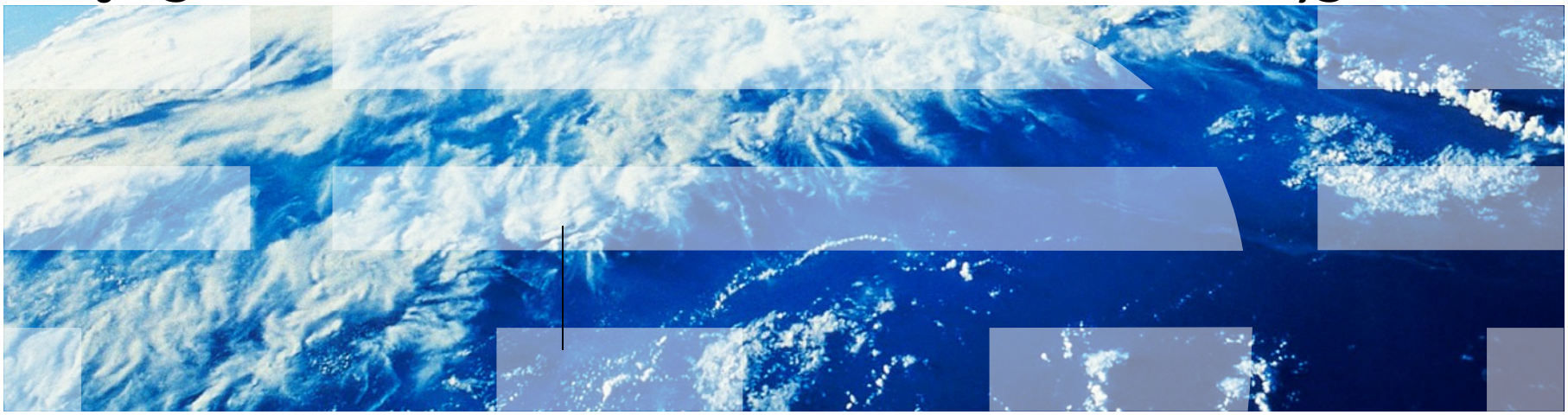
z/VM Single System Image and Live Guest Relocation Overview

Emily Kate Hugenbruch
ekhugen@us.ibm.com

John Franciscovich
francisj@us.ibm.com



Trademarks

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

z/VM® z10™ z/Architecture® zEnterprise™ System z196 System z114

Not all common law marks used by IBM are listed on this page. Failure of a mark to appear does not mean that IBM does not use the mark nor does it mean that the product is not actively marketed or is not significant within its relevant market.

Those trademarks followed by ® are registered trademarks of IBM in the United States; all others are trademarks or common law marks of IBM in the United States.

For a complete list of IBM Trademarks, see www.ibm.com/legal/copytrade.shtml:

The following are trademarks or registered trademarks of other companies.

Adobe, the Adobe logo, PostScript, and the PostScript logo are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States, and/or other countries.
Cell Broadband Engine is a trademark of Sony Computer Entertainment, Inc. in the United States, other countries, or both and is used under license therefrom.
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.
Intel, Intel logo, Intel Inside, Intel Inside logo, Intel Centrino, Intel Centrino logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium, and Pentium are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.
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.
ITIL is a registered trademark, and a registered community trademark of the Office of Government Commerce, and is registered in the U.S. Patent and Trademark Office.
IT Infrastructure Library is a registered trademark of the Central Computer and Telecommunications Agency, which is now part of the Office of Government Commerce.

* 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
- Major Attributes of a z/VM SSI Cluster
- z/VM SSI Cluster Operation
- Planning and Creating a z/VM SSI Cluster

Introduction

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

- VMSSI Feature of z/VM 6.2

- 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

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 a CSE cluster requires manual management of links
    - No automatic link protection or cache management

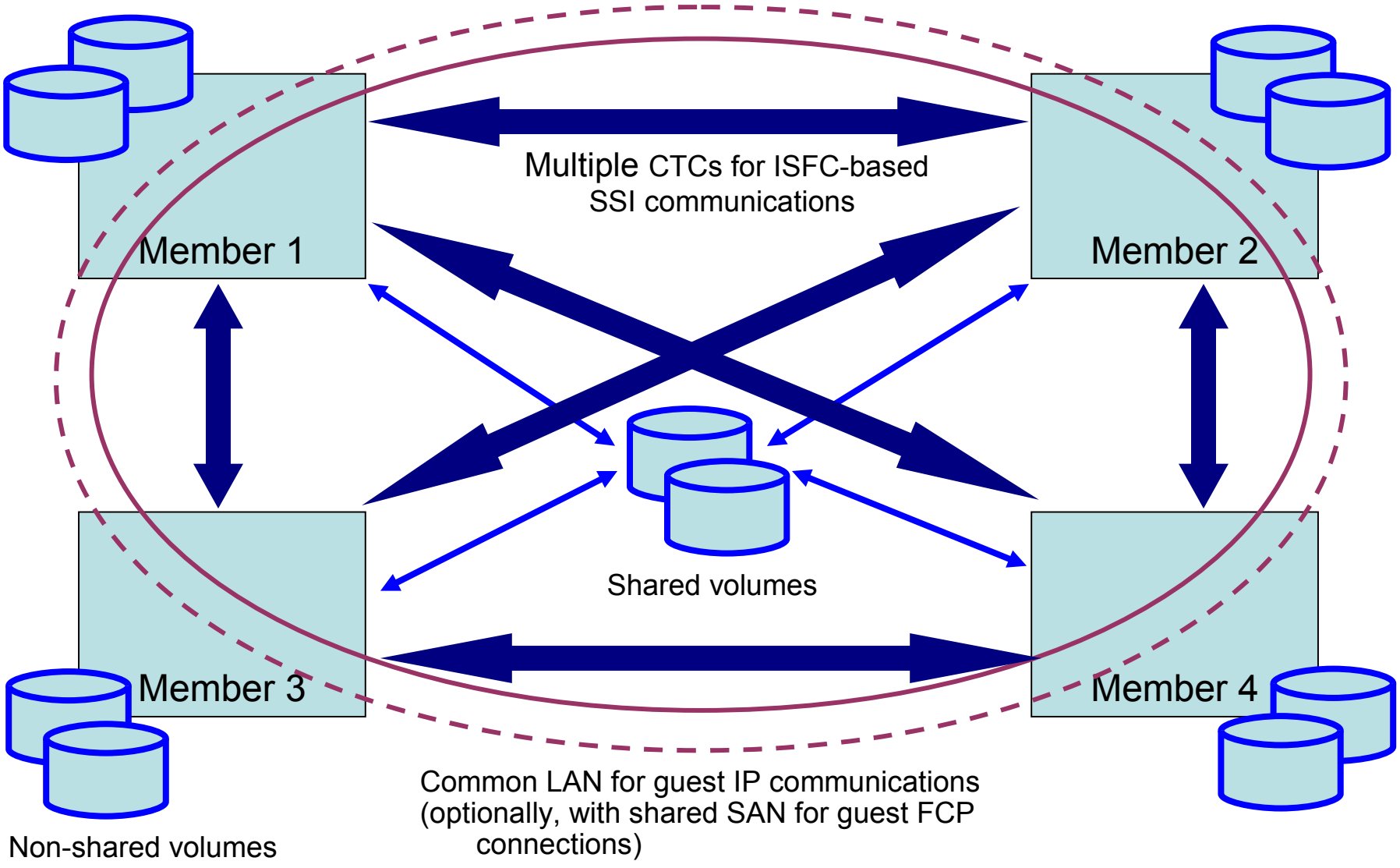
Benefits of a z/VM SSI Cluster

- Facilitates horizontal growth of z/VM workloads
- Reduce effect of planned outages for z/VM and hardware maintenance
  - Less disruptive to virtual server workloads
- Simplifies system management of a multi-z/VM environment
  - Concurrent installation of multiple-system cluster
  - Single maintenance stream
- Enhances workload balancing

SSI Cluster Considerations

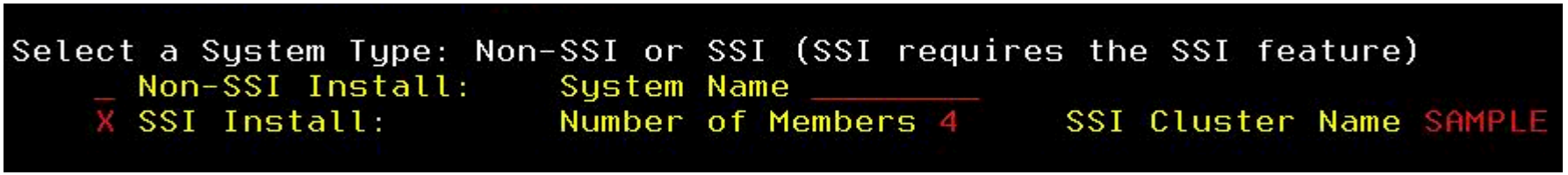
- Physical systems must be close enough to allow
  - FICON CTC connections
  - Shared DASD
  - Common network and disk fabric connections
- Installation to SCSI devices is not supported
  - Guests may use SCSI devices
- If using RACF, the database must reside on a fullpack 3390 volume
  - Single RACF database shared by all members of the cluster
- Live Guest Relocation is only supported for Linux on System z guests

z/VM SSI Cluster



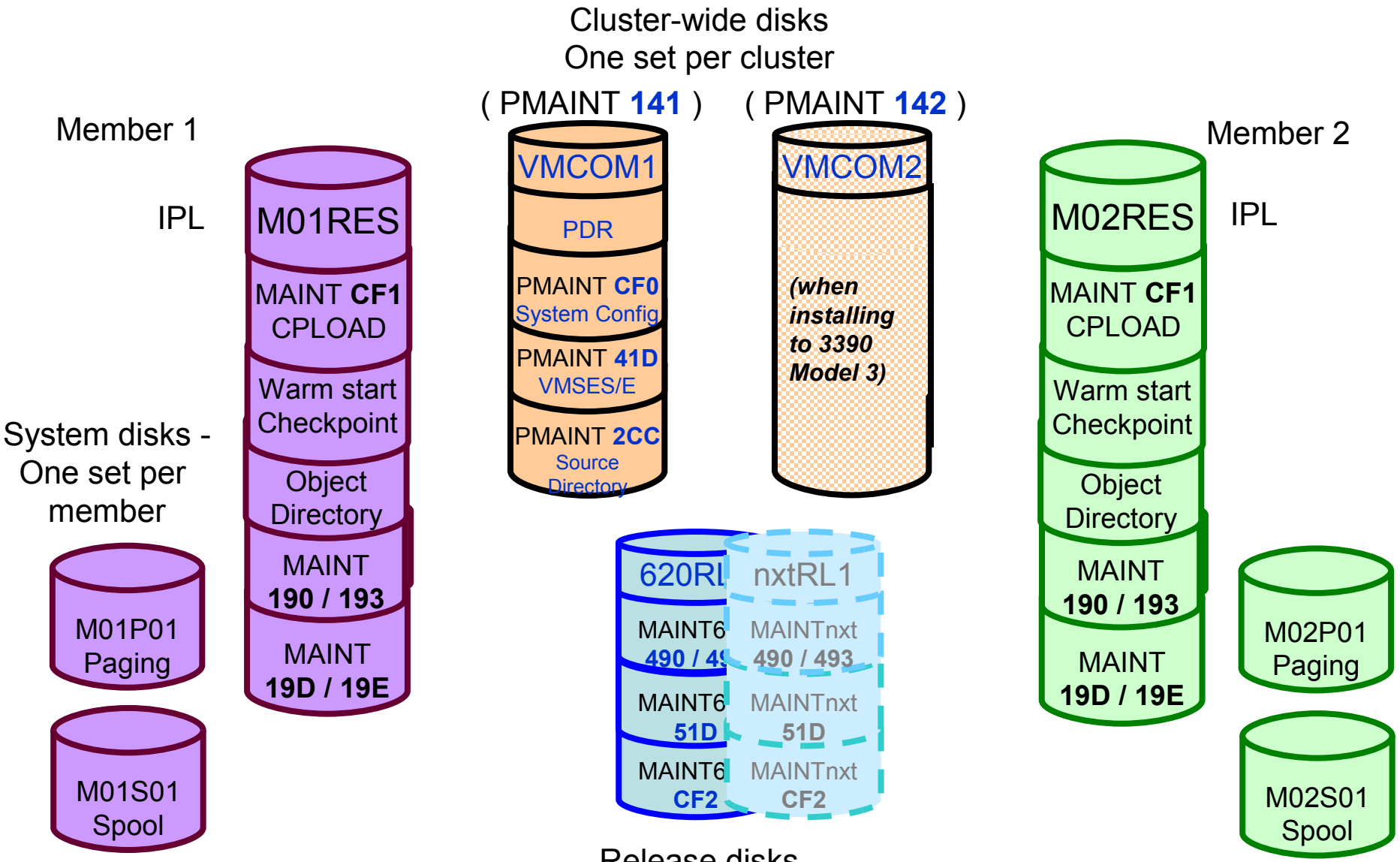
Major Attributes of a z/VM SSI Cluster

Multisystem Installation



- 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

DASD Volumes and Minidisks



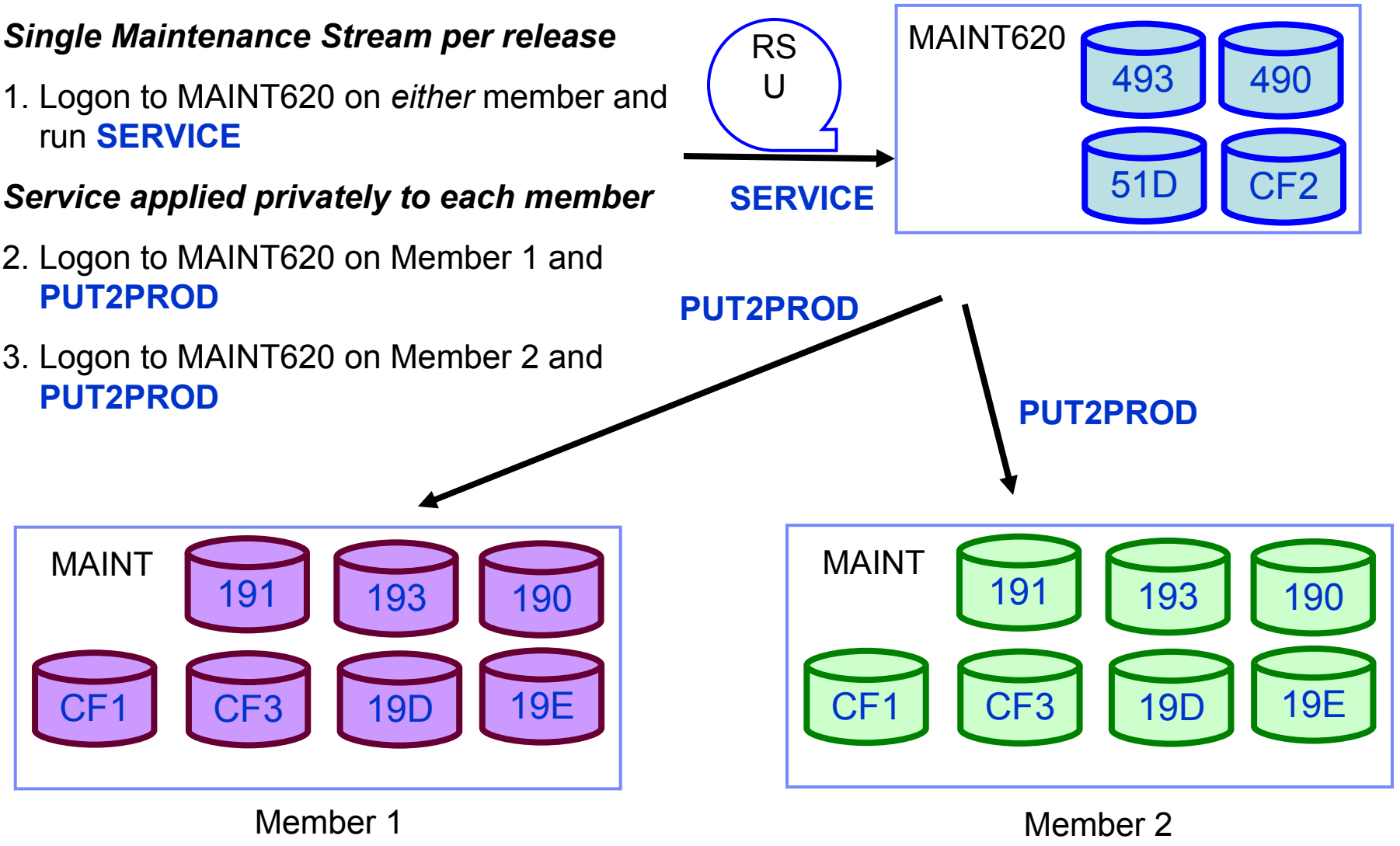
Applying Service

Single Maintenance Stream per release

1. Logon to MAINT620 on *either* member and run **SERVICE**

Service applied privately to each member

2. Logon to MAINT620 on Member 1 and **PUT2PROD**
3. Logon to MAINT620 on Member 2 and **PUT2PROD**



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

- Different for both SSI and non-SSI installs
  - Install and service tasks distributed among different "maint" userids
  - 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)
  - Directory
    - New structure and statements
  - System configuration file
    - New structure and statements
- Installation and service programs restructured
  - If you use customized programs, make sure you understand new structure

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

```
/************************************************************/
/*                                    SYSRES  VOLUME        */
/************************************************************/
VMSYS01: CP_Owned   Slot   1  M01RES
VMSYS02: CP_Owned   Slot   1  M02RES
VMSYS03: CP_Owned   Slot   1  M03RES
VMSYS04: CP_Owned   Slot   1  M04RES

/************************************************************/
/*                                    COMMON VOLUME         */
/************************************************************/
CP_Owned   Slot   5  VMCOM1

/************************************************************/
/*                                  DUMP & SPOOL VOLUMES    */
/* Dump and spool volumes begin with slot 10 and are        */
/* assigned in ascending order, without regard to the       */
/* system that owns them.                                   */
/************************************************************/
CP_Owned   Slot  10  M01S01
CP_Owned   Slot  11  M02S01
CP_Owned   Slot  12  M03S01
CP_Owned   Slot  13  M04S01
```

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

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

Ownership Checking – CP-Owned Volumes…

```
cpfmtxa
ENTER FORMAT, ALLOCATE, LABEL, OWNER OR QUIT:
owner
ENTER THE VDEV TO BE PROCESSED OR QUIT:
3001
ENTER THE VOLUME LABEL FOR DISK E000:
m01s01
ENTER THE OWNING SSI NAME (OR NOSSI) FOR DISK E000:
clustera
ENTER THE OWNING SYSTEM NAME (OR NOSYS) FOR DISK E000:
vmsys01
```

```
query cpowned

SLOT VOL-ID RDEV TYPE  STATUS                    SSIOWNER SYSOWNER
                           .
                           .
                           .
10   M01S01 3001 OWN   ONLINE AND ATTACHED      CLUSTERA VMSYS01
```

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)

USER statement

Same definitions and resources on all members

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

Multiconfiguration Virtual Machine (new)

IDENTITY statement

Definitions and resources common to all members

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

SUBCONFIG statement for member 1

Definitions and resources unique to member 1

SUBCONFIG statement for member 2

Definitions and resources unique to member 2

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

- Virtual machines on other members can be the target of some CP commands
  - Single-configuration virtual machines are usually found wherever they are logged on
  - Multiconfiguration virtual machines require explicit targeting
- **AT *sysname*** operand for the following commands
  - MESSAGE (MSG)
  - MSGNOH
  - SEND
  - SMSG
  - WARNING

```
MSG userid AT sysname
```

  - CMS TELL and SENDFILE commands require RSCS in order to communicate with multiconfiguration virtual machines on other members
- **AT** command can be used to issue most privileged commands on another active member

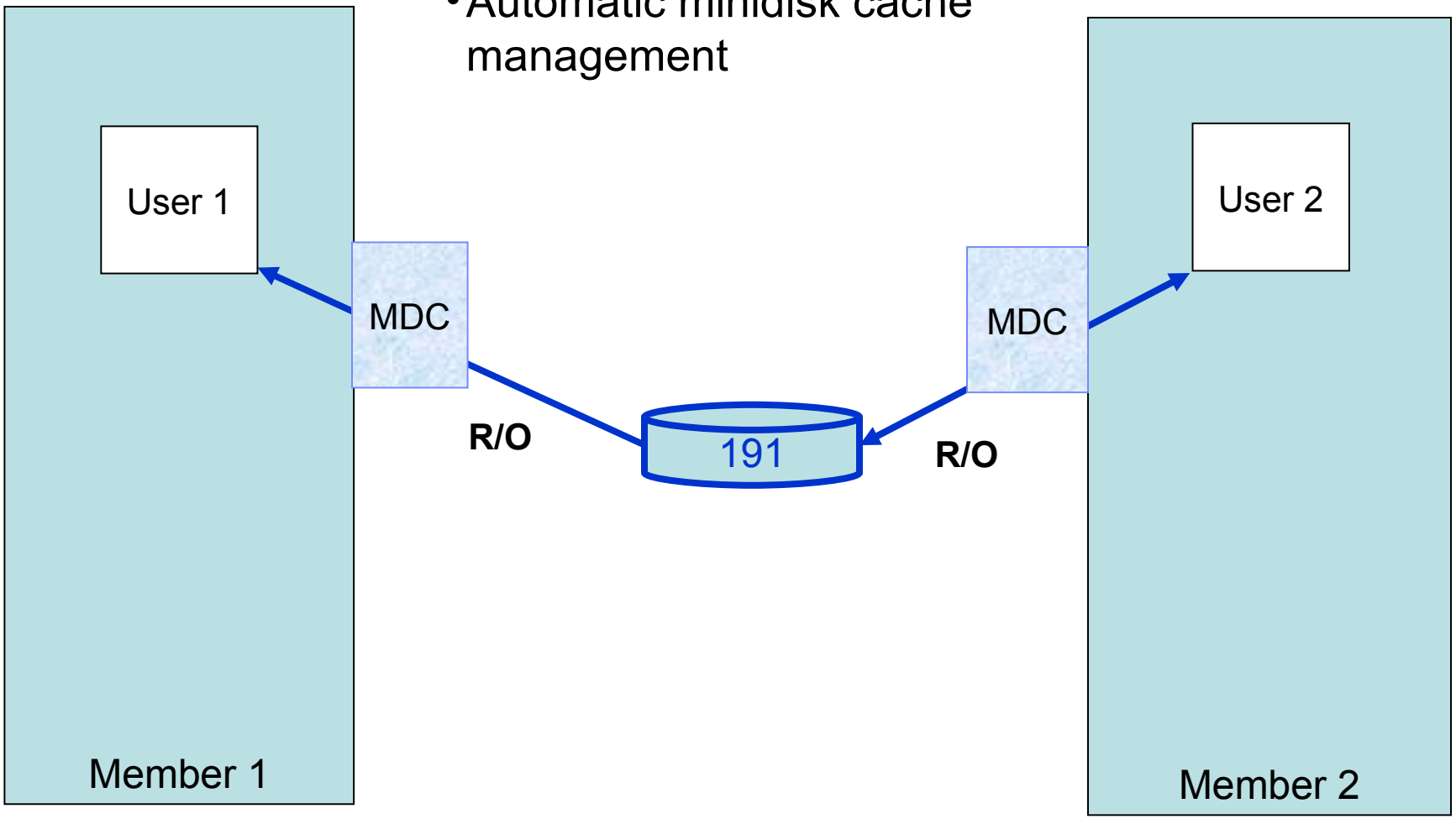
```
AT sysname CMD cmdname
```

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**

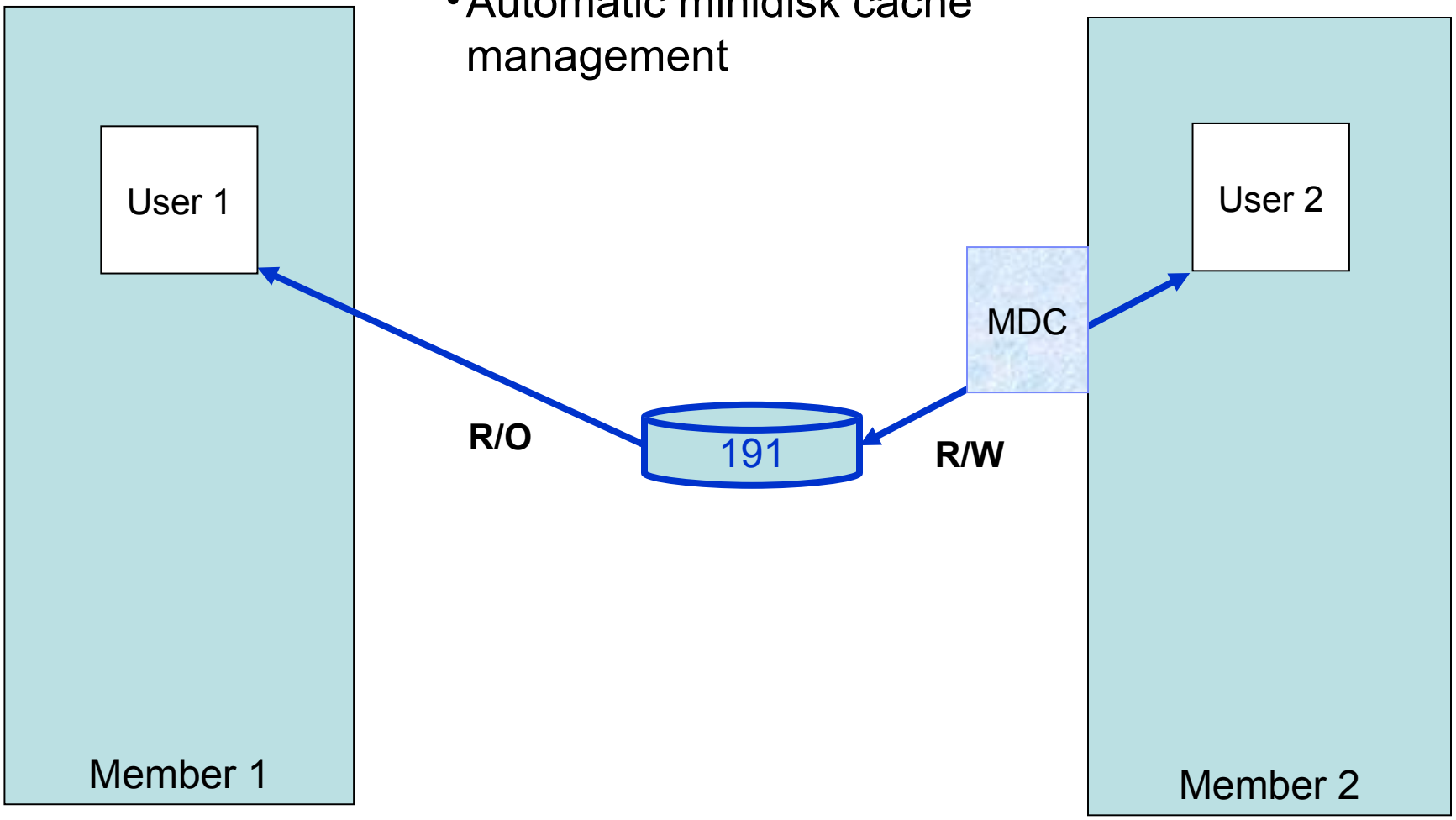
Cross-System Minidisk Management…

- Automatic minidisk cache management



Cross-System Minidisk Management…

- Automatic minidisk cache management



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

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
- Relocations are **NOT** done automatically by the system
  - New **VMRELOCATE** command
- Relocation capacity is determined by various factors (e.g. system load, ISFC bandwidth, etc.)
- Relocating guests continue to run on source member until destination is ready
  - Briefly quiesced
  - Resumed on destination member
- A guest to be relocated must meet eligibility requirements, including:
  - It must be logged on but disconnected
  - Architecture and functional environment on destination must be comparable
  - Destination member must have capacity to accommodate the guest
  - Devices and resources needed by guest must be shared and available on destination
- Relocation domains define a set of members among which virtual machines can relocate freely

Live Guest Relocation – VMRELOCATE Command

- New **VMRELOCATE** command initiates and manages live guest relocations
  - Several operands to control and monitor relocations, including:
    - **TEST** – determine if guest is eligible for specified relocation
    - **MOVE** – relocates guest
    - **STATUS** – display information about relocations that are in progress
    - **CANCEL** – stop a relocation
    - **MAXQUIESCE** – maximum quiesce time (relocation is cancelled if exceeded)
    - **MAXTOTAL** – maximum total time (relocation is cancelled if exceeded)
- Guest continues to run on originating member if a relocation fails or is cancelled

z/VM SSI Cluster Operation

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

```
HCPPLM1644I The following is the current status of the SSI member
HCPPLM1644I systems according to the PDR:
 SSI Name: JFSSIA
 SSI Persistent Data Record (PDR) device: JFEFE0 on EFE0
 SLOT SYSTEMID STATE     CONNECT TYPE          HOPS
    1 JFSSIA1  Joined    Not connected          -
    2 JFSSIA2  Down      Local                  -
    3 JFSSIA3  Down      Not connected          -
    4 JFSSIA4  Down      Not connected          -

 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
```

SSI Cluster Operation

- 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

HCPKDM2719E Link device 912A was stopped by the remote node.
HCPKDL2716I Link device 912A is stopped.
HCPALN2701I Link JFSSIA2 went down.
HCPKCB2715I Link device 912A removed from link JFSSIA2.
HCPFDL2706I Link JFSSIA2 deactivated by user SYSTEM.
HCPKCB2703I Node JFSSIA2 deleted from collection.
```

Protecting Integrity of Shared Data and Resources

- Normal operating mode
  - All members communicating and sharing resources
  - Guests have access to same resources on all members
- Unexpected failure causes automatic "safing" of the cluster
  - Existing running workloads continue to run
  - New allocations of shared resources are "locked down" until failure is resolved
    - Communications failure between any members
    - Unexpected system failure of any member
- Most failures are resolved automatically
  - Manual intervention may be required
    - **SET SSI membername DOWN** command
    - **REPAIR** IPL parameter

SSI Cluster Status – Example 1

```
query ssi status

SSI Name: CLUSTERA
SSI Mode: Influx
Cross-System Timeouts: Enabled
SSI Persistent Data Record (PDR) device: VMCOM1 on EFE0

SLOT SYSTEMID STATE      PDR HEARTBEAT       RECEIVED HEARTBEAT
   1 VMSYS01  Joined     2010-07-11 21:22:00 2010-07-11 21:22:00
   2 VMSYS02  Joined     2010-07-11 21:21:40 2010-07-11 21:21:40
   3 VMSYS03  Joining    2010-07-11 21:21:57 None
   4 VMSYS04  Down (not IPLed)
```

SSI Cluster Status – Example 2

```
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
```

Summary

- An SSI cluster gives you
  - Workload balancing (move work to system resources)
  - Maintenance on your schedule (not the application owner)
  - Easier operation and management of multiple z/VM images

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

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

More Information

z/VM 6.2 resources
http://www.vm.ibm.com/zvm620/

z/VM Single System Image Overview
http://www.vm.ibm.com/ssi/

Redbook – An Introduction to z/VM SSI and LGR
http://publib-b.boulder.ibm.com/redpieces/abstracts/sg248006.html?Open

Thanks!

Contact Information:

Emily Hugenbruch
IBM
z/VM Development
Endicott, NY

ekhugen@us.ibm.com