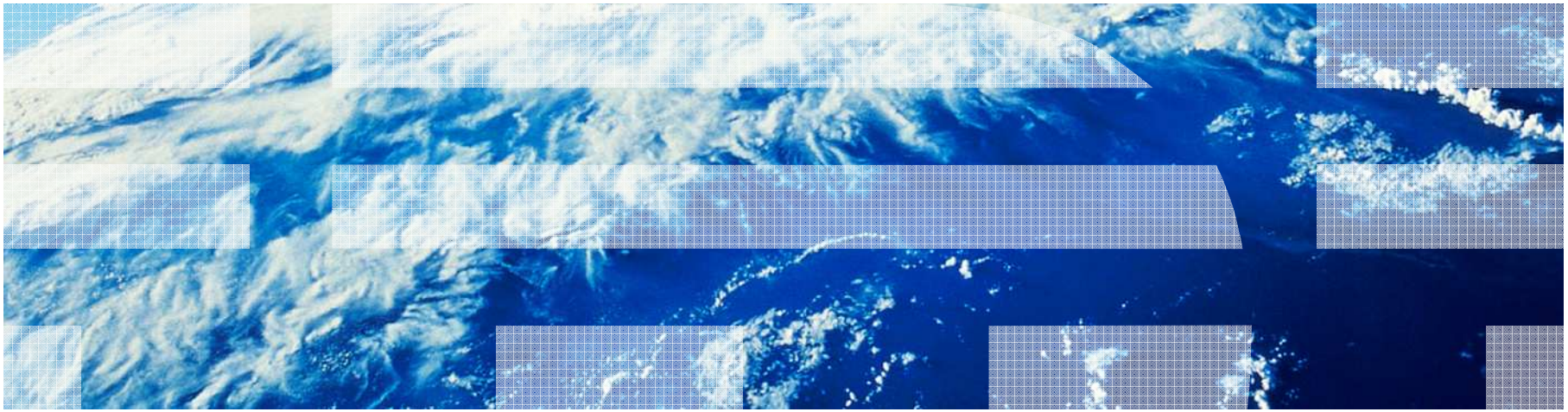
Getting Ready for z/VM Single System Image (SSI)

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™

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

IBM Statement of Direction – July 22, 2010

- **z/VM Single System Image with Live Guest Relocation**

  IBM intends to provide capabilities that permit multiple z/VM systems to collaborate in a manner that presents a single system image to virtual servers. An integrated set of functions will enable multiple z/VM systems to share system resources across the single system image cluster. Among those functions will be Live Guest Relocation, the ability to move a running Linux virtual machine from one member of the cluster to another. This virtual server mobility technology is intended to enhance workload balancing across a set of z/VM systems and to help clients avoid planned outages for virtual servers when performing z/VM or hardware maintenance.

Note: All statements regarding IBM's plans, directions, and intent are subject to change or withdrawal without notice, and represent goals and objectives only.

Topics

- Resource and Capacity Planning for SSI
- Installation Planning - Getting to SSI
- Updating your Directory for SSI
- Planning for Live Guest Relocation

Resource and Capacity Planning for SSI

Cluster Topography

1. How many members in your cluster?

2. Production configuration
   - How many CECs?
   - How many LPARS/CEC?
     - *Preferred configuration for 4-member cluster is 2 LPARs on each of 2 CECs*

3. Test configuration
   - VM guests?
   - LPARs?
   - Mixed?

4. Virtual server (guest) distribution
   - Each guest's "resident" member?
   - Where will each guest be relocated to?
   - *Distribute workload so each member has capacity to receive relocated guests*
     - CPU
     - Memory

Memory Requirements for Live Guest Relocation

- A relocating guest's current memory size ***must*** fit in available space on destination member
  - ***Current memory size*** assumes virtual memory is fully populated, including:
    - Private V-disks
    - Estimated size of supporting CP structures
  - ***Available space*** includes the sum of available memory:
    - Central
    - Expanded
    - Paging disk
- Additional memory checks
  - Does the guest's current memory size exceed the paging capacity on the destination?
  - Does the guest's maximum memory size exceed available space on the destination?
  - Does the guest's maximum memory size exceed paging capacity on the destination?
    - *These checks may be overridden if you are certain that they are not applicable to your environment*

Memory Requirements for Live Guest Relocation…

- Include standby and reserved storage settings when calculating maximum memory size for a guest

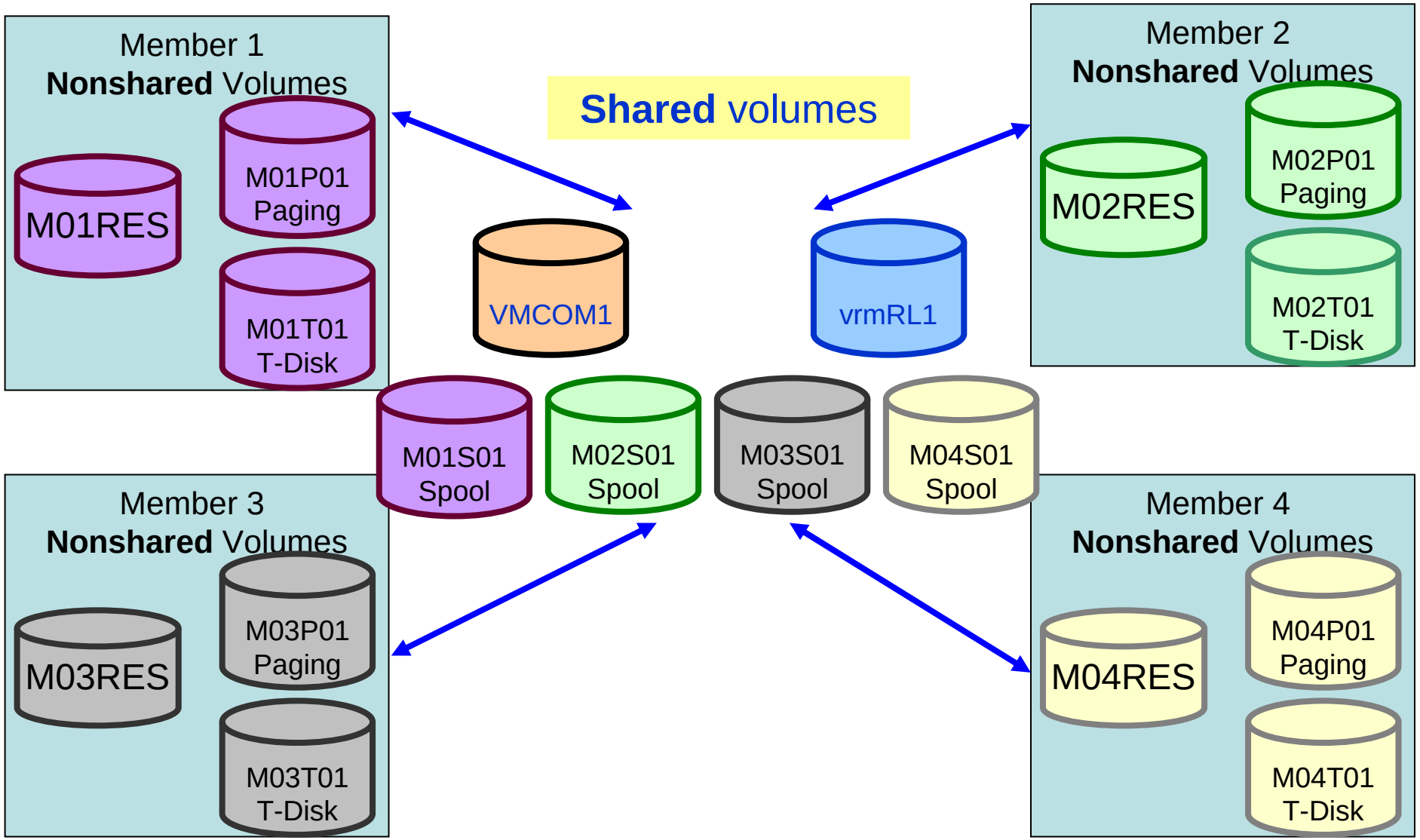
- Relocations may increase paging demand
  - Available paging space should be at least 2x total virtual memory of all guests
    - Including guests to be relocated to this member
  - Avoid allocating more than 50% of available paging space
    - If size of guests to be relocated increase in-use amount to > 50%, system performance could be impacted

```
q alloc page

                EXTENT     EXTENT  TOTAL  PAGES   HIGH    %
VOLID  RDEV      START        END  PAGES IN USE   PAGE USED
------ ---- ---------- ---------- ------ ------ ------ ----
L24B66 4B66           0       3338 601020 252428 252428   42%
```

DASD Planning

- Determine which DASD volumes will be used for
  - Cluster-wide volume
  - Release volumes
  - System volumes
    - Shared
    - Nonshared
  - User data (minidisks)
    - Shared
    - Nonshared
- Determine which member owns each CP-Owned volume

DASD Planning – Non-Shared and Shared System Volumes



DASD Planning – CP Volume Ownership

- CP-Owned volumes are marked with ownership information (CPFMTXA)
  - Cluster name
  - System name of owning member

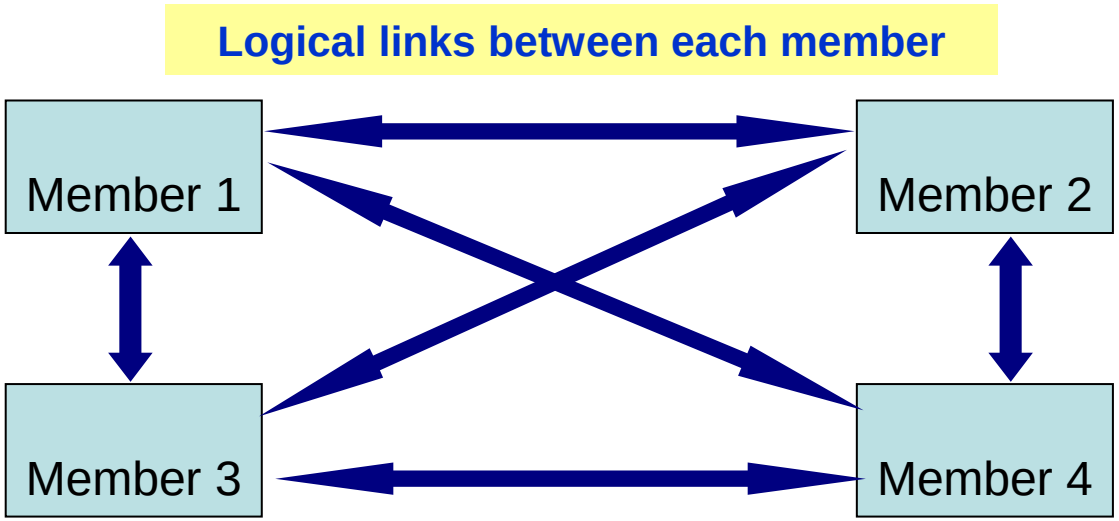
CP-Owned areas brought online in an SSI cluster

Cluster Name on Volume	System Name on Volume	SPOL Extents (Owner or Shared)	DRCT, PAGE, and TDSK Extents and Checkpoint and Warm Start Areas (Nonshared)
None	None	No	No
None	Name of this member	Yes (owner, single-member cluster only)	Yes
None	Not the name of this member	No	No
Name of this cluster	None	No	No
Name of this cluster	Name of this member	Yes (owner)	Yes
Name of this cluster	Name of another member	Yes (shared)	No
Name of this cluster	Not the name of a member (probable configuration error)	No	No
Not the name of this cluster	Any value	No	No

- Ownership information may also be used on non-SSI systems
  - System name but no cluster name
  - Default on non-SSI installs

CTC Connections

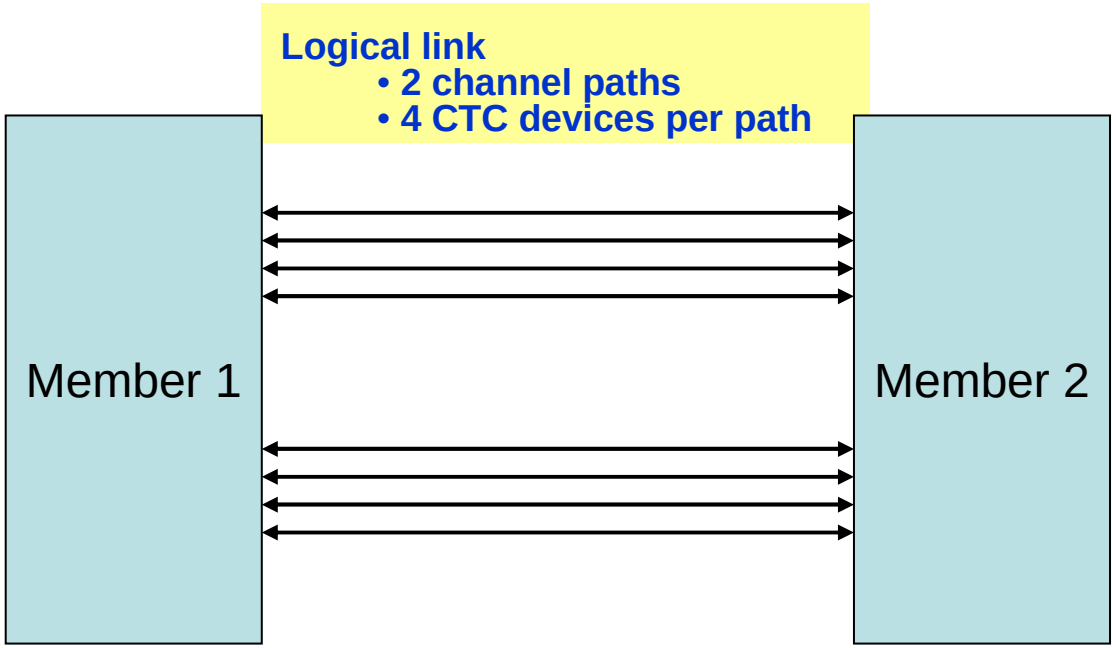
- Each member of an SSI cluster must have a direct ISFC connection to every other member (logical link)
- Logical links are composed of 1-16 CTC connections
  - FICON channel paths
  - May be switched or unswitched
- Use multiple CTCs distributed on multiple FICON channel paths between each member
  - Avoids write collisions that impact link performance
  - Avoids severing of logical link if one channel path is disconnected or damaged
- *Preferred practice:* Use same real device number for same CTC on each member



CTC Connections – How Many Do I Need?

- 4 CTCs per FICON channel path provides most efficient ISFC data transfer
- For large guests, relocation time and quiesce time can be improved with more channel paths*
  - Up to 4, with 4 CTCs each

* *Based on early performance measurements; there are additional factors that affect relocation and quiesce times*



Network Planning

- All members must have identical network connectivity
  - Connected to same physical LAN segments
  - Connected to same SAN fabric

- Assign equivalency identifiers (EQIDs) to all network devices
  - Devices assigned same EQID on each member must be of same type, have the same capabilities, and have connectivity to the same destinations

Network Planning – Virtual Switches

- Define virtual switches with same names on each member
- For relocating guests:
  - Source and destination virtual switch guest NIC and port configurations must be equivalent
    - Port type
    - Authorizations (access, VLAN, promiscuous)
  - Source and destination virtual switches must be equivalent
    - Name and type
    - VLAN settings
    - Operational UPLINK port with matching EQID
    - Device and port numbers don't need to match, but connectivity to the same LAN segment is required

Network Planning – MAC Addresses

- MAC address assignments are coordinated across an SSI cluster
  - VMLAN statement
    - MACPREFIX must be set to different value for each member
      - Default is 02-xx-xx where xx-xx is "system number" of member
    - USERPREFIX must be set for SSI members
      - Must be identical for all members
      - Must not be equal to any member's MACPREFIX value
      - Default is 02-00-00
    - MACIDRANGE is ignored in an SSI cluster
    - Example:

```
VMSYS01: VMLAN MACPREFIX 021111 USERPREFIX 02AAAA
VMSYS02: VMLAN MACPREFIX 022222 USERPREFIX 02AAAA
VMSYS03: VMLAN MACPREFIX 023333 USERPREFIX 02AAAA
VMSYS04: VMLAN MACPREFIX 024444 USERPREFIX 02AAAA
```

SSI Planning Worksheet

Linux server user ID	Memory	Virtual processors	DASD	Networking devices	Cryptographic requirements	Member 1	Member 2	Member 3	Member 4

Maximum number of resident and relocated virtual servers:				
Maximum memory for normally resident and relocated virtual servers:				
Memory for z/VM:				
Total memory requirement:				
Central storage estimate (Total memory × .75):				
Expanded storage estimate (Total memory ×.25):				
Number of real CPUs:				
DASD paging space (Total memory × 2):				

Installation
... or ...
How Do I Get to SSI?

Planning Your SSI Installation

What kind of Installation should I select?

- SSI installation
  - Single installation for multiple z/VM images
  - Installed and configured as an SSI cluster
    - Single source directory
    - Shared system configuration file
    - Creates Persistent Data Record (PDR) on Common volume

- Non-SSI installation
  - Single z/VM image
  - Can be converted to initial member of an SSI cluster later
  - Builds DASD layout, directory, and configuration file the same as SSI installation

- Documented migration scenarios require non-SSI installation
  - SSI installation primarily for new or "from scratch" installs

Migrating to SSI

- "Use case" scenarios (CP Planning and Administration)
  - Migration procedures for existing z/VM environments
    - Converting a z/VM System to a Single-Member z/VM SSI Cluster
    - Adding a Member to a z/VM SSI Cluster by Cloning an Existing Member
    - Combining Two Non-SSI z/VM Systems to Create a z/VM SSI Cluster
    - Moving a Second-Level z/VM SSI Cluster to First-Level
    - Converting a CSE Complex to a z/VM SSI Cluster
    - Decommissioning a Member of a z/VM SSI Cluster
- Review documented procedures before deciding whether to do SSI or non-SSI install

Non-SSI Installation

Select installation type

```
Select a System Type: Non-SSI or SSI (SSI requires the SSI feature)
    X Non-SSI Install:    System Name SYSTEM1
    _ SSI Install:        Number of Members _     SSI Cluster Name ________
```

Identify CP-Owned and Release volumes

```
       VOLUME    DASD        DASD                          FORMAT DASD
        TYPE     LABEL      ADDRESS                           Y/N
       =======  ========   =========                      ==========
       COMMON    VMCOM1      0111                            N
       RELVOL    620RL1      0222
       RELVOL2   620RL2      0333

       VOLUME    DASD        DASD
        TYPE     LABEL      ADDRESS
       =======  ========   =========
SYSTEM1
       RES       M01RES      0444
       SPOOL     M01S01      0666
       PAGE      M01P01      0888
       WORK      M01W01      AAAA
```

SSI Installation

Select installation type

```
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 SSICLUST
```

Identify SSI member systems

```
SSI Member Name(s):

SLOT #      MEMBER NAME     IPL LPAR/USERID
======      ===========     ===============
  1          MEM1             LPAR1
  2          MEM2             LPAR2
  3          MEM3             LPAR3
  4          MEM4             LPAR4
```

SSI Installation (cont.)

Define CP-Owned and Release volumes for all members

	VOLUME TYPE	DASD LABEL	DASD ADDRESS				FORMAT DASD Y/N
	COMMON	VMCOM1	0111				N
	RELVOL	620RL1	0222				
	RELVOL2	620RL2	0333				
	VOLUME TYPE	DASD LABEL	DASD ADDRESS		VOLUME TYPE	DASD LABEL	DASD ADDRESS
MEM1				MEM2			
	RES	M01RES	0444		RES	M02RES	0995
	SPOOL	M01S01	0666		SPOOL	M02S01	3945
	PAGE	M01P01	0888		PAGE	M02P01	A345
	WORK	M01W01	AAAA		WORK	M02W01	3345
MEM3				MEM4			
	RES	M03RES	2224		RES	M04RES	4556
	SPOOL	M03S01	1345		SPOOL	M04S01	0234
	PAGE	M03P01	0ACF		PAGE	M04P01	0FCD
	WORK	M03W01	033D		WORK	M04W01	0DD3

SSI Installation (cont.)

Define Common Volume and CTC Device addresses

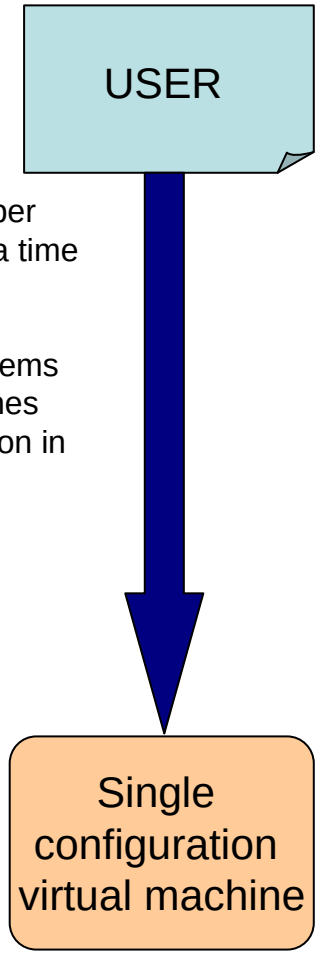
```
Real addresses for the common volume on each member LPAR:

  VOLUME     DASD        MEM1         MEM2         MEM3         MEM4
   TYPE      LABEL       ADDRESS      ADDRESS      ADDRESS      ADDRESS
  =======   ========    =========    =========    =========    =========
   COMMON    VMCOM1       0111         0212         0122         0111

CTC device addresses:

  From: MEM1                              From: MEM2
        To: MEM1          N/A                   To: MEM1     003D 000D
        To: MEM2     0993 0032                  To: MEM2        N/A
        To: MEM3     0335 0992                  To: MEM3     0223 ____
        To: MEM4     0944 ____                  To: MEM4     DDF1 FFF3

  From: MEM3                              From: MEM4
        To: MEM1     0AAA DFE7                  To: MEM1     0334 ____
        To: MEM2     0AFD ____                  To: MEM2     3334 DFA7
        To: MEM3        N/A                     To: MEM3     DDDD AAF2
        To: MEM4     DDDD AAF2                  To: MEM4        N/A
```

Updating Your Directory For SSI

Shared Source Directory – Virtual Machine Definition Types

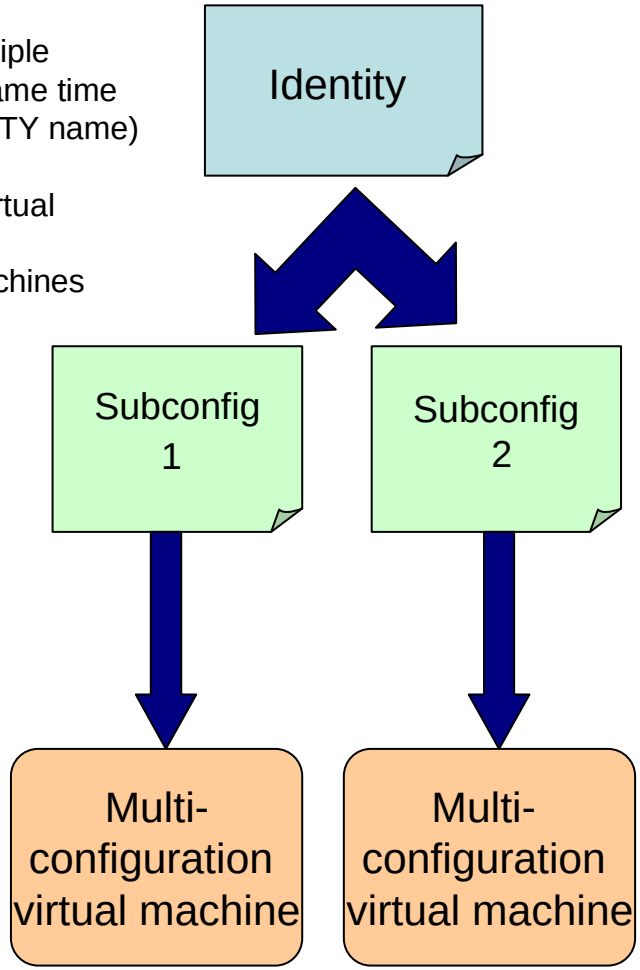
Traditional Definition

- 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



New Definition

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



New Directory Layout

- IBM-supplied directory will be significantly different than previous releases
  - Both SSI and non-SSI installations
  - Directory for non-SSI installations will be in "SSI-ready" format
    - Facilitate future SSI deployment
- Many of the IBM-supplied userids will be multiconfiguration virtual machines
- Determine if any of your users should be defined as multiconfiguration virtual machines
  - Most will be single-configuration virtual machines
  - Userids defined on SYSTEM_USERIDS statements will usually be multiconfiguration virtual machines
- Merge your user definitions into the IBM-supplied directory

Multiconfiguration Virtual Machine Definition

```
 <identity> MAINT    MAINTPAS     128M 1000M ABCDEFG

 <use> MAINT-1 <when on> SSIMEMB1
 <use> MAINT-2 <when on> SSIMEMB2
 <use> MAINT-3 <when on> SSIMEMB3
 <use> MAINT-4 <when on> SSIMEMB4

 CONSOLE 009 3215 T
 SPOOL 00C 2540 READER *
 SPOOL 00D 2540 PUNCH  A
 SPOOL 00E 1403 A
 LINK    USER1     2CC 2CC RR
 LINK    USER1     551 551 RR

<Entry> MAINT-1
 MDISK 0191 3390 1000 20 MNTVL1 WR
 MDISK CF1  3390 100  20 M01RES RR
* END OF MAINT-1

<Entry> MAINT-2
 MDISK 0191 3390 1000 20 MNTVL2 WR
 MDISK CF1  3390 100  20 M02RES RR
* END OF MAINT-2

<Entry> MAINT-3
 MDISK 0191 3390 1000 20 MNTVL3 WR
 MDISK CF1  3390 100  20 M03RES RR
* END OF MAINT-3

<Entry> MAINT-4
 MDISK 0191 3390 1000 20 MNTVL4 WR
 MDISK CF1  3390 100  20 M04RES RR
* END OF MAINT-4
```

These statements apply to all instances of MAINT on all members

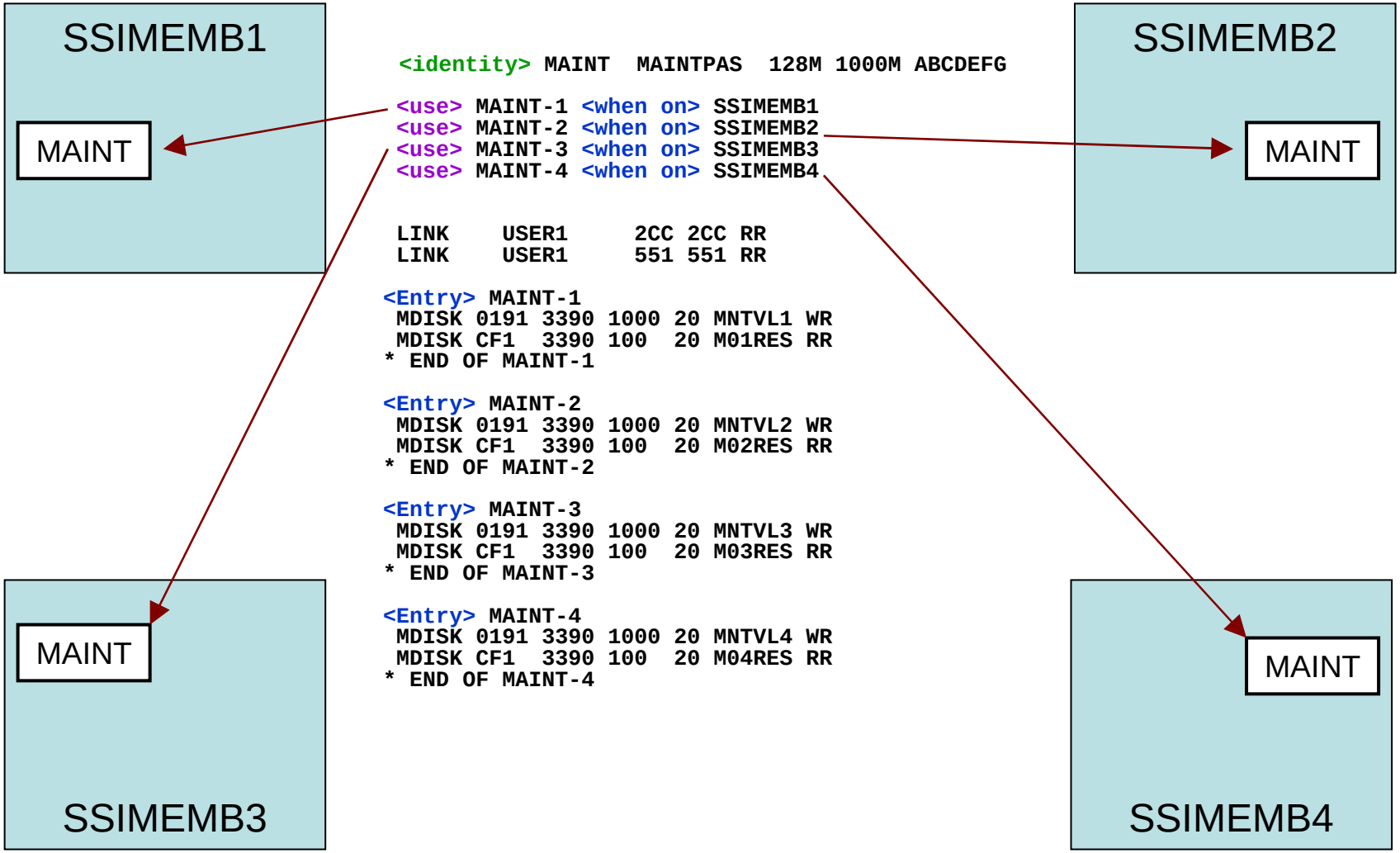
These statements only apply to MAINT on member SSIMEMB1

These statements only apply to MAINT on member SSIMEMB2

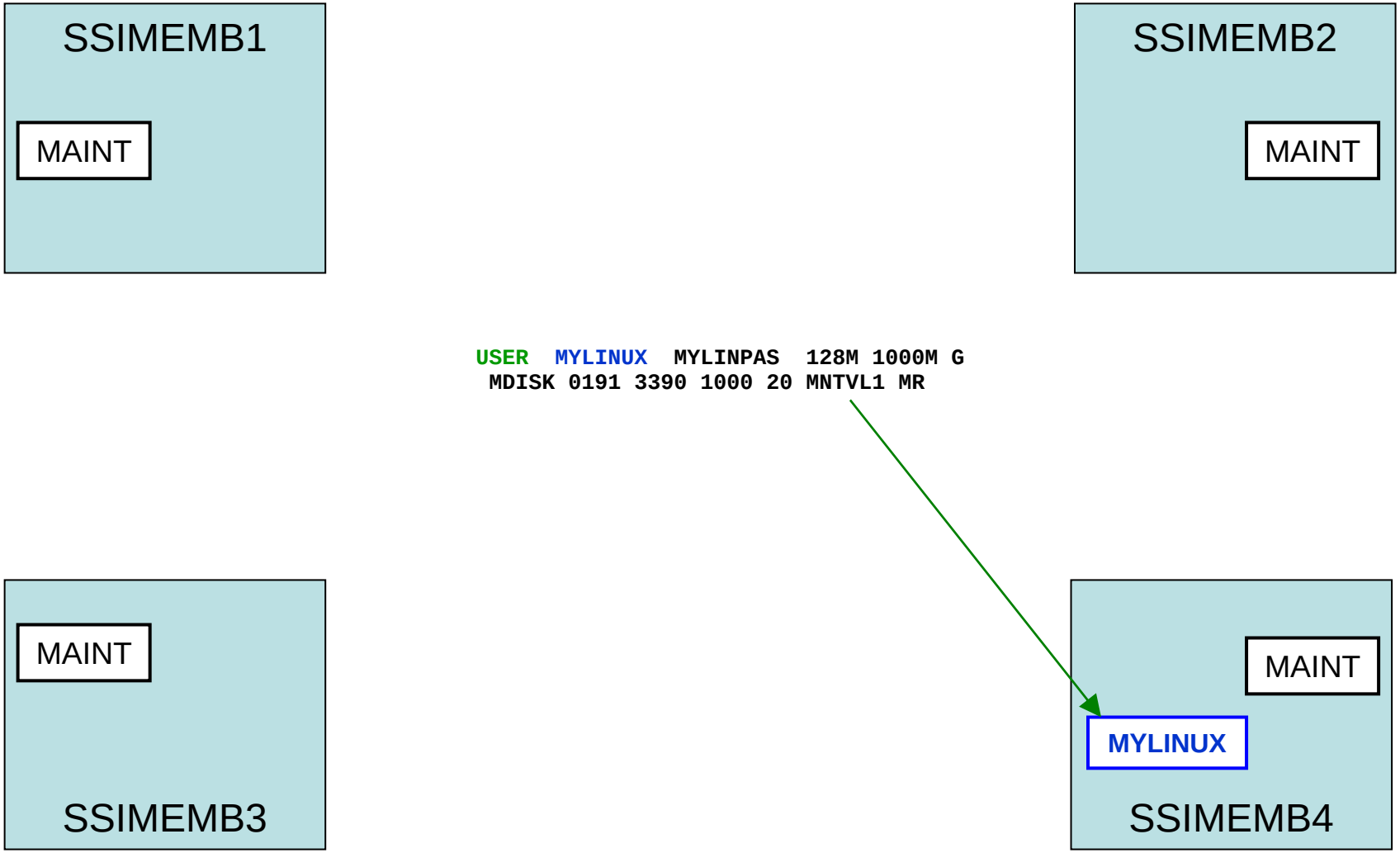
These statements only apply to MAINT on member SSIMEMB3

These statements only apply to MAINT on member SSIMEMB4

Multiconfiguration Virtual Machines



Single Configuration Virtual Machines



Planning for Live Guest Relocation

Guest Configuration for Live Guest Relocation

- In order to be eligible to relocate, a Linux guest must be:
  - Defined as a single configuration virtual machine
  - Running in an ESA or XA virtual machine running ESA/390 or z/Architecture mode
  - Logged on but disconnected
  - Running only type CP or type IFL virtual processors
  - IPLed from either a
    - Device
    - Named saved system (NSS)

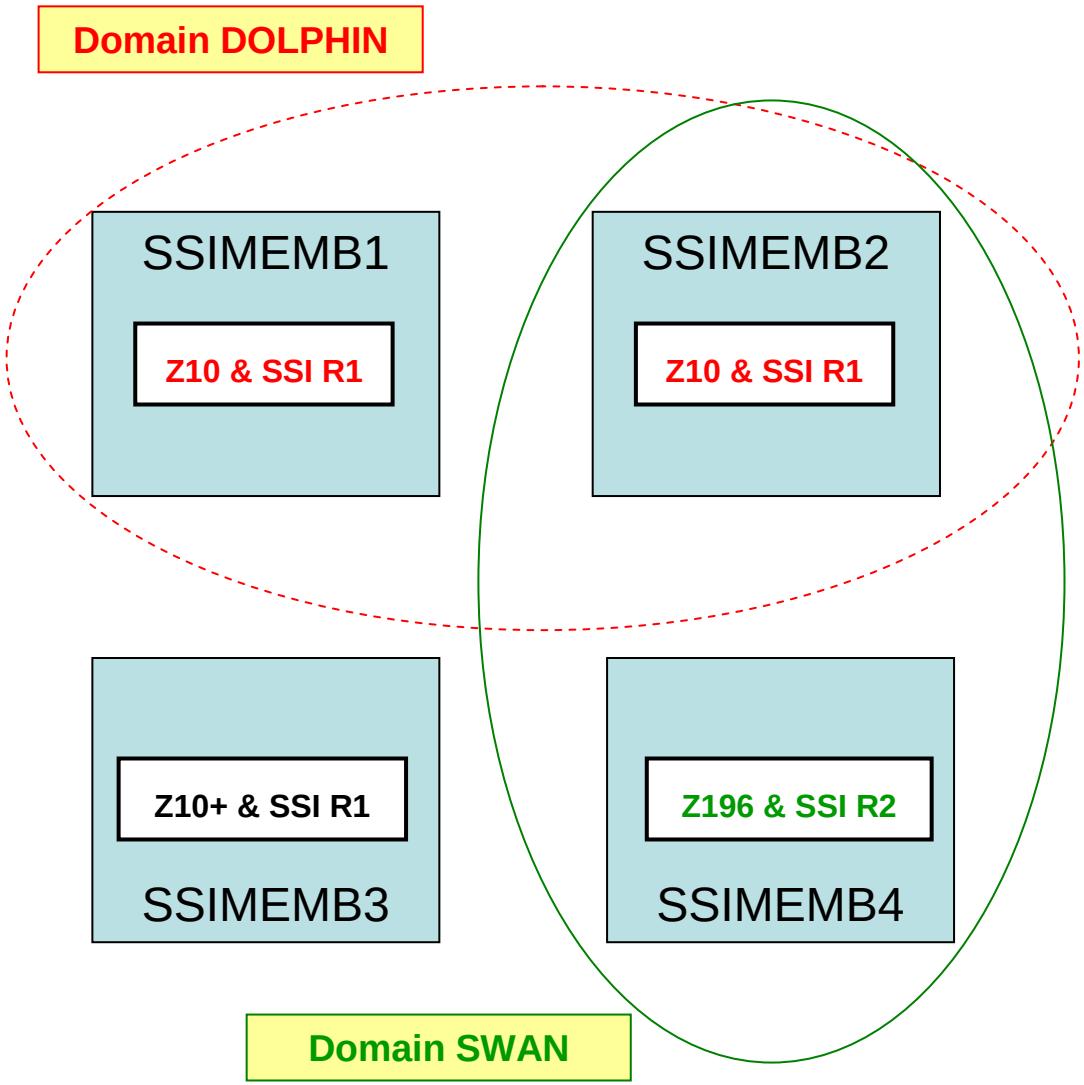
- If a guest is using a DCSS or NSS:
  - Identical NSS or DCSS must be available on the destination member
  - It cannot have the following types of page ranges
    - SW (shared write)
    - SC (shared with CP)
    - SN (shared with no data)

Guest Configuration for Live Guest Relocation (cont.)

- A guest can relocate if it has any of the following:
  - Dedicated devices
    - Equivalent devices and access must be available on destination member
  - Private v-disks
  - No open spool files other than console files
  - VSWITCHes
    - Equivalent VSWITCH and network connectivity must be available on destination

- A relocating guest can be using any of the following facilities:
  - Cryptographic adapter
    - Crypto cards for shared domains on source and destination must be same AP type
  - Virtual machine time bomb (Diag x'288')
  - IUCV connections to *MSG and *MSGALL CP system services
  - Application monitor record collection
    - If guest buffer is not in a shared DCSS
  - Single Console Image Facility
  - Collaborative Memory Management Assit (CMMA)

Relocation Domains



- Default domains:
  - SSI
    - includes all members
  - SSIMEMB1
  - SSIMEMB2
  - SSIMEMB3
  - SSIMEMB4

- "Customized" domains
  - DOLPHIN includes members
    - SSIMEMB1
    - SSIMEMB2
  - SWAN includes members
    - SSIMEMB2
    - SSIMEMB4

Relocation Domains (cont.)

- Set of members among which guests can relocate freely
  - Destination member does not need to support architecture or CP facilities that are available on the source member
  - "Maximal Common Subset"
    - Each domain is assigned a "virtual architecture" based on the least capable system in the domain
    - Guests have facilities available in their domain's "virtual architecture"
    - Guests can relocate to any system in their domain without losing capabilities

- Relocation domains are defined in the system configuration file or dynamically
  - Default (built-in) domains:
    - SSI (includes all members)
    - Single-member domains for each member

- Single-configuration virtual machines
  - Assigned to a relocation domain in directory or dynamically
  - Default domain is entire cluster

- Multiconfiguration virtual machines
  - Permanently assigned to single member domain for each member it can log on to

Summary

- New way to deploy z/VM images and resources
  - Benefit from clustering and virtual server mobility

- Planning and thought required
  - Capacity and equipment
  - Resource sharing
  - Virtual networks
  - Installation
    - SSI cluster configuration
    - Migrating from your current z/VM environment
  - User directory
    - Virtual machine (guest) definition and distribution
  - Live Guest Relocation

- New documentation to assist with
  - SSI Planning
  - Migrating to an SSI cluster

Thanks!

Contact Information:

John Franciscovich
IBM
z/VM Development
Endicott, NY

francisj@us.ibm.com