SSI part 1 – Configuration and Setup

Jacob Gagnon
z/VM CP Development
jpgagnon@us.ibm.com

Agenda

Background

- Single System Image
- Live Guest Relocation

Major Attributes of a z/VM SSI Cluster

- Installation and service
- Configuration
- Persistent Data Record (PDR)
- Shared source directory
- Shared system resources and data

Brief Overview of z/VM SSI Cluster Management

Introduction to SSI and LGR

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

- VMSSI was a priced feature prior to z/VM 7.1, now included for free
- 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 (CTC) connections between members
  - No virtual servers required

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 workloads

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

- 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
- 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

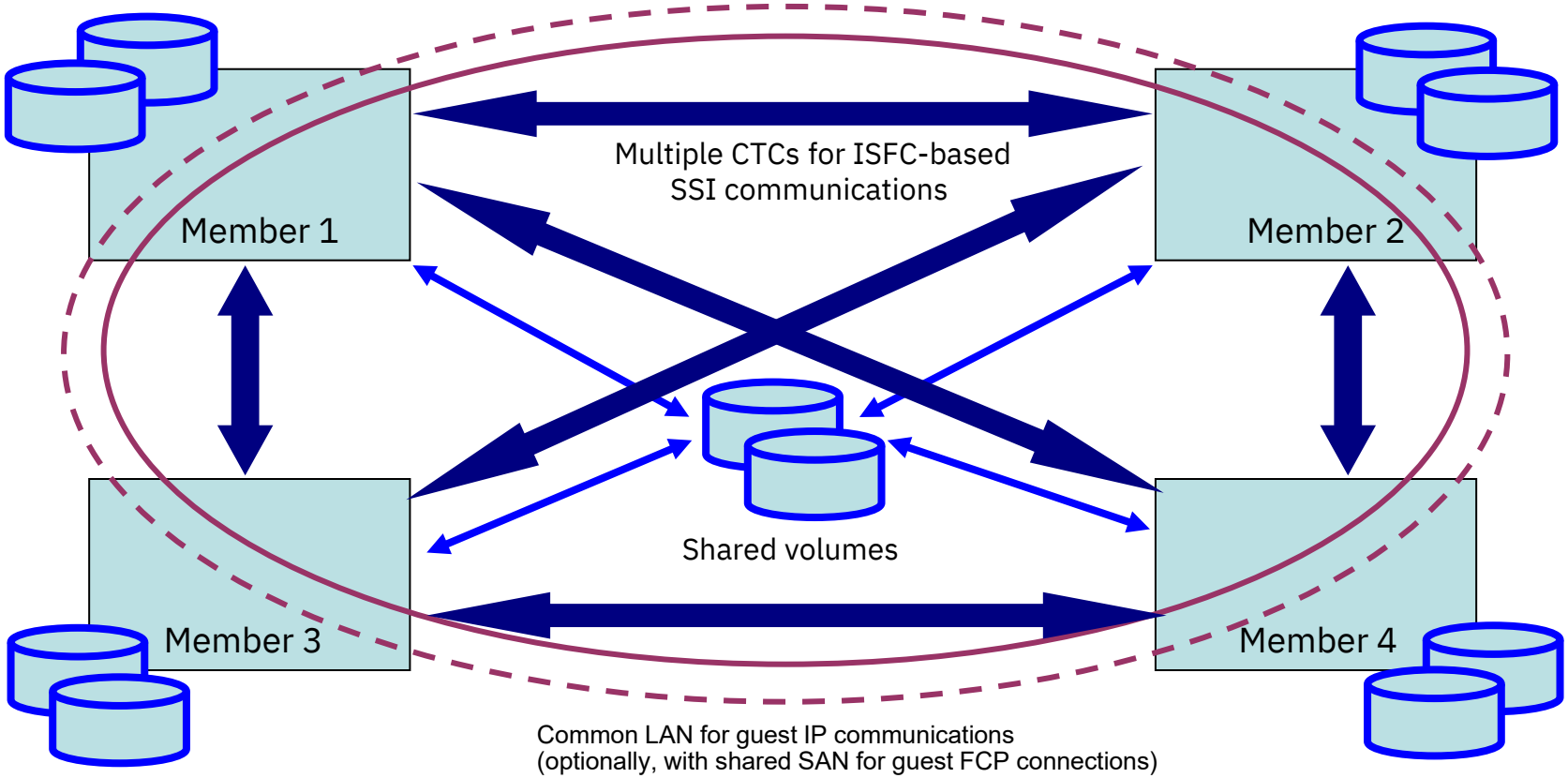
Live Guest Relocation...

- Relocation capacity is determined by various factors (e.g. system load, ISFC bandwidth, etc)

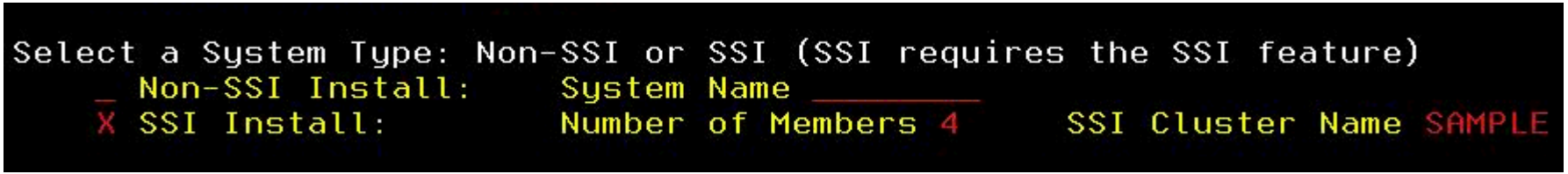
- In order to be relocated, a guest must meet eligibility requirements, including:
  - The architecture and functional environment on destination member must be comparable
    - Relocation domains can be used to define sets of members among which guests can relocate freely.
  - Devices and resources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

Major Attributes of a z/VM SSI Cluster

z/VM SSI Cluster

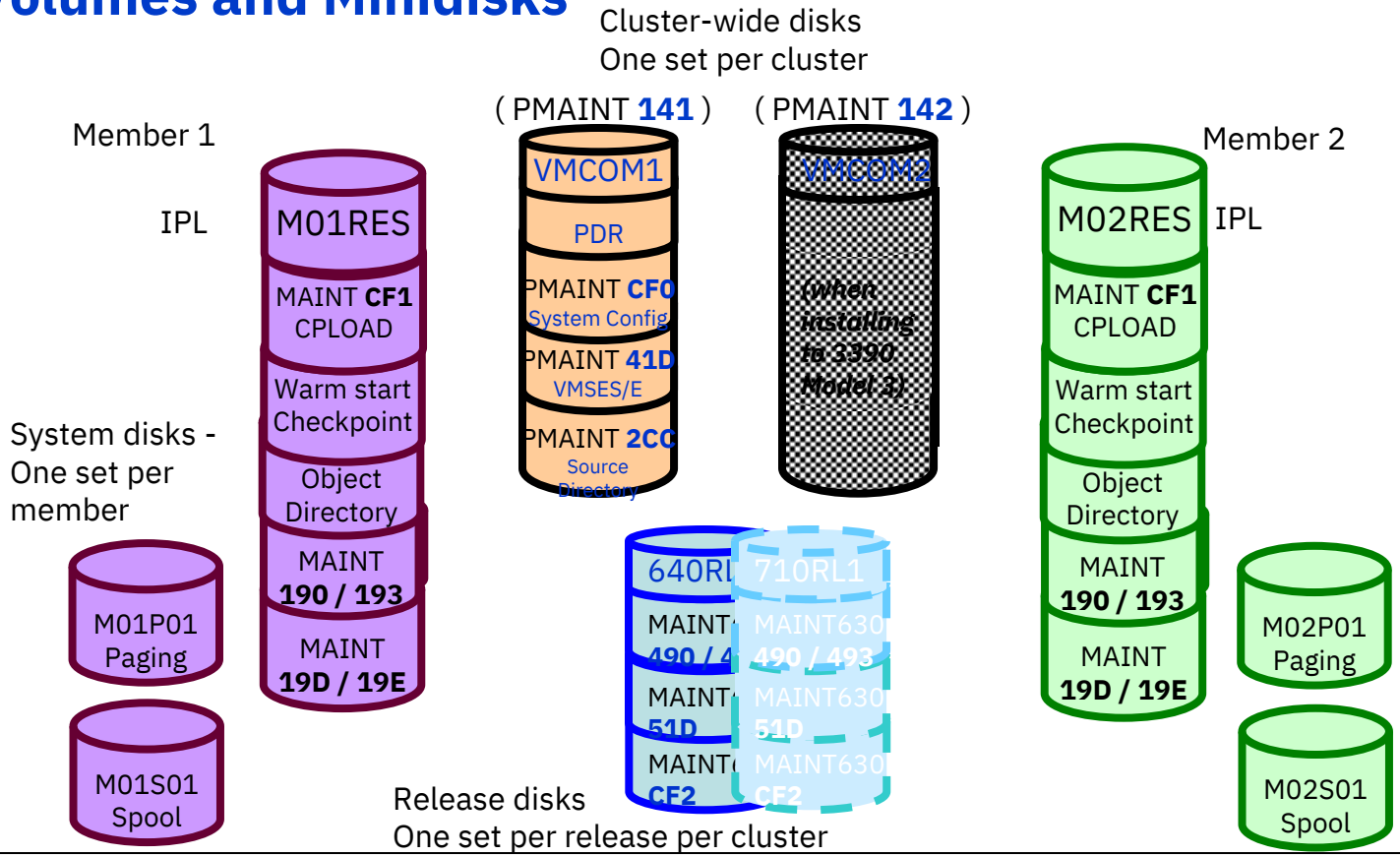


Multisystem Installation



- 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



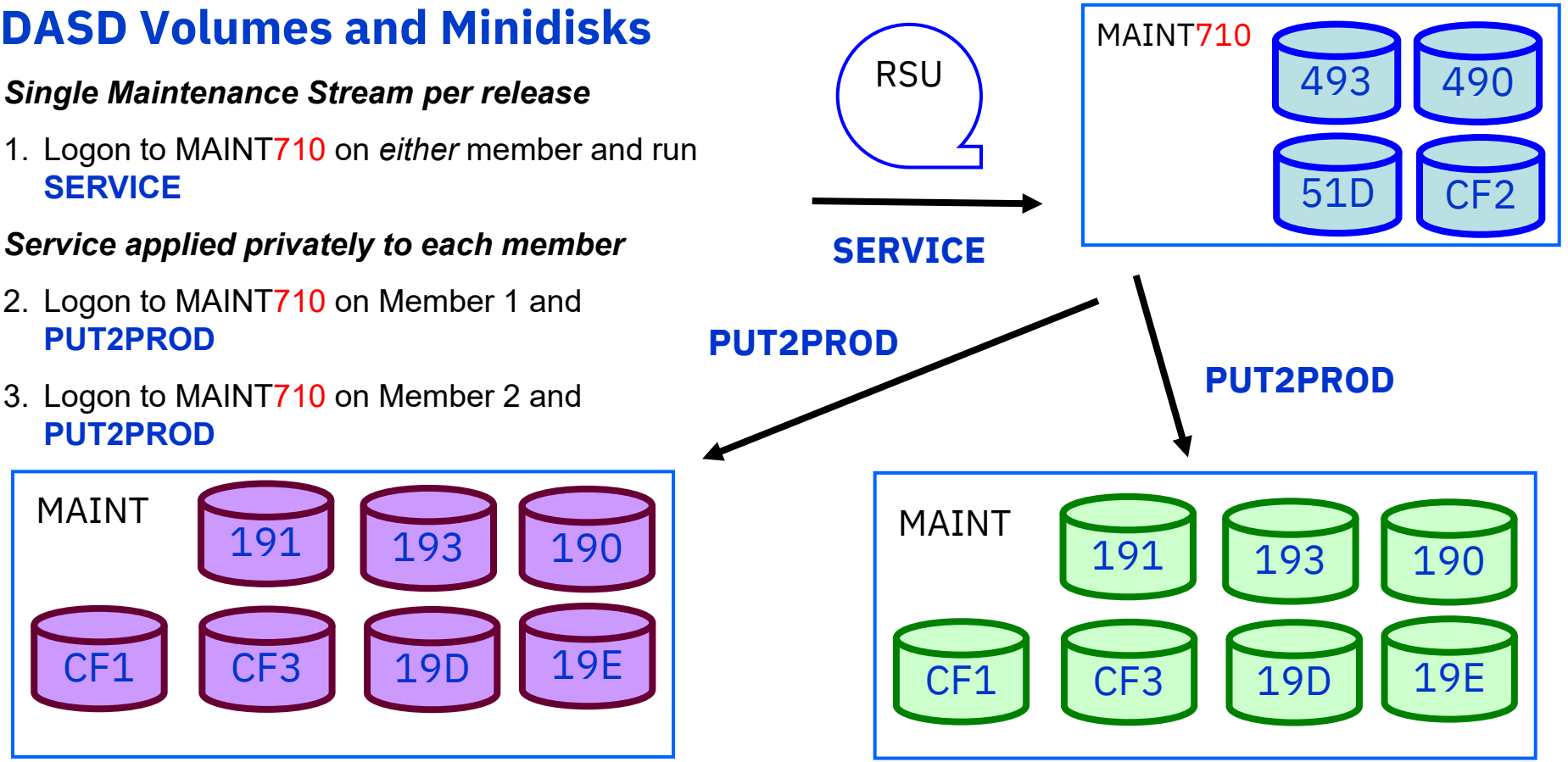
DASD Volumes and Minidisks

Single Maintenance Stream per release

1. Logon to MAINT710 on *either* member and run **SERVICE**

Service applied privately to each member

2. Logon to MAINT710 on Member 1 and **PUT2PROD**
3. Logon to MAINT710 on Member 2 and **PUT2PROD**



Shared System Configuration File

- Resides on 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 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
- QUERY CPOWNED displays 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

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

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.
    - Secondary user will not receive responses to SEND commands or other output from primary user
    - Output from secondary user will only be received by primary user on the same member

Primary User or Observer	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

- **AT** *command* can be used to issue most privileged commands

```
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

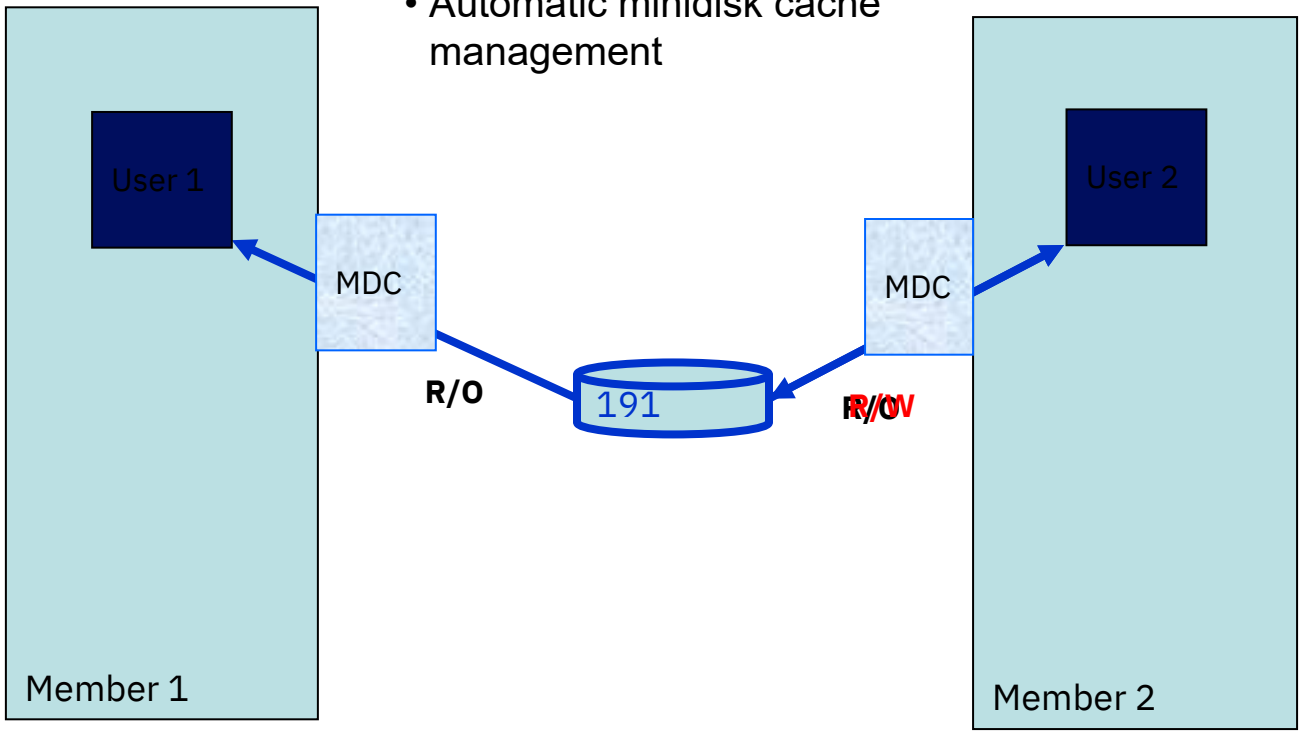
- Minidisks can either be shared across all members or restricted to a single member
  - CP Checks for conflicts throughout the cluster when a link is requested

- 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



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 communicate 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 are “locked down” until failure is resolved
- Most failures are resolved automatically
  - Manual intervention may be required
    - **SET SSI membername DOWN** command
    - **REPAIR** IPL parameter

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

R-219
G-38
B-99

R-105
G-166
B-255

R-0
G-100
B-255