

z/VM Single System Image & Live Guest Relocation – Overview

Session 14585

John Franciscovich
francisj@us.ibm.com



Trademarks

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

IBM*	System z10*	System z196
IBM Logo*	Tivoli*	System z114
DB2*	z10 BC	System zEC12
DS8000*	z9*	System zBC12
Dynamic Infrastructure*	z/OS*	
FICON*	z/VM*	
GDPS*	z/VSE	
HiperSockets	zEnterprise*	
HyperSwap*		
Parallel Sysplex*		
PR/SM		
RACF*		
System z*		

* Registered trademarks of IBM Corporation

The following are trademarks or registered trademarks of other companies.

OpenSolaris, Java and all Java-based trademarks are trademarks of Sun Microsystems, Inc. in the United States, other countries, or both.
 Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.
 INFINIBAND, InfiniBand Trade Association and the INFINIBAND design marks are trademarks and/or service marks of the INFINIBAND Trade Association.
 UNIX is a registered trademark of The Open Group in the United States and other countries.
 Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.

All other products may be trademarks or registered trademarks of their respective companies.

Notes:

Performance is in Internal Throughput Rate (ITR) ratio based on measurements and projections using standard IBM benchmarks in a controlled environment. The actual throughput that any user will experience will vary depending upon considerations such as the amount of multiprogramming in the user's job stream, the I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve throughput improvements equivalent to the performance ratios stated here.

IBM hardware products are manufactured from new parts, or new and serviceable used parts. Regardless, our warranty terms apply.

All customer examples cited or described in this presentation are presented as illustrations of the manner in which some customers have used IBM products and the results they may have achieved. Actual environmental costs and performance characteristics will vary depending on individual customer configurations and conditions.

This publication was produced in the United States. IBM may not offer the products, services or features discussed in this document in other countries, and the information may be subject to change without notice. Consult your local IBM business contact for information on the product or services available in your area.

All statements regarding IBM's future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only.

Information about non-IBM products is obtained from the manufacturers of those products or their published announcements. IBM has not tested those products and cannot confirm the 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.

Prices subject to change without notice. Contact your IBM representative or Business Partner for the most current pricing in your geography.

Disclaimer

The information contained in this document has not been submitted to any formal IBM test and is distributed on an "AS IS" basis without any warranty either express or implied. The use of this information or the implementation of any of these techniques is a customer responsibility and depends on the customer's ability to evaluate and integrate them into the operational environment. While each item may have been reviewed by IBM for accuracy in a specific situation, there is no guarantee that the same or similar results will be obtained elsewhere. Customers attempting to adapt these techniques to their own environments do so at their own risk.

In this document, any references made to an IBM licensed program are not intended to state or imply that only IBM's licensed program may be used; any functionally equivalent program may be used instead.

Any performance data contained in this document was determined in a controlled environment and, therefore, the results which may be obtained in other operating environments may vary significantly. Users of this document should verify the applicable data for their specific environments.

All statements regarding IBM's plans, directions, and intent are subject to change or withdrawal without notice, and represent goals and objectives only. This is not a commitment to deliver the functions described herein

Topics

- Introduction
 - z/VM Single System Image (SSI) Clusters
 - Live Guest Relocation (LGR)

- Major Attributes of a z/VM SSI Cluster
 - Installation and service
 - Configuration
 - Shared source directory
 - Shared system resources and data

- Brief Overview of z/VM SSI Cluster Management

Sessions 14601, 14602, 14603:
z/VM Installation or Migration or Upgrade Hands-on Lab
Tuesday, 1:30, 3:00, and 4:30
Platinum Ballroom, Salon 7 (Richard Lewis)

***Introduction
to
SSI and LGR***

Multi-system Virtualization with z/VM Single System Image (SSI)

- VMSSI Feature of z/VM 6.2 and 6.3

- Up to 4 z/VM instances (members) in a single system image (SSI) cluster
 - Same or different CECs

- Provides a set of shared resources for the z/VM systems and their hosted virtual machines
 - Managed as a single resource pool

- **Live Guest Relocation** provides virtual server mobility
 - Move Linux virtual servers (guests) non-disruptively from one from one member of the cluster to another

- A single SSI cluster can have members running both z/VM 6.2 and 6.3

z/VM Single System Image (SSI) Cluster

- Common resource pool accessible from all members
 - Shared disks for system and virtual server data
 - Common network access

- All members of an SSI cluster are part of the same ISFC collection

- CP validates and manages all resource and data sharing
 - Uses ISFC messages that flow across channel-to-channel connections between members
 - No virtual servers required

- **NOT** compatible with CSE (Cross System Extensions)
 - Cannot have SSI and CSE in same cluster
 - Disk sharing between an SSI cluster and CSE cluster requires manual management of links
 - No automatic link protection or cache management

 - *Support of most CSE function has been withdrawn in z/VM 6.3*
 - XLINK remains supported

Benefits and Uses of z/VM SSI Clusters

- Horizontal growth of z/VM workloads
 - Increased control over server sprawl
 - Distribution and balancing of resources and workload

- Flexibility for planned outages for service and migration
 - z/VM
 - Hardware
 - Less disruptive to virtual server workloads

- Workload testing
 - Different service/release levels
 - Various environments (stress, etc.)
 - New/changed workloads and applications can be tested before moving into production

- Simplified system management of a multi-z/VM environment
 - Concurrent installation of multiple-system cluster
 - Single maintenance stream
 - Reliable sharing of resources and data

SSI Cluster Considerations

- Physical systems must be close enough to allow
 - FICON CTC connections
 - Shared DASD
 - Common network and disk fabric connections

- SSI Installation to SCSI devices is not supported
 - Guests may use SCSI devices

- If using RACF, the database must reside on a fullpack 3390 volume
 - Single RACF database shared by all members of the cluster

- Live Guest Relocation is only supported for Linux on System z guests

Live Guest Relocation

- Relocate a running Linux virtual server (guest) from one member of an SSI cluster to another
 - Load balancing
 - Move workload off a member requiring maintenance

- **VMRELOCATE** command initiates and manages live guest relocations
 - Check status of relocations in progress
 - Cancel a relocation in progress(relocations are **NOT** automatically done by the system)

- Guests continue to run on source member while they are being relocated
 - Briefly quiesced
 - Resumed on destination member

- If a relocation fails or is cancelled, the guest continues to run on the source member

Live Guest Relocation ...

- Relocation capacity and duration is determined by various factors including:
 - ISFC bandwidth
 - Guest size and activity
 - Destination system's memory size and availability
 - Load on destination system

- In order to be relocated, a guest must meet eligibility requirements, including:
 - The architecture and functional environment on destination member must be comparable
 - Relocation domains can be used define sets of members among which guests can relocate freely

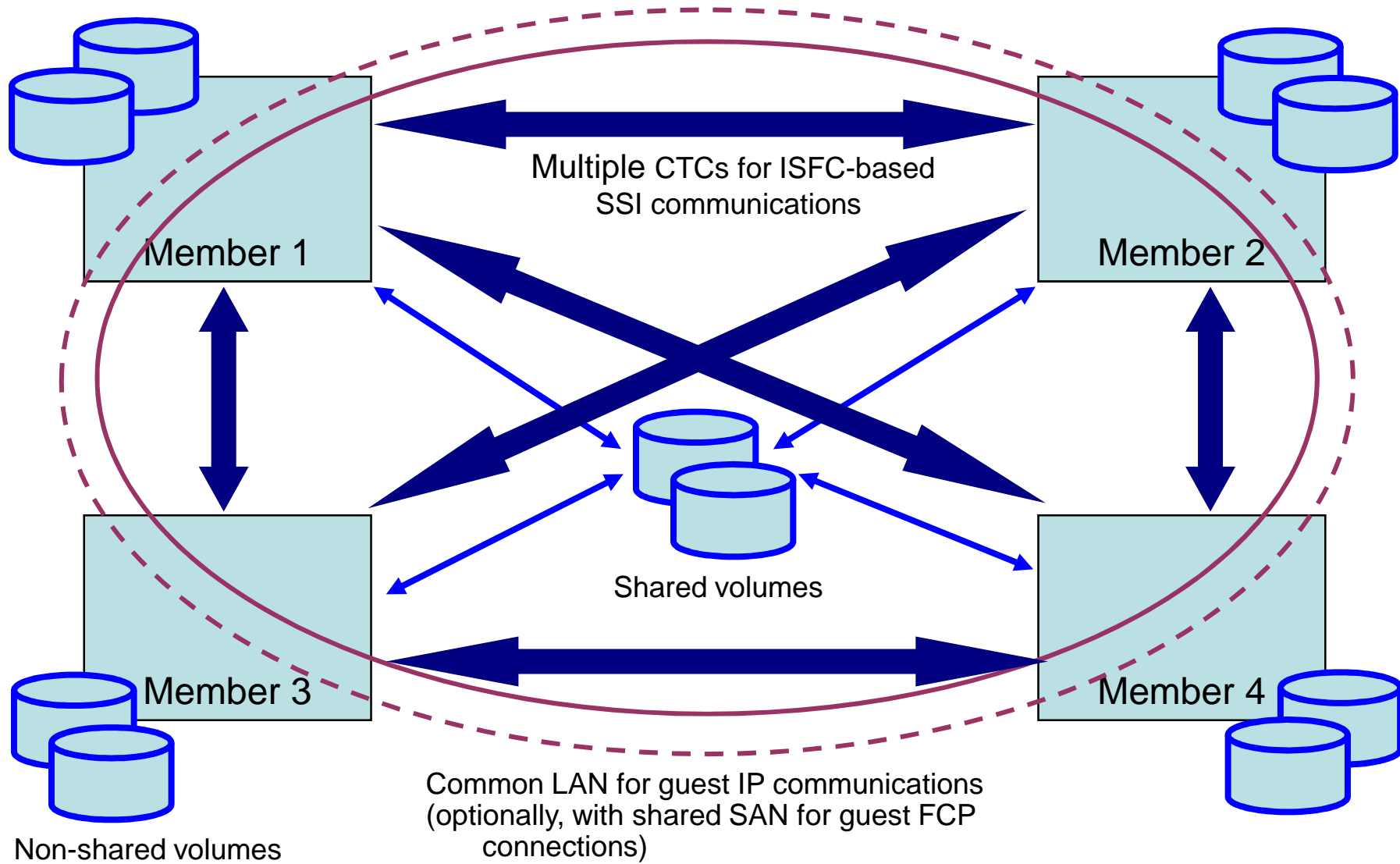
 - Devices and resources used by the guest must be shared and available on the destination member

 - Is it "safe" to move the guest?

 - Eligibility checks are repeated multiple times during a relocation

***Major Attributes of a
z/VM SSI Cluster***

z/VM SSI Cluster



Multisystem Installation

```
Select a System Type: Non-SSI or SSI (SSI requires the SSI feature)
  _ Non-SSI Install:      System Name _____
  X SSI Install:         Number of Members 4      SSI Cluster Name SAMPLE
```

- SSI cluster can be created with a single z/VM install
 - Cluster information is specified on installation panels
 - Member names
 - Volume information
 - Channel-to-channel connections for ISFC
 - Specified number of members are installed and configured as an SSI cluster
 - Shared system configuration file
 - Shared source directory
- Non-SSI single system installation also available
 - System resources defined in same way as for SSI
 - Facilitates later conversion to an SSI cluster

Upgrade Installation

- New technique for upgrading from z/VM 6.2 to 6.3
 1. Install new release system as temporary second level guest of system being upgraded
 2. Move new code level to current system

- Current (to be upgraded) system can be either:
 - Non-SSI
 - Only member of single-member SSI cluster
 - Any member of a multi-member SSI cluster

- In a multi-member SSI cluster, only one member at a time is upgraded
 - Minimum impact to the cluster and other members
 - Can thoroughly test new release on one member before upgrading other members

Sessions 14586: z/VM Upgrade In Place Installation

Wednesday, 8:00 - Platinum Ballroom, Salon 5 (Richard Lewis)

Installation and Service is Different Starting with z/VM 6.2 !!

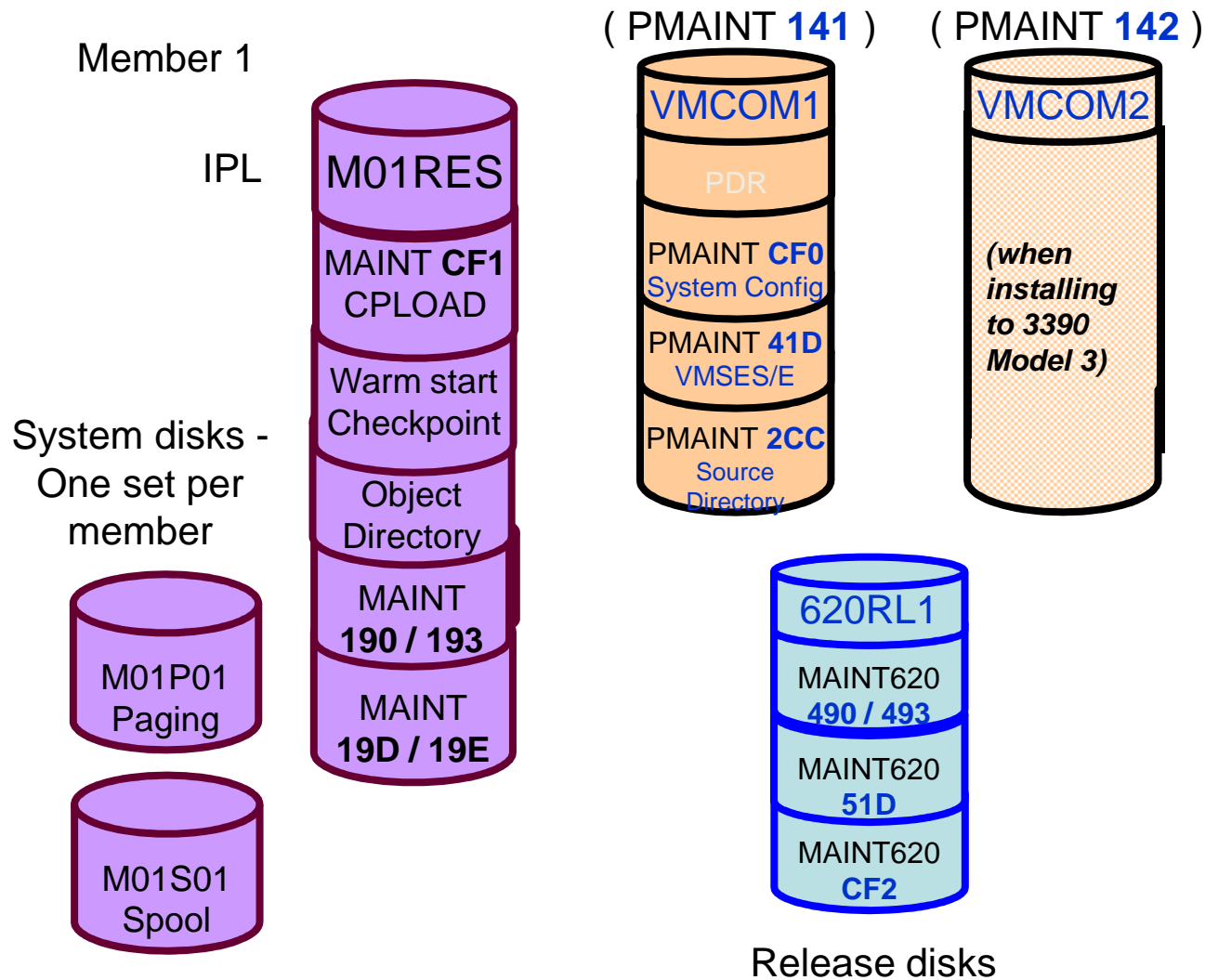
- Different for both SSI and non-SSI installs
 - Install and service tasks distributed among different "maint" userids
 - MAINT
 - PMAINT
 - MAINT620 or MAINT630 (depending on z/VM release)

 - Disks
 - New disk volumes
 - Owning userids and volumes of parm disks and various minidisks are changed
 - New CF0 parm disk now contains system configuration file
 - Source directory (2CC)
 - VMSES/E (41D)
 - CF2 parm disk (for applying service)

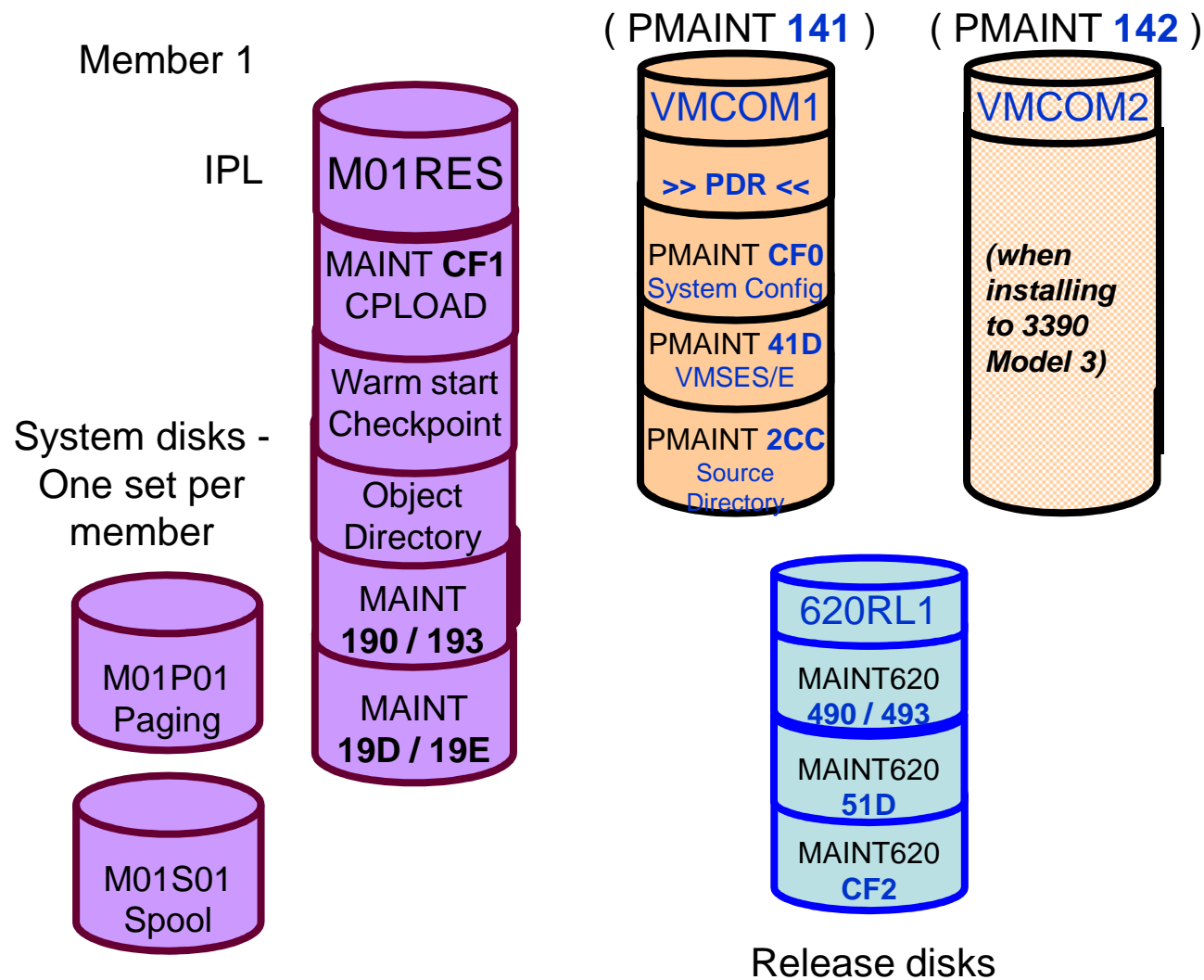
 - New structure and statements for:
 - System configuration file
 - Directory

- Installation and service programs restructured
 - If you use customized programs, make sure you understand new structure

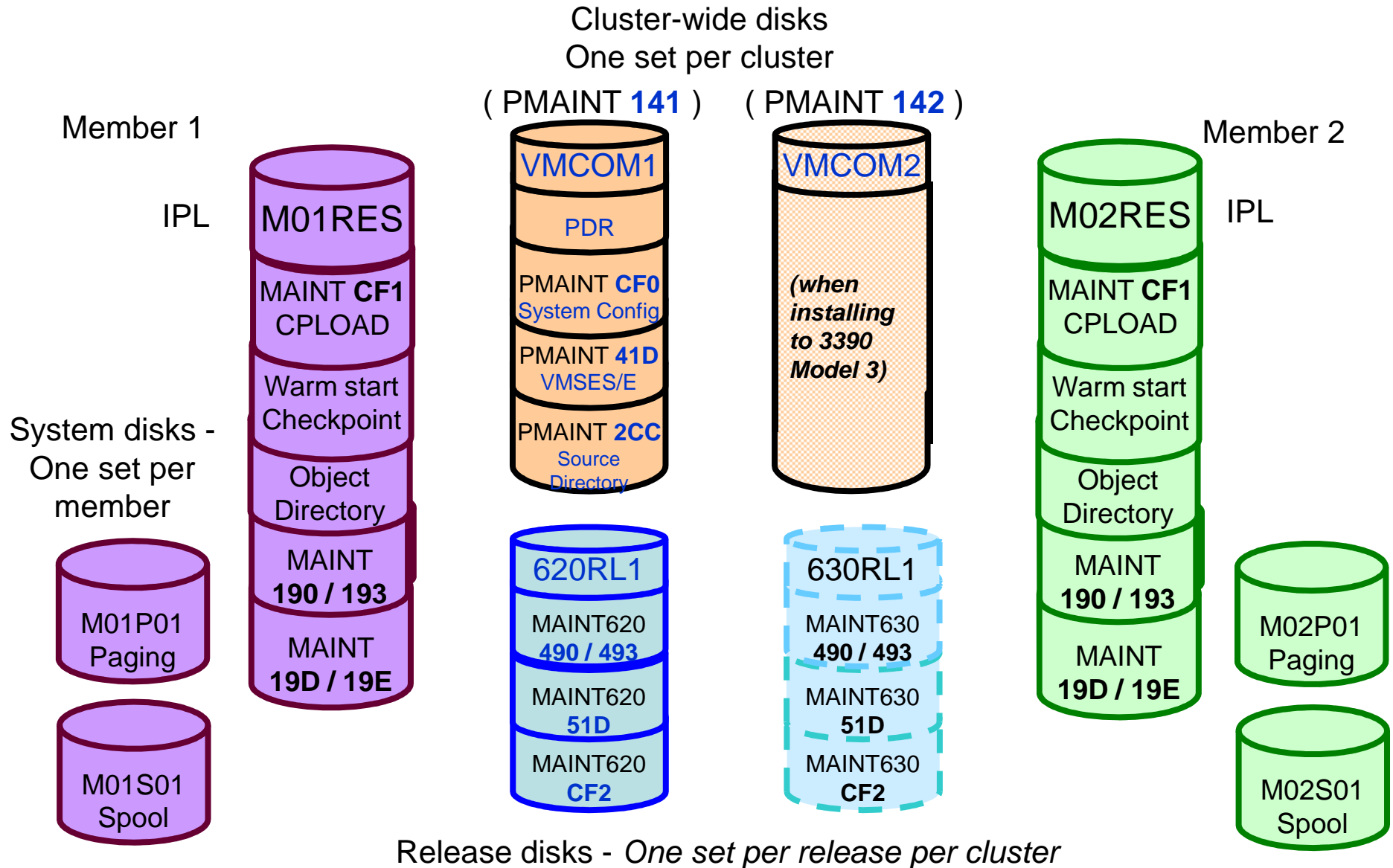
System Volumes and Minidisks – *Non-SSI Install*



System Volumes and Minidisks – SSI Single Member



System Volumes and Minidisks – SSI Multiple Members

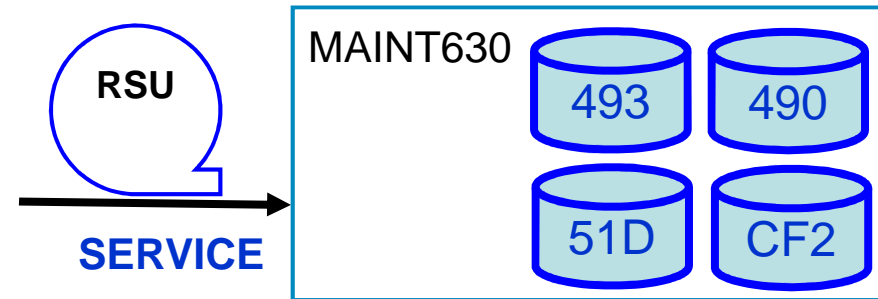


Applying Service

Single Maintenance Stream per release

For z/VM 6.3 members:

1. Logon to MAINT630 on *either* member and run **SERVICE**

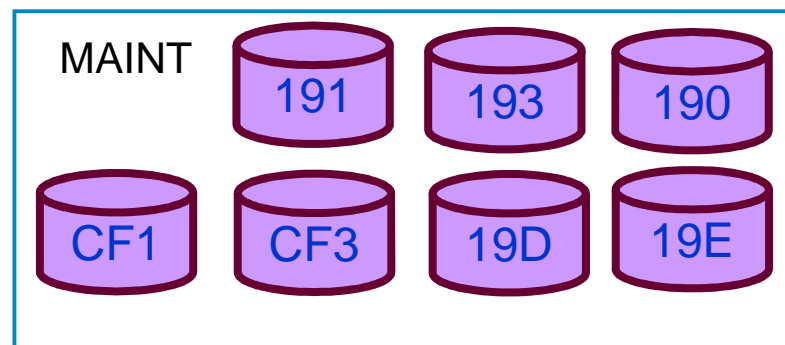


Service applied privately to each member

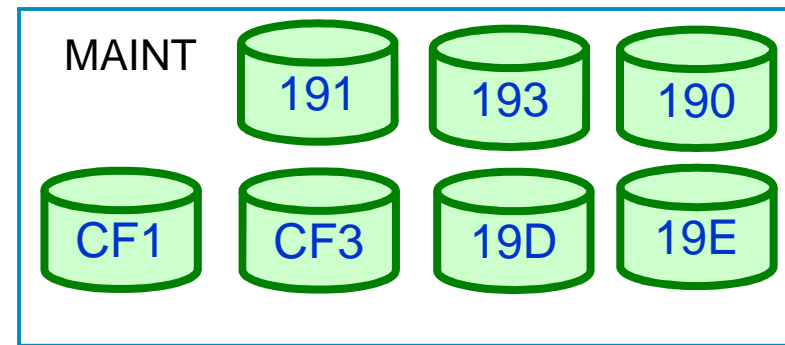
2. Logon to MAINT630 on Member 1 and **PUT2PROD**
3. Logon to MAINT630 on Member 2 and **PUT2PROD**

PUT2PROD

PUT2PROD



Member 1



Member 2

Shared System Configuration File

- Resides on new shared parm disk
 - PMAINT CF0

- Can include member-specific configuration statements
 - Record qualifiers
 - New BEGIN/END blocks

- Define each member's system name
 - Enhanced SYSTEM_IDENTIFIER statement
 - LPAR name can be matched to define system name
`System_Identifier LPAR LP1 VMSYS01`

 - System name can be set to the LPAR name
`System_Identifier LPAR * &LPARNAME`

- Define cluster configuration (cluster name and member names)
`SSI CLUSTERA PDR_VOLUME VMCOM1,
SLOT 1 VMSYS01,
SLOT 2 VMSYS02,
SLOT 3 VMSYS03,
SLOT 4 VMSYS04`

Shared System Configuration File...

- Identify direct ISFC links between members
 - One set of statements for each member

```
VMSYS01: BEGIN
          ACTIVATE ISLINK 912A    /* Member 1 TO Member 2 */
          ACTIVATE ISLINK 913A    /* Member 1 TO Member 3 */
          ACTIVATE ISLINK 914A    /* Member 1 TO Member 4 */
VMSYS01: END
```

- Define CP Owned volumes
 - Shared
 - SSI common volume
 - Spool
 - Private
 - Sysres
 - Paging
 - Tdisk

Shared System Configuration File – CP-Owned Volumes...

```
/* **** */
/*          PAGE & TDISK VOLUMES */
/* To avoid interference with spool volumes and to          */
/* automatically have all unused slots defined as          */
/* "Reserved", begin with slot 255 and assign them in      */
/* descending order.                                       */
/* **** */

VMSYS01: BEGIN
        CP_Owned Slot 254 M01T01
        CP_Owned Slot 255 M01P01
VMSYS01: END

VMSYS02: BEGIN
        CP_Owned Slot 254 M02T01
        CP_Owned Slot 255 M02P01
VMSYS02: END

VMSYS03: BEGIN
        CP_Owned Slot 254 M03T01
        CP_Owned Slot 255 M03P01
VMSYS03: END

VMSYS04: BEGIN
        CP_Owned Slot 254 M04T01
        CP_Owned Slot 255 M04P01
VMSYS04: END
```


Persistent Data Record (PDR)

- Cross-system serialization point on disk
 - Must be a shared 3390 volume (VMCOM1)
 - Created and viewed with new FORMSSI utility
- Contains information about member status
 - Used for health-checking
- Heartbeat data
 - Ensures that a stalled or stopped member can be detected

```
formssi display efe0
HCPPDF6618I Persistent Data Record on device EFE0 (label VMCOM1) is for CLUSTERA
HCPPDF6619I PDR                                state: Unlocked
HCPPDF6619I                                time stamp: 07/11/10 21:22:03
HCPPDF6619I      cross-system timeouts: Enabled
HCPPDF6619I PDR  slot 1                        system: VMSYS01
HCPPDF6619I                                state: Joined
HCPPDF6619I      time stamp: 07/11/10 21:22:00
HCPPDF6619I                                last change: VMSYS01
HCPPDF6619I PDR  slot 2                        system: VMSYS02
HCPPDF6619I                                state: Joined
HCPPDF6619I      time stamp: 07/11/10 21:21:40
HCPPDF6619I      last change: VMSYS02
HCPPDF6619I PDR  slot 3                        system: VMSYS03
HCPPDF6619I                                state: Joining
HCPPDF6619I      time stamp: 07/11/10 21:21:57
HCPPDF6619I      last change: VMSYS03
HCPPDF6619I PDR  slot 4                        system: VMSYS04
HCPPDF6619I                                state: Down
HCPPDF6619I      time stamp: 07/02/10 17:02:25
HCPPDF6619I      last change: VMSYS02
```

Ownership Checking – CP-Owned Volumes

- Each CP-owned volume in an SSI cluster will be marked with ownership information
 - Cluster name
 - System name of the owning member
 - The marking is created using CPFMTXA

- Ensures that one member does not allocate CP data on a volume owned by another member
 - Warm start, checkpoint, spool, paging, temporary disk, directory

- No need to worry about OWN and SHARED on CP_OWNED definitions
 - Ignored on SSI members

- QUERY CPOWNED enhanced to display ownership information

Defining Virtual Machines – Shared Source Directory

- All user definitions in a single shared source directory

- Run DIRECTXA on each member

- No system affinity (SYSAFFIN)

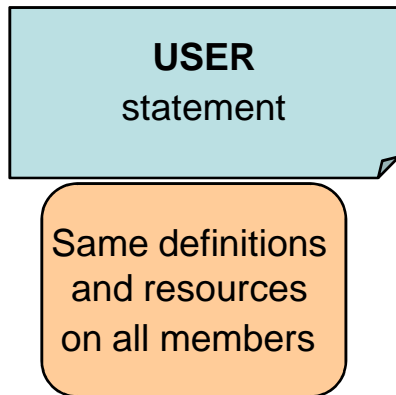
- Identical object directories on each member

- Single security context
 - Each user has same access rights and privileges on each member

Using a directory manager is strongly recommended!

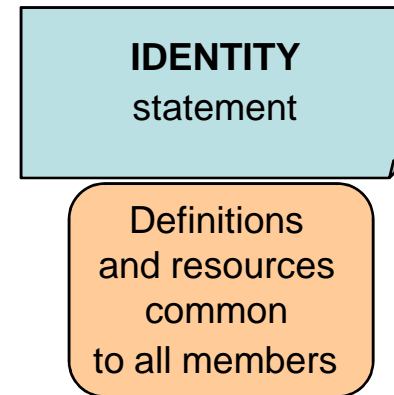
Shared Source Directory – Virtual Machine Definition Types

Single Configuration Virtual Machine (traditional)

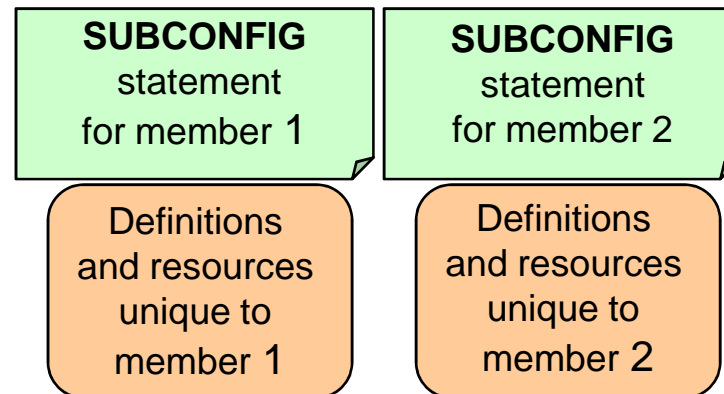


- May log on to any member
- Only one member at a time
- General Workload
 - Guest Operating Systems
 - Service virtual machines requiring only one login in the cluster

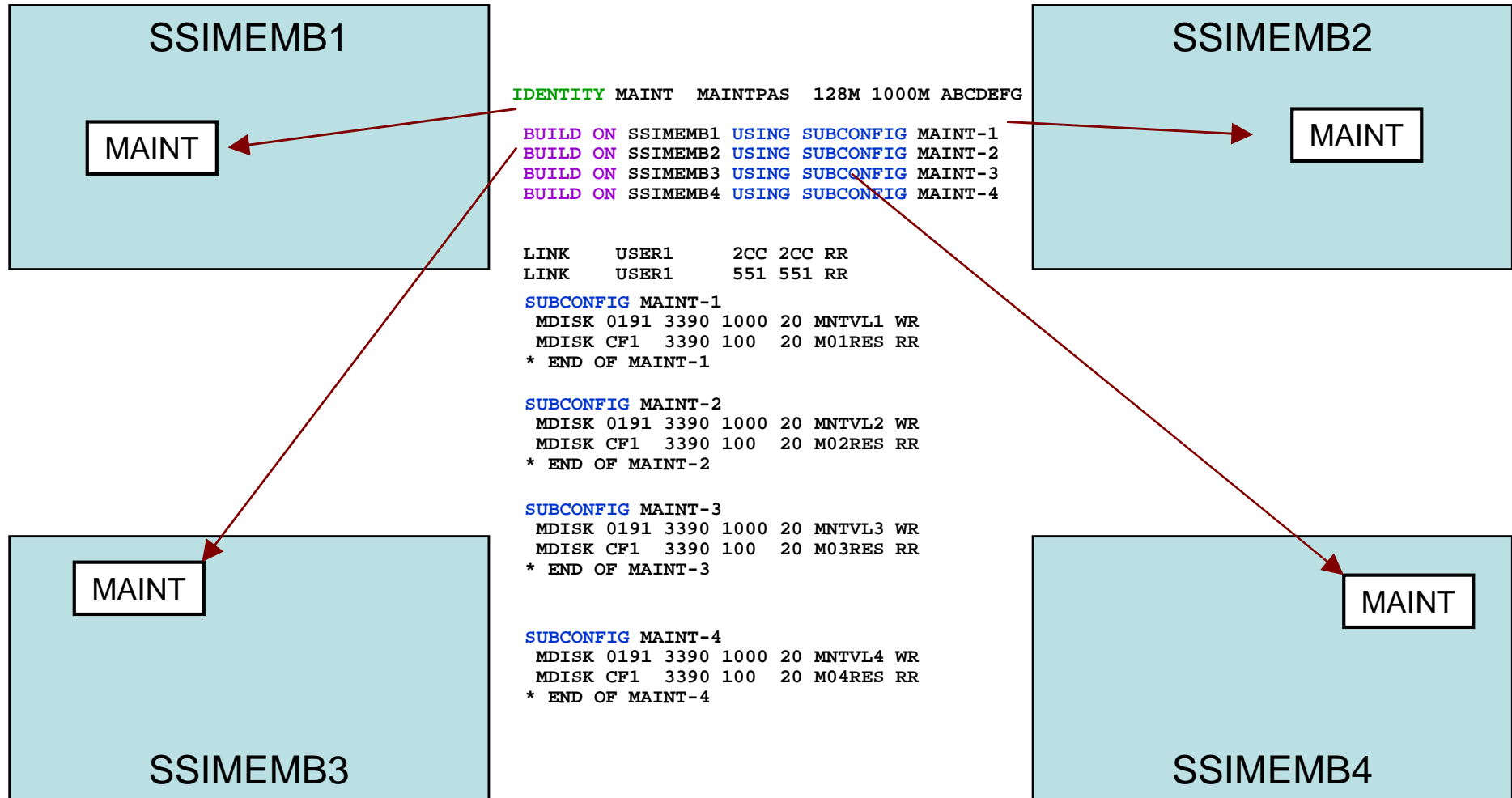
Multiconfiguration Virtual Machine (new)



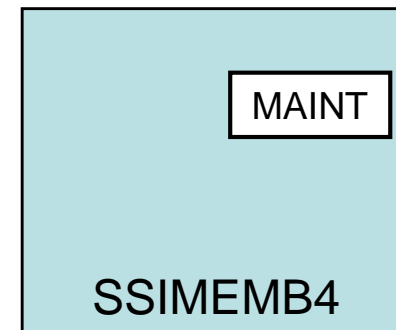
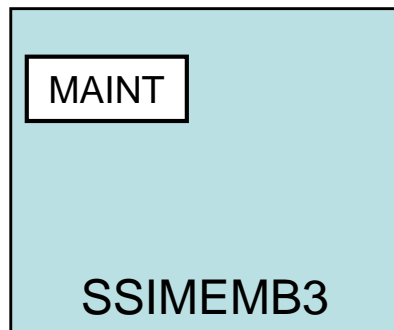
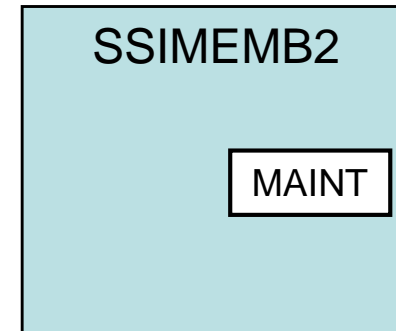
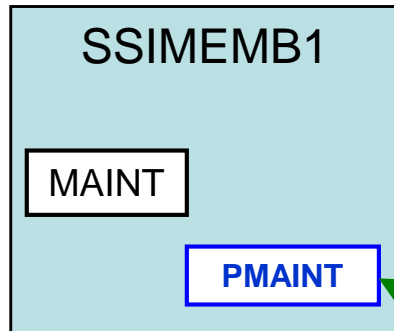
- May log on to multiple members at the same time (known by IDENTITY name)
- System support virtual machines
- Service virtual machines



Shared Source Directory – Multiconfiguration Virtual Machines



Shared Source Directory – Single Configuration Virtual Machines



```

USER PMAINT    P           128M 1000M G

MDISK 141 3390 000 END PXA2D0 MR
MDISK 142 3390 000 END PXA2D1 MR
MDISK 2CC 3390 121 010 PXA2D0 MR
MDISK CF0 3390 001 120 PXA2D0 RR
MDISK 191 3390 0351 175 PXA2D0 MR
MDISK 41D 3390 0526 026 PXA2D0 MR
MDISK 550 3390 0552 020 PXA2D0 MR
MDISK 551 3390 0572 040 PXA2D0 MR
    
```

Cross-System Spool

- Spool files are managed cooperatively and shared among all members of an SSI cluster
- Single-configuration virtual machines (most users) have a single logical view of all of their spool files
 - Access, manipulate, and transfer all files from any member where they are logged on
 - Regardless of which member they were created on
- Multiconfiguration virtual machines do not participate in cross-system spool
 - Each instance only has access to files created on the member where it is logged on
- All spool volumes in the SSI cluster are shared (R/W) by all members
 - Each member creates files on only the volumes that it owns
 - Each member can access and update files on all volumes

SLOT	VOL-ID	RDEV	TYPE	STATUS	SSIOWNER	SYSOWNER
10	M01S01	C4A8	OWN	ONLINE AND ATTACHED	CLUSTERA	VMSYS01
11	M02S01	C4B8	SHARE	ONLINE AND ATTACHED	CLUSTERA	VMSYS02
12	M01S02	C4A9	OWN	ONLINE AND ATTACHED	CLUSTERA	VMSYS01
13	M02S02	C4B9	SHARE	ONLINE AND ATTACHED	CLUSTERA	VMSYS02
14	M01S03	C4AA	DUMP	ONLINE AND ATTACHED	CLUSTERA	VMSYS01
15	M02S03	C4BA	DUMP	ONLINE AND ATTACHED	CLUSTERA	VMSYS02
16	-----	----	-----	RESERVED	-----	-----

Cross-System SCIF (Single Console Image Facility)

- Allows a virtual machine (secondary user) to monitor and control one or more disconnected virtual machines (primary users)
- If both primary and secondary users are single configuration virtual machines (SCVM)
 - Can be logged on different members of the SSI cluster
- If either primary or secondary user is a multiconfiguration virtual machine (MCVM)
 - Both must be logged on to the same member in order for secondary user to function in that capacity
 - If logged on different members and primary user is a MCVM
 - SEND commands can be issued to primary user with **AT sysname** operand (new)
 - Secondary user will not receive responses to SEND commands or other output from primary user
 - Output from secondary user will be only be received by primary user on the same member

Primary User or Observee	SECUSER or Observer	If Local	If Remote
SCVM	SCVM	Yes	Yes
SCVM	MCVM	Yes	No
MCVM	SCVM	Yes	No
MCVM	MCVM	Yes	No

Cross-System CP Commands

- New **AT** *command* can be used to issue most privileged commands on a different active member

AT sysname CMD cmdname

- **AT sysname** *operand* can be used to target virtual machines on different active member(s)
 - MESSAGE (MSG)
 - MSGNOH
 - SEND
 - SMSG
 - WARNING

MSG userid AT sysname

- Single-configuration virtual machines are usually found wherever they are logged on
- Multiconfiguration virtual machines require explicit targeting

- CMS TELL and SENDFILE commands require RSCS in order to communicate with multiconfiguration virtual machines on other members

Cross-System Minidisk Management

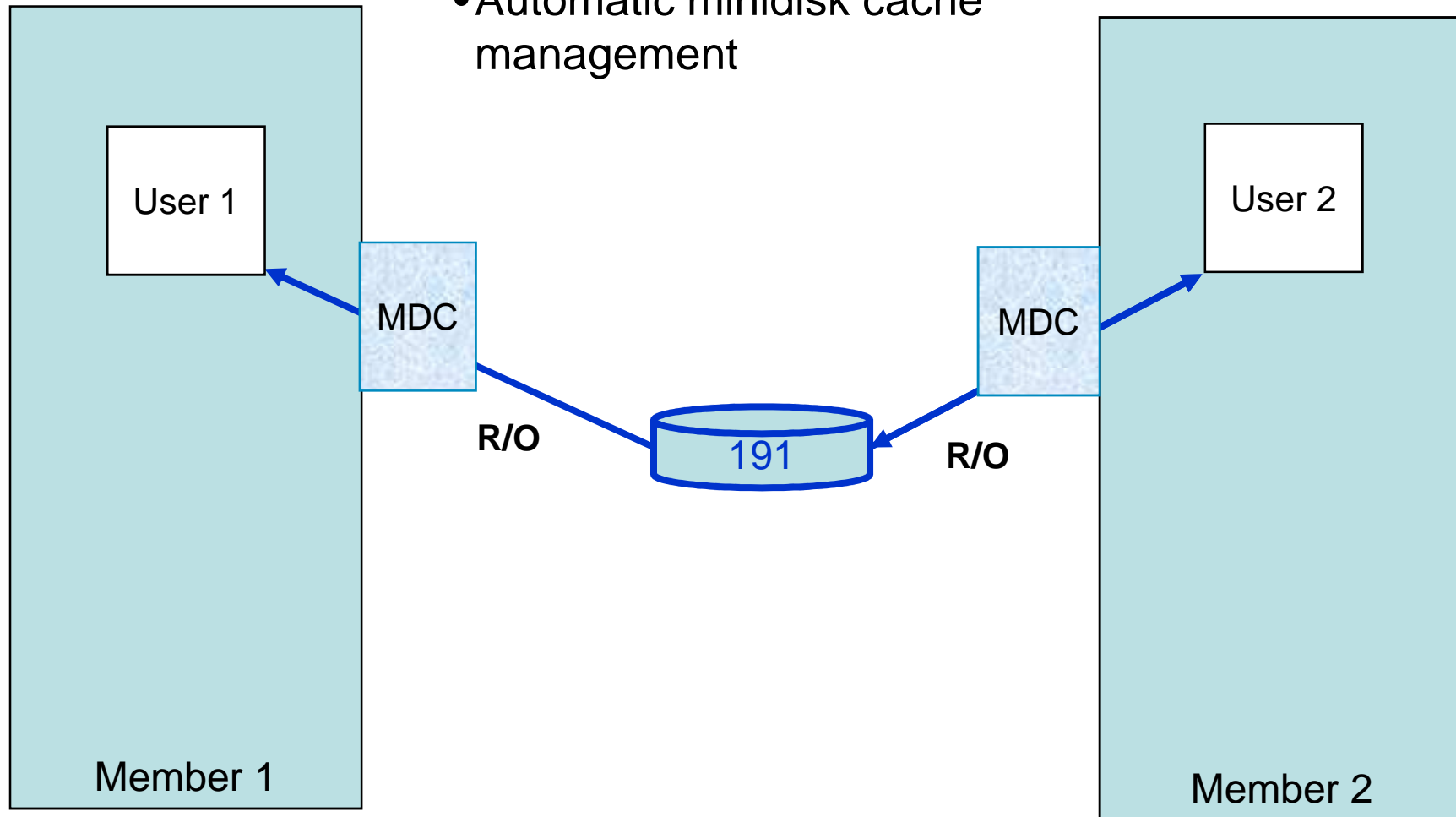
- Minidisks can either be shared across all members or restricted to a single member
 - CP checks for conflicts throughout the cluster when a link is requested

- Virtual reserve/release for fullpack minidisks is supported across members
 - Only supported on one member at a time for non-fullpack minidisks

- Volumes can be shared with systems outside the SSI cluster
 - **SHARED YES** on RDEVICE statement or SET RDEVICE command
 - **Link conflicts must be managed manually**
 - Not eligible for minidisk cache
 - **Use with care**

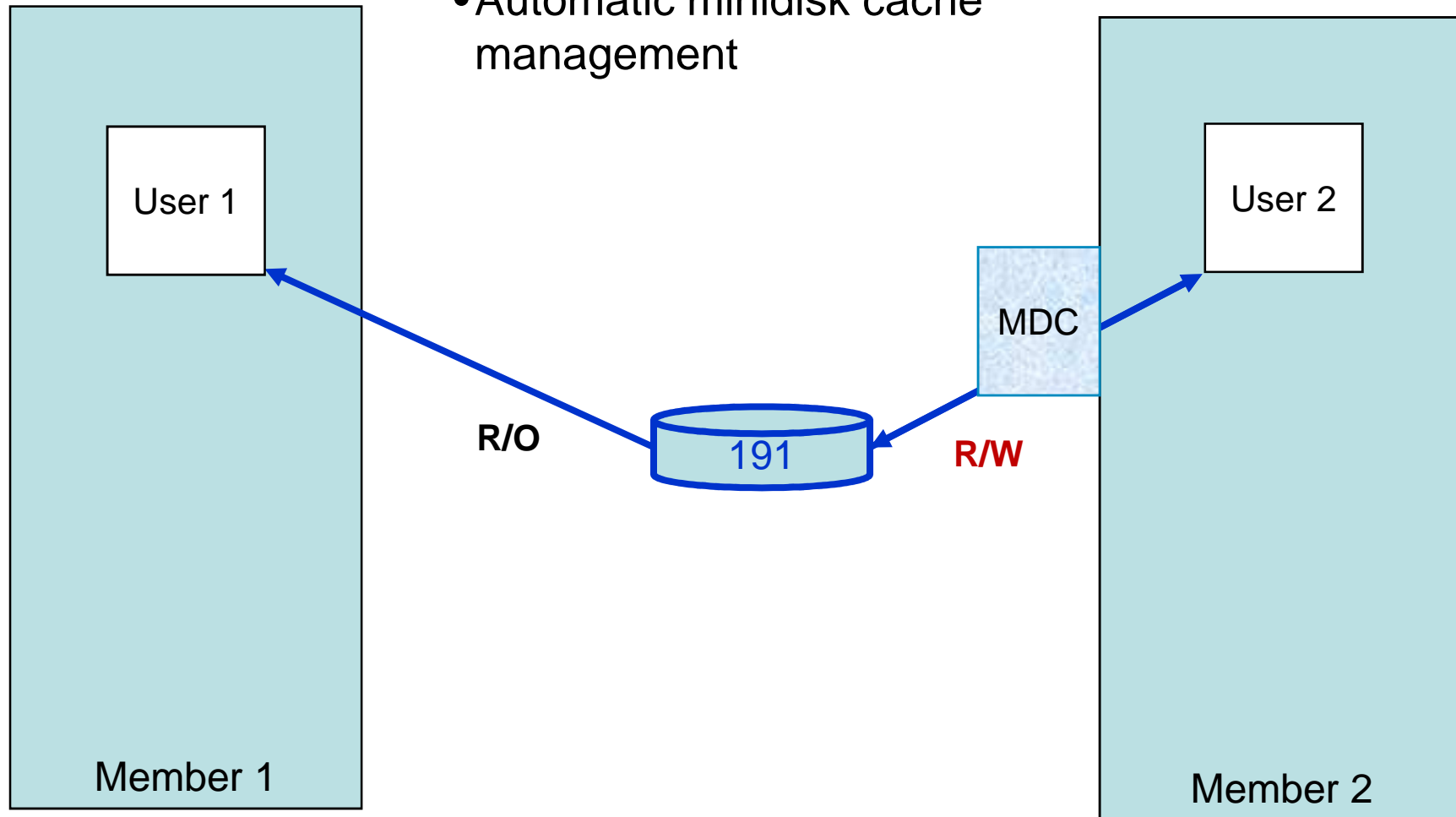
Cross-System Minidisk Management...

- Automatic minidisk cache management



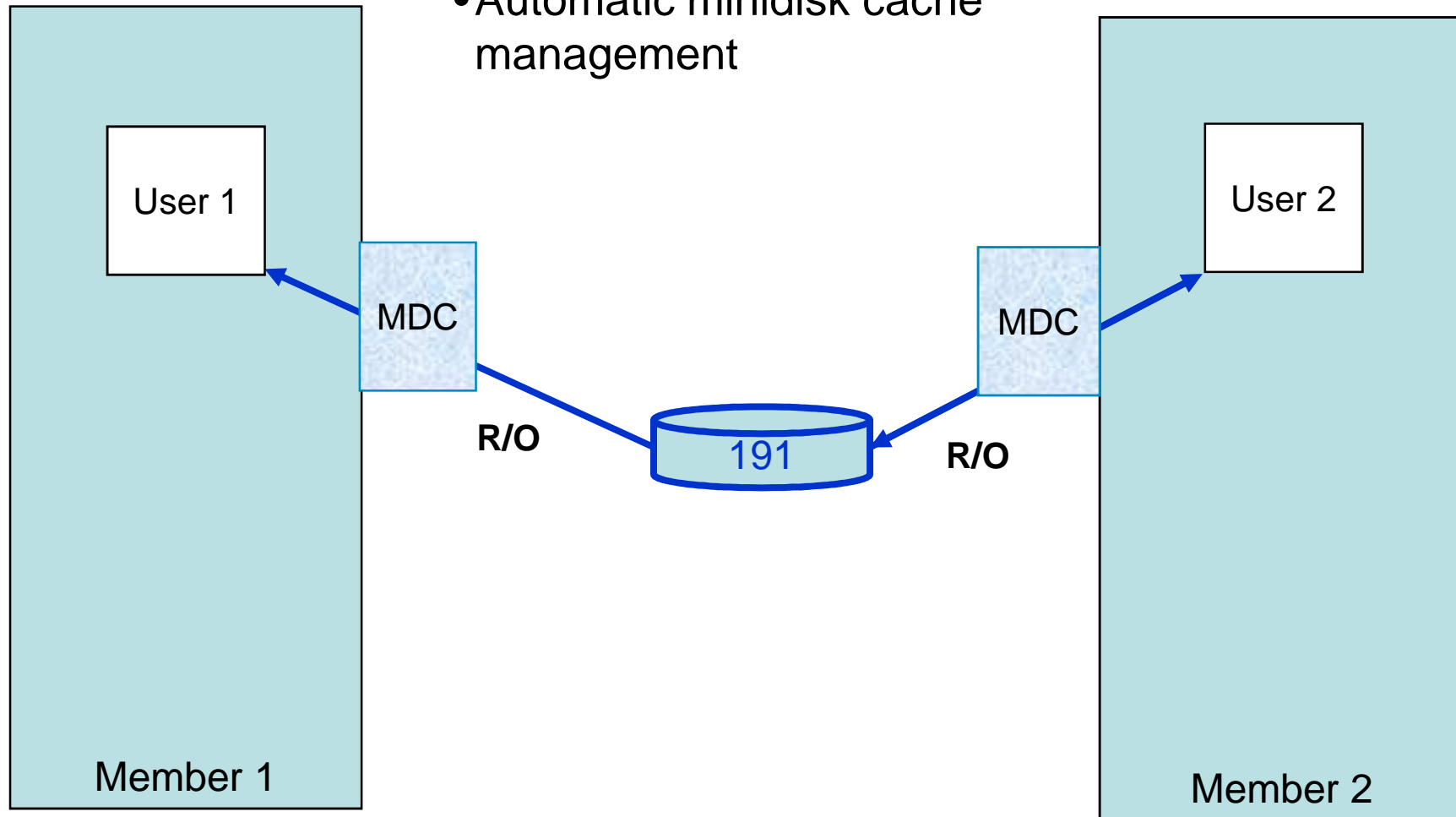
Cross-System Minidisk Management...

- Automatic minidisk cache management



Cross-System Minidisk Management...

- Automatic minidisk cache management



Real Device Management

- Unique identification of real devices within an SSI cluster
 - Ensures that all members are using the same physical devices where required

- CP generates an equivalency identifier (EQID) for each disk volume and tape drive
 - Physical device has same EQID on all members

- EQID for network adapters (CTC, FCP, OSA, Hipersockets) must be defined by system administrator
 - Connected to same network/fabric
 - Conveying same access rights

- EQIDs used to select equivalent device for live guest relocation and to assure data integrity

Virtual Networking Management

- Assignment of MAC addresses by CP is coordinated across an SSI cluster
 - Ensure that new MAC addresses aren't being used by any 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

***SSI Cluster
Management***

SSI Cluster Operation

- A system that is configured as a member of an SSI cluster joins the cluster during IPL
 - Verifies that its configuration is compatible with the cluster
 - Establishes communication with other members

```
HCPPLM1669I Waiting for ISFC connectivity in order to join the SSI cluster.  
HCPFCA2706I Link JFSSIA1 activated by user SYSTEM.  
HCPKCL2714I Link device 921A added to link JFSSIA1.  
HCPALN2702I Link JFSSIA1 came up.  
HCPACQ2704I Node JFSSIA1 added to collection.
```

```
HCPPLM1697I The state of SSI system JFSSIA2 has changed from DOWN to JOINING  
HCPPLM1698I The mode of the SSI cluster is IN-FLUX  
HCPXHC1147I Spool synchronization with member JFSSIA1 initiated.  
HCPPLM1697I The state of SSI system JFSSIA2 has changed from JOINING to JOINED  
HCPPLM1698I The mode of the SSI cluster is IN-FLUX  
HCPXHC1147I Spool synchronization with member JFSSIA1 completed.  
HCPNET3010I Virtual machine network device configuration changes are permitted  
HCPPLM1698I The mode of the SSI cluster is STABLE
```

- Members leave the SSI cluster when they shut down

```
HCPPLM1697I The state of SSI system JFSSIA2 has changed from JOINED to LEAVING  
HCPPLM1698I The mode of the SSI cluster is IN-FLUX  
HCPPLM1697I The state of SSI system JFSSIA2 has changed from LEAVING to DOWN  
HCPPLM1698I The mode of the SSI cluster is IN-FLUX  
HCPPLM1698I The mode of the SSI cluster is STABLE
```

Reliability and Integrity of Shared Data and Resources

- Normal operating mode
 - All members communicating and sharing resources
 - Guests have access to same resources on all members

- Cluster-wide policing of resource access
 - Volume ownership marking
 - Coordinated minidisk link checking
 - Automatic minidisk cache management
 - Single logon enforcement

- Unexpected failure causes automatic "safing" of the cluster
 - Communications failure between any members
 - Unexpected system failure of any member
 - Existing running workloads continue to run
 - New access to shared resources is "locked down" until failure is resolved

- Most failures are resolved automatically
 - "Manual intervention" may be required

"Manual Intervention" for Communication or Configuration Failures

- **SET SSI membername DOWN** command
 - May be used when a member of the cluster has gone down without notifying other members
 - Member **membername** must be inactive and non-responsive
 - Allows active members to return to STABLE mode

- **REPAIR** IPL parameter
 - To be used only when a running system is required to correct a problems such as:
 - PDR initialization
 - System or SSI cluster configuration errors

 - ***No other members of the SSI cluster may be active!***
 - Operator confirmation prompt issued during IPL

 - All SSI initialization and management functions are bypassed
 - No sharing of data or resources

 - NOAUTOLOG and DISABLE IPL options automatically invoked

Summary

- An SSI cluster makes it easier to:
 - Manage and balance resources and workloads (move work to resources)
 - Schedule maintenance without disrupting key workloads
 - Test workloads in different environments
 - Operate and manage multiple z/VM images
 - Reliable sharing of resources and data

- Allow sufficient time to plan for an SSI cluster
 - Migration from current environment
 - Configuration
 - Sharing resources and data

- Plan for extra
 - CPU capacity
 - Disk capacity
 - Memory
 - CTC connections

More Information

z/VM 6.3 resources

<http://www.vm.ibm.com/zvm630/>

<http://www.vm.ibm.com/events/>

z/VM Single System Image Overview

<http://www.vm.ibm.com/ssi/>

Live Virtual Classes for z/VM and Linux

<http://www.vm.ibm.com/education/lvc/>

Redbooks

– An Introduction to z/VM SSI and LGR

<http://publib-b.boulder.ibm.com/redpieces/abstracts/sg248006.html?Open>

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

<http://publib-b.boulder.ibm.com/abstracts/sg248039.html?Open>

– DB2 10 for Linux on System z Using z/VM v6.2, Single System Image Clusters and Live Guest Relocation

<http://www.redbooks.ibm.com/abstracts/sg248036.html?Open>

Whitepaper

– z/VM Migration: Migrating the User Directory and RACF Environment

<http://public.dhe.ibm.com/common/ssi/ecm/en/zsw03246usen/ZSW03246USEN.PDF>

Thanks!

John Franciscovich
IBM
z/VM Design and Development
Endicott, NY

francisj@us.ibm.com

Session 14585

