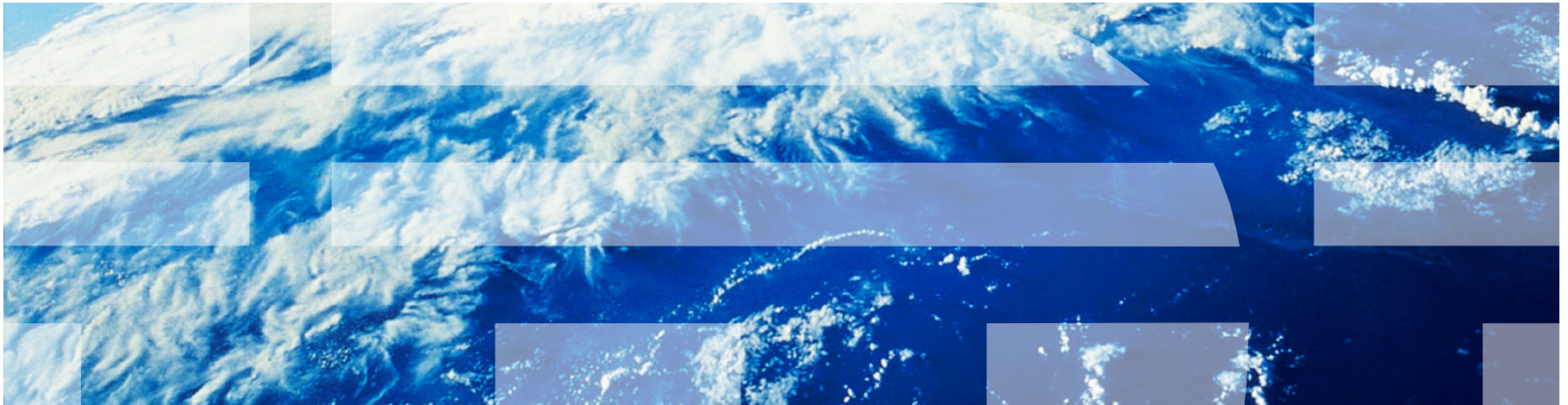# z/VM Single System Image and Live Guest Relocation Preview

Steve Wilkins

wilkinss@us.ibm.com

10/20/2011

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™

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.

IBM

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 virtual servers (guests) non-disruptively from one from one member of the cluster to another
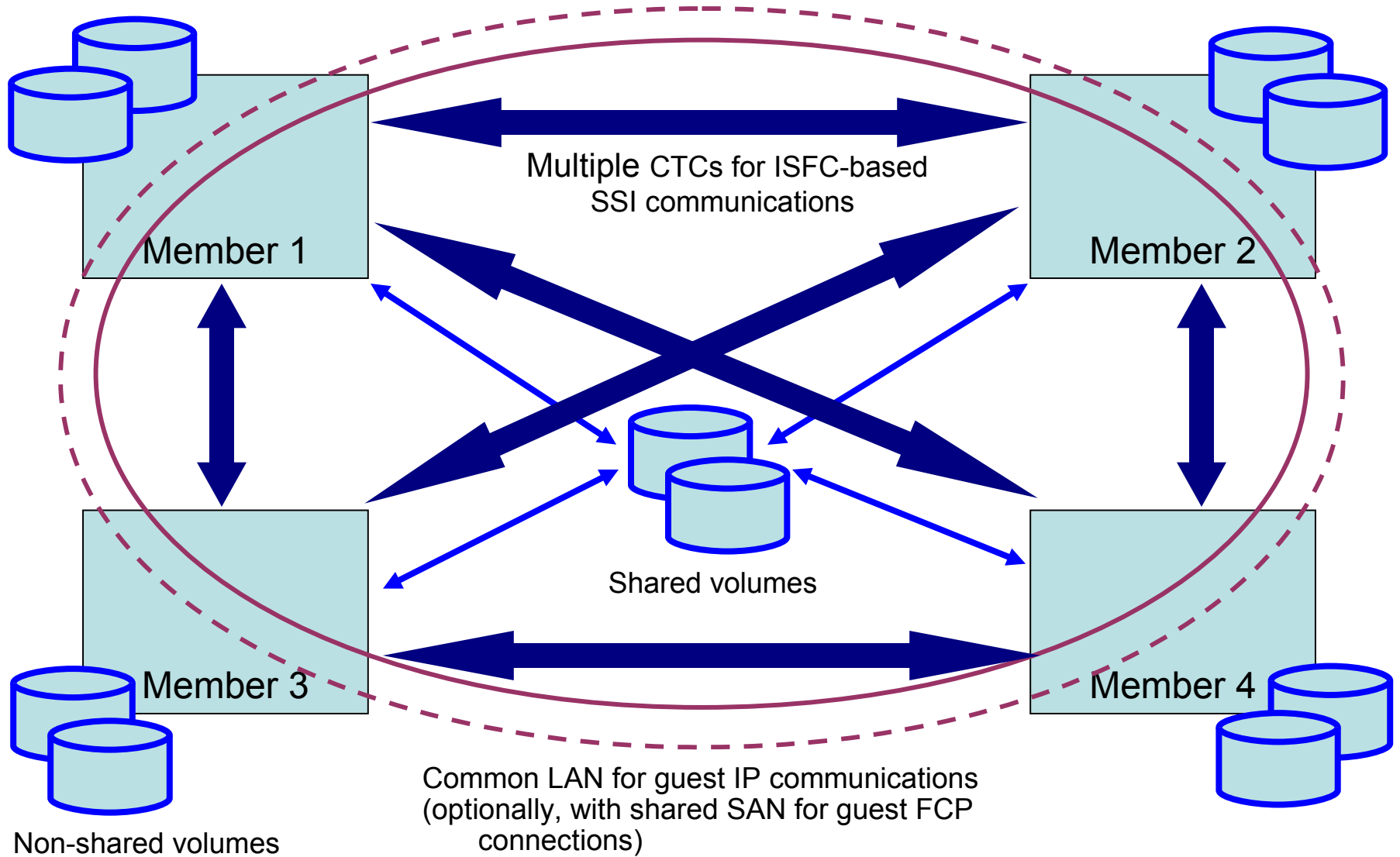
6

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

7

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

8

z/VM SSI Cluster

Member 1

Member 2

Member 3

Member 4

Multiple CTCs for ISFC-based
SSI communications

Shared volumes

Common LAN for guest IP communications
(optionally, with shared SAN for guest FCP
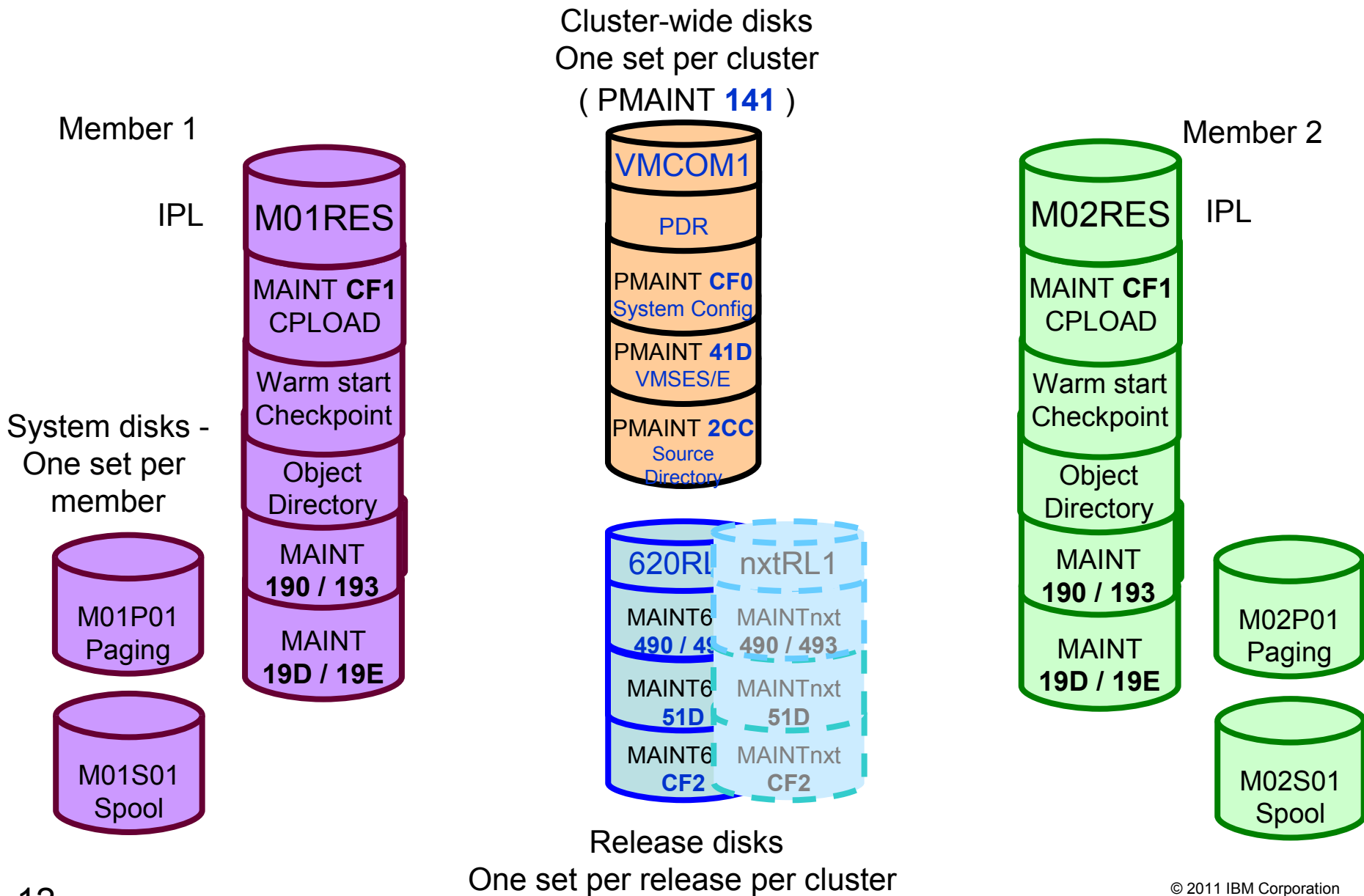connections)

Non-shared volumes

9

Major Attributes of a 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
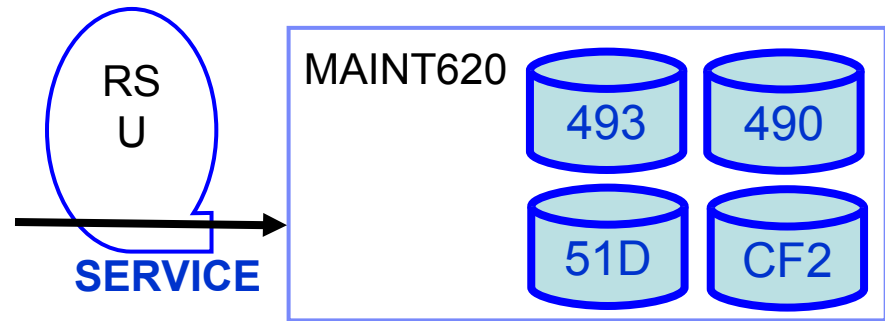
11

DASD Volumes and Minidisks

Cluster-wide disks
One set per cluster
(PMAINT 141)

Member 1

IPL

M01RES

MAINT **CF1**
CPLOAD

Warm start
Checkpoint

Object
Directory

MAINT
190 / 193

MAINT
19D / 19E

System disks -
One set per
member

M01P01
Paging

M01S01
Spool

VMCOM1

PDR

PMAINT **CF0**
System Config

PMAINT **41D**
VMSES/E

PMAINT **2CC**
Source
Directory

620RL | nxtRL1

MAINT6 | MAINTnxt
490 / 49 | **490 / 493**

MAINT6 | MAINTnxt
51D | **51D**

MAINT6 | MAINTnxt
CF2 | **CF2**

Member 2

IPL

M02RES

MAINT **CF1**
CPLOAD

Warm start
Checkpoint

Object
Directory

MAINT
190 / 193

MAINT
19D / 19E

M02P01
Paging

M02S01
Spool

Release disks
One set per release per cluster

12

Applying Service

Single Maintenance Stream per release

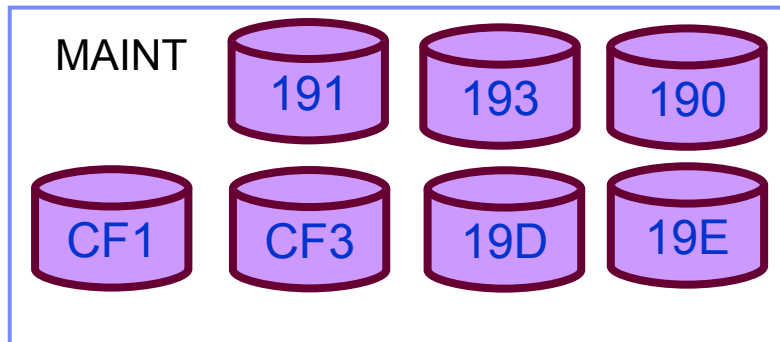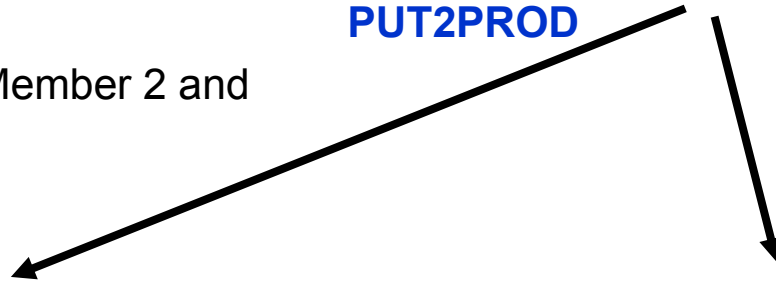1. Logon to MAINT620 on *either* member and run **SERVICE**

Service applied privately to each member

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

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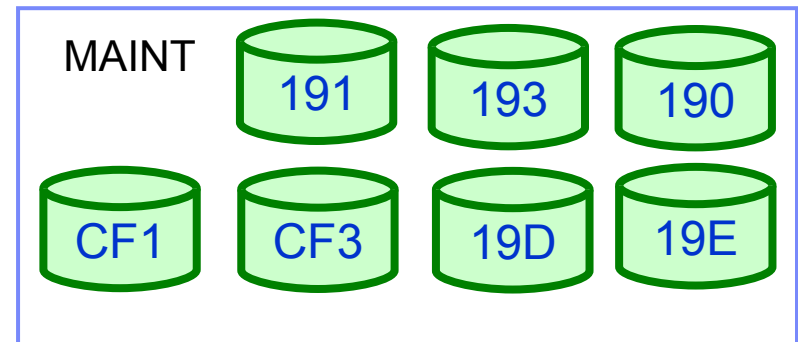
RSU

SERVICE

MAINT620

| 493 | 490 |
| 51D | CF2 |

PUT2PROD

PUT2PROD

MAINT

| 191 | 193 | 190 |
| CF1 | CF3 | 19D | 19E |

Member 1

MAINT

| 191 | 193 | 190 |
| CF1 | CF3 | 19D | 19E |

Member 2

13

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
  ```

14

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

15

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
```

17

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

18

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

19

Ownership Checking – CP-Owned Volumes…

```
cpfmtxa
ENTER FORMAT, ALLOCATE, LABEL, OWNER OR QUIT:
owner
ENTER THE VDEV TO BE PROCESSED OR QUIT:
c4a8
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 C4A8 OWN    ONLINE AND ATTACHED       CLUSTERA VMSYS01
```

20

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
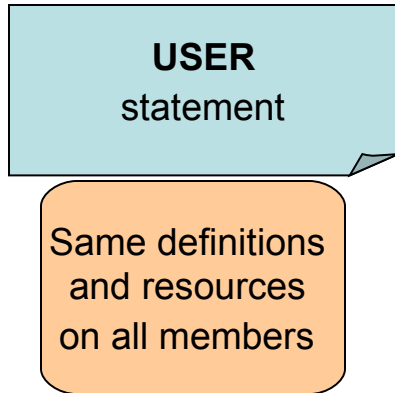
• May log on to any member
  • Only one member at a time

• 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
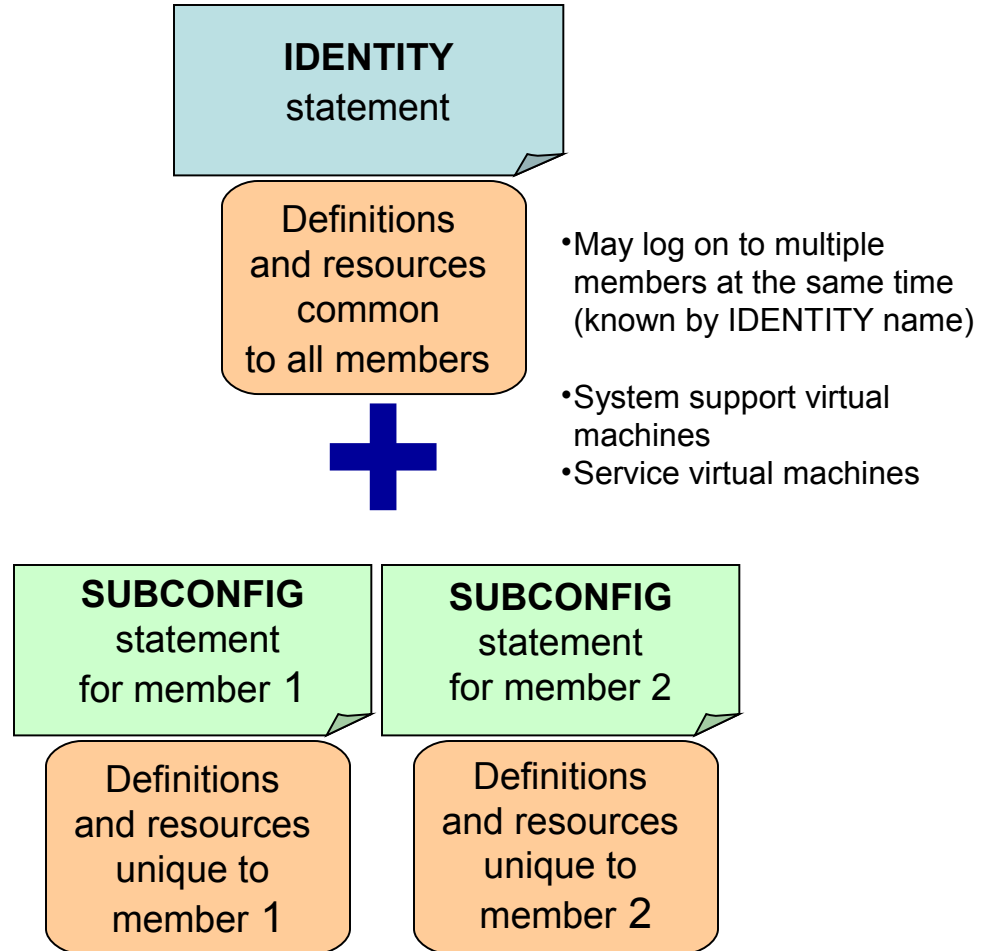
• May log on to multiple members at the same time (known by IDENTITY name)

• System support virtual machines
• Service virtual machines

+

SUBCONFIG statement for member 1

SUBCONFIG statement for member 2

Definitions and resources unique to member 1

Definitions and resources unique to member 2

22

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

- Cross-System SCIF (Single Console Image Facility)
  - Allows one virtual machine (secondary user) to monitor and control one or more disconnected virtual machines (primary users)
    - CONSOLE statement in directory
    - SET SECUSER command
    - SET OBSERVER command

  - Secondary and primary users can be logged on different members of an SSI cluster

- Some restrictions for multiconfiguration virtual machines (MCVM):

Primary User or Observee	SECUSER or Observer	If Local	If Remote
SCVM	**SCVM**	**Yes**	**Yes**
SCVM	**MCVM**	**Yes**	**Yes**
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

Member 1

User 1

MDC

R/O

191

Member 2

User 2

MDC

R/W R/O
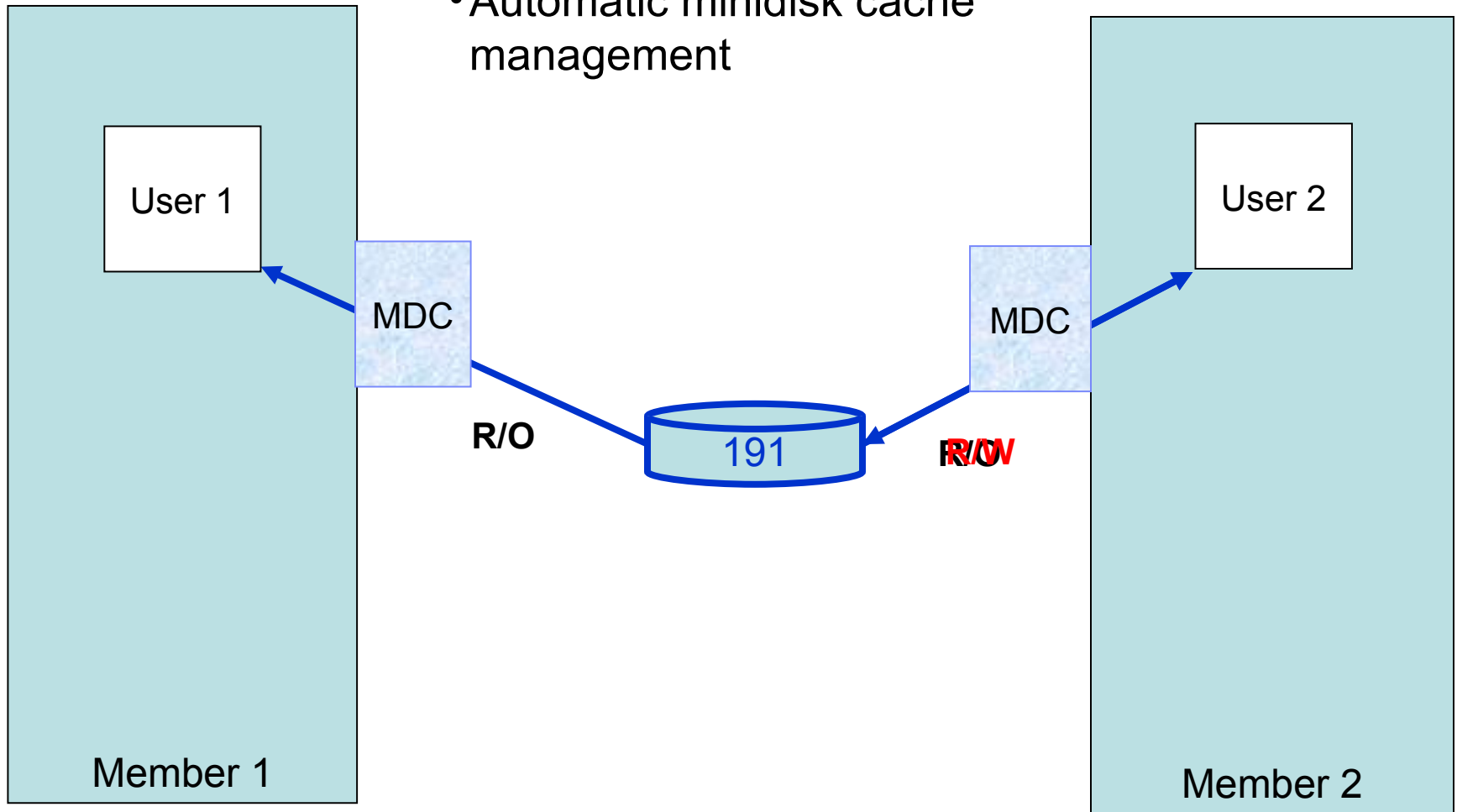
Real Device Management

- 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

28

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
  - Moving workload off a member requiring maintenance

- Relocating guests continue to run on source member until destination is ready
  - Briefly quiesced
  - Resumed on destination member

- New **VMRELOCATE** command will initiate and manage guest relocations
  - Relocation capacity determined by various factors (e.g. system load, ISFC bandwidth, etc.)

- 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 can be used to define a set of members among which virtual machines can relocate freely

LGR Restrictions (1 of 2)

GUEST STATE RESTRICTIONS

- Relocation of this guest is already in progress
- User is still logging on or is being autologged on
- User is logging off or is being forced off the system
- Guest terminating
- Guest is IPLing
- Guest system reset is in progress
- Guest is primary system operator
- Guest is defined as a "restricted" user
- Guest has an architecture incompatibility
- Guest has SET VMRELOCATE OFF
- Guest is an XC-mode virtual machine

DEVICE STATE RESTRICTIONS

- I/O is active to a device that will result in a delayed response.
- Guest has a single-user TAPE device defined in its configuration and a real assign exists for the device
- Guest has a multi-user TAPE device defined in its configuration and it owns the assign for the TAPE
- Guest has an open SPOOL file (other than a console file)

DEVICE RESTRICTIONS

- Guest not running disconnected
- Guest has real terminal attached
- The ASCII console is attached to the guest
- Guest has T-disk
- Guest has a shared V-disk
- Guest has a dedicated device that has no device with a matching device identifier on the destination system
- Guest has unsupported devices in its configuration
- Guest has dialed GRAF devices in its configuration
- Guest has dialed in LINE devices in its configuration
- Guest has virtual CTCA devices in its configuration
- Guest has a link to a restricted user's minidisk
- Guest has FCP, OSA, IQD, and/or CTC real devices defined to its configuration and associated equivalency IDs (EQIDs) for these devices were not available
- Guest has a real device defined to its configuration and an associated Device Unique ID (DUID) or EQIDs does not exist
- Guest has FBA DASD in its configuration and associated equivalency IDs (EQIDs) for these devices were not created
- Guest has a device defined that is a simulated NIC and is coupled with a VSWITCH that is without a real OSA port configured
- Guest has a device defined that is a VSWITCH that is without a real OSA port configured

LGR Restrictions (2 of 2)

VIRTUAL FACILITY RESTRICTIONS

- Guest configuration includes access to a virtual Coupling
- Guest is coupling facility virtual machine
- Guest has expanded storage attached
- Guest is using VM data spaces
- Guest has an SPXTAPE operation in progress
- User has active CP traces
- Guest is using TRSOURCE
- Guest is using diagnose x'E0' to write data into a system trace file
- Guest hosting logical devices
- Guest is a worker machine (using diagnose x'D4' function)
- Guest using TOD Clock Accounting (diagnose x'70')
- Guest using VMCF
- Guest is using CMS POSIX facilities
- Guest using Diag244, MAPMDISK
- Guest has directory authorization to use diagnose x'98'
- Guest has IUCV connections to any system service other than *MSG or *MSGALL, or to another virtual machine
- Guest has an APPC/VM connection

CONFIGURATION RESTRICTIONS

- Guest has a dedicated CPU
- Guest has dedicated crypto capability
- Guest is using a VMGROUP NSS or is using an NSS or DCSS of type SW, SC or SN, or is using one that does not exist on the destination system
- Guest has a non-homogeneous CPU type configuration
- Guest is defined with all type CP CPUs and destination is an all IFL environment

RESOURCE LIMIT RESTRICTIONS

- Guest will not "fit" into storage on the destination system
- Guest's V-disk usage would violate usage restrictions on the destination system

MISCELLANEOUS

- A customized CP control block field appears to be in use.
- Fields in the VMDBK (VMDUSER1-VMDUSER8, VMDEBUG1-VMDEBUG8) and VDEV (VDEVUSR1-VDEVUSR4) control blocks are reserved for customer and vendor use. If any of those are other than their initial values, the guest cannot be relocated.

32

z/VM SSI Cluster Operation

SSI Cluster Management

- 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
```

34

SSI Cluster Management

- 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.
```

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
```

SSI Cluster Management - Features for Greater Reliability

- Cross-checking of configuration details as members join cluster and as resources are used
  - SSI membership definition and identity
  - Consistent definition of shared spool volumes
  - Compatible virtual network configurations (MAC address ranges, VSwitch definitions)

- Cluster-wide policing of resource access
  - Volume ownership marking to prevent dual use
  - Coordinated minidisk link checking
  - Autonomic minidisk cache management
  - Single logon enforcement

- Communications failure "locks down" future resource allocations until resolved

- Comprehensive checking for resource and machine feature compatibility during relocation
  - Relocation domains can be defined to support customer relocation policy

38

Planning and Creating a z/VM SSI Cluster

SSI Cluster Requirements

- Servers must be IBM System z10 or later (z/VM Version 6)

- Shared and non-shared DASD
  - 3390 volume required for the PDR
  - All volumes should be cabled to all members
    - Makes non-shared disks accessible to other members to fix configuration problems

- LPARs
  - 1-16 FICON CTC devices between LPARs
    - Provide direct ISFC links from each member to all other members
  - FICON channels to shared DASD
  - OSA access to the same LAN segments
  - FCP access to same storage area networks (SANs) with same storage access rights

- Shared system configuration file for all members

- Shared source directory containing user definitions for all members

- Capacity planning for each member of the SSI cluster
  - Ensure sufficient resources are available to contain shifting workload
    - Guests that will relocate
    - Guests that logon to different members

40

SSI Cluster Restrictions

- 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

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

41

SSI Cluster Setup – Suggested Practices

- Use the same real device numbers across LPARs to simplify cloning of z/VM systems
  - DASD volumes
  - Ranges for OSA and hipersockets subchannels connected to same network
  - Ranges for FCP subchannels connected to the same fabric

- Install no more than 2 members of an SSI cluster on the same server

- Maintain parallel volume layouts for each member (again, simplifies cloning)

- Allocate object directory (DRCT) extents only on the system residence volume for each member

- Do not place user data on the installation volumes
  - Simplifies release-to-release migration

- Keep member-specific data and SSI cluster data on separate volumes
  - Simplifies cloning and release-to-release migration

- Use a directory manager

IBM

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

43

IBM

Thanks!

Contact Information:	Stephen (Steve) G. Wilkins
	IBM
	z/VM Development
	Endicott, NY
	wilkinss@us.ibm.com

A PDF of this presentation will be available at:
http://www.vm.ibm.com/devpages/wilkinss