.18-238.el5/kernel/drivers/md/dm-raid45.ko
/sbin/nash -> /tmp/initrd.i10142/bin/nash
/sbin/insmod.static -> /tmp/initrd.i10142/bin/insmod
copy from `/lib/modules/2.6.18-238.el5/kernel/fs/jbd/jbd.ko' [elf64-s390] to
`/tmp/initrd.i10142/lib/jbd.ko' [elf64-s390]
...
copy from `/lib/modules/2.6.18-238.el5/kernel/drivers/md/dm-region_hash.ko'
[elf64-s390] to `/tmp/initrd.i10142/lib/dm-region_hash.ko' [elf64-s390]
copy from `/lib/modules/2.6.18-238.el5/kernel/drivers/md/dm-message.ko'
[elf64-s390] to `/tmp/initrd.i10142/lib/dm-message.ko' [elf64-s390]
copy from `/lib/modules/2.6.18-238.el5/kernel/drivers/md/dm-raid45.ko'
[elf64-s390] to `/tmp/initrd.i10142/lib/dm-raid45.ko' [elf64-s390]
/sbin/dmraid.static -> /tmp/initrd.i10142/bin/dmraid
/sbin/kpartx.static -> /tmp/initrd.i10142/bin/kpartx
```

```

Adding module jbd
Adding module ext3
Adding module dasd_mod with options dasd=201-203
Adding module dasd_eckd_mod
Adding module dasd_fba_mod
Adding module dm-mod
Adding module dm-log
Adding module dm-mirror
Adding module dm-zero
Adding module dm-snapshot
Adding module multipath
Adding module scsi_mod
Adding module sd_mod
Adding module scsi_dh
Adding module dm-multipath
Adding module dm-round-robin
Adding module dm-mem-cache
Adding module dm-region_hash
Adding module dm-message
Adding module dm-raid45

```

---

Run the **zipl** command shown in Example 4-20 to ensure that the newly created ram disk is going to be used for boot up.

*Example 4-20 Command to make sure the correct ram disk will be used at boot time*

---

```

[root@itsolnx4 ~]# zipl -V
Using config file '/etc/zipl.conf'
Target device information
  Device.....: 5e:04
  Partition.....: 5e:05
  Device name.....: dasdb
  Device driver name.....: dasd
  DASD device number.....: 0202
  Type.....: disk partition
  Disk layout.....: ECKD/compatible disk layout
  Geometry - heads.....: 15
  Geometry - sectors.....: 12
  Geometry - cylinders.....: 9016
  Geometry - start.....: 24
  File system block size.....: 4096
  Physical block size.....: 4096
  Device size in physical blocks..: 1622856
Building bootmap in '/boot/'
Building menu 'rh-automatic-menu'
Adding #1: IPL section 'Linux' (default)
  kernel image.....: /boot/vmlinuz-2.6.18-238.el5
  kernel parmline...: 'root=LABEL=/'
  initial ramdisk...: /boot/initrd-2.6.18-238.el5.img
  component address:
    kernel image.....: 0x00010000-0x005b2fff
    parmline.....: 0x00001000-0x00001fff
    initial ramdisk.: 0x01800000-0x01ab4fff
    internal loader.: 0x0000a000-0x0000afff
Preparing boot device: dasdb (0202).
Preparing boot menu

```

```
Interactive prompt.....: enabled
Menu timeout.....: 15 seconds
Default configuration...: 'Linux'
Syncing disks...
Done.
```

---

## 4.7.1 Setting up multipath

When using SCSI disks, multipathing is a standard practice. Multipathing provides redundant access to your storage devices and provides a way of increasing throughput via load balancing. Use the following steps to set up multipathing:

1. The package needed for multipathing is called `device-mapper-multipath`. To verify that the package is installed, run the command shown in Example 4-21.

*Example 4-21 Verify that the required package is installed*

---

```
[root@itsolnx4 ~]# rpm -qa | grep device-mapper-multipath
device-mapper-multipath-0.4.7-42.el5
```

---

2. After installing the required package, load the module to the kernel. The module name is `dm-multipath`. It can be loaded into the kernel using the `modprobe` command shown in Example 4-22.

*Example 4-22 Add the dm-multipath module to the kernel using modprobe command*

---

```
[root@itsolnx4 ~]# modprobe dm-multipath
```

---

3. By default, all the devices in the system are blacklisted. You can enable multipathing on certain devices by editing the `/etc/multipath.conf` file and commenting out the `devnode_blacklist` directive. Add a “#” to the lines you want to comment out (Example 4-23).

*Example 4-23 /etc/multipath.conf file commented to allow device multipathing*

---

```
[root@itsolnx4 ~]# cat /etc/multipath.conf
defaults { user_friendly_names yes
}

#blacklist {
    # devnode "^(ram|raw|loop|fd|md|dm-|sr|scd|st|sd)[0-9]*"
    # devnode "^(hd|xvd|vd|sd)[a-z]*"
    # wwid "*"
# devnode '*'
#}
```

---

4. Enable the multipath daemon to execute at boot time. Do this by issuing a series of `chkconfig` commands, as shown in Example 4-24.

*Example 4-24 Add multipathd to the boot process*

---

```
[root@itsolnx4 ~]# chkconfig --add multipathd
[root@itsolnx4 ~]# chkconfig multipathd on
[root@itsolnx4 ~]# chkconfig --list | grep multipathd
multipathd    0:off  1:off  2:on   3:on   4:on   5:on   6:off
```

---

After everything is set up correctly, start the multipath daemon. This must be done only one time because on each subsequent reboot it will occur automatically. See Example 4-25.

*Example 4-25 Start the multipath daemon manually*

---

```
[root@itsolnx4 ~]# /etc/init.d/multipathd start
Starting multipathd daemon:
```

---

To view the multipath devices, issue the multipath command with the **-l** option, as shown in Example 4-26.

*Example 4-26 Multipath command shows the two paths to the SCSI device*

---

```
[root@itsolnx4 ~]# multipath -l
36005076305ffc74c0000000000001000 dm-0 IBM,2107900
[size=10G][features=1 queue_if_no_path][hwhandler=0][rw]
\_ round-robin 0 [prio=0][active]
  \_ 0:0:0:1 sda 8:0 [active][undef]
  \_ 1:0:0:1 sdb 8:16 [active][undef]
```

---

To use the multipathed device, the system now has a logical representation under `/dev/mapper/` which can be used like any other device on Linux. See Example 4-27.

*Example 4-27 The logical device that represents the multipathed devices of sda and sdb*

---

```
[root@itsolnx4 ~]# ls /dev/mapper/36005076305ffc74c0000000000001000
/dev/mapper/36005076305ffc74c0000000000001000
```

---

## 4.7.2 Define the volume group

You must initialize the devices to create the volume group. Do this by issuing the **pvcreate** command with the corresponding device path, as shown in Example 4-28.

*Example 4-28 This will initialize the multipathed device to be part of a volume group.*

---

```
[root@itsolnx4 ~]# pvcreate /dev/mapper/36005076305ffc74c0000000000001000
Physical volume "/dev/mapper/36005076305ffc74c0000000000001000" successfully
created
```

---

Create a volume group and specify the devices that are going to be part of the volume group. This can be done using the **vgcreate** command. Enter the name of the volume group and specify the devices that are to be included, as shown in Example 4-29. This example creates a volume group with the name `scsivolgrp` and the multipathed device is used as part of the newly created volume group.

*Example 4-29 Create a volume group*

---

```
[root@itsolnx4 ~]# vgcreate scsivolgrp
/dev/mapper/36005076305ffc74c0000000000001000
Volume group "scsivolgrp" successfully created
```

---

Create logical volumes inside the volume group. Do this using the **lvcreate** command. Enter the size of the logical volume, the name of the logical volume, and the name of the volume group it should belong to.

*Example 4-30 Creating logical volumes in a volume group*

---

```
[root@itsolnx4 ~]# lvcreate -L5G -nWAS_INSTALL scsivolgrp
Logical volume "WAS_INSTALL" created
[root@itsolnx4 ~]# lvcreate -l1279 -n DB2_LOGS scsivolgrp
```

---

Logical volume "DB2\_LOGS" created

---

In Example 4-30, the first **lvcreate** command specified the amount of space required for the logical volume using the **-L** option, whereas the second **lvcreate** command used the **-l** option to specify the number of extents for the logical volume.

Now that the volume group and the logical volumes are created, they should be available for use under the **/dev** directory. The **/dev** directory will contain a directory with the volume group name. That directory will have references to the logical volumes that were created, as shown in Example 4-31.

*Example 4-31 The device handle to the logical volumes we created*

---

```
[root@itsolnx4 ~]# ls /dev/scsivolgrp/  
DB2_LOGS  WAS_INSTALL
```

---

### 4.7.3 Creating a file system

After defining the logical reference to your device, install a file system on it. A common file system used in current Linux systems is **ext3**. Install a file system using the **mkfs** command, as shown in Example 4-32.

*Example 4-32 Create ext3 file systems on the respective disks*

---

```
[root@itsolnx4 ~]# mkfs -t ext3 /dev/scsivolgrp/DB2_LOGS  
[root@itsolnx4 ~]# mkfs -t ext3 /dev/scsivolgrp/WAS_INSTALL
```

---

After creating the file systems on the respective devices, create the directories where the disks will be mounted (Example 4-33).

*Example 4-33 Make the directories for mount points*

---

```
[root@itsolnx4 ~]# mkdir /DB2LOGS  
[root@itsolnx4 ~]# mkdir /WASINSTALL
```

---

Create labels to the devices so that the label can be used in **/etc/fstab**. This would allow the changing of the underlying device seamlessly, without having to edit the **/etc/fstab** file again. To create labels, use the **e2label** command, as shown in Example 4-34.

*Example 4-34 Create the label for the devices*

---

```
[root@itsolnx4 /]# e2label /dev/mapper/scsivolgrp-WAS_INSTALL /WASINSTALL  
[root@itsolnx4 /]# e2label /dev/mapper/scsivolgrp-DB2_LOGS /DB2LOGS
```

---

Create the entries in **/etc/fstab** so that the disks will be automounted on system reboot (Example 4-35 on page 39).

*Example 4-35 Create entries in /etc/fstab for persistent mounting after reboot*

---

```
[root@itsolnx4 ~]# cat /etc/fstab  
LABEL=/ / ext3 defaults 1 1  
tmpfs /dev/shm tmpfs defaults 0 0  
devpts /dev/pts devpts gid=5,mode=620 0 0  
sysfs /sys sysfs defaults 0 0  
proc /proc proc defaults 0 0  
LABEL=SWAP-dasda1 swap swap defaults 0 0
```

```
LABEL=/DB2LOGS /DB2LOGS      ext3 rw,auto,user,exec 0 0  
LABEL=/WASINSTALL /WASINSTALL ext3 rw,auto,user,exec 0 0
```

---

The final step is to mount the file systems. This can be done using the **mount** command, as shown in Example 4-36.

*Example 4-36 Mount the file systems*

---

```
[root@itsolnx4 /]# mount -a
```

---



## Relocation domains

Relocation domains were discussed extensively in *An introduction to z/VM Single System Image (SSI) and Live Guest Relocation (LGR)*, SG24-8006.

This chapter expands that discussion by first describing the assignment of relocation domains to a virtual machine and then examining some relocation domain examples:

- ▶ Using the default relocation domain (called SSI)
- ▶ Using relocation domains to reflect the architecture level
- ▶ Relocating guests to a different architecture domain
- ▶ Using relocation domains with special hardware or software features
- ▶ Using relocation domains with special business purposes such as creating a development and production relocation domain
- ▶ Using relocation domains during upgrade of hardware

As a reminder, the ITSO lab environment consists of an IBM System z10® and a System z196 with three LPARs on the System z10 and three LPARs on the System z196.

## 5.1 Review of relocation domain concepts

A *relocation domain* defines a set of members of an SSI cluster among which Linux guests can be relocated. A relocation domain can be used to define all or a subset of members of an SSI cluster to which a particular guest can be relocated. These domains can be defined for business or technical reasons.

For example, a domain can be defined that has all of the architectural facilities necessary for a particular application, or a domain can be defined to allow access only to systems with a particular software tool. Whatever the reason for the definition of a domain, the z/VM control program (CP) allows relocation among the members of the domain without any change to architectural characteristics or CP functionality as seen by the guest. Regardless of differences in the facilities of the individual members, a domain has a common architectural level set to all members. Information about the common architectural level is held in the Persistent Data Record (PDR).

Several default domains are automatically defined by the z/VM control program, specifically:

- ▶ A single member domain for each member of the SSI cluster
- ▶ An SSI domain that has the features and facilities common to all members of the SSI cluster

Defining your own domains is useful in an SSI cluster with three or more members. In a one or two member cluster, all possible domains are defined by default.

If no domain is specified for a USER then, by default, it is put into the SSI domain. An IDENTITY by default is put into a single member domain for each member. A USER or IDENTITY can be assigned to only one domain.

## 5.2 Assigning relocation domains to a virtual guest

There are two ways to assign a relocation domain to a virtual guest:

- ▶ Dynamically, using the `define relocation` command

A running virtual guest can be dynamically reassigned to a domain with the same or greater facilities provided that the SSI member where the virtual guest is running has access to those facilities. For example, a guest might be in the SSI domain, but relocated to a member with access to more facilities, so you might want to reassign this virtual guest to a domain with higher facilities. Assigning a virtual guest to a domain with less facilities can have unpredictable results.

- ▶ Permanently, by adding the `vmrelocate` command to the user directory

Upon successful logon, the virtual guest is assigned to a virtual architecture level, according to its dynamically defined domain or the architectural level specified in its directory entry. This virtual guest can use only the features or facilities defined within its domain, even if it is logged on to an SSI member that has more features available. This restriction is known as *fencing* or a *fenced response*. A fenced response means that the virtual guest cannot use facilities or features that are not included in the domain even if the members that are in the domain have access to these features.

Some commands or instructions that have fenced responses are:

- ▶ **query cpuid** - Displays the doubleword processor identifier used by your virtual machine. This processor identifier will always reflect the virtual architecture level and the processor number set at logon and is not affected by relocation or relocation domain changes.
- ▶ Diagnose x'00' - Used to examine the z/VM extended-identification code. This will show the virtual control program attributes and is used mainly in programs (such as REXX EXECs).
- ▶ Store Facility List Extended (STFLE) - A hardware instruction to list processor facilities in the System z box.

The easiest way to determine the CPU ID is by using the **query cpuid** command. If you issue the command from a USER, the command result shows the common architecture level of all members of the SSI cluster. As shown in Example 5-1, in our case it is a 2097 (z10).

*Example 5-1 QUERY CPUID issued by a USER*

---

```
id
MAINT620 AT ITS0SSI4 VIA RSCS      04/20/12 11:34:12 EDT      FRIDAY
Ready; T=0.01/0.01 11:34:12
q cpuid
CPUID = FF2B3BD520978000
```

---

If the Q CPUID command is issued on an SSI member from an IDENTITY, the command result shows the architecture level for that member. As shown in Example 5-2, in our case it is a 2817 (z196).

*Example 5-2 QUERY CPUID issued by an IDENTITY*

---

```
id
MAINT      AT ITS0SSI4 VIA RSCS      04/20/12 11:37:06 EDT      FRIDAY
q cpuid
CPUID = FF2B3BD528178000
Ready; T=0.01/0.01 11:36:55
```

---

All USERS that are in the default relocation domain named SSI will show the common level of architecture for all the members within this cluster. In our environment this is a System z z10.

The same information is obtained from the Linux on System z guest running in the same SSI member because it is a USER using the default relocation domain SSI (Example 5-3).

*Example 5-3 CPU information from Linux guest in the default domain*

---

```
itsolnx1:~ # cat /proc/cpuinfo
vendor_id      : IBM/S390
# processors    : 1
bogomips per cpu: 14367.00
features        : esan3 zarch stfle msa ldisp eimm dfp etf3eh highgrps
processor 0: version = FF, identification = 2A3BD5, machine = 2097
```

---

But if you assign any of the Linux guests or USERS to a specific domain, it will show the information about that specific domain. We defined ITSOLNX1 to a domain that only included the System z z196. The output from the display is shown in Example 5-4.

```
itsolnx1:~ # cat /proc/cpuinfo
vendor_id      : IBM/S390
# processors   : 1
bogomips per cpu: 14367.00
features       : esan3 zarch stfle msa ldisp eimm dfp etf3eh highgrs
processor 0: version = FF, identification = 2A3BD5, machine = 2817
```

## 5.3 Using the default relocation domain named SSI

As described in Chapter 2, “Lab environment” on page 9, our four member cluster consists of two members located on a System z z10 and two members located on a System z z196. In this section, we describe examples using the default relocation domain named SSI.

### 5.3.1 All four members active in the cluster

Figure 5-1 shows our default SSI domain with all of the Linux guests defined in that domain.

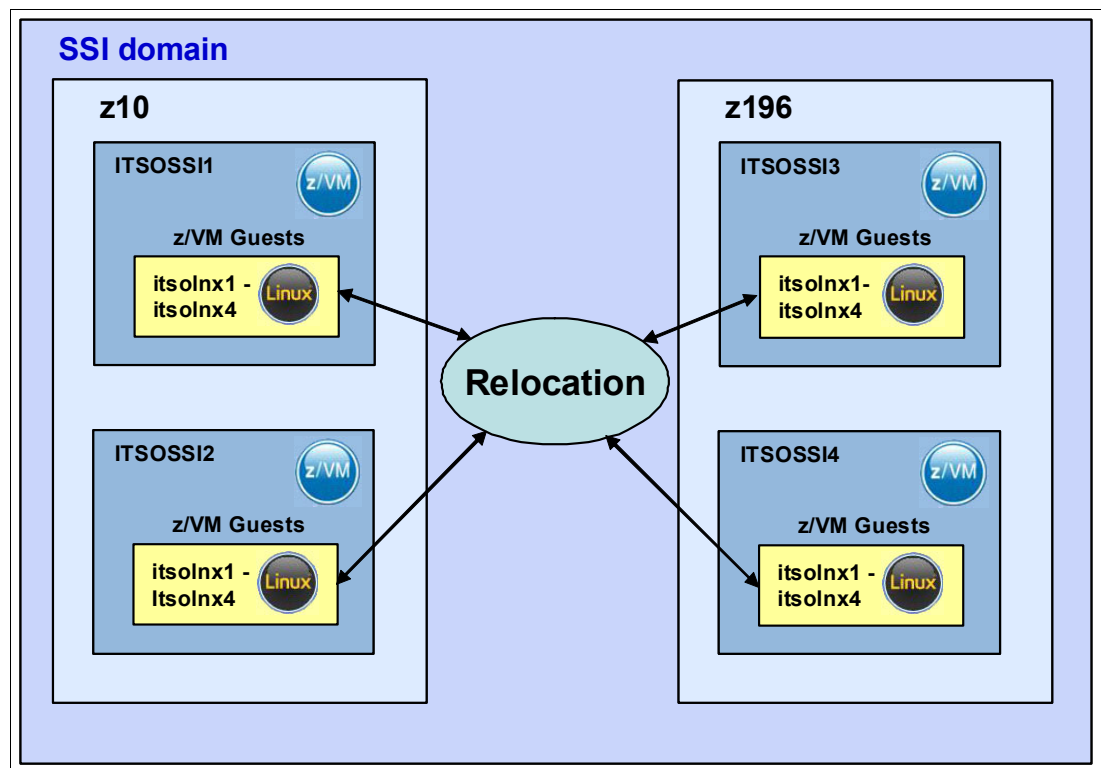


Figure 5-1 Default relocation domain 'SSI'

The Linux guest named ITSOLNX1 was started in this environment, and the **query vmrelocate** command was issued to show that the Linux guest is running on the default domain named SSI, as shown in Example 5-5.

*Example 5-5 Linux guest in default relocation domain*

---

```
q vmrelocate itsolnx1
Running on member ITS0SSI3
Relocation enabled in Domain SSI
Ready; T=0.01/0.01 11:01:06
```

---

If a z/VM cluster has members with different architecture levels (for example, members on a System z z10 and members on a System z z196), the Linux guest is automatically downgraded to the architecture level of the z10. The `/proc/cpuinfo` file from Linux system ITSOLNX1 (Example 5-6), shows the machine parameter to be 2097, which indicates the architecture level of the z10 system.

*Example 5-6 cpuinfo for the Linux guest in the default domain, 'SSI'*

---

```
itsolnx1:~ # cat /proc/cpuinfo
vendor_id       : IBM/S390
# processors    : 1
bogomips per cpu: 14367.00
features        : esan3 zarch stfle msa ldisp eimm dfp etf3eh highgrps
processor 0: version = FF, identification = 2A3BD5, machine = 2097
```

---

The relocation of our Linux guest works without the need to use the force option for relocation domains or architecture features, as shown in the `vmrelocate test` command (Example 5-7).

*Example 5-7 vmrelocate test in default domain*

---

```
vmrelocate test itsolnx1 itsossi1
User ITSOLNX1 is eligible for relocation to ITS0SSI1
Ready; T=0.01/0.01 11:01:43
```

---

**Explanation:** If you use the default relocation domain SSI for Linux guests in a cluster with z/VM members of different architecture levels, the following occurs:

- ▶ The Linux guest is downgraded to the common architecture level of all z/VM members in the SSI cluster. So if you have members on z/10 and members on z196 in your cluster, your Linux guest never will use features of the z196.
- ▶ Relocation to other members works without any issues because they are all set to the common architecture level.

### 5.3.2 Only members of a System z z196 are active in the cluster

Figure 5-2 on page 46 shows our environment when we shut down members ITS0SSI1 and ITS0SSI2 so that the default domain only included the System z z196 processor.

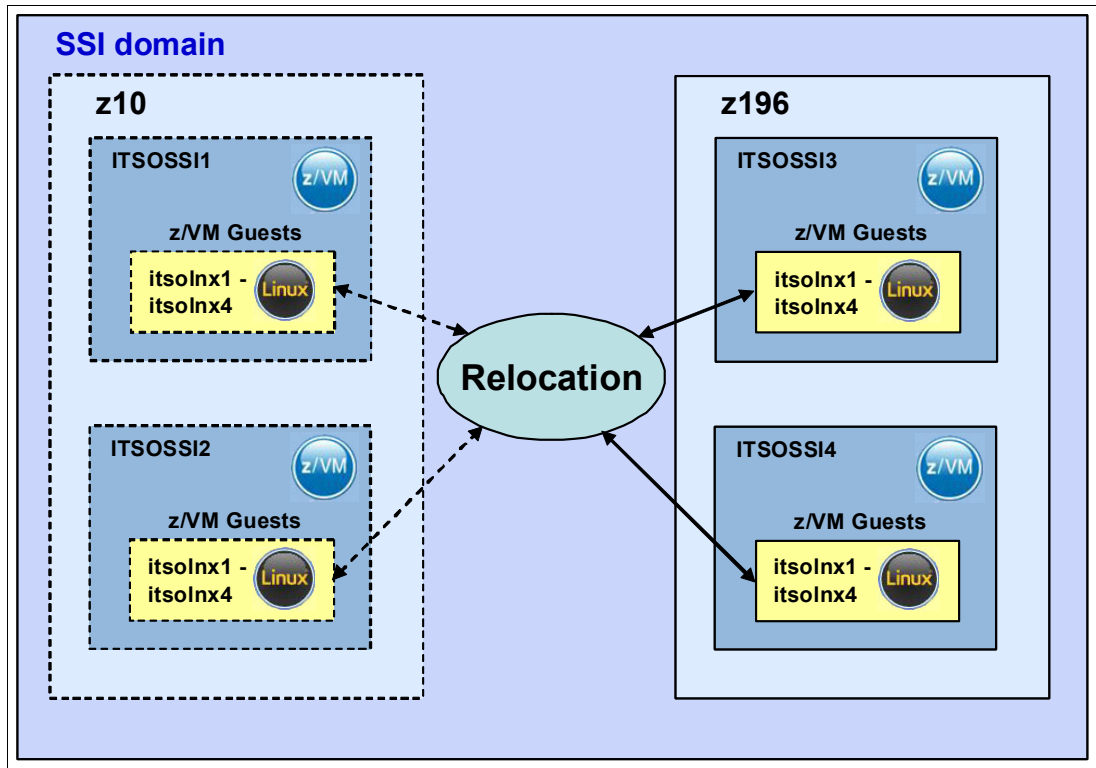


Figure 5-2 Default relocation domain SSI if only z196 members are running

Example 5-8 shows that ITSOSSI1 and ITSOSSI2 are not running.

*Example 5-8 Default relocation domain if only System z z196 members are running*

```
q ssi
SSI Name: ITSOSSIA
SSI Mode: Stable
Cross-System Timeouts: Enabled
SSI Persistent Data Record (PDR) device: SSIAC2 on 9E20
SLOT SYSTEMID STATE      PDR HEARTBEAT      RECEIVED HEARTBEAT
  1 ITSOSSI1 Down (shut down successfully)
  2 ITSOSSI2 Down (shut down successfully)
  3 ITSOSSI3 Joined    04/19/12   11:03:51 04/19/12   11:03:51
  4 ITSOSSI4 Joined    04/19/12   11:03:45 04/19/12   11:03:45
Ready; T=0.01/0.01 11:04:07
```

The guest is running in the default relocation domain, as shown in Example 5-9.

*Example 5-9 Guest running in default relocation domain*

```
q vmrelocate itsolnx1
Running on member ITSOSSI3
Relocation enabled in Domain SSI
Ready; T=0.01/0.01 11:05:16
```

You would expect that the Linux guest would use features of the System z z196, but this is not the case. The Linux guest still only uses the features of the System z z10, as shown in Example 5-10 on page 47.

*Example 5-10 Results of `cpuinfo` command if only z196 members are running*

---

```
itsolnx1:~ # cat /proc/cpuinfo
vendor_id      : IBM/S390
# processors    : 1
bogomips per cpu: 14367.00
features       : esan3 zarch stfle msa ldisp eimm dfp etf3eh highgprs
processor 0: version = FF, identification = 2A3BD5, machine = 2097
```

---

**Explanation:** If a cluster has different architectures, such as a System z z10 and a System z z196, and you are working with the default relocation domain SSI, which contains all members, the running Linux guests are downgraded to the common architecture level from z10 even if the z10 members are shut down. The reason for this is that the architecture levels for the relocation domains are stored in the PDR (Persistent Data Record) of the SSI cluster and we do not want the virtual architectures changing just because one of the systems happens to be down.

## 5.4 Using relocation domains to reflect the architecture level

If you want to use System z architecture features specific to the z196 for your Linux guests, then you cannot use it in an environment with different hardware levels, such as the default relocation domain named SSI.

The solution is to define relocation domains according to the hardware configuration. We defined the following members:

DOMAIN 1 - DMNZ10, which is the System z z10 processor

DOMAIN 2 - DMNZ196, which is the System z z196 processor

Domain DMNZ10 contains the members ITSOSI1 and ITSOSI2, domain DMNZ196 contains the members ITSOSI3 and ITSOSI4, as shown in Figure 5-3 on page 48.

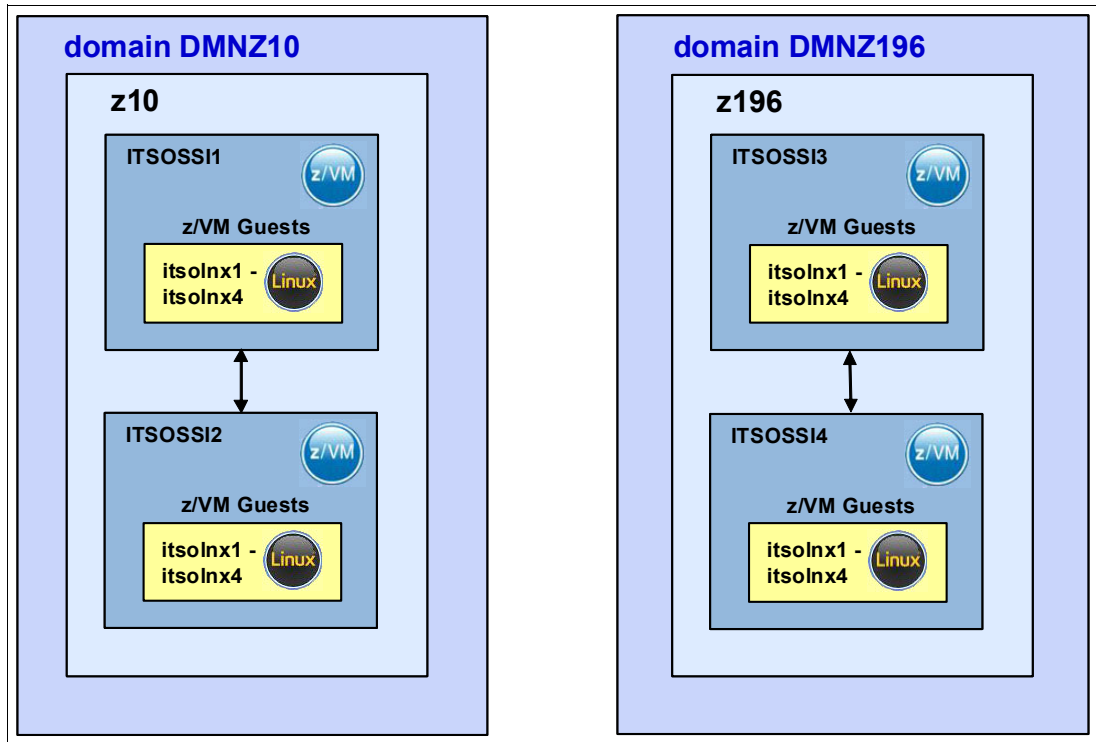


Figure 5-3 Relocation domains reflecting architecture level

We used the **define relodomain** command to define relocation domains dynamically. The command is issued on one member of the SSI cluster. Example 5-11 shows our define command for setting up the domain architecture level which reflects this.

*Example 5-11 Define RELODOMAIN command*

```
define relodomain DMNZ10 member ITS0SSI1 ITS0SSI2
define relodomain DMNZ196 member ITS0SSI3 ITS0SSI4
```

To make the definition permanent, place the RELOCATION\_DOMAIN statement in the SYSTEM CONFIG file (Example 5-12).

*Example 5-12 RELOCATION\_DOMAIN statement*

```
Relocation_Domain DMNZ10 MEMBER ITS0SSI1 ITS0SSI2
Relocation_Domain DMNZ196 MEMBER ITS0SSI3 ITS0SSI4
```

Use the **set vmrelocate** command to dynamically set the relocation domain for the virtual guest (Example 5-13).

*Example 5-13 Set relocation domain for a Linux guest*

```
set vmrelocate itsolnx1 on domain DMNZ196
```

Assign the virtual guest permanently to its appropriate relocation domain in its directory entry so that its virtual architecture level is determined at logon time. Example 5-14 on page 49 shows the **dirmaint** command to do this.

---

*Example 5-14 Permanent setting of a relocation domain for a Linux guest*

---

```
dirmaint for itsolnx1 vmrelocate on domain DMNZ196
```

---

If the domain is changed dynamically at a later time, the server's virtual architecture level might be changed during the operation of the server, with unpredictable results.

The output from `/proc/cpuinfo` indicates that the guest is now running with the features of the System z z196.

---

*Example 5-15 `cpuinfo` of Linux guests in the DMNZ196 domain*

---

```
itsolnx1:~ # cat /proc/cpuinfo
vendor_id      : IBM/S390
# processors   : 1
bogomips per cpu: 14367.00
features       : esan3 zarch stfle msa ldisp eimm dfp etf3eh highgprs
processor 0: version = FF, identification = 2A3BD5, machine = 2817
```

---

The relocation of the Linux guest inside of the domain DMNZ196 (for example, from ITSOSI3 to ITSOSI4) works without any additional force options, as shown in Example 5-16.

---

*Example 5-16 `VMRELOCATE` in DMNZ196 domain*

---

```
vmrelocate test itsolnx1 itsossi4
User ITSOLNX1 is eligible for relocation to ITSOSI4
Ready; T=0.01/0.01 14:03:12
```

---

**Note:** If you have a cluster of SSI members with different architecture levels, consider the following issues:

- ▶ If the Linux guest must use the architecture features of the newest level of architecture, create a separate relocation domain and set this as the relocation domain for the Linux guest.
- ▶ Performing a relocation without losing any architecture features is only possible if you have at least two members in the cluster with the same architecture level in the same relocation domain.

## 5.5 Relocation to a different architecture domain

This example uses the same environment described in 5.4, “Using relocation domains to reflect the architecture level” on page 47. The only difference is that the Linux guest is relocated out of its defined architecture domain, as shown in Figure 5-4 on page 50.

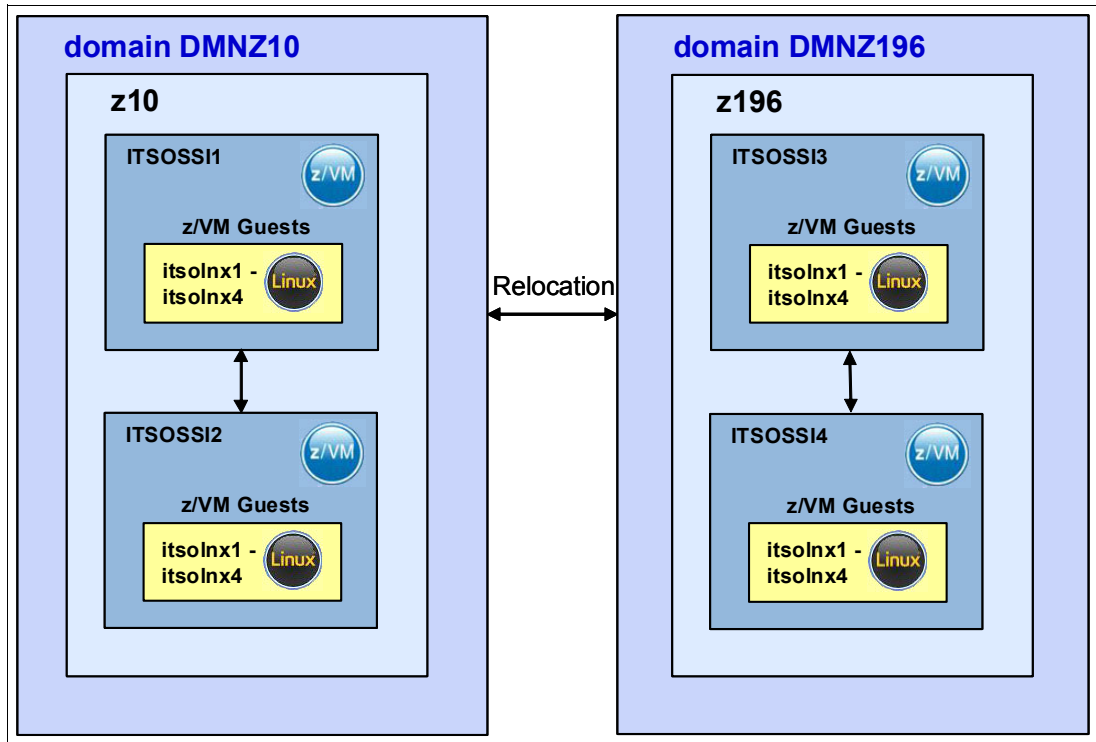


Figure 5-4 Relocation out of the architecture domain

For ITSOLNX1, the relocation domain DMNZ196 is enabled (Example 5-17).

*Example 5-17 Relocation domain*

```
q vmrelocate itsolnx1
Running on member ITSOSI4
Relocation enabled in Domain DMNZ196
Ready; T=0.01/0.01 13:31:13
```

If we try to relocate our Linux guest ITSOLNX1 to the DMNZ10 domain, we get the relocation messages shown in Example 5-18.

*Example 5-18 Relocation out of the domain failed*

```
vmrelocate move itsolnx1 itsossi1
Relocation of user ITSOLNX1 from ITSOSI4 to ITSOSI1 did not complete. Guest has not been moved
HCPRLH1940E ITSOLNX1 is not relocatable for the following reason(s):
HCPRL1944I ITSOLNX1: Architecture incompatibility
Ready(01940); T=0.01/0.01 14:50:42
```

If we use the force architecture or force domain option, we get the messages shown in Example 5-19.

*Example 5-19 Relocation with force architecture or force domain option*

```
vmrelocate move itsolnx1 itsossi1 force architecture
Relocation of user ITSOLNX1 from ITSOSI4 to ITSOSI1 did not complete. Guest has not been moved
HCPRLH1940E ITSOLNX1 is not relocatable for the following reason(s):
```

```
HCPRL1944I ITSOLNX1: Architecture incompatibility
Ready(01940); T=0.01/0.01 14:52:38
vmrelocate move itsolnx1 itsossi1 force domain
Relocation of user ITSOLNX1 from ITSOSI4 to ITSOSI1 did not complete. Guest has not been moved
HCPRLH1940E ITSOLNX1 is not relocatable for the following reason(s):
HCPRL1944I ITSOLNX1: Architecture incompatibility
Ready(01940); T=0.01/0.01 14:52:50
```

---

The solution is to use the force domain architecture option, as shown in Example 5-20.

*Example 5-20 Relocation with the force domain architecture option*

---

```
vmrelocate move itsolnx1 itsossi1 force domain architecture
Relocation of ITSOLNX1 from ITSOSI4 to ITSOSI1 started with FORCE ARCHITECTURE
DOMAIN in effect
User ITSOLNX1 has been relocated from ITSOSI4 to ITSOSI1
Ready; T=0.01/0.01 14:57:09
```

---

**Important:**

- ▶ An out-of-domain relocation to a target member that is running on another architecture level enforces the relocation option force domain architecture.
- ▶ *Use this option with caution.* It can cause unpredictable results in the Linux guest.

## 5.6 Using relocation domains with special hardware or software features

Using relocation domains with licensed hardware, such as Floating Point Extension Facility (FLOAT-PT), or third-party software that requires a license for a CPUID, can cause a problem in relocation.

Define domains for members with the same characteristics. When trying to relocate the Linux guest, ensure that all the features that are used in one member will be present in the other member, and that the relocation can be done without using the **force** option.

**Note:** It is best to avoid relocation using the force architecture or domain option. The Linux guest would not be able to use these hardware or software features.

Licensing of products and features must be enabled in all SSI members which belong to the same relocation domain.

**Note:** Ensure that each product is licensed for all members of the relocation domain.

## 5.7 Using relocation domains for different business purposes

In this example, we are defining our relocation domains according to the business purpose served by the Linux guest. We separated the environment into a development domain and a production domain:

- ▶ DOMAIN 1 - DMNDEV - Development environment
- ▶ DOMAIN 2 - DMNPROD - Production environment

We defined the following members to domain DMNDEV:

- ▶ ITSOSI1 running on a System z z10
- ▶ ITSOSI3 running on a System z z196

We defined the following members to domain DMNPROD:

- ▶ ITSOSI2 running on a System z z10
- ▶ ITSOSI4 running on a System z z196

This environment is depicted in Figure 5.5.

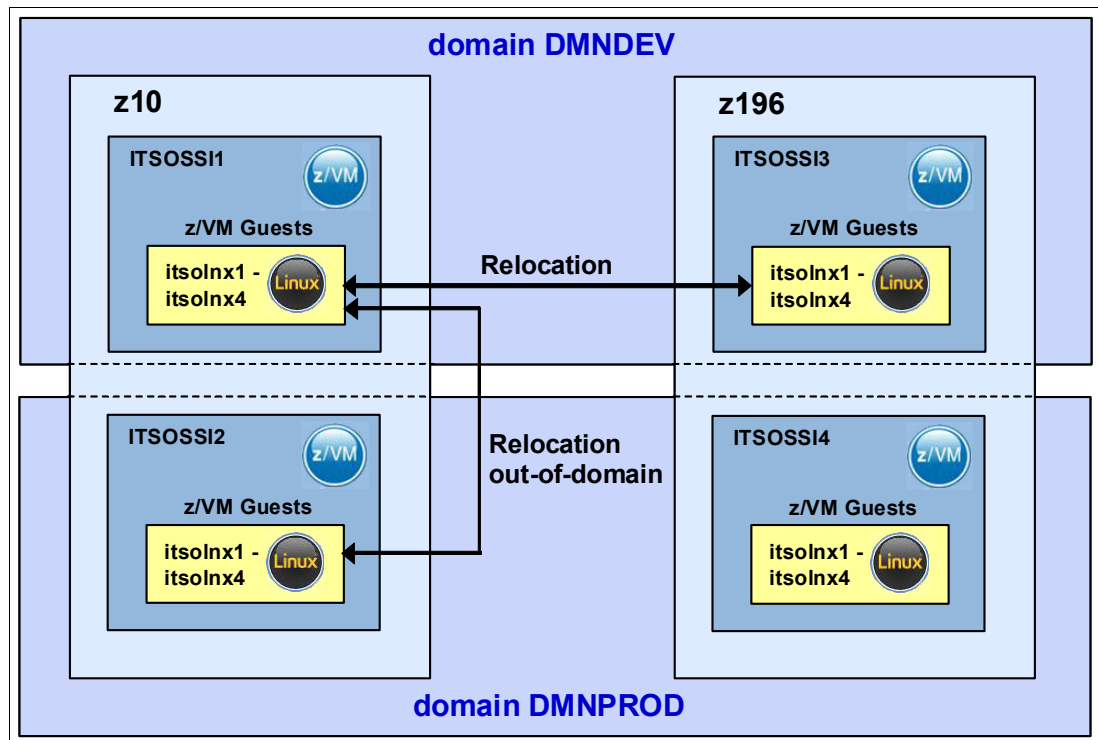


Figure 5-5 Relocation domains for different business purposes

1. We dynamically defined the new relocation domains using the commands shown in Example 5-21.

*Example 5-21 Defining DEV and PROD relocation domains*

```
define relodomain DMNDEV member ITSOSI1 ITSOSI3
Ready; T=0.01/0.01 12:16:45
define relodomain DMNPROD member ITSOSI2 ITSOSI4
Ready; T=0.01/0.01 12:17:05
```

2. We displayed the domains we have defined, as shown in Example 5-22.

*Example 5-22 QUERY RELODOMAIN*

---

```
q reلودomain
DOMAIN DMNPROD  MEMBERS: ITS0SSI2 ITS0SSI4
DOMAIN DMNDEV   MEMBERS: ITS0SSI1 ITS0SSI3
DOMAIN DMNZ196  MEMBERS: ITS0SSI3 ITS0SSI4
DOMAIN DMNZ10   MEMBERS: ITS0SSI1 ITS0SSI2
DOMAIN ITS0SSI1 MEMBERS: ITS0SSI1
DOMAIN ITS0SSI2 MEMBERS: ITS0SSI2
DOMAIN ITS0SSI3 MEMBERS: ITS0SSI3
DOMAIN ITS0SSI4 MEMBERS: ITS0SSI4
DOMAIN SSI      MEMBERS: ITS0SSI1 ITS0SSI2 ITS0SSI3 ITS0SSI4
```

---

3. We dynamically added the guests to each domain, as shown in Example 5-23.

*Example 5-23 SET VMRELOCATE*

---

```
set vmrelocate user itsolnx1 domain DMNDEV
Running on member ITS0SSI1
Relocation enabled in Domain DMNDEV
```

---

**Hint:** In z/VM, you can always define the domain dynamically by issuing the commands, or define it permanently in the user directory. If z/VM is IPLed, the dynamic definition is lost.

Example 5-24 shows our directory entry for ITSOLNX1.

*Example 5-24 User ITSOLNX1 directory definition*

---

```
USER ITSOLNX1 XXXXXXXX 4G 4G G
DVHRXV3355I The following records are included from profile: LINDFLT
  PROFILE LINDFLT
  CLASS G
  COMMAND CP SPOOL CONSOLE START TO VMLOGS
  COMMAND CP TERM LINEND %
  COMMAND CP SET PF12 RETRIEVE
  COMMAND CP TERM MORE 1 0
  COMMAND CP SET RUN ON
  COMMAND CP TERM HOLD OFF
  COMMAND CP SET OBSERVER OPERATOR
  OPTION CHPIDV ONE
VMRELOCATE ON DMNDEV
  CONSOLE 0009 3215 T
  NICDEF C200 TYPE QDIO LAN SYSTEM ITS0VSW1
  SPOOL 000C 2540 READER *
  SPOOL 000D 2540 PUNCH A
  SPOOL 000E 1403 A
```

---

4. We issued the relocation command to move the Linux guest from ITS0SSI1 running in a System z z10 to ITS0SSI3 in a System z z196, both of which are in the same DMNDEV domain (Example 5-25 on page 54).

*Example 5-25 VMRELOCATE MOVE*

---

**vmrelocate move itsolnx1 to itsossi3**

Relocation of ITSOLNX1 **from ITS0SSI1 to ITS0SSI3** started

User ITSOLNX1 has been relocated from ITS0SSI1 to ITS0SSI3

---

**Note:** Although both LPARs are in different System z processors with different architecture levels, DMNDEV relocation domain sets the common architecture level between the two SSI members.

5. If we try to relocate the Linux guest from ITS0SSI1 (z10) in DMNDEV domain to ITS0SSI2 (z10) in DMNPROD domain, the message shown in Example 5-26 is displayed.

*Example 5-26 Relocate from DMNDEV to DMNPROD failed*

---

**vmrelocate move itsolnx1 itsossi2**

Relocation of user ITSOLNX1 **from ITS0SSI1 to ITS0SSI2** did not complete. Guest has not been moved

HCPRLH1940E ITSOLNX1 is not relocatable for the following reason(s):

HCPRL1944I ITSOLNX1: **Architecture incompatibility**

---

6. If we want to force the guest to relocate, we can use the force domain option, as shown in Example 5-27.

*Example 5-27 Relocate with the force domain option*

---

**vmrelocate move itsolnx1 itsossi2 force domain**

Relocation of ITSOLNX1 **from ITS0SSI1 to ITS0SSI2** started with FORCE DOMAIN in effect

User ITSOLNX1 has been relocated from ITS0SSI1 to ITS0SSI2

---

7. If we try to relocate a Linux guest from ITS0SSI1 (z10) in the DMNDEV domain to ITS0SSI4 (z196) in the DMNPROD domain and they have different architecture levels, we get the message shown in Example 5-28.

*Example 5-28 Relocation failing due to architecture incompatibility*

---

**vmrelocate move itsolnx1 to itsossi4**

Relocation of user ITSOLNX1 **from ITS0SSI1 to ITS0SSI4** did not complete. Guest has not been moved

HCPRLH1940E ITSOLNX1 is not relocatable for the following reason(s):

HCPRL1944I ITSOLNX1: **Architecture incompatibility**

---

8. To perform the relocation from ITS0SSI1 (z10) in the DMNDEV domain to DMNPROD (z196), we used the force domain option because they are in different domains. See Example 5-29

*Example 5-29 Relocate from DMNDEV to DMNPROD with force domain option*

---

**vmrelocate move itsolnx1 to itsossi4 force domain**

Relocation of ITSOLNX1 **from ITS0SSI1 to ITS0SSI4** started with FORCE DOMAIN in effect

User ITSOLNX1 has been relocated from ITS0SSI1 to ITS0SSI4

---

**Explanation:** If you have a mixture of System z hardware, such as a zEnterprise or a z10 in your relocation domain, you will have the architecture level set to the level of z10 architecture. Because each domain defined had one System z z10 and one System z z196, both domains were set at the common architecture level of a z10.

## 5.8 Upgrading your environment

We did not build this example in our environment. Nevertheless, this section describes how to upgrade your environment using the SSI and LGR functionality of z/VM 6.2. In our example, we explain the move from an SSI cluster with a combined System z z10 and System z z196 environment to a pure System z z196 environment. For details about adding members to an SSI cluster or removing members from an SSI cluster, see *z/VM CP Planning and Administration version 6 release 2*, SC24-6178 found at:

<http://publibz.boulder.ibm.com/epubs/pdf/hcsg0c10.pdf>

To move from an SSI cluster with a combined System z z10 and System z z196 environment to a pure System z z196 environment, follow these steps:

1. We start with following environment: one member on System z z10 and one member on System z z196. The domain SSI supports the common architecture level of both guests, which means it supports the architecture level of the System z z10 (Figure 5-6).

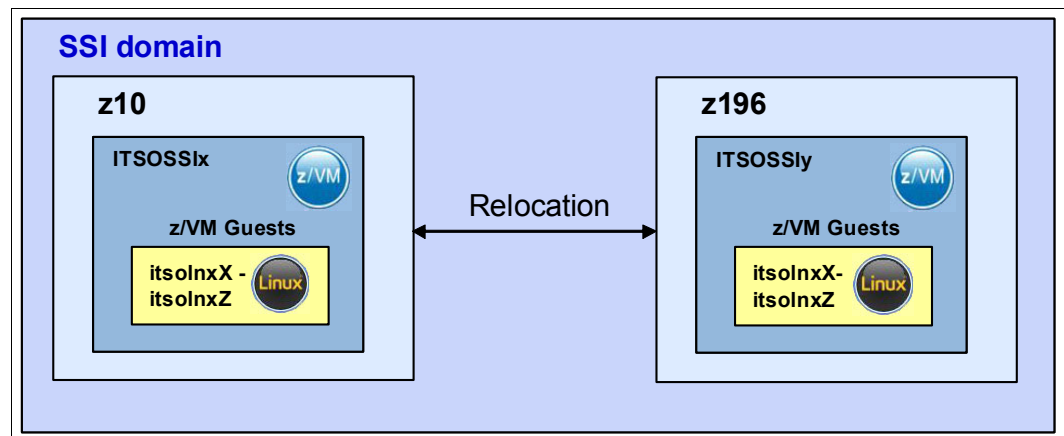


Figure 5-6 Combined z10 and z196 environment

2. Add the new System z z196 member to the SSI cluster. The new member will also support the architecture level of System z z10 (Figure 5-7 on page 56).

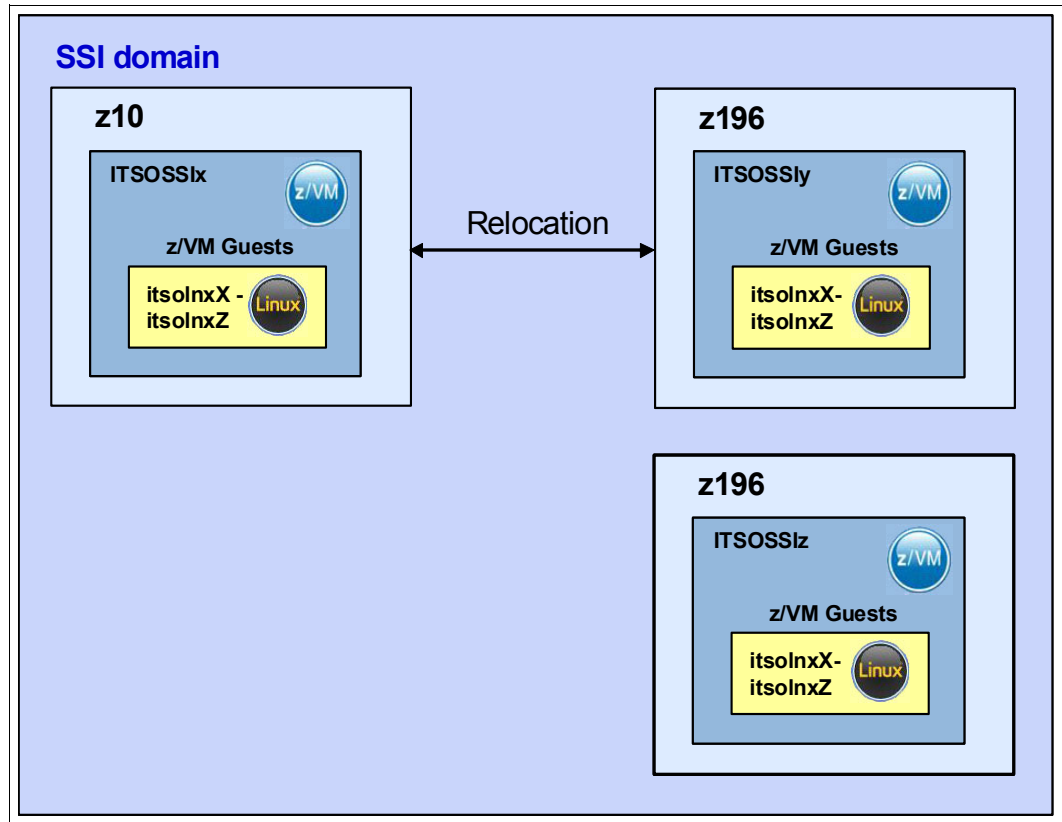


Figure 5-7 Add z196 member to cluster

3. Prepare guests for running on the System z z196 hardware (Figure 5-8 on page 57).

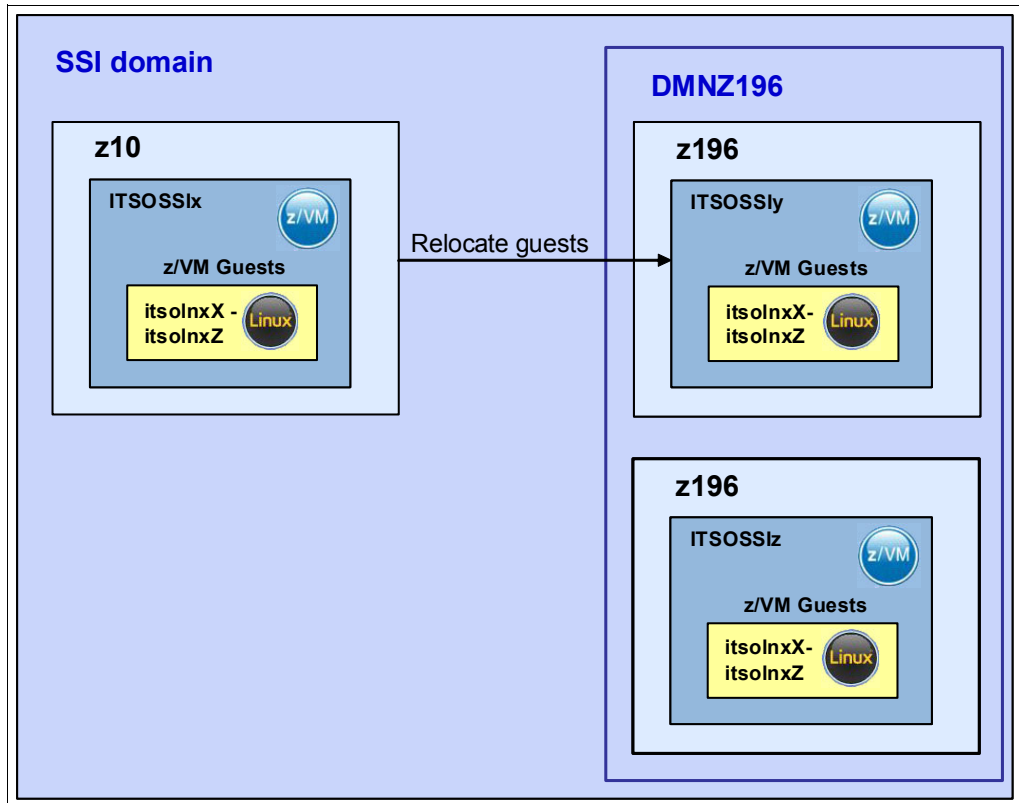


Figure 5-8 Prepare relocation environment for running on z196 environment

- a. Relocate all of the guests to the z196 members using the command in Example 5-30.

*Example 5-30 VMRELOCATE command*

```
vmrelocate move itsolnxX ITS0SSIy
```

- b. Create a DMNZ196 relocation domain containing both z196 members using the command in Example 5-31. This domain only contains z196 members and therefore it supports the z196 architecture.

*Example 5-31 Create the new relocation domain DMNZ196*

```
define relodomain DMNZ196 member ITS0SSIy ITS0SSIz
```

- c. Define the Linux guests in the new to DMNZ196 relocation domain using the command in Example 5-32. This step is done to prevent a relocation back to the System z z10 member.

*Example 5-32 SET VMRELOCATE command*

```
set vmrelocate itsolnxX on domain DMNZ196
```

4. Remove the System z z10 member from the cluster, as shown in Figure 5-9 on page 58. As soon as you remove this member, the PDR will automatically be updated so that the domain SSI will now also support the System z z196 architecture features.

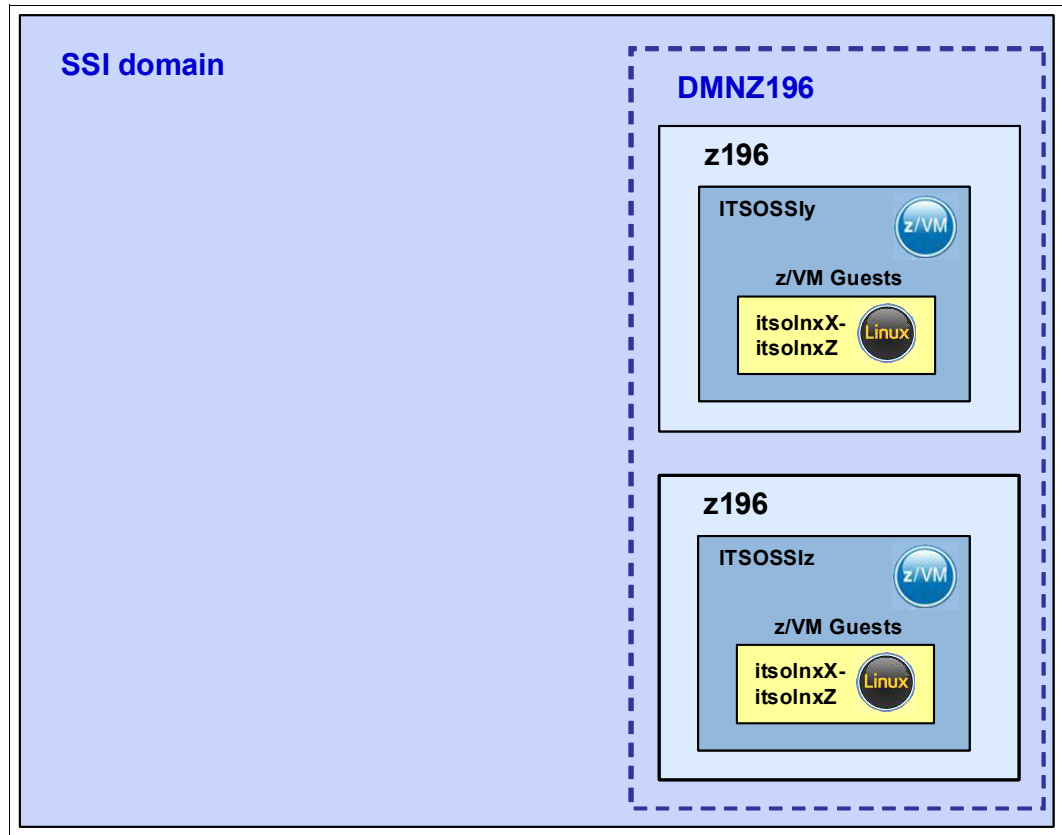


Figure 5-9 Environment without z10 member

5. Optional: Define the Linux guests back to the default relocation domain, SSI, and remove the DMNZ196 domain. This is possible because the new SSI domain also supports the System z z196 architecture.
6. Optional: Reboot the Linux Guests to use the z196 architecture features. If you really want to use the architecture features from z196, we suggest rebooting the Linux guests.



## Performance topics

In this chapter, we describe SSI-specific performance topics, including:

- ▶ Installation and setup of the IBM Performance Toolkit for VM feature offered with z/VM and used in an SSI environment
- ▶ New data screens for SSI in the IBM Performance Toolkit for VM
- ▶ An introduction to performance aspects of LGR and a description of how relocation time can be monitored

## 6.1 Install and set up the IBM Performance Toolkit for VM in an SSI environment

As described in *An introduction to z/VM Single System Image (SSI) and Live Guest Relocation (LGR)*, SG24-8006, the IBM Performance Toolkit for VM was installed using the instructions in the *Program Directory for Performance Toolkit for VM* (found at <http://www.ibm.com/eserver/zseries/zvm/library/>) and *z/VM Getting Started with Linux on System z*, SC24-6194 (<http://publib.boulder.ibm.com/infocenter/zvm/v6r1/topic/com.ibm.zvm.v610.hcp10/hc5x0c00.htm>).

For performance tasks, we set up the following IBM Performance Toolkit functionality:

- ▶ Performance Toolkit web interface
- ▶ Monitoring Linux virtual servers with the IBM Performance Toolkit for VM
- ▶ Monitoring the different z/VM members of the SSI cluster from one central GUI interface

### 6.1.1 Activate the IBM Performance Toolkit web interface

Use the following steps to activate the IBM Performance Toolkit web interface:

1. Open PORT 81 in TCPIP CONFIG (Example 6-1).

*Example 6-1 Open PORT 81 for Performance Toolkit*

---

```
...  
81    TCP PERFSVM                ; Performance Toolkit  
...
```

---

2. The FCONX \$PROFILE of user PERFSVM is used to customize the Performance Toolkit operation. The FCONX \$PROFILE contains the statements that must be activated to establish the Performance Toolkit web interface. We used the LOCALMOD process, as instructed in the *Program Directory for Performance Toolkit for VM for use with z/VM version 6 release 2*, GI11-4351-00, to make the changes in the FCONX \$PROFILE. With the LOCALMOD process, changes to the FCONX \$PROFILE will automatically become common to all members of the cluster.

For changes in the FCONX \$PROFILE, see Example 6-2.

*Example 6-2 Activate Performance Toolkit web interface in FCONX \$PROFILE*

---

```
*    Following command activates VMCF data retrieval interface  
FC MONCOLL VMCF ON  
*    Define the maximum allowed number of Internet connections  
FC MONCOLL WEBSERV MAXCONN 100  
*    Define the timeout of inactive Internet connections in minutes  
FC MONCOLL WEBSERV TIMEOUT 30  
*    Following command activates Internet interface  
FC MONCOLL WEBSERV ON TCPIP TCPIP 81  
*    Following command activates Internet interface with SSL  
*C MONCOLL WEBSERV ON SSL TCPIP TCPIP 81
```

---

3. Start the Performance Toolkit web interface using a host name or TCPIP address. In our example we used `http://9.12.4.23x:81`.

## 6.1.2 Activate Linux guest monitoring by IBM Performance Toolkit

Follow these steps to activate the monitoring of Linux guests:

1. Activate the FCONX statement for data retrieval from the LINUX RMF™ DDS interface in the FCONX \$PROFILE using the LOCALMOD process, as shown in Example 6-3.

*Example 6-3 Activate TCP/IP interface for data retrieval from LINUX RMF DDS interface*

---

```
* Following command activates TCP/IP interface for data retrieval
* from LINUX RMF DDS interface
FC MONCOLL LINUXUSR ON TCPIP TCPIP
```

---

2. Update FCONX LINUXUSR on the PERFSVM.191 disk with all the Linux guests that are to be monitored, as shown in Example 6-4.

*Example 6-4 FCONX LINUXUSR definition*

---

```
ITSOLNX1 9.12.4.141:8803
ITSOLNX2 9.12.4.140:8803
ITSOLNX3 9.12.4.228:8803
ITSOLNX4 9.12.4.229:8803
```

---

3. Provide a modular data gatherer for Linux. The gathered data can be analyzed using the RMF PM client application or the z/OS Management Facility. The performance data is accessible through XML over HTTP, so you can easily exploit it in your own applications. As of October 2011, support for the Linux **rmfpms** agent has been withdrawn, but continues to be available on an as-is basis. It can be found at:

<ftp://public.dhe.ibm.com/eserver/zseries/zos/rmf/>

- a. Using PuTTY, log into your Linux guest.
- b. Change the directory to /opt.  

```
# cd /opt/
```
- c. Extract rmfpms\_s390x\_kernel26.tgz.  

```
# tar -zxvf rmfpms_s390x_kernel26.tgz
```
- d. Change directories to the rmfpms directory.  

```
# cd rmfpms/
```
- e. Before you can start **rmfpms**, you must configure it using the file named .rmfpms\_config found in /opt/rmfpms. Note the “.” preceding the file name.

```
#vi .rmfpms_config
```

Change \$HOME to /opt in two of the environment variables in the file, as shown in Example 6-5.

*Example 6-5 Modify the configuration file*

---

```
export IBM_PERFORMANCE_REPOSITORY=$HOME/rmfpms/.rmfpms <===replace $HOME with /opt
export IBM_PERFORMANCE_HOME=$HOME/rmfpms/bin/ <===replace $HOME with /opt
export IBM_PERFORMANCE_MINTIME=60
export LD_LIBRARY_PATH=$IBM_PERFORMANCE_HOME:$LD_LIBRARY_PATH
export APACHE_ACCESS_LOG=/var/log/httpd/access_log
export APACHE_SERVER=localhost
export APACHE_SERVER_PORT=80
```

---

- f. You should now be able to start **rmfpms** in the /bin directory with the following commands. The output is shown in Example 6-6.

```
# cd bin
#./rmfpms start
```

*Example 6-6 Starting rmfpms*

---

```
Starting performance gatherer backends ...
DDSRV: RMF-DDS-Server/Linux-Beta (Sep 8 2007) started.
DDSRV: Functionality Level=2.339
DDSRV: Reading exceptions from gpmexsys.ini and gpmexusr.ini.
DDSRV: Server will now run as a daemon process.
done!
```

---

4. As soon as **rmfpms** is running, you can view the performance data in either of the following ways:
  - Point your browser to the Linux guest IP address and port 8803, for example:  
`http://9.12.4.x:8803`
  - Go to the ip address of the IBM Performance Toolkit web interface:  
`http://9.12.4.x:81/`  
Then select Option 29 → RMF PM system selection menu → ITSOLNX1

### 6.1.3 Monitoring multiple z/VM members from a central monitor machine

In our environment, we used the z/VM member ITS0SSI4 (9.12.4.232) for the four member cluster and ITS0SSI6 (9.12.4.237) for the two member cluster as the central monitor interface for the Performance Toolkit. From these two central monitoring points, we can monitor the related z/VM members and their Linux guests.

We used the steps described in this section to configure monitoring in our four member cluster. Similar steps were required to configure monitoring in the two member cluster, but those steps are not detailed here.

The following configuration was done on ITS0SSI4, the central point of view for performance data on our four member cluster:

1. Create the file **FCONRMT SYSTEMS A** on the central monitor machine ITS0SSI4 with a list of all the remote systems to be monitored, as shown in Example 6-7.

*Example 6-7 FCONRMT SYSTEMS for the central monitor machine*

---

*Node-id	Userid	VM-Type	Append	Nickname
ITS0SSI1	PERFSVM	z/VM6.2	N	FCXC1R01
ITS0SSI2	PERFSVM	z/VM6.2	N	FCXC1R02
ITS0SSI3	PERFSVM	z/VM6.2	N	FCXC1R03
ITS0SSI4	PERFSVM	z/VM6.2	N	FCXC1R04

---

2. Set up a file called **FCONRMT AUTHORIZ A** on the central monitor machine on ITS0SSI4. This is the authorization file for the remote data retrieval facility. The relevant entries in the **FCONRMT AUTHORIZ** file must be made using the **SYSTEMID** of the target system. Our **FCONRMT AUTHORIZ** file on ITS0SSI4 is shown in Example 6-8.

*Example 6-8 FCONTRMT AUTHORIZ for the central monitor machine*

```
*****
* AUTHORIZATION FILE FOR LOCAL AND REMOTE PERFORMANCE DATA *
* RETRIEVAL AND COMMAND EXECUTION *
*****
ITSOSS14 MAINT DATA CMD EXCPMSG
*
*****
* ALLOWS ANYONE FROM A SPECIFIC SYSTEM TO REQUEST DATA FROM THE *
* ID RUNNING PERFKIT *
*****
ITSOSS14 PERFSVM S&FSERV DATA
ITSOSS14 PERFSVM CMD DATA
ITSOSS13 PERFSVM CMD DATA
ITSOSS12 PERFSVM CMD DATA
ITSOSS11 PERFSVM CMD DATA
```

3. Update the file FCONX LINUXUSR on PERFSVM.191 disk with all the Linux guests that are to be monitored. Example 6-4 on page 61 shows the file we used. (Skip this step if you updated the file previously.)
4. Activate the communications interface by uncommenting the FC MONCOLL VMCF ON entry in the FCONX \$PROFILE (Example 6-9).

*Example 6-9 Activate communication interface in FCONX \$PROFILE.*

```
* Following command activates VMCF data retrieval interface
FC MONCOLL VMCF ON
```

For further information regarding the implementation of a central monitor interface, see *z/VM V6R2.0 Performance Toolkit Guide*, SC24-6209-02.

Figure 6-1 shows the Performance Toolkit web interface overview panel that is displayed when the configuration is completed successfully.

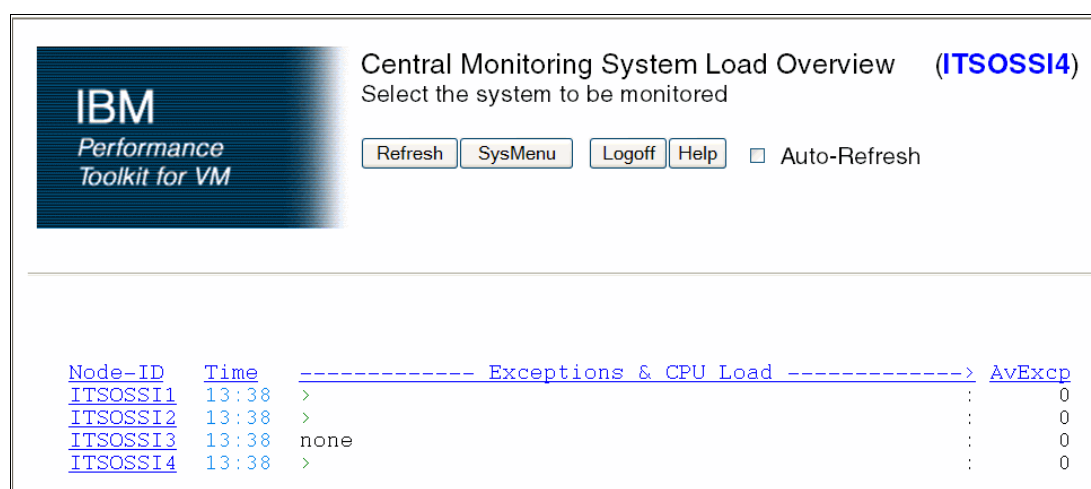


Figure 6-1 Central Monitoring System Load Overview

In our environment, we prepared ITSOSS14 as the central monitor point, so if we choose ITSOSS14, we get the panel shown in Figure 6-2 on page 64.

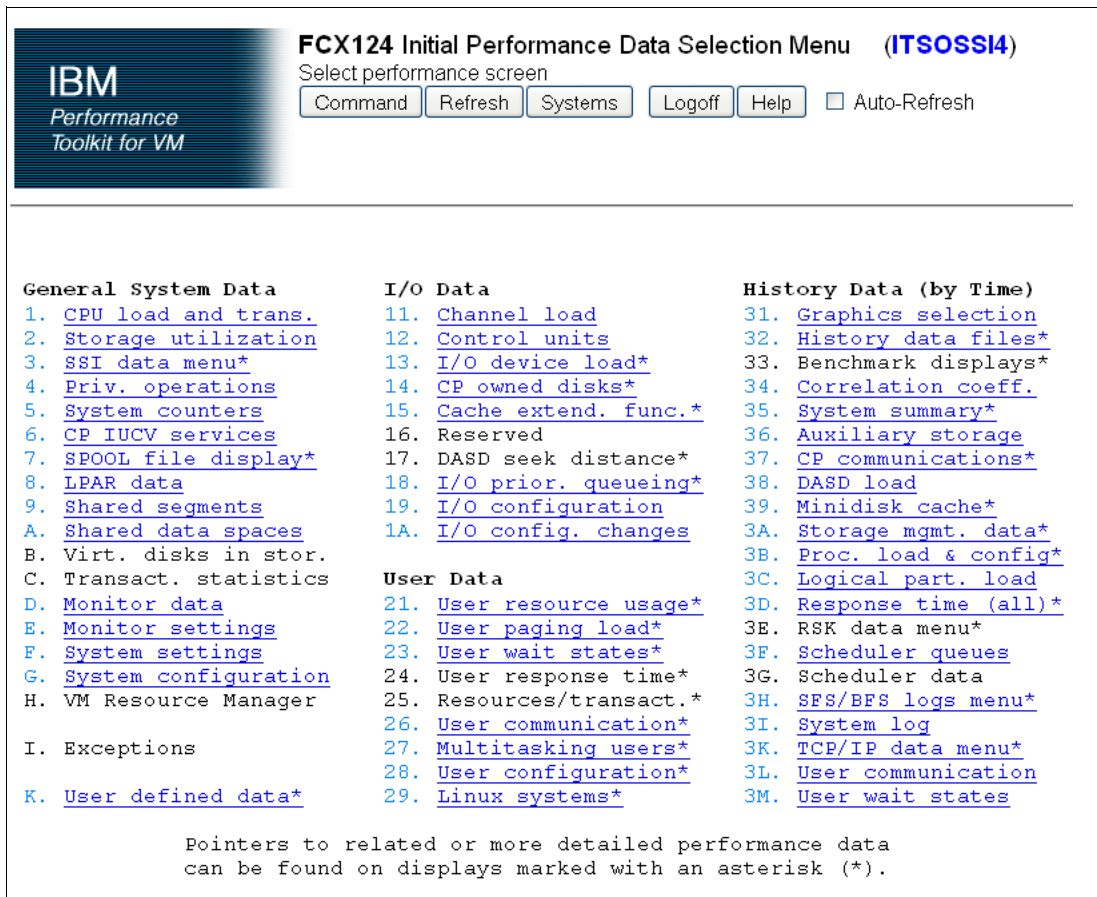


Figure 6-2 Initial Performance Data Selection Menu for ITSOSI4

The Linux Performance Data Selection Menu shown in Figure 6-3 was displayed when we selected option 29.

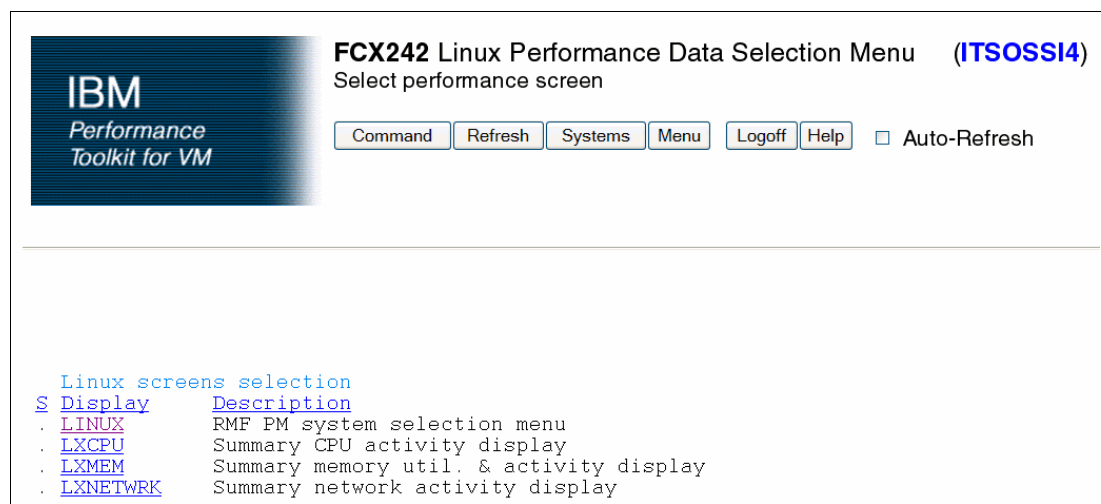


Figure 6-3 Linux Performance Data Selection Menu

We selected RMF PM system selection menu to get the performance data for the various Linux guests, as shown in Figure 6-4 on page 65.

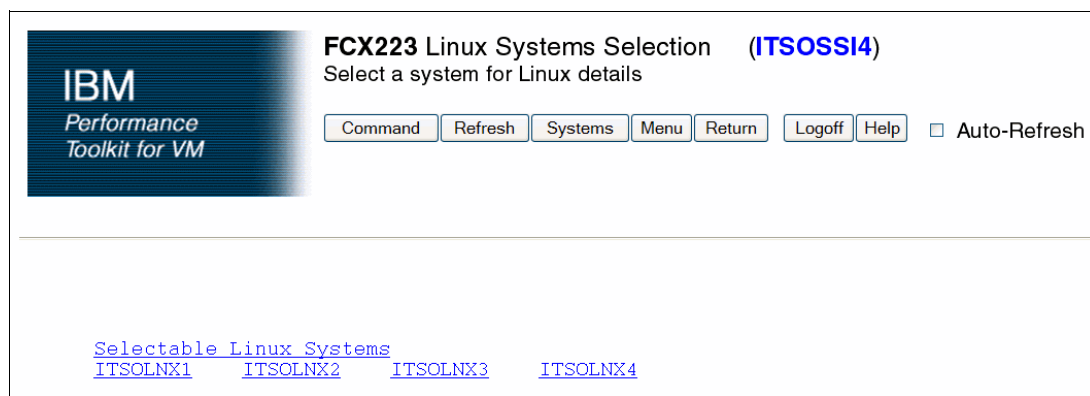


Figure 6-4 Linux Systems Selection menu

## 6.2 Monitoring SSI-relevant data in z/VM 6.2

The z/VM v6.2 IBM Performance Monitor Toolkit for VM provides SSI-relevant information. In this section, we describe the activation and the data screens that were built for SSI data.

### 6.2.1 Activate monitoring of SSI data

To see the complete set of SSI data provided in the IBM Performance Monitor Toolkit for VM, you must enable EVENT DATA collection for the SSI domain and also SAMPLE DATA collection. Example 6-10 shows the required changes to the PROFILE EXEC of the PERFSVM userid.

Example 6-10 Additional entries in PROFILE EXEC of PERFSVM userid

```
'CP MONITOR SAMPLE ENABLE SSI'
'CP MONITOR EVENT  ENABLE SSI'
```

After these changes are made, the SSI monitor settings are displayed, as shown in Example 6-11, and found in the Performance Monitor, Option E Monitor Settings. Note that Example 6-11 shows the display obtained using the **MONITOR** command from the PEFSVM userid, not from the web interface.

Example 6-11 PERFSVM monitor settings

FCX149 Monitor Settings: Initial and Changed				(ITS0SSI4)		
Initial Settings		<----- Active ----->		<--- Seconds --->		
2012/04/16 14:00			High	Sample	HF	
Nr	Domain	Event	Sample	Frequency	Interval	Rate
0	SYSTEM	---	YES	YES	60	2.00
1	MONITOR	YES	YES	---	60	---
2	SCHEDULER	NO	---	---	---	---
3	STORAGE	YES	YES	---	60	---
4	USER	YES	YES	YES	60	2.00
5	PROCESSOR	YES	YES	YES	60	2.00
6	I/O	YES	YES	YES	60	2.00
7	SEEK	NO	---	---	---	---
8	NETWORK	NO	YES	YES	60	2.00

9	ISFC	YES	YES	---	60	---
10	APPLDATA	YES	YES	---	60	---
11	SSI	YES	YES	---	60	---

Changed Monitor Settings  
Date Time Command Line  
..... No commands entered

---

## 6.2.2 New performance data screens in z/VM 6.2 supporting the SSI function

New performance data screens supporting the single system image function are included with z/VM v6.2. The data is accessed from Option 3, SSI data menu\* on the Initial Performance Data Selection Menu.

The SSI data menu (FCX271) shown in Example 6-12 is obtained by issuing the **MONITOR** command from the PEFSVM userid (again, not from the web interface).

*Example 6-12 SSI data menu*

---

<b>FCX271</b>	<b>CPU nnnn</b>	<b>SER nnnnn</b>	<b>SSI Data Menu</b>	<b>Perf. Monitor</b>
<b>SSI performance reports</b>				
<b>S Command Description</b>				
_ SSICONF SSI configuration				
_ SSISCHLG SSI State Change Synchronization Activity log				
_ SSISMILG SSI State/Mode Information log				
<b>ISFC performance reports</b>				
<b>S Command Description</b>				
_ ISFECONF ISFC End Point configuration				
_ ISFEACT ISFC End Point activity				
_ ISFLCONF ISFC Logical Link configuration				
_ ISFLACT ISFC Logical Link activity state				
_ ISFLALOG ISFC Logical Link activity log				
<b>Command ==&gt; _</b>				
<b>F1=Help F4=Top F5=Bot F7=Bkwd F8=Fwd F12=Return</b>				

---

We do not describe all the data panels in detail here. For further information, see *z/VM Performance Toolkit Reference*, SC24-6210-02.

The following list is a short description of the new data panels and some examples that result from a shutdown (relPL) of an SSI member.

1. ISFC End Point Configuration (panel FCX272)  
The ISFC End Point Configuration panel (FCX272) displays the ISFC end points present on the system. There is one row for each endpoint.
2. ISFC End Point activity (panel FCX273)  
The ISFC End Point Activity panel displays the traffic on ISFC Transport, by EndPoint. There is one row for each endpoint.
3. ISFC Logical Link activity state (FCX274)  
The ISFC Logical Link Activity panel displays ISFC logical link transport activity. There is one row for each ISFC logical link.

4. ISFC Logical Link Configuration (panel FCX275)

The ISFC Logical Link Configuration panel displays the configuration of ISFC logical links and corresponding changes in the configuration status. There is one row for each ISFC logical link. See Figure 6-7 on page 68.

5. SSI configuration (panel FCX276)

The SSI configuration panel displays the SSI configuration of the system and contains information about any changed configuration status. Figure 6-5 shows the layout of this panel from the web interface.

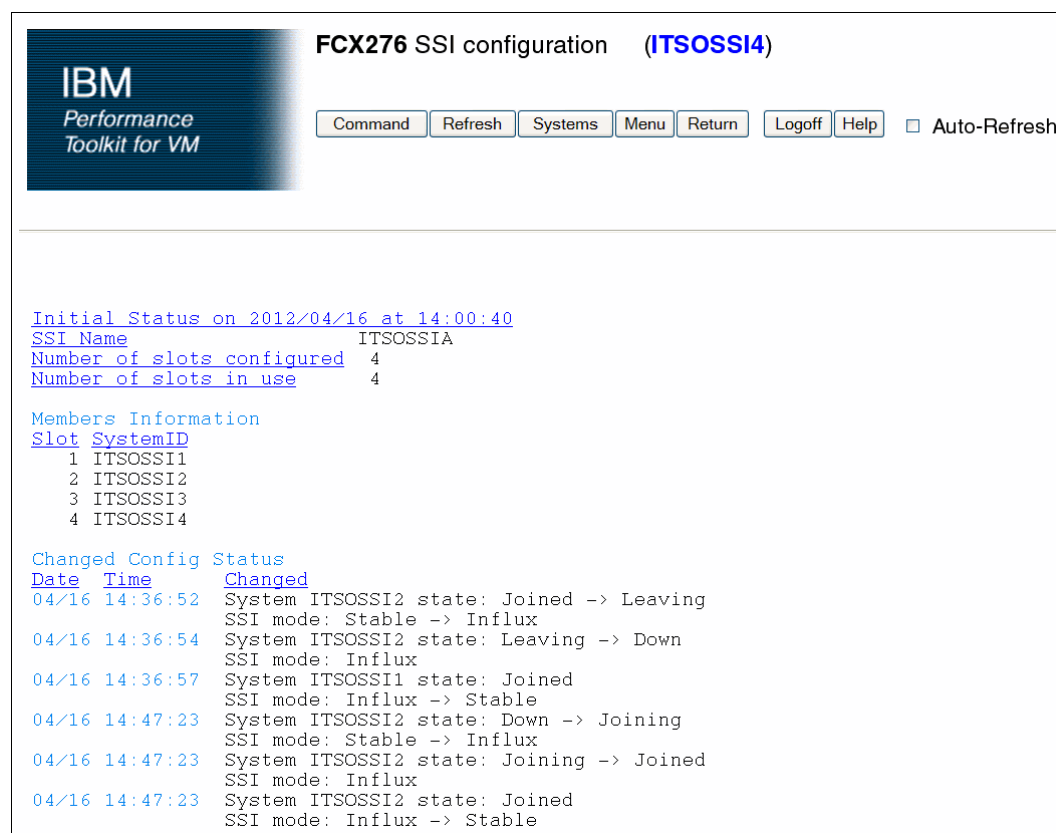


Figure 6-5 SSI configuration data, after shutdown (re-IPL)

6. SSI State Change Synchronization Activity Log (panel FCX277)

The SSI State Change Synchronization Activity Log panel displays the current SSI state change synchronization activity. If you are using the **MONITOR** command in PERSVM, the subcommand is **SSISCHLG**.

7. SSI State/Mode Information Log (panel FCX278)

The SSI State/Mode Information Log panel displays the SSI configuration of the system by time. Figure 6-6 on page 68 shows this panel on the web interface.

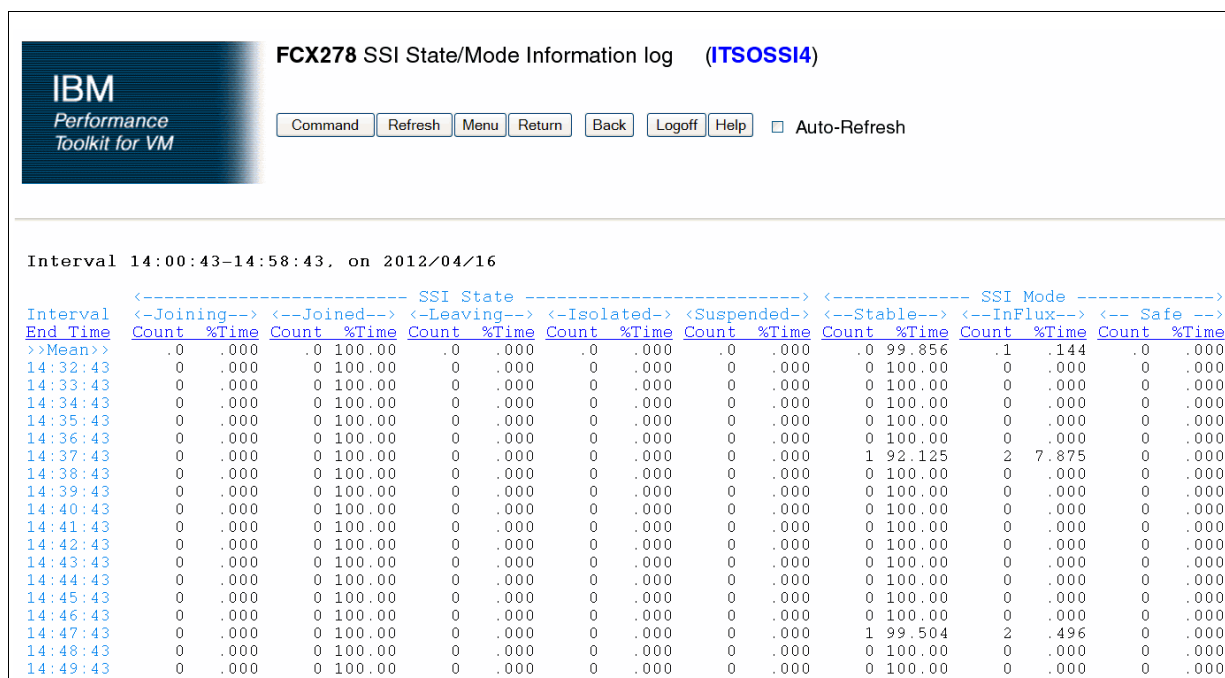


Figure 6-6 State/Mode information log

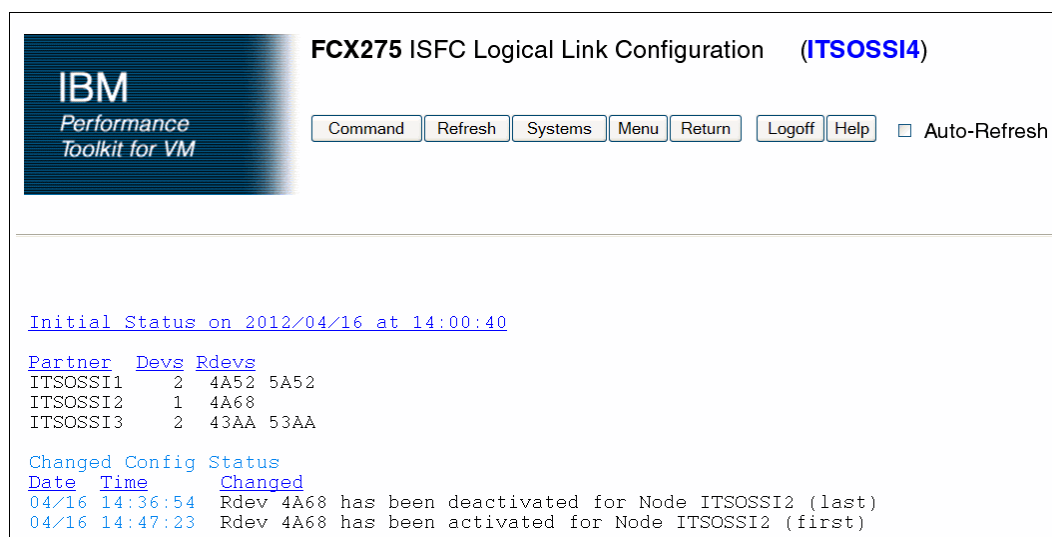


Figure 6-7 ISFC Logical Link Configuration with status changes

## 8. ISFC Logical Link activity log (FCX281)

The ISFC Logical Link Activity By Time Log displays overall performance data for all ISFC Logical Links that exist in the system, by time. Each entry consists of a group of lines for every ISFC Logical Link per interval.

## 6.3 Introduction to performance aspects of LGR

For live guest relocation (LGR), two key values of relocation performance are relevant: quiesce time (QT) and relocation time (RT).

- ▶ *Quiesce time* is the amount of time a relocating guest virtual machine is stopped. Quiesce occurs during the final two passes through storage to move all the guest's pages that have changed since the previous pass. It is important to minimize the quiesce time because the guest is not running for this length of time. Certain applications might have a limit on the length of quiesce time that they can tolerate and still resume running normally after relocation.
- ▶ *Relocation time* is the amount of elapsed time between issuing the VMRELOCATE command and the successful restart of the guest virtual machine on the destination system. The relocation time represents the total time required to complete the relocation of a guest virtual machine. Relocation time can be important because a whole set of relocations might be required during a fixed period of time, such as a maintenance window.

As described in *An introduction to z/VM Single System Image (SSI) and Live Guest Relocation (LGR)*, SG24-8006, the following factors can influence relocation time:

- ▶ Virtual server memory size  
The more memory a Linux guest has, the longer the relocation and quiesce time. During memory movement, CP attempts to relocate all the virtual server's memory in a series of passes, sending on each pass only the memory changed since the last pass. This process continues until internal algorithms determine that no more progress can be made, at which point CP quiesces the virtual server. During quiesce time, CP must relocate the final guest state, I/O information, and changed pages.
- ▶ Virtual machine page change rate  
The rate at which a Linux guest changes its pages in memory has a direct effect on the total relocation time and, possibly, quiesce time. A Linux guest changing pages rapidly has more pages to relocate in each memory pass so the memory move stage lasts longer. In general, the higher a virtual server's page change rate, the greater its relocation time.
- ▶ ISFC setup  
Faster CTC speeds and the number of CTCs defined will increase throughput and result in shorter relocation and quiesce time. All SSI clusters must have direct logical links between all systems. All SSI clusters use ISFC for intra-cluster communication and live guest relocation. ISFC uses CTC devices. Follow the suggestions in *Guidelines for planning your network in an SSI cluster* found at:  
<http://publib.boulder.ibm.com/infocenter/zvm/v6r2/topic/com.ibm.zvm.v620.hcp10/isfcnetw.htm#isfcnetw>
- ▶ Relocation options  
Relocation options can influence the relocation and quiesce time. Options that affect relocation and quiesce times include:
  - SYNC - Synchronous - Only one relocation can be issued at a time by a particular user. Quiesce time is shortest when relocations are done synchronously because the relocations are not competing with each other for system resources. Using the default (SYNC) option on the VMRELOCATE command can help ensure relocations are done serially, especially if you are issuing the VMRELOCATE command in an EXEC.
  - ASYNCH - Asynchronous - Using the ASYNC option on the VMRELOCATE command and issuing multiple relocations at one time might cause undesirably long quiesce times.

- IMMEDIATE - Causes quiesce time to occur after only one pass through memory. This option usually results in shorter overall relocation times, but longer quiesce times because the first pass through memory usually takes the longest, and the virtual server might have changed many pages, which then need to be moved during quiesce time.
- MAXTOTAL - Maximum total relocation time. The default value for MAXTOTAL is no limit.
- MAXQUIESCE - Maximum quiesce time. The default for MAXQUIESCE is 10 seconds.
- Other non-relocation activity
 

Other non-relocation activity on source and destination systems might increase relocation or quiesce time. Constraints on the source or target system might make relocation or quiesce time longer and the addition of more virtual servers might have undesirable effects on the destination system.

### 6.3.1 Monitor records for relocation information

The default z/VM system is installed with a identity user named MONWRITE. This userid is normally used for running the MONWRITE utility, which allows writing the contents of the MONDCSS segment to a file. We also used this userid to collect monitor data and especially the relocation data. The relocation data is stored in the USER domain of the monitor data, so we have to check whether the event data for the USER domain is collected. Therefore, the PROFILE EXEC of user MONWRITE should be similar to that shown in Example 6-13.

The MONITOR commands must be issued from a privileged userid which has PRIVCLASS A or E, so we changed the PRIVCLASS of our MONWRITE user accordingly.

*Example 6-13 PROFILE EXEC of MONWRITE identity*

---

```
/* */
'CP SET RUN ON'
'CP MONITOR SAMPLE CONFIG SIZE 500'
'CP MONITOR SAMPLE ENABLE ALL'
'CP MONITOR EVENT  ENABLE STORAGE'
'CP MONITOR EVENT  ENABLE PROCESSOR'
'CP MONITOR EVENT  ENABLE I/O ALL'
'CP MONITOR EVENT  ENABLE APPLDATA ALL'
''CP MONITOR EVENT  ENABLE USER ALL'
```

---

After recording the monitor data, the MONWSTOP command stops the MONWRITE program from processing and closes the output file.

The IBM Performance Toolkit for VM does not show any information regarding relocation or quiesce times. The Performance Toolkit data area is intended for use by a formatted output collector (such as IBM OMEGAMON® XE on z/VM and Linux). However, if you need to extract data from the monitor data, you can write your own interface program. Information about monitor records can be found in *z/VM V6R2.0 Monitor Records* available at:

<http://www.vm.ibm.com/pubs/mon620/index.html>

Another article to help with this task can be found at:

<http://www.vm.ibm.com/pubs/int620.html>

After extraction of the monitor data, control block contents with information about relocation data can be found. The start of guest relocation can be found in Domain 4, Record 11 of the monitor record; the end of guest relocation can be found in Domain 4, Record 12. Example 6-14 is a sample of our control block information.

*Example 6-14 Control block contents containing the relocation data*

---

```

0 header length      = 010C
2 header zeros       = 0000
4 header domain      = 04
5 header reserved    = 00
6 header record      = 000C
8 header timestamp   = C96F1095C5FA3B8E 10:28:28.861347
16 header reserved   = 00000000
20 USERLE_RLOISSUER  = MASEN
28 USERLE_RLOUSER    = LNXSU11
36 USERLE_RLOSRCSYS  = ITS0SSI5
44 USERLE_RLODSTSYS  = ITS0SSI6
52 USERLE_RLOSTARTM  = C96F1089CD66D8EE 10:28:16.308845
60 USERLE_RLOMAXT    = 0
64 USERLE_RLOMAXQ    = 10
68 USERLE_LCLFLAGS   = 10000000
69 USERLE_RLOMVOPT   = 00011000
70 USERLE_VMDSTRLO   = 00000000
71 USERLE_RLOFINCD   = 00
72 USERLE_RLOVDXCT   = 27
76 USERLE_RLOAIOCT   = 4
80 USERLE_RLONQDCT   = 0
84 USERLE_RLOQDCT    = 4
88 USERLE_RLOMEMPS   = 4
92 USERLE_RLOPASSA   = 603983
100 USERLE_RLOPSAVG  = 42240
108 USERLE_RLOPASSY  = 1730
116 USERLE_RLOPCNT   = 14575
124 USERLE_RLOCONTM  = C96F1089CD920DC0 10:28:16.309536
132 USERLE_RLOELGTM  = C96F1089CE697742 10:28:16.312983
140 USERLE_RLOCRETM  = C96F1089D1189084 10:28:16.323977
148 USERLE_RLOSETTM  = C96F1089D1D93714 10:28:16.327059
156 USERLE_RLOMEMTM  = C96F10943690B130 10:28:27.225355
164 USERLE_RLOFCPTM  = C96F1094369D1EF2 10:28:27.225553
172 USERLE_RLOQUITM  = C96F109436A014FE 10:28:27.225601
180 USERLE_RLOIOCTM  = C96F109495C91A92 10:28:27.615377
188 USERLE_RLOSTATM  = C96F10948C519D2A 10:28:27.576601
196 USERLE_RLOCRYTM  = C96F10948C51ABA0 10:28:27.576602
204 USERLE_RLOVSETM  = C96F1094577A8148 10:28:27.360168
212 USERLE_RLOSMETM  = C96F10945826BEB4 10:28:27.362923
220 USERLE_RLOPENTM  = C96F10949F657FB0 10:28:27.654743
228 USERLE_RLOLSTTM  = C96F1095C31F0A14 10:28:28.849648
236 USERLE_RLOIOETM  = C96F1095C4B20DE6 10:28:28.856096
244 USERLE_RLORESTM  = C96F1095C5F86FEE 10:28:28.861318
252 USERLE_RLOCLNTM  = C96F1095C5FA0B0A 10:28:28.861344
260 USERLE_RLOSRCRSV = 00000000
264 USERLE_RLODSTRSV = 00000000

```

---

With this information and the description of the monitor records in *z/VM V6R2.0 Monitor Records*, we are able to create relocation data, as shown in Example 6-15.

*Example 6-15 Relocation data*

---

----- Relocation Step Durations -----	
Time to establish connection	691 microseconds
Initial eligibility checks	3447 microseconds
Create skeleton	10994 microseconds
Storage management set up	3082 microseconds
Memory transfer (pre-quiesce passes)	10898296 microseconds
FCP I/O quiesce	199 microseconds
Time to quiesce	47 microseconds
I/O relocation	389776 microseconds
Final storage management eligibility checks	137323 microseconds
Penultimate memory pass	291820 microseconds
Final VSIM eligibility checks	134567 microseconds
Machine state relocation	216434 microseconds
Crypto relocation	1 microseconds
Last memory pass	1194905 microseconds
Final I/O eligibility checks	6448 microseconds
Resume guest	5222 microseconds
Clean-up	26 microseconds
Total quiesce time	1635718 microseconds
Total relocation time	12552499 microseconds

---

We used this information for our performance benchmarks in Chapter 7, “Benchmarks for relocating Linux on System z guests using LGR” on page 73.

## 6.4 Sources of additional information

For further performance-related documentation regarding SSI and LGR, see:

- ▶ <http://www.vm.ibm.com/perf/reports/zvm/html/620lgr.html> - Performance aspects of LGR, specifically, quiesce time and the total relocation time, for Linux virtual servers that are relocated within an SSI cluster.
- ▶ <http://www.vm.ibm.com/perf/reports/zvm/html/620isfc.html> - Offers some insight into the inner workings of ISFC and provides some guidance on ISFC logical link capacity estimation.
- ▶ <http://www.vm.ibm.com/pubs/mon620/index.html> - Monitor records, including those that contain information about LGR relocation times.



## Benchmarks for relocating Linux on System z guests using LGR

This chapter describes the scenarios that we used to test the relocation of applications.

We ran tests in the following environments:

- ▶ A two cluster system using relocation options  
See 7.1, “Relocation benchmark dependent on relocation options” on page 74 for these results.
- ▶ A four cluster system with different LPARs and processors  
See 7.3, “Two Linux guests in a four cluster SSI system dependent on different LPARs and processors” on page 78 for these results.
- ▶ On two and four cluster systems  
See 7.4, “Two Linux guests in two and four cluster systems dependent on SCSI and non-SCSI” on page 81 for these results.

The actual applications running on the Linux guests are described in Chapter 3, “Applications setup” on page 15. We set up the applications so that we could conduct a variety of tests and monitor the performance of relocating those guests using LGR.

For our performance tests of the relocation of Linux guests, we used our two member cluster ITSOSIB with the two Linux guests LNXSU11 and LNXRH56. See 2.2, “Overview of our two member cluster” on page 12 for details about this configuration.

We examined the following performance scenarios:

1. Quiesce time and relocation time according to the relocation options SYNCHRONOUS, IMMEDIATE, and ASYNCHRONOUS.
2. Quiesce time and relocation time according to the number of CTCs that were defined in the ISFC setup.

## 7.1 Relocation benchmark dependent on relocation options

We executed the relocations of the two Linux guests with the options SYNCHRONOUS, SYNCHRONOUS IMMEDIATE, and ASYNCHRONOUS, and compared the quiesce times and relocation times. We did the performance benchmarks in an environment with two, four, and eight activated channel-to-channel (CTC) connections. The relocation options have the following effects:

<b>SYNCHRONOUS</b>	This is the default option. It causes the VMRELOCATE MOVE command to complete only after the relocation is completed. This option is used to serialize relocations.
<b>IMMEDIATE</b>	This option causes the quiesce to occur after only one pass through memory. It usually results in shorter overall relocation times, but longer quiesce times.
<b>ASYNCHRONOUS</b>	This option causes the VMRELOCATE MOVE command to return as soon as the initial eligibility check is done. It can be used to do relocations in parallel.

The following figures are the performance benchmarks for the different relocation options that we tested, shown for two CTCs (Figure 7-1 on page 74), four CTCs (Figure 7-2 on page 75) and eight activated CTCs (Figure 7-3 on page 75). “QT” represents the Quiesce Time and “RT” is the relocation time for the guests.

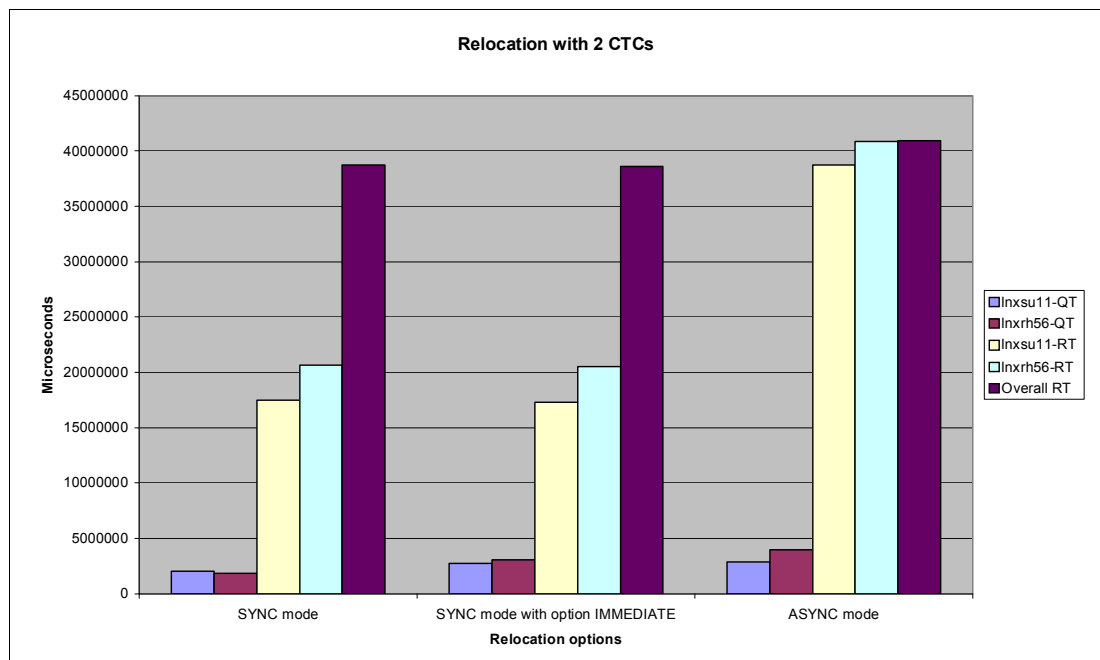


Figure 7-1 Relocation benchmark with two CTCs

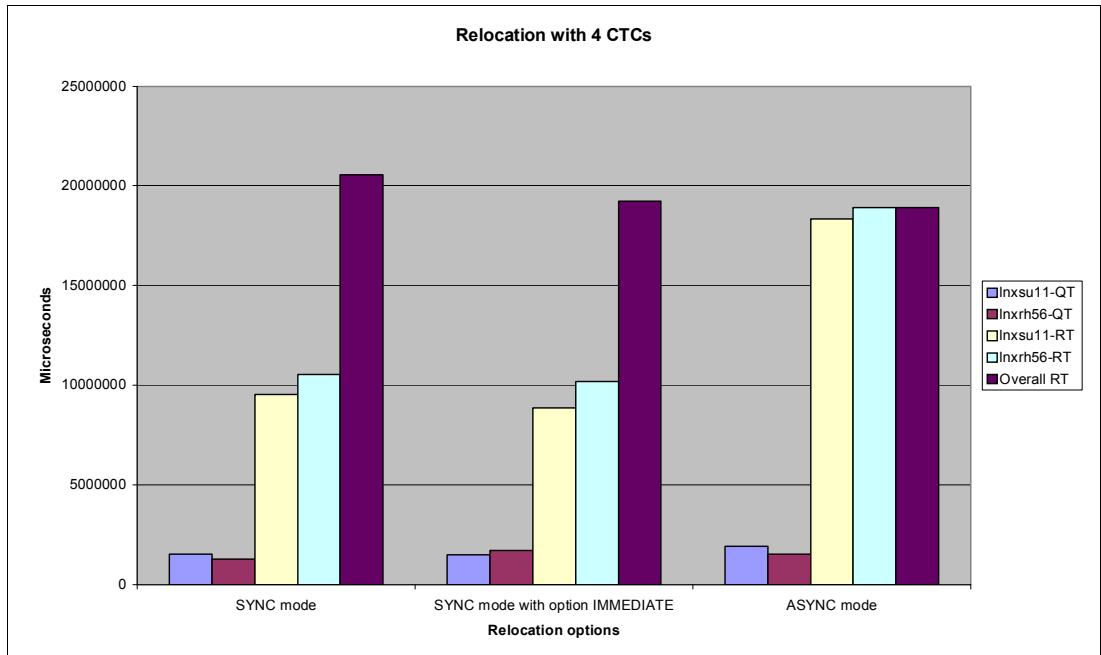


Figure 7-2 Relocation benchmark with four CTCs

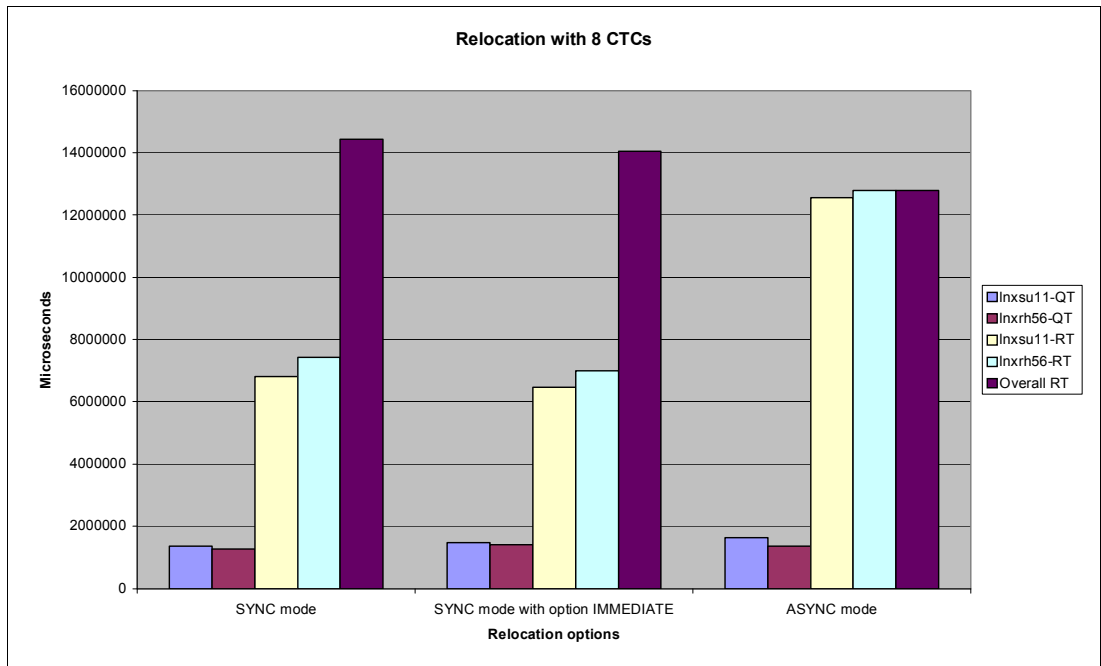


Figure 7-3 Relocation benchmark with eight CTCs

## Benchmark results

- Quiesce time and individual relocation time improves substantially when relocations are done serially in synchronous mode. In our scenarios, the quiesce times and relocation times were almost doubled when the relocation was done in asynchronous mode. When more than two guests are relocated at the same time, we assume that the time difference is much higher.

- Overall relocation time for several Linux guests in asynchronous (parallel) mode enhances the overall relocation time only slightly (if at all).
- Relocation option IMMEDIATE resulted in slightly longer quiesce times and slightly shorter relocation times.

## Summary

The combinations of relocation options resulted in different relocation and quiesce times. Table 7-1 shows the combinations that did best on the success measures considered. No single combination was the best in all categories.

Table 7-1 Success measures and vmrelocate option combinations

Success measures	Synchronous	Synchronous immediate	Asynchronous immediate	Asynchronous
Best total relocation time for all users			X	
Best individual relocation times		X		
Best individual quiesce times	X			
Least number of memory move passes		X		
Best response times for PING		X		

## 7.2 Relocation benchmark dependent on the number of CTCs

Our tests involved relocating two Linux guests with first two, then four, and finally eight CTC connections defined in the ISFC setup. We used IBM FICON® CTCs via switches. The switches limited the connections to two gigabit because we had older switches. We defined four addresses on each CHPID.

We dynamically activated each ISLINK on available CTCs with the command shown in Example 7-1.

*Example 7-1 Activate ISLINK command on first system*

---

```
activate islink 4291 node itsossi6
Link device 4291 activated.
Ready; T=0.01/0.01 17:39:55
```

---

If you also do this on the second member to get the connection, as shown in Example 7-2, you get the results from a query to the ISLINK as shown in Example 7-3.

*Example 7-2 Activate ISLINK command on second system*

---

```
activate islink 5051 node itsossi5
Link device 5051 activated.
Ready; T=0.01/0.01 17:42:15
```

---

*Example 7-3 Connected ISLINKs*

---

```
q islink
Node ITS0SSI6
...
Link device: 4291      Type:  FCTC
Node:      ITS0SSI6  Bytes Sent:      8646904610
```

---

State: Up Bytes Received: 16133728269  
Status: Idle  
Remote link device: 5051

...

To permanently set the CTC connections that are used for relocation, add the additional ISLINKs to the SYSTEM config. As shown in Example 7-4, there are now four connections, shown in bold print, defined for the next IPL.

*Example 7-4 Permanent setting of ISLINKs in SYSTEM CONFIG*

```
ITS0SSI5: ACTIVATE ISLINK 4290 5290 4291 5291 NODE ITS0SSI6  
ITS0SSI6: ACTIVATE ISLINK 5050 4050 5051 4051 NODE ITS0SSI5
```

Figure 7-4 shows the overall relocation times for both Linux guests by number of CTCs.

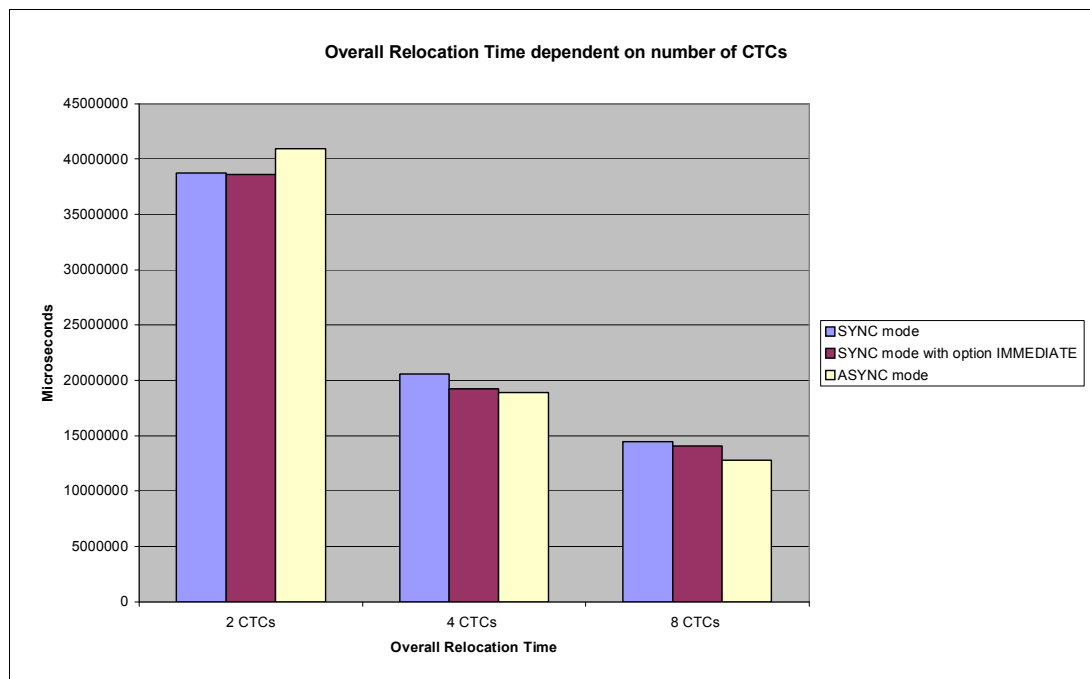


Figure 7-4 Overall relocation time by number of CTCs

Figure 7-5 on page 78 shows the total quiesce times of both Linux guests by number of CTCs.

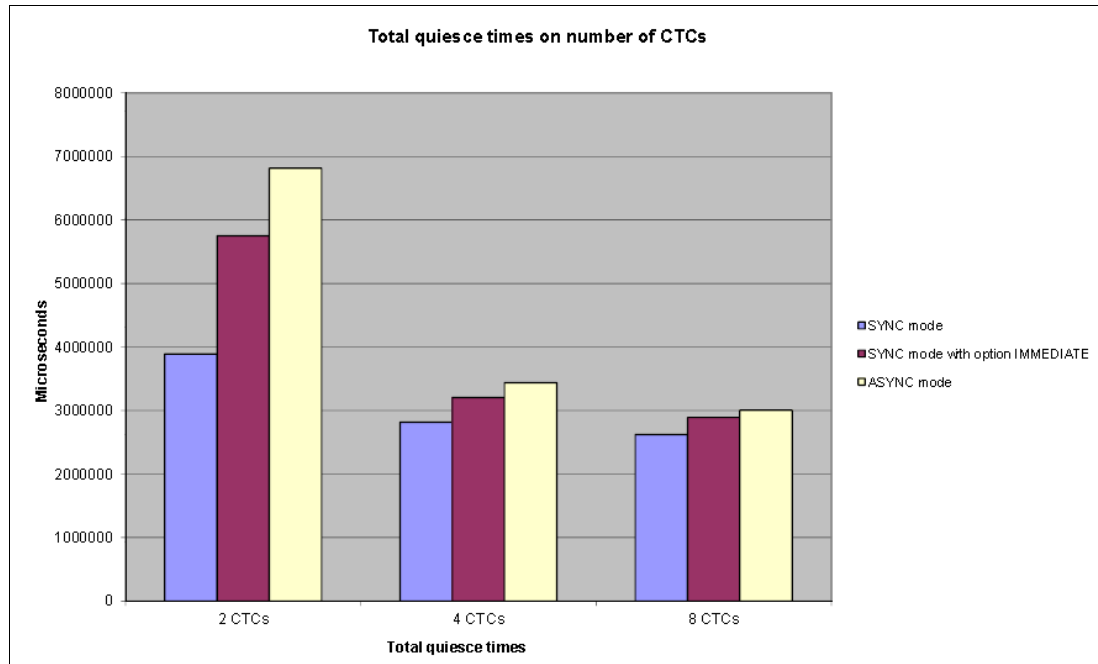


Figure 7-5 Total quiesce times by number of CTCs

#### Benchmark results:

- The number of ISFC channels (CTCs) shortens the quiesce time and relocation time. The relocation times show significant improvements, as shown in Figure 7-4 on page 77.

## 7.3 Two Linux guests in a four cluster SSI system dependent on different LPARs and processors

We relocated two Linux guests in a four cluster SSI system based in different LPARs and different processors. For our tests, we established two CTC connections between the LPARs in the four cluster system. QT is quiesce time and RT is the relocation time.

As in the previous benchmarks, we relocated the guests SYNCHRONOUSLY (SYNC), which is serially, and ASYNCHRONOUSLY (ASYNC), which is in parallel.

**Note:** The total relocation time when the guests are moved synchronously is not the sum of the relocation times of the two separate guests. It includes the time between one guest finishing relocation and the next guest starting relocation.

Figure 7-6 on page 79 shows the benchmark results for the relocation of the two Linux guests to an LPAR on the same processor.

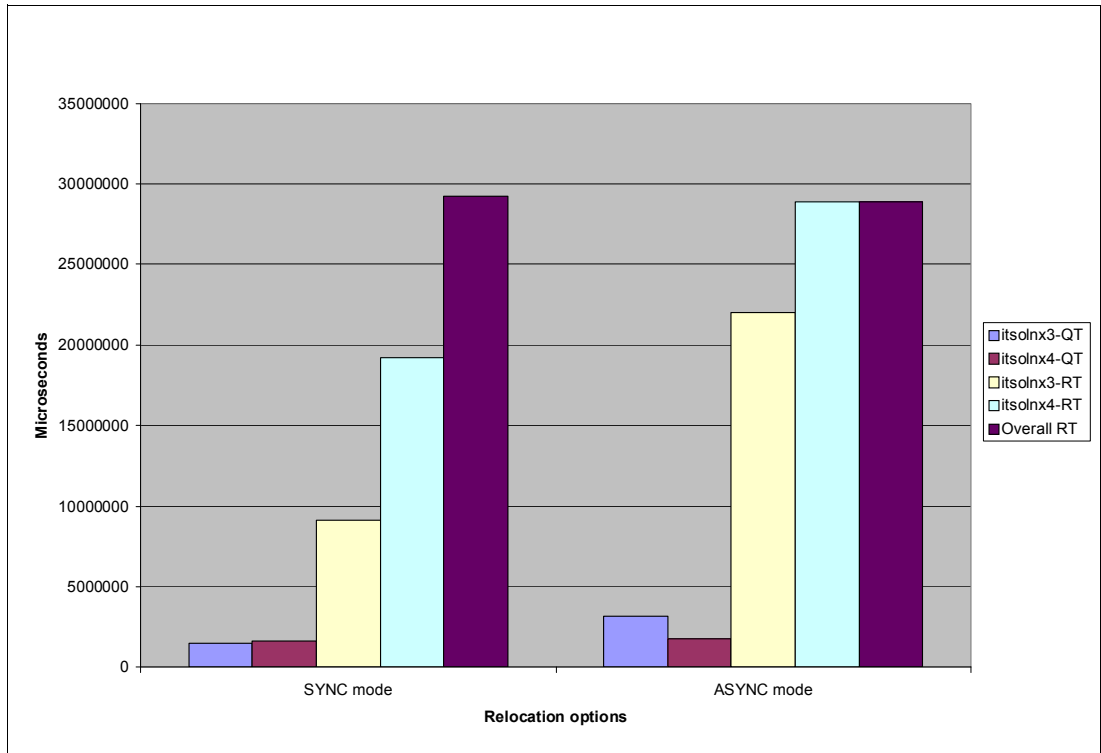


Figure 7-6 Two Linux guests relocated to an LPAR on the same processor

Figure 7-7 shows the benchmark results for two Linux guests relocated to an LPAR on a different processor.

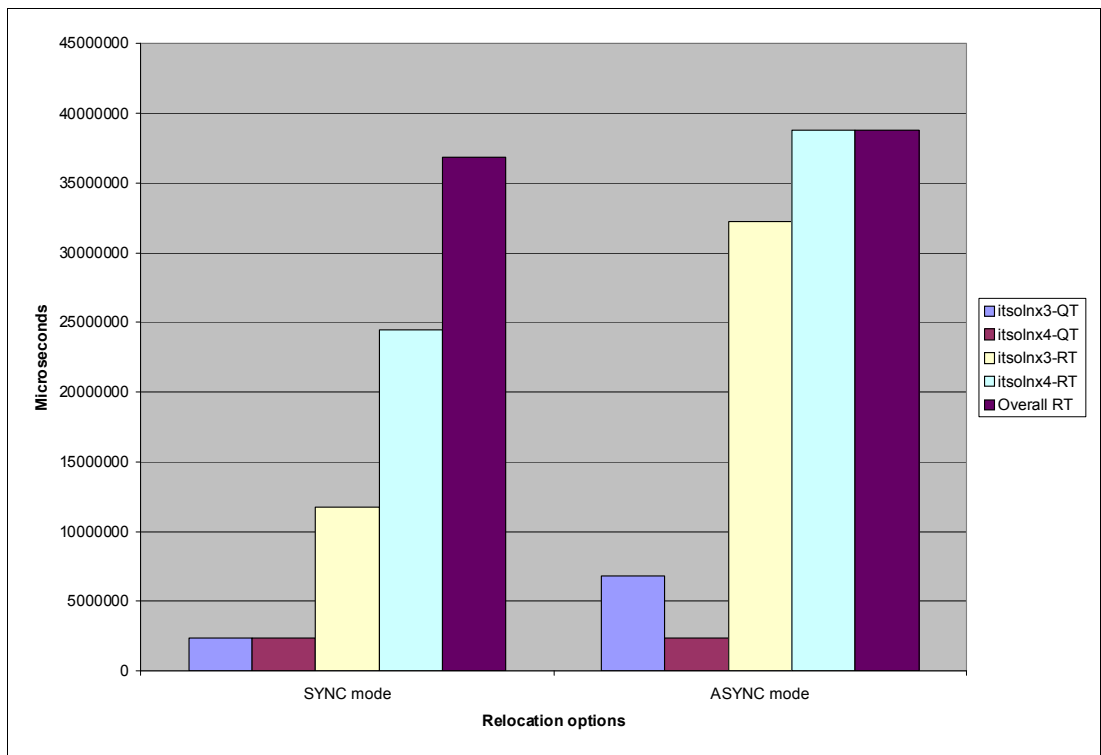


Figure 7-7 Two Linux guests relocated to an LPAR on a different processor

Figure 7-8 shows the benchmark results of two Linux guests relocating to two different LPARs that are running on a different processor from where the guests were originally located.

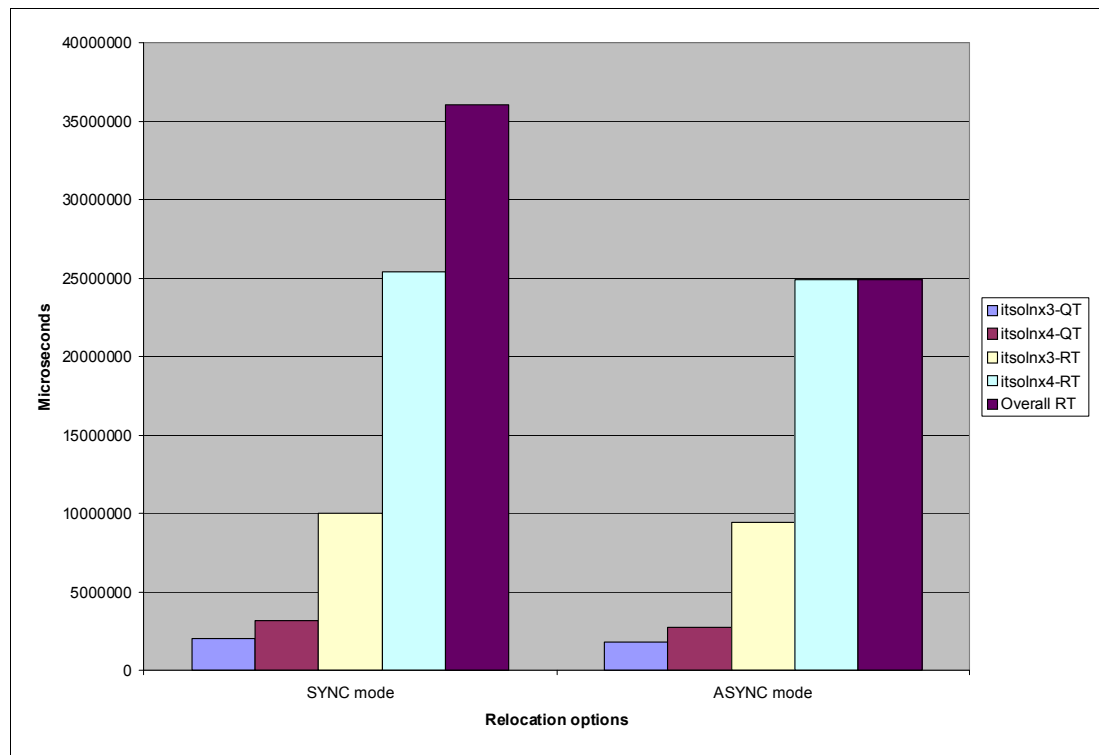


Figure 7-8 Two Linux guests relocated to two different LPARs

In this case the time taken to relocate each guest is the same whether they are relocated synchronously or asynchronously. We believe that this is because we were relocating to two different LPARs and different CTCs were used for each of the relocations.

Figure 7-9 on page 81 shows the benchmark results of one Linux guest relocated to an LPAR on the same processor and the other Linux guest relocated to an LPAR on a different processor.

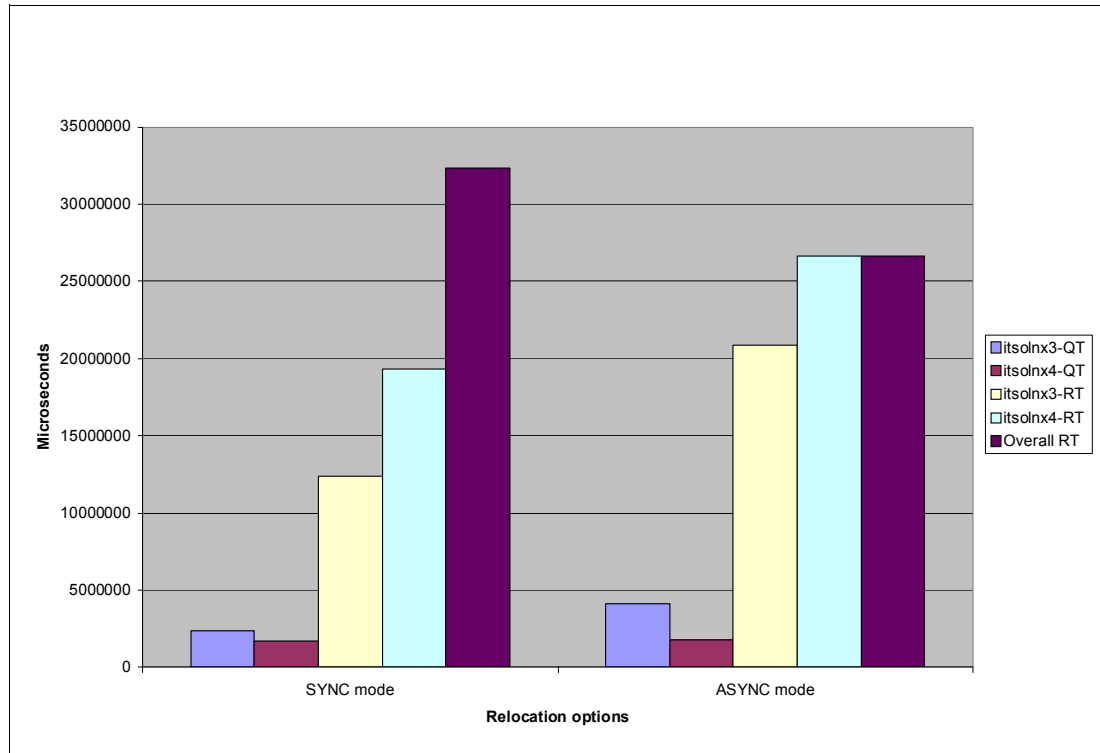


Figure 7-9 The Linux guests relocated on one LPAR on the same processor and one LPAR on a different processor

### Benchmark results

In some of our tests, the quiesce time and individual relocation time improves when relocations are done serially in synchronous mode. We were only able to use a relatively light load on our systems and we did not carry out multiple tests for each scenario. However, we found that relocating the guests serially provided the best results.

## 7.4 Two Linux guests in two and four cluster systems dependent on SCSI and non-SCSI

For our final benchmark test, we defined two Linux guests, one in a two cluster system and the other in a four cluster system. In both the two cluster and four cluster systems, two ISFC channels (CTCs) were defined between the LPARs within the clusters.

Figure 7-10 on page 82 shows the benchmark results of a two Linux guest relocation. One Linux guest is in a two cluster system, with applications based on an IBM DS8300 disk subsystem, and one Linux guest is in a four cluster system, with applications based on SCSI attached disks.

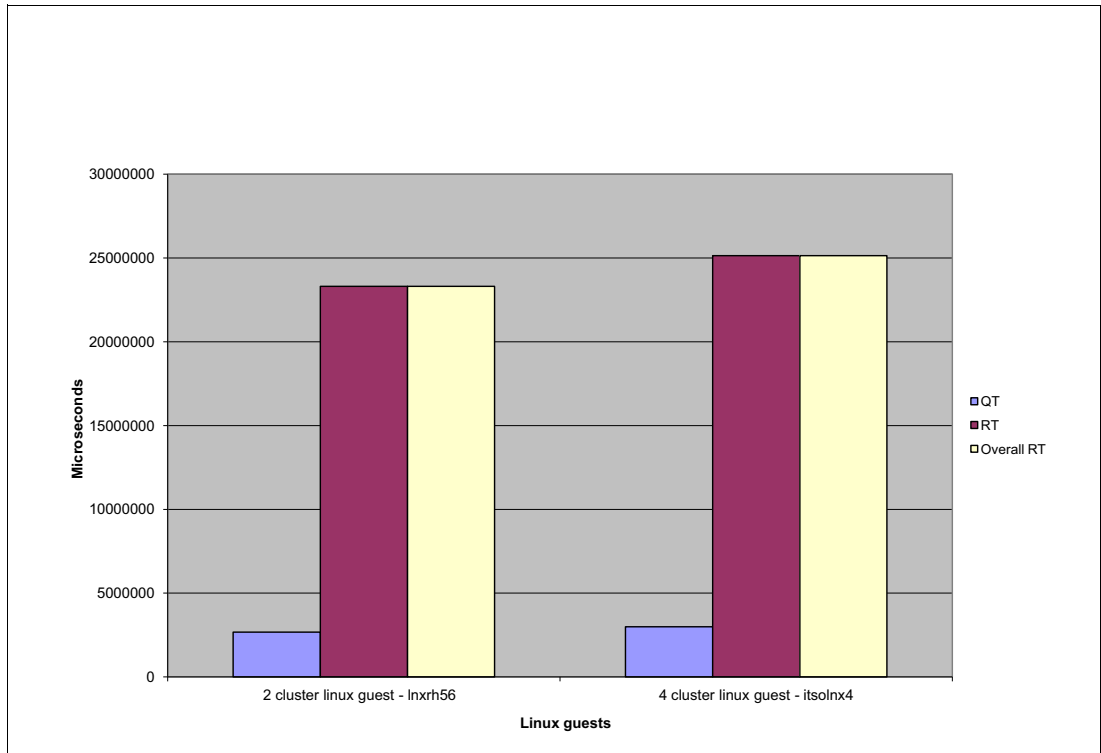


Figure 7-10 Benchmark test results for one Linux guest in a two cluster system and the other Linux guest in a four cluster system

## Benchmark results

The Linux guest on the two cluster system, based on a DS8300 disk device, has a slightly improved relocation time over the Linux guest four cluster relocation, based on a SCSI attached device. There could be other factors that affected the relocation. We tried to eliminate these factors as much as possible by relocating the guests between the same two processors.



## IBM Backup and Restore Manager for z/VM

Backing up and restoring data are essential components of data storage management. Backing up your data on a regular basis helps protect your system against the loss of data in the event of a major disaster, or when data is accidentally deleted or becomes corrupted.

In this chapter, we provide an overview of the IBM Backup and Restore Manager for z/VM and describe how we installed and configured it in our lab. We also describe the changes that must be made for IBM Backup and Restore Manager for z/VM to operate in an SSI cluster.

## 8.1 Overview of the IBM Backup and Restore Manager for z/VM

The IBM Backup and Restore Manager for z/VM enables system administrators and operators to efficiently and effectively back up and restore files and data on z/VM systems. Source files and data can be CMS or non-CMS format and the target media can be DASD or tape. The flexibility of Backup and Restore Manager is apparent in its ability to do full physical and logical backup and restore operations with support for inclusion and exclusion of files and user IDs.

Backup and Restore Manager for z/VM is a powerful tool with many possibilities. In this book, we focus on the new or changed installation and configuration steps for SSI clusters.

Figure 8-1 shows how we set up Backup and Restore Manager for z/VM in our lab environment. While this is just one way to configure the backup and restore environment, it provides the most efficient and effective method for back up and restoration.

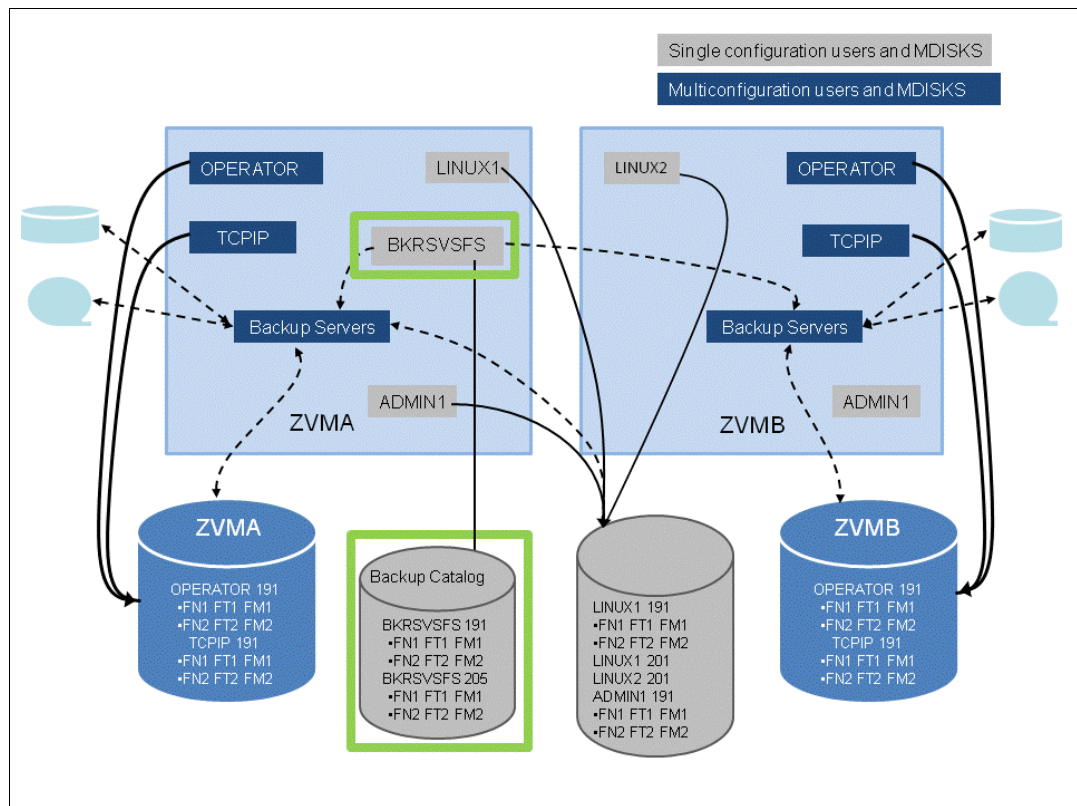


Figure 8-1 Our Backup and Restore Manager for z/VM configuration

Here is a brief summary of the back up steps:

- ▶ Create Backup Manager service machines as IDENTITY users on each member of the cluster
- ▶ Create one single configuration user for the SFS server/filepool, which is used for the backup catalog. There will be just one catalog (and therefore one SFS server) that will serve all Backup Manager servers on all members of the cluster. This allows single configuration users to restore their own data when logged onto any member of the cluster.
- ▶ Create multiple backup jobs to support your environment, using the following as a guide or minimum:

- For all single configuration users, create just one job. Always run this from the same member
- For all multiconfiguration (IDENTITY) users create one job per member. Use a unique job name on each member of the cluster, and always run the member specific job on that member's backup server.

## 8.2 Installation of IBM Backup and Restore Manager for z/VM

In this section we describe the changes to the installation steps for installing the Backup and Restore Manager in an SSI cluster. For more information on this, see the *Getting Started* presentation on the Library page of the Backup and Restore Manager web site which can be found at:

<http://www.ibm.com/software/stormgmt/zvm/backup/library.html>

### 8.2.1 Prerequisite: Create a Shared File System server and file pool

We created a Shared File System (SFS), which is where the catalog for the Backup and Restore Manager will reside. Create a separate file pool for this purpose so that it does not get mixed up with the other file pools the system is using and when you upgrade z/VM, you can easily move this separate file pool to the new release (the file pools provided by z/VM are replaced in new releases). Depending on the size of your installation, the catalog data can be large. The product documentation suggests starting with a minimum of 3000 cylinders. See *CMS File Pool Planning, Administration, and Operation*, SC24-6074, for more details about this.

In our example, we also put the Backup Manager TEMPLATE and DISKPOOL files in this SFS file pool. We use BKRSFS as the file pool name and BKRSVSFS as the file pool server (as shown in Figure 8-1). We created the user BKRSVSFS as shown in Example 8-1.

*Example 8-1 Directory entry BKRSVSFS*

---

```

USER BKRSVSFS BKRSVSPW 64M 64M BG
  IPL CMS
  IUCV ALLOW
  IUCV *IDENT RESANY GLOBAL
  MACHINE XC
  OPTION MAXCONN 2000 NOMDCFS APPLMON QUICKDSP SVMSTAT
  SHARE REL 1500
  CONSOLE 0009 3215 T
  SPOOL 000C 2540 READER *
  SPOOL 000D 2540 PUNCH A
  SPOOL 000E 1403
  LINK MAINT 0190 0190 RR
  LINK MAINT 0193 0193 RR
  LINK MAINT 019D 019D RR
  LINK MAINT 019E 019E RR
  MDISK 0191 3390 140 2 LX6032 MR
  MDISK 0250 3390 142 80 LX6032 MR
  MINIOPT NOMDC
  MDISK 0405 3390 222 10 LX6032 MR
  MINIOPT NOMDC
  MDISK 0406 3390 232 10 LX6032 MR
  MINIOPT NOMDC

```

```
MDISK 0260 3390 242 50 LX6032 MR
MDISK 0310 3390 292 750 LX6032 MR
MDISK 0311 3390 1042 750 LX6032 MR
MDISK 0312 3390 1792 750 LX6032 MR
MDISK 0313 3390 2542 750 LX6032 MR
```

---

We created a PROFILE EXEC for this userid on the 191 disk and restarted the userid (see Example 8-2).

*Example 8-2 Profile exec for BKRSVSFS*

---

```
/* */
'ACCESS 193 C'
'CP SET EMSG ON'
'CP SET PF11 RETRIEVE FORWARD'
'CP SET PF12 RETRIEVE'
EXIT 0
```

---

We defined the startup parameters for BKRSVSFS by creating a file called BKRSVSFS DMSPARMS on BKRSVSFS 191 disk. This file defines the userids that can operate as administrators for backup and restore. It also gives some startup parameters. See Example 8-3.

*Example 8-3 BKRSVSFS DMSPARMS file*

---

```
ADMIN 5697J06B
ADMIN BKRADMIN
ADMIN BKRBKUP
ADMIN BKRCATLG
ADMIN BKRWRK01
ADMIN BKRWRK02
ADMIN BKRWRK03
ADMIN BKRWRK04
ADMIN MAINT620
NOBACKUP
FILEPOOLID BKRSFS
NOCRR
NOLUNAME
SSI
SAVESEGID CMSFILES
USERS 700
CATBUFFERS 5000
```

---

We then created the file pool BKRSFS by logging on to user BKRSVSFS and issuing the command:

```
fileserv generate
```

When prompted for \$\$TEMP \$POOLDEF we deleted the lines shown in Example 8-4.

*Example 8-4 \$\$ TEMP \$POOLDEF displayed*

---

```
MAXUSERS=1000
MAXDISKS=500
DDNAME=CONTROL          VDEV=301
DDNAME=LOG1              VDEV=302
DDNAME=LOG2              VDEV=303
```

DDNAME=BACKUP	DISK	FN=CONTROL	FT=BACKUP	FM=A
DDNAME=MDK00001		VDEV=304	GROUP=1	BLOCKS=0
DDNAME=MDK00002		VDEV=305	GROUP=2	BLOCKS=0
DDNAME=CRR1		VDEV=306		
DDNAME=CRR2		VDEV=307		

---

We included the statements shown in Example 8-5.

---

*Example 8-5 Definition for file pool*

---

```

MAXUSERS=4000
MAXDISKS=500
DDNAME=CONTROL VDEV=250
DDNAME=LOG1 VDEV=405
DDNAME=LOG2 VDEV=406
DDNAME=MDK00001 VDEV=260 GROUP=1 BLOCKS=0
DDNAME=MDK00002 VDEV=310 GROUP=2 BLOCKS=0
DDNAME=MDK00003 VDEV=311 GROUP=2 BLOCKS=0
DDNAME=MDK00004 VDEV=312 GROUP=2 BLOCKS=0
DDNAME=MDK00005 VDEV=313 GROUP=2 BLOCKS=0

```

---

When we saved this file, z/VM began to format the disk. In our case, this took a few minutes.

Finally, we added the line shown in Example 8-6 to the PROFILE EXEC of userid BKRSVSFS and restarted the userid.

---

*Example 8-6 Additional command in profile exec*

---

```
'EXEC FILESERV START'
```

---

From the installation userid, 5697J06B, we had to authorize several users to BKRSFS, as shown in Example 8-7.

---

*Example 8-7 Authorize Users*

---

```

enroll user bkradmin bkrsfs (blocks 4000 storgroup 2
enroll user bkrbkup bkrsfs (blocks 4000 storgroup 2
enroll user bkrcatlg bkrsfs (blocks 500000 storgroup 2
enroll user bkrwrk01 bkrsfs (blocks 20000 storgroup 2
enroll user bkrwrk02 bkrsfs (blocks 20000 storgroup 2
enroll user bkrwrk03 bkrsfs (blocks 20000 storgroup 2
enroll user bkrwrk04 bkrsfs (blocks 20000 storgroup 2

```

---

We created the required BKRSFS directory structures shown in Figure 8-8.

---

*Example 8-8 Create Directory entries for BKRSFS*

---

```

create directory bkrsfs:bkradmin.workarea
create directory bkrsfs:bkradmin.jobdefs
create directory bkrsfs:bkrcatlg.workarea
create directory bkrsfs:bkrbkup.workarea
create directory bkrsfs:bkrwrk01.workarea
create directory bkrsfs:bkrwrk02.workarea
create directory bkrsfs:bkrwrk03.workarea
create directory bkrsfs:bkrwrk04.workarea

```

---

We authorized MAINT620 as an additional user to create and update the backup job templates, as shown in Example 8-9.

*Example 8-9 Authorize MAINT620*

---

```
grant auth bkrsfs:bkraadmin.jobdefs to maint620 (write newwrite  
grant auth * * bkrsfs:bkraadmin.jobdefs to maint620 (write
```

---

Additional users can be added using the same commands.

**Note:** You might get a warning message when you enter the second command shown in Example 8-9 because you are granting access to a shared file system that is empty. This warning can safely be ignored.

## 8.2.2 Prerequisite: Install REXX library

You need a REXX library to run the IBM Backup and Restore Manager for z/VM. We used the IBM Alternate Library for REXX on zSeries®. It is available as a no-charge download from:

<http://www-01.ibm.com/software/awdtools/rexx/rexxzseries/index.html>

We followed the installation steps described in the MAKEALTV.README.TXT file that comes with the download. All the steps that are necessary for one system installation, such as copying files to MAINT 19E disk and changing the PROFILE EXEC file on a userid, had to be performed for *each* member of our SSI cluster. After finishing the installation we issued a **put2prod** command on each member.

**Note:** If you decide to copy files to MAINT 19E disk, make sure all files have a filemode number of 2 (not the default of 1). This is required for loading into the CMS saved segment.

We did not install the Operations Manager or the Tape Manager because there is nothing new in these products relating to an SSI cluster.

## 8.2.3 Userids used for Backup and Restore Manager for z/VM

In this section we describe the various userids required to run the Backup and Restore Manager for z/VM.

### BKRAADMIN

This is the default administration userid and has administrative authority over many things such as submitting backup and restore requests to the BKRBKUP users in each member of the SSI cluster and receives the console output from the BKRWRKnn userid to see the result of the jobs. It is a unique userid, so there is only one interface to the whole SSI cluster. Example 8-10 shows our directory entry for BKRAADMIN.

BKRAADMIN is defined as a USER in the directory, which means it can only be logged on to one member of the SSI cluster at any one time.

*Example 8-10 Directory entry for BKRAADMIN*

---

```
USER BKRAADMIN itsossi 128M 128M BG  
MACHINE ESA  
IPL CMS  
OPTION LNKNOPAS
```

---

```
CONSOLE 01F 3215
SPOOL 00C 2540 READER A
SPOOL 00D 2540 PUNCH A
SPOOL 00E 1403 A
LINK 5697J06B 198 198 MR
LINK 5697J06B 591 591 RR
LINK 5697J06B 199 199 MR
LINK 5697J06B 592 592 RR
LINK MAINT 190 190 RR
LINK MAINT 19D 19D RR
LINK MAINT 19E 19E RR
```

---

In order to submit a backup job and see the results from the worker, you must be logged into the same system where the BKRBKUP user that you want to submit the job for is running.

## BKRCATLG

The catalog service virtual machine manages the backup catalog, which represents data that is being managed by the IBM Backup and Restore Manager for z/VM.

BKRCATLG is set up in the directory as an IDENTITY, so it has one userid for each member of the SSI cluster.

This has to be an IDENTITY because each SSI cluster member needs to have access to the catalog of the backup data. Example 8-11 shows our directory entry for BKRCATLG.

*Example 8-11 Directory entry for BKRCATLG*

---

```
IDENTITY BKRCATLG ITSOSI 128M 512M BEG
  BUILD ON ITSOSI1 USING SUBCONFIG BKRCAT-1
  BUILD ON ITSOSI2 USING SUBCONFIG BKRCAT-2
  BUILD ON ITSOSI3 USING SUBCONFIG BKRCAT-3
  BUILD ON ITSOSI4 USING SUBCONFIG BKRCAT-4
  IPL CMS
  MACHINE ESA
  OPTION LKNOPAS
  CONSOLE 001F 3215
  SPOOL 000C 2540 READER A
  SPOOL 000D 2540 PUNCH A
  SPOOL 000E 1403 A
  LINK 5697J06B 0198 0198 RR
  LINK 5697J06B 0199 0199 RR
  LINK 5697J06B 0591 0591 RR
  LINK MAINT 0190 0190 RR
  LINK MAINT 019D 019D RR
  LINK MAINT 019E 019E RR
  LINK MAINT 0193 0193 RR
SUBCONFIG BKRCAT-1
  MDISK 0191 3390 3297 1 LX6032 MR
SUBCONFIG BKRCAT-2
  MDISK 0191 3390 3298 1 LX6032 MR
SUBCONFIG BKRCAT-3
  MDISK 0191 3390 3299 1 LX6032 MR
SUBCONFIG BKRCAT-4
  MDISK 0191 3390 3299 1 LX6032 MR
```

---

## BKRBKUP

BKRBKUP is the master backup service virtual machine. Among other things, it receives backup request from administrators, receives restore requests from users and administrators, and assigns those requests to worker service virtual machines. There should be a separate master backup user in each member of the SSI cluster that controls the workers that are running in this SSI cluster.

Therefore, BKRBKUP has to be set up as an IDENTITY, as shown in Example 8-12.

*Example 8-12 Directory entry for BKRBKUP*

---

```
IDENTITY BKRBKUP ITSOSI 128M 256M ABDEG
  BUILD ON ITSOSI1 USING SUBCONFIG BKRBK-1
  BUILD ON ITSOSI2 USING SUBCONFIG BKRBK-2
  BUILD ON ITSOSI3 USING SUBCONFIG BKRBK-3
  BUILD ON ITSOSI4 USING SUBCONFIG BKRBK-4
  IPL CMS
  MACHINE ESA
  OPTION LNKNOPAS
  CONSOLE 001F 3215
  SPOOL 000C 2540 READER A
  SPOOL 000D 2540 PUNCH A
  SPOOL 000E 1403 A
  LINK 5697J06B 0198 0198 RR
  LINK 5697J06B 0199 0199 RR
  LINK 5697J06B 0591 0591 RR
  LINK MAINT 0190 0190 RR
  LINK MAINT 019D 019D RR
  LINK MAINT 019E 019E RR
  LINK MAINT 0193 0193 RR
SUBCONFIG BKRBK-4
  MDISK 0191 3390 3301 1 LX6032 MR
SUBCONFIG BKRBK-3
  MDISK 0191 3390 3302 1 LX6032 MR
SUBCONFIG BKRBK-2
  MDISK 0191 3390 3303 1 LX6032 MR
SUBCONFIG BKRBK-1
  MDISK 0191 3390 3303 1 LX6032 MR
```

---

## BKRWRKnn

The worker IDs are BKRWRK01 - BKRWRKxx. You can set up as many worker IDs as you need for the amount of disks that you want to back up and restore. The default number is four. They are started by BKRBKUP in parallel when the restore and backup requests come in from users and administrators. Because you want to back up data from each member of your SSI cluster, you need to set up enough worker identities on each member of the SSI cluster.

The workers also must be IDENTITIES. Example 8-13 provides an example of just one worker ID.

*Example 8-13 Directory entry for BKRWRK01*

---

```
IDENTITY BKRWRK01 ITSOSI 128M 512M ABEG
  BUILD ON ITSOSI1 USING SUBCONFIG BKRWR1-1
  BUILD ON ITSOSI2 USING SUBCONFIG BKRWR1-2
  BUILD ON ITSOSI3 USING SUBCONFIG BKRWR1-3
  BUILD ON ITSOSI4 USING SUBCONFIG BKRWR1-4
```

---

```

IPL CMS
MACHINE ESA
OPTION LNKNOPAS DEVMAINT
CONSOLE 001F 3215
SPOOL 000C 2540 READER A
SPOOL 000D 2540 PUNCH A
SPOOL 000E 1403 A
LINK 5697J06B 0198 0198 RR
LINK 5697J06B 0199 0199 MR
LINK 5697J06B 0591 0591 RR
LINK MAINT 0190 0190 RR
LINK MAINT 019D 019D RR
LINK MAINT 019E 019E RR
LINK MAINT 0193 0193 RR
SUBCONFIG BKRWR1-1
MDISK 0191 3390 3305 1 LX6032 MR
SUBCONFIG BKRWR1-2
MDISK 0191 3390 3306 1 LX6032 MR
SUBCONFIG BKRWR1-3
MDISK 0191 3390 3307 1 LX6032 MR
SUBCONFIG BKRWR1-4
MDISK 0191 3390 3307 1 LX6032 MR

```

---

**Note:** Depending on how many members you have in your installation and how many workers you want per system, we found it helpful to define a PROTODIR for the BKRWRKxx IDENTITY and SUBCONFIG statements. See *An introduction to z/VM Single System Image (SSI) and Live Guest Relocation (LGR)*, SG24-8006, for details.

The setup of a PROTODIR for an IDENTITY and a PROTODIR for a SUBCONFIG are shown in Example 8-14 and Example 8-15. For more detailed information about setting up PROTODIRs see *An introduction to z/VM Single System Image (SSI) and Live Guest Relocation (LGR)*, SG24-8006.

---

*Example 8-14 IDENT PROTODIR file*

---

```

IDENTITY IDENT ITSOSI 128M 512M B
MACHINE ESA
IPL CMS
OPTION LNKNOPAS
CONSOLE 01F 3215
SPOOL 00C 2540 READER A
SPOOL 00D 2540 PUNCH A
SPOOL 00E 1403 A
LINK 5697J06B 198 198 RR
LINK 5697J06B 199 199 MR
LINK 5697J06B 591 591 RR
LINK MAINT 190 190 RR
LINK MAINT 19D 19D RR
LINK MAINT 19E 19E RR

```

---



---

*Example 8-15 SUBCON-1 PROTODIR file*

---

```

SUBCONFIG subcon-1
MDISK 0191 3390 AUTOV 001 LX6032 MR

```

---

## BKUPDISK

If you are planning to back up to disk (instead of or in addition to tape), then BKUPDISK is another user that must be defined. This user owns the minidisks that actually hold the backup data. The more workers you have doing backups to disk concurrently, the more disks you need to define. We had three minidisks: 300, 310, and 320. See Example 8-16.

*Example 8-16 Directory entry for BKUPDISK*

---

```
USER BKUPDISK ITSOSI 16M 16M G
  ACCOUNT SYSTEMS
  IPL CMS
  MACH ESA
  CONSOLE 0009 3215
  SPOOL 000C 2540 READER *
  SPOOL 000D 2540 PUNCH A
  SPOOL 000E 1403 A
  LINK MAINT 0190 0190 RR
  LINK MAINT 019E 019E RR
  LINK MAINT 019D 019D RR
  MDISK 0191 3390 3336 5 LX6032 MR
  MDISK 0300 3390 3341 100 LX6032 MR
  MDISK 0310 3390 3441 100 LX6032 MR
  MDISK 0320 3390 3541 100 LX6032 MR
```

---

## 8.2.4 Set up service machines

Create the PROFILE EXEC for the service machines. We logged on to the installation userid 5697J06B. On the minidisk 2C2 are sample profiles for the userids BKRADMIN, BKRCATLG, BKRBKUP, and BKWRKxx. We copied these sample configuration files to the 191 disk of the corresponding userids and made the following changes in the sample files:

- ▶ Take out the /\* and \*/ before and after MINIDISK-based installation.
- ▶ Change:  
Job\_Template = '199' to Job\_Template = 'BKRSFS:BKRADMIN.JOBDEFS'
- ▶ Take out the /\* and \*/ before and after the section for work areas in SFS.
- ▶ Change:  
Work\_Area = 'VMSYS:'USERID()''.WORKAREA to Work\_Area = 'BKRSFS:'()''.WORKAREA

## 8.2.5 Final installation steps

This section details the steps used to complete the installation of Backup and Recovery Manager for z/VM.

1. Define additional special users to IBM Backup and Recovery Manager for z/VM.

From the installation userid 5697J06B minidisk 2C2, copy the file BKRUSERS NAMESAMP to minidisk 198 as BKRUSERS NAMES. We added additional authorized users BKRBKUP and MAINT620. Remove the workers that you do not need. Then copy the file to the MAINT 19E disk in filemode number 2.

**Reminder:** Be sure to copy the files onto MAINT 19E disk for every member of the SSI cluster.

2. Update the configuration file BKRSYSTEM CONFIG.

From the installation userid 5697J06B minidisk 2C2, copy the file BKRSYSTEM CONFSAMP to minidisk 198 as BKRSYSTEM CONFIG. This is where we changed our local options. The comments in the file describe which options you can change and what effect these changes will have in your environment. We used the default settings.

Like BKRUSERS NAMES, this file also has to be put on the MAINT 19E disk in filemode number 2.

**Reminder:** Be sure to copy the files onto MAINT 19E disk for every member of the SSI cluster.

3. Complete SFS configuration and authorization.

From userid 5697J06B enter:

```
enroll public bkrsfs
```

This command gives all users access to the catalog for restore requests. User access is limited to catalog directories for their own data.

From userid MAINT620, rebuild your CMS NSS. Issue the command:

```
put2prod savecms
```

On each member of the SSI cluster, issue the command:

```
put2prod
```

## 8.3 Set up the configuration file

The configuration file is BKSYSTEM CONFIG. It is located on userid 5697J06B minidisk 198. After you have modified it to your needs, you should put it on the MAINT 19E disk on each member of your cluster, so BKRADMIN can log on to any member and submit backup jobs.

We used the defaults as they are defined in the file.

## 8.4 Back up and restore a single configuration user (a USER directory entry)

This section describes how to back up and restore the minidisks attached to a USER. Backup and restore of an IDENTITY is described in 8.5, “Back up and restore a multi-configuration user (an IDENTITY directory entry)” on page 97

### 8.4.1 How to back up a single configuration user (a USER directory entry)

The backup of any userid must be done by an authorized Backup Administrator. Usually BKRADMIN is used to do this task.

The job templates are on the SFS disk BKRSFS:BKRADMIN.JOBDEFS. We used a unique job name so we could identify our backups later.

We used the name PLAY1, as shown in Example 8-17.

### Example 8-17 PLAY1 Template

```
* * * Top of File * * *
* IBM Backup and Restore Manager for z/VM - 5697-J06 - 1.2.0
*
Config BKR_Output_Spec = CMSFILE TESTFULL DISKPOOL *

CP_Command SPOOL CONSOLE TO $$ADMIN$$ CLASS T TERM START NAME PLAY1
CP_Command TERM LINES 255

Config BKR_Job_Workers = 1
Config BKR_Job_Name      = PLAY1
Config BKR_Job_Instance = $$INST$$
Config BKR_Job_Owner    = $$ADMIN$$
Config BKR_Job_Master    = $$MASTER$$
Config BKR_Job-Token     = $$SDATE$$

Config BKR_Job_Tape_Retention = 30

Config BKR_Job_CMS_FileMask = * * *
Config BKR_Job_SFS_PathMask = *

Config BKR_Job_Catalog      = Enabled
Config BKR_Catalog_Verbose = Disabled
Config BKR_Catalog_Master   = $$CATALOG$$
Config BKR_Catalog_Granule_FN = PLAY1
Config BKR_Catalog_Granule_FT = GRANULE
Config BKR_Catalog_Granule_FM = D1

Config BKR_EDF_Incr_Toggle = Off
Config BKR_SFS_Incr_Toggle = Off
Config BKR_Out_EDF_Verbose = Disabled
Config BKR_Out_Tape_Verbose = Disabled

FUNCTION  MEDIATYPE  OWNER      VDEV VOLUME DEVTYPE  START      END      SIZE  RESERVED
|-----|-----|-----|---|----|-----|-----|---|-----|---|-----|-----|
Exclude   Minidisk   *          = *    *    *      = *      = *      = *      *
Include   Minidisk   MASEN      = 0191 *    *      = *      = *      = *      *
```

EOJ

In the line **Config BKR\_Output\_Spec** (in bold in Example 8-17) we specify where the backup data will be stored. We back up to disk (not tape), so this is the CMS file name of the file that specifies the SFS filepools or CMS minidisks to which the data will be backed up. The TESTFULL DISKPOOL file resides on the same shared disk that our templates are on. In this file, we specified the minidisks owned by the BKUPDISK userid (see Example 8-16 on page 92).

In the line **Backup\_Job\_Workers** (the next bold line in Example 8-17) we specify the number of workers we want to use for this job. The workers are defined for each member of the SSI cluster (see Example 8-13 on page 90).

In the lines **Exclude** and **Include** we specify the userid and the minidisk we want to back up. Actually we first excluded all users and minidisks and then specified just the ones we want to back up. To keep the example easy, we just used one userid for our backup.

The job is submitted to the backup userid BKRBACKUP with the command:

```
MSG BKRBACKUP SUBMIT PLAY1
```

BKRBKUP will send the job to a worker, and because we specified to use only one worker, it will not be divided into segments and sent to several workers.

The console logs from the worker are sent back to the Administrator, in our case BKRADMIN, to check for any error that might have occurred.

Backup and restore commands like BKRJOB, BKRUSER and BKRVOL can be used to see the list of jobs that have been submitted by BKRADMIN. The F11 key shows the details of the job selected. If you use different job names for different backups, it is easier to locate the job you are interested in. See 8.6, “Backup and restore commands” on page 98 for more details about these commands.

## 8.4.2 How to restore a data for a single configuration user (a USER directory entry)

Restore can be done by the Administrator, and each userid can see and restore its own files. Each userid can only see the backup of its own files.

To restore, you first must link to the 5697J06B 592 disk, which is where the command executables (EXECs) for the Backup and Restore Manager reside. If you put the executables on MAINT 19E, then you don't need to link to 5697J06B 592.

The command **BKRLIST** shows all the files that have been backed up for your userid (Figure 8-2 on page 95).

Files for owner(s): *						
Selection: Name: * Type: * Mode: * 81 of 81 shown						
Current filters: Name: * Type: * Mode: * Owner: *						
Owner	Filename	Filetype	Fm	Date	Time	Device or Path
MASEN	AUTHFOR	CONTROL	1	12/04/10	16:24:23	0191
MASEN	BKRADMIN	DIRECT	0	12/04/20	14:08:13	0191
MASEN	BKRWRK01	DIRECT	0	12/04/09	16:42:03	0191
MASEN	BKRWRK02	DIRECT	0	12/04/09	16:41:52	0191
MASEN	BRK120	BETA	1	12/04/12	13:17:52	0191
MASEN	LASTING	GLOBALV	1	12/04/20	14:12:52	0191
MASEN	LINDFLT	DIRECT	0	12/04/13	11:47:33	0191
MASEN	MASEN	DIRECT	0	12/04/10	08:41:23	0191
MASEN	MASEN	NETLOG	0	12/04/20	14:08:23	0191
MASEN	PROFILE	EXEC	1	12/04/09	11:33:58	0191
MASEN	SHEISSER	DIRECT	0	12/04/10	08:41:54	0191
MASEN	SSI4	TEST	1	12/04/20	13:52:36	0191
MASEN	USER	NOPASS	1	12/04/19	10:18:48	0191
MASEN	5697J06B	DIRECT	0	12/04/20	14:09:02	0191

Figure 8-2 Output of BKRLIST command

Go to the file you want to restore and press the F10 key. The panel shown in Figure 8-3 is displayed.

CMS EDF Minidisk Restore Specifications		
From MASEN 0191 date 12/04/10 time 16:24:23 (job PLAY1 00000001 ).		
To EDF minidisk, userid: _____	and virtual address: _____	
FORMAT: OK if needed? NO	FORMAT regardless? NO	
Or to RDR of userid: _____	node: _____	(defaults to this node).
Or to SFS filepool: _____	and file space: _____	
and path: _____		
File filters: Filename: AUTHFOR Filetype: CONTROL mode number: 1		
Master backup userid: BKR BKUP Options: _____		

Figure 8-3 Restore a single file

Fill in the data about where you want the file restored (minidisk, reader, or SFS filepool) then press F10 again and the restore request is sent to a BKRWRKxx userid to run the restore job. In our example, the file will be sent to the reader (RDR) together with the console output from the worker.

Another way to restore your files is to use the command BKRUSER, which only shows your own userid. See Figure 8-4.

Catalog: BKRSFS: BKRCATLG. USERCAT.	
Ownerid filter: *	1 of 1 ownerids displayed
	Ownerids
MASEN	

Figure 8-4 BKRUSER output for MASEN

Press the F11 key for details; you will arrive at the panel shown in Figure 8-3 and be able to restore the files.

### 8.4.3 Known issues and workaround

We received the following error message when we tried to back up a USER that had previously been backed up on a different member of the SSI cluster:

```
ITS0LNK1 0333 -- EDF was previously backed up by LNK1FULL 00000004
```

The workaround for this problem is to always back up single configuration users from the same member of the SSI cluster regardless of which member of the SSI cluster the userid is currently logged on to.

## 8.5 Back up and restore a multi-configuration user (an IDENTITY directory entry)

This section describes how backup and restore of an IDENTITY differs from backup and restore of a USER. An IDENTITY has to be backed up from each of the members of the SSI cluster because it has different minidisks on each z/VM system it logs onto.

We set up template WRK4FULL to back up TCPMAINT, which is defined as an Identity. Example 8-18 shows the Include and Exclude statements we used to back up the TCPMAINT 191 minidisk. The rest of the template is the same as the one used for backing up a USER except that we changed all occurrences of PLAY1 to WRK4FULL.

Example 8-18 WRK4FULL Template

FUNCTION	MEDIATYPE	OWNER	VDEV	VOLUME	DEVTYPE	START	END	SIZE	RESERVED
Exclude	Minidisk	*	=	*	*	=	*	=	*
Include	Minidisk	TCPMAINT	=	0191	*	=	*	=	*

### 8.5.1 How to back up a multi-configuration user (an IDENTITY directory entry)

The same commands are used to back up an IDENTITY as are used to back up a USER. The only difference is that the backup has to be run on all members of the SSI cluster. If the same job name is used to back up the IDENTITY on all members of the SSI cluster, it is difficult to tell which backup relates to which member of the SSI cluster without using the BKR VOL command and looking at the actual extents of the minidisk being backed up.

It is suggested that the backup job on each member of the SSI cluster used for backing up an IDENTITY has a different job name to readily identify the member of the SSI cluster where the backup is run. We used WRK4FULL for backing up TCPMAINT on ITSOSS4 and WRK1FULL for backing up TCPMAINT on ITSOS11. Our Include and Exclude statements were the same in both templates.

### 8.5.2 How to restore an IDENTITY

The main difference between restoring a USER and an IDENTITY is that an IDENTITY can see all the backups for the IDENTITY regardless of which member of the SSI cluster the backup was run on. As shown in Figure 8-5, we logged on to TCPMAINT and issued the BKRJOB command to show the list of backup jobs for TCPMAINT.

Catalog: BKRSFS: BKRCATLG.JOBCAT							
Filters:							
Job: *	Instance: *	Owner: *	Type: *	Object: *	4 of 4 selected		
Command	Job	Instance	Owner	Type	Object	Date	Time
	WRK1FULL	00000013	TCPMAINT	EDF	£DEV0191		
	WRK4FULL	00000006	TCPMAINT	EDF	£DEV0191		
	WRK4FULL	00000007	TCPMAINT	EDF	£DEV0191		
	WRK4FULL	00000008	TCPMAINT	EDF	£DEV0191		

Figure 8-5 TPCMAINT Backup Jobs,

The list of jobs contains both WRK4FULL, which is the backup taken from ITSOSI4, and WRK1FULL, the backup taken from ITSOSI1. This is why it is suggested to have different job names for the backup jobs on the different members of the SSI cluster. TCPMAINT does not have access to the BKR VOL command, only Backup and Restore Administrators have access to that command.

Restore of an IDENTITY is the same as restore of a USER.

## 8.6 Backup and restore commands

This section gives a brief description of some of the more useful commands for managing backup and restore jobs.

### 8.6.1 BKR VOL

BKR VOL can only be issued by a userid that is defined as a backup and restore administrator. It displays details about the volumes that are backed up. Figure 8-6 shows the output from the BKR VOL command.

```

Catalog: BKRSFS: BKRCATLG. EXTENTBYDASD.
2 of 2 volumes displayed
Volume filter: *
Volumes
LX6032 SSI1I2
  
```

Figure 8-6 BKR VOL output

PF11 Details can be used to display details about the backups of users whose minidisks reside on that volume. Figure 8-7 on page 98 shows the details about the users that were backed up on volume SSI1I2.

```

Catalog: BKRSFS: BKRCATLG. EXTENTBYDASD.
2 of 2 devices displayed
Devices on volume SSI1I2
Owner filter: * Device filter: *
BKR8920E Restore is not available in this view.
Owner Dev Start Size Instances in catalog
MAINT 0191 0000000494-0000000175 1 instances
TCPMAINT 0191 0000005957-0000000007 4 instances
  
```

Figure 8-7 BKR VOL User Details

Pressing PF11 again yields details about the job names, as show in Figure 8-5.

### 8.6.2 BKRJOB

BKRJOB is used to get a list of all the backup jobs run. If it is issued from a backup and restore administrator, then a list of all backup jobs will be displayed. Figure 8-8 shows the list of jobs when the BKRJOB command is issued from BKRADMIN.

Catalog: BKRSFS: BKRCATLG. JOBCAT							
Filters:							
Job: *	Instance: *	Owner: *	Type: *	Object: *	13 of 13 selected		
Command	Job	Instance	Owner	Type	Object	Date	Time
	PLAY1	00000001	MASEN	EDF	£DEV0191		
	PLAY1	00000002	MASEN	EDF	£DEV0191		
	PLAY1	00000007	MASEN	EDF	£DEV0191		
	PLAY1	00000009	MASEN	EDF	£DEV0191		
	PLAY1	00000010	SHEISSER	EDF	£DEV0191		
	PLAY1	00000011	MASEN	EDF	£DEV0191		
	PLAY1	00000012	MASEN	EDF	£DEV0191		
	WRK1FULL	00000013	TCPMAINT	EDF	£DEV0191		
	WRK4FULL	00000004	BKRWRK01	EDF	£DEV0191		
	WRK4FULL	00000005	MAINT	EDF	£DEV0191		
	WRK4FULL	00000006	TCPMAINT	EDF	£DEV0191		
	WRK4FULL	00000007	TCPMAINT	EDF	£DEV0191		
	WRK4FULL	00000008	TCPMAINT	EDF	£DEV0191		

Figure 8-8 BKRJOB output for all jobs

Figure 8-9 shows the output from the BKRJOB command issued from TCPMAINT. Even though some of the backups were run on different members of the SSI cluster, TCPMAINT can see them all.

Catalog: BKRSFS: BKRCATLG. JOBCAT							
Filters:							
Job: *	Instance: *	Owner: *	Type: *	Object: *	4 of 4 selected		
Command	Job	Instance	Owner	Type	Object	Date	Time
	WRK1FULL	00000013	TCPMAINT	EDF	£DEV0191		
	WRK4FULL	00000006	TCPMAINT	EDF	£DEV0191		
	WRK4FULL	00000007	TCPMAINT	EDF	£DEV0191		
	WRK4FULL	00000008	TCPMAINT	EDF	£DEV0191		

Figure 8-9 BKRJOB output from TCPMAINT

### 8.6.3 BKRUSER

BKRUSER shows a list of all backup jobs by User. Figure 8-10 shows the output from the BKRUSER command issued from BKRADMIN.

Catalog: BKRSFS: BKRCATLG. USERCAT							
Ownerid filter: *							
5 of 5 ownerids displayed							
Ownerids							
BKRWRK01	MAINT	MASEN	SHEISSER	TCPMAINT			

Figure 8-10 BKRUSER output for all jobs

If the same command is issued by TCPMAINT, then only TCPMAINT is listed.

### 8.6.4 BKRLIST

BKRLIST is used to display details about every file on every minidisk that has been backed up. It is sorted by owner and there are filter criteria at the top of the panel that can be used to find specific files. Figure 8-11 show as example of the BKRLIST output with all files selected.

Files for owner(s): *						
Selection: Name: *		Type: *	Mode: *	181 of 181 shown		
Current filters: Name: *		Type: *	Mode: *	Owner: *		
Owner	Filename	Filetype	Fm	Date	Time	Device or Path
MASEN	ITSOLNX2	DIRECT	0	12/04/20	18:00:19	0191
MASEN	LASTING	GLOBALV	1	12/04/23	11:29:03	0191
MASEN	LINDFLT	DIRECT	0	12/04/13	11:47:33	0191
MASEN	MASEN	DIRECT	0	12/04/10	08:41:23	0191
MASEN	MASEN	NETLOG	0	12/04/20	18:00:03	0191
MASEN	PROFILE	EXEC	1	12/04/09	11:33:58	0191
MASEN	SHEISSER	DIRECT	0	12/04/10	08:41:54	0191
MASEN	SSI4	TEST	1	12/04/20	13:52:36	0191
MASEN	USER	NOPASS	1	12/04/19	10:18:48	0191
MASEN	5697J06B	DIRECT	0	12/04/20	14:09:02	0191
SHEISSER	ITSOLNX4	DIRECT	0	12/04/17	08:42:13	0191
SHEISSER	LASTING	GLOBALV	1	12/04/17	12:49:47	0191
SHEISSER	PLAY1	TEMPLATE	2	12/04/20	09:01:47	0191
SHEISSER	PROFILE	EXEC	1	12/04/10	13:48:02	0191
SHEISSER	SAMPFULL	TEMPSAMP	2	11/04/11	15:44:10	0191
SHEISSER	SAMPINCR	TEMPSAMP	2	11/04/11	15:44:10	0191
SHEISSER	SAMPLNX	TEMPSAMP	2	11/04/11	15:44:11	0191
SHEISSER	SCSIDISC	LOG	1	12/04/12	14:53:34	0191
SHEISSER	SCSIDISC	OUT	1	12/04/12	14:53:34	0191
SHEISSER	SHEISSER	DIRECT	0	12/04/13	11:08:53	0191
SHEISSER	SHEISSER	NETLOG	0	12/04/20	09:59:05	0191
TCPMAINT	PROFILE	EXEC	2	11/09/23	12:38:58	0191
TCPMAINT	PROFILE	XEDIT	1	98/11/18	12:26:20	0191
TCPMAINT	SSI1	FILE	1	12/04/20	14:06:46	0191
TCPMAINT	SSI4	FILE	1	12/04/20	13:54:06	0191

Figure 8-11 BKRLIST output for all jobs

If a user issues the BKRLIST command, they will only see a list of their own files.

**Note:** BKRLIST is not intended to be used by Backup and Restore Manager administrators. The amount of data returned by an administrator will often be too large to display.



# A

## Hints and tips

This appendix provides some processes and procedures that we found helpful when using z/VM 6.2 SSI and LGR.

## SCSI connection

The setup for the SCSI may require the following consideration.

### Redundancy considerations

For redundancy reasons, it would be best to have two FCP channels defined in each Linux, configured with the multipath option. Connect the 2 FCP channels over two different switches to each control unit of the DS8300. Thus if any hardware fails, there is an alternate path to reach your SCSI attached volume.

## SSI and IBM VM Backup and Restore Manager

To prepare the environment for the IBM VM Backup and Restore software testing, the following tasks were carried out:

### Removing queued VM backup jobs before starting a backup

If a backup or restore job fails it can leave a file in the reader of the BKRWRKnn users. The BKRWRKnn users only process one reader file when a job is submitted. So if there are files left in the reader queue and another backup or restore job is submitted, the existing reader file will be processed instead of the one just submitted. After a backup or restore job failure we checked the reader queue of the BKRWRKnn users and purged all existing files.

### Use different templates for each SSI member

To effectively back up and restore the data of a member in an SSI cluster, it is suggested to set up a backup template for each member system of an SSI. Each template would include the IDENTITIES defined to that system.

## LGR and performance

This section describes our observations on relocating Linux guests.

### Preferred relocation options for LGR

Quiesce time and relocation time improves substantially when relocations are done in serial mode. The overall relocation time for several Linux guests shows almost no difference when the relocation is done in serial mode (option SYNC) instead of parallel mode (option ASYNC). Therefore, use the default option SYNC if possible.

With option IMMEDIATE you can improve the relocation times at the expense of the quiesce times. Chapter 7, “Benchmarks for relocating Linux on System z guests using LGR” on page 73 shows the combinations that did best with specific success measures considered.

### Number of CTCs in ISFC setup improves the relocation times

If you are not satisfied with your relocation times, the numbers of CTCs in the ISFC setup can improve relocation times substantially. z/VM allows the definition up to 16 CTCs. For further details, see section 7.2, “Relocation benchmark dependent on the number of CTCs” on page 76.

# Related publications

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

## IBM Redbooks publications

The following IBM Redbooks publications provide additional information about the topic in this document. Note that some publications referenced in this list might be available in softcopy only.

- ▶ *An introduction to z/VM Single System Image (SSI) and Live Guest Relocation (LGR)*, SG24-8006
- ▶ *Where are the LUN numbers on a DS8000?*, TIPS0598

You can search for, view, download, or order these documents and other Redbooks, Redpapers, Web Docs, drafts, and additional materials, at the following website:

[ibm.com/redbooks](http://ibm.com/redbooks)

## Other publications

These publications are also relevant as further information sources:

- ▶ *z/VM CP Planning and Administration version 6 release 2*, SC24-6178
- ▶ *z/VM Getting Started with Linux on System z version 6 release 2*, SC24-6194
- ▶ *z/VM Installation Guide version 6 release 2*, GC24-6246
- ▶ *z/VM CMS File Pool Planning, Administration and Operation*, SC24-6167
- ▶ *z/VM Migration Guide*, GC24-6201
- ▶ *Program Directory for Performance Toolkit for VM for use with z/VM version 6 release 2*, GI11-4351-00

## Online resources

These websites are also relevant as further information sources:

- ▶ Introduction to SCSI over FCP for Linux on System z  
<http://www.vm.ibm.com/education/lvc/lvc1020c.pdf>
- ▶ Backing Up and Restoring a z/VM Cluster and Linux on System z Guests  
<ftp://ftp.software.ibm.com/software/stormgmt/zvm/backup/BackupScenariosforzVMzLinux20120325.pdf>
- ▶ Live Guest Relocation  
<http://www.vm.ibm.com/perf/reports/zvm/html/620lgr.html>

- ▶ ISFC Improvements  
<http://www.vm.ibm.com/perf/reports/zvm/html/620isfc.html>
- ▶ z/VM V6R2.0 Monitor Records  
<http://www.vm.ibm.com/pubs/mon620/index.html>

## Help from IBM

IBM Support and downloads

[ibm.com/support](http://ibm.com/support)

IBM Global Services

[ibm.com/services](http://ibm.com/services)

# Index

## Numerics

5697J06B 92–93

## A

architecture domain 49

## B

Backup\_Job\_Workers 94  
BKR\_Output\_Spec 94  
BKRAADMIN 88  
BKRBKUP 90  
BKRCATLG 89  
BKRJOB 95, 98  
BKRLIST 95, 99  
BKRSFS  
    BKRAADMIN.JOBDEFS 93  
BKRSVSFS 86  
BKRSYSTEM CONFIG 93  
BKUSER 96, 99  
BKUSERS NAMES 92–93  
BKUSERS NAMESAMP 92  
BKRVOL 98  
BKRWRK01 90  
BKSYSTEM CONFIG 93  
BKUPDISK 92

## C

catalog service 89  
chccwdev command 33  
chkconfig command 37  
CHPID 25  
    channel path identifier 25  
CTC channels 2

## D

Data Storage Control Line Interface 30  
data storage management 83  
DEDICATE statement 33  
default relocation domain SSI 44  
DIRMAINT 33  
Dirmaint 1  
DS8300 24–25, 27  
    GUI interface 29  
    storage controller 28  
DSCLI 30

## E

e2label command 39  
EQID 3, 31–32

## F

FCONX \$PROFILE 60  
FCP 24  
    channel paths 25  
    connection in an SSI cluster 24  
    FCP channels 25  
    Hardware prerequisites 26  
    Hint 26  
    query fcp 27  
    querying FCP to see if they channel is active 25  
FCP adapters 26  
FCP card 26  
FCP channels 28, 32  
    attach 25  
    attached 25  
FCP devices 26  
fibre channel protocol 24  
file pool server 85

## H

HMC 27  
    CPC Configuration 27  
    FCP Configuration 27  
    Hardware Management Console 27  
    Support Element 27  
host connections 29

## I

IBM Backup and Restore Manager for z/VM 83  
IBM Performance Toolkit  
    activate Linux guest monitoring 61  
    web interface activation 60  
IBM Performance Toolkit for VM 59–60  
    how relocation time can be monitored 59  
    New data screens for SSI 59  
    performance aspects of LGR 59  
identity entry 6  
Inter-System Facility for Communications 2  
IOCDS 26  
    Input/Output Configuration Data Set 26

## J

Job\_Template 92

## L

Live guest relocation  
    vmrelocate command 25  
live guest relocation 1  
Logical Unit Number 26  
LUN 26  
    Logical Unit Number 26  
LUN ID 27, 34

LUN mapping 27  
LUN masking 28  
lunmap 31

## M

MAC address 4  
MAC address assignment 4  
master backup service 90  
mkfs command 39  
Monitor Records 70  
MONWRITE 70  
mount command 40  
Multipathing 37

## N

NPIV 28  
    N\_Port ID Virtualization 27  
    NPIV mode 27  
    N-Port Id Virtualization 28

## O

Open System Volume 30  
Overview four-member cluster 10  
Overview two-member cluster 12

## P

Performance Toolkit 1  
persistent data record 5  
PMAINT 32  
port\_add file 34  
Programmable Operator 1  
PROTODIR 91  
put2prod command 88  
pvcreate command 38

## Q

quiesce time 69

## R

RACF® 1  
Redbooks website 103  
    Contact us xi  
redundancy 26  
relocation domain 3  
relocation time 69  
REXX 88  
RSCS 1

## S

s390utils package 33  
SAN 24–25  
    restricting access to 28  
    Storage Area Netork 24  
    Storage Area Network 25  
SAN zoning 28  
SCSI 23

Defined to the Linux on System z guest 24  
definition

    hardware prerequisites 26  
    example setup 24  
    Hardware prerequisites 24  
    setup 23  
    Small Computer System Interface 23  
Shared File System 85  
Shared File System server and file pool 85  
single 1  
single system image 1  
single system image (SSI) 3  
SSI data menu 66  
SSI Cluster 1  
start the multipath daemon 38  
Storage Controller 29  
storage controller 24  
SUBCONFIG 91  
System z  
    z10 24  
    z196 24

## T

TCPIP 1

## U

unit\_add file 34  
user entry 6

## V

vgcreate command 38  
volume group 29

## W

Work\_Area 92  
WWPN 24  
    define 26  
    define a mapping 26  
    identification 27  
    mapping list 28  
World Wide Port Name 24

## Z

z/VM Single System Image 1  
z/VM system 3  
zipl command 36

## Using z/VM v 6.2 Single System Image (SSI) and Live Guest Relocation (LGR)

(0.2"spine)  
0.17"<->0.473"  
90<->249 pages







# Using z/VM v 6.2 Single System Image (SSI) and Live Guest Relocation (LGR)

**LGR performance comparisons in different SSI environments**

**Application workloads and relocation domains**

**Backup and restore in an SSI environment**

In this IBM Redbooks publication, we expand upon the concepts and experiences described in *An introduction to z/VM Single System Image (SSI) and Live Guest Relocation (LGR)*, SG24-8006. An overview of that book is provided in Chapter 1, "Overview of SSI and LGR" on page 1.

In writing this book, we re-used the same lab environment used in the first book, but expanded it to include IBM DB2 v10 on Linux on System z, two IBM WebSphere Application Server environments, and added a WebSphere application, used for performance benchmarking, which provided a workload that allowed us to observe the performance of the WebSphere Application Server during relocation of the z/VM 6.2 member that was hosting the application server.

Additionally, this book examines the use of small computer system interface (SCSI) disks in the z/VM v6.2 environment and the results of using single system images (SSI) and live guest relocation (LGR) in this type of environment.

In the previous book, a detailed explanation of relocation domains was provided. In this book, we expand that discussion and provide use cases of relocation domains in different situations.

Finally, because the ability to back up and restore your data is of paramount importance, we have provided a discussion about how to use one tool, the IBM Backup and Restore Manager for z/VM, which can be used in the new z/VM6.2 environment. We provide a brief overview of the tool and describe the changes in the installation process as a result of using single system image clusters. We also demonstrate how to set up the configuration file, and how to back up and restore both a user and an identity.

This publication is intended for IT architects who will be responsible for designing the system and IT specialists who will have to build the system.

## 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](http://ibm.com/redbooks)

SG24-8039-00

ISBN 0738437042