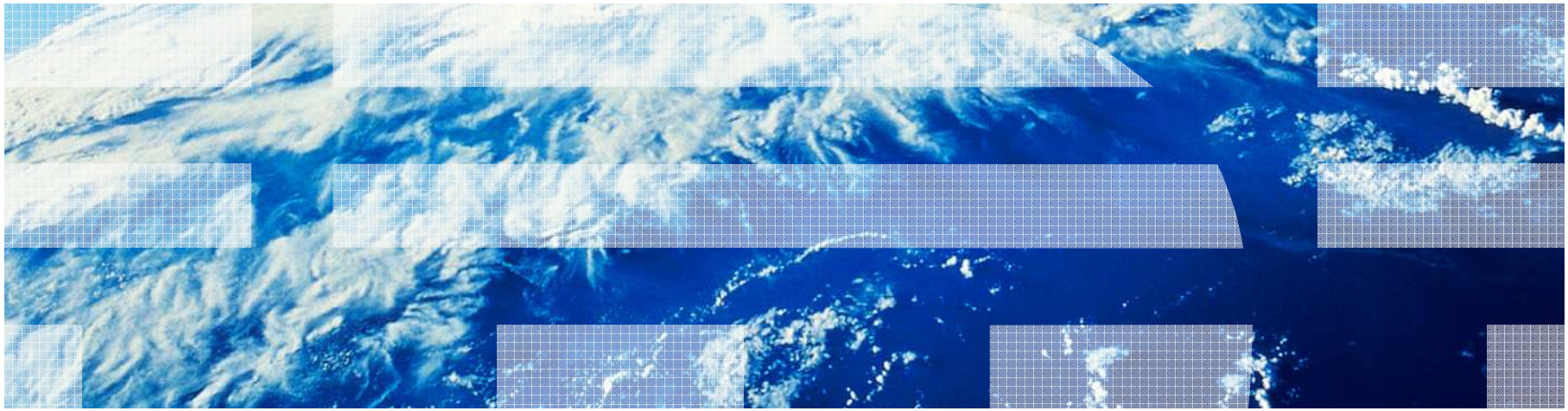
John Franciscovich – IBM
March 2011

z/VM Single System Image and Guest Mobility Preview

SHARE – Anaheim, CA
Session 8453



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

- 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

z/VM Single System Image (SSI) Cluster

- Up to 4 z/VM systems (*members)* in an ISFC collection
  - Provides a set of shared resources for the z/VM systems and their virtual machines
    - Managed as a single resource pool
  - Recommend 2 CECs, 2 LPARs on each

- 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

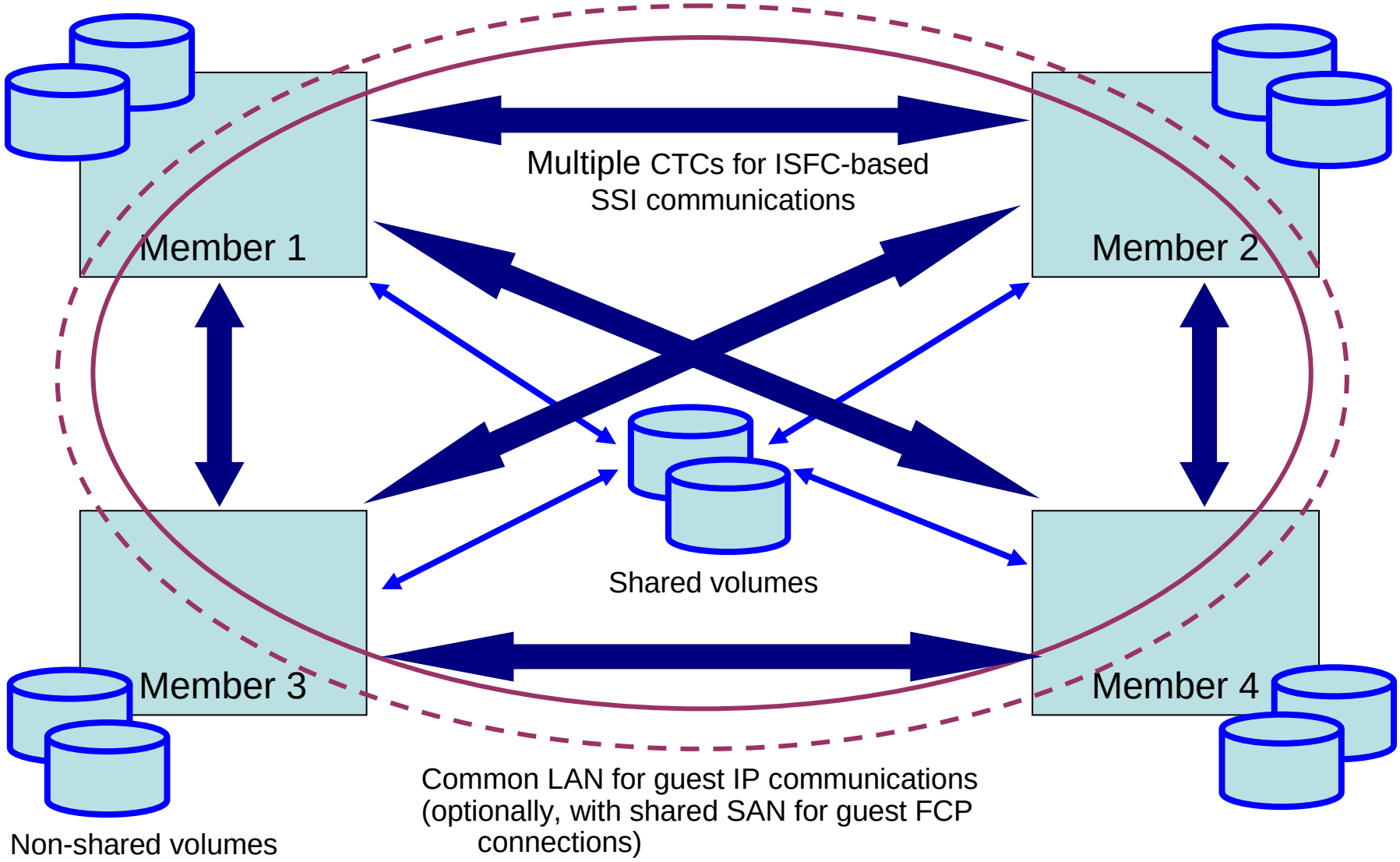
- Each member can access common resources
  - Shared DASD volumes
  - Same Ethernet LAN segments
  - Same storage area networks (SANs)

- **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 z/VM planned outages
  - z/VM and hardware maintenance are less disruptive to workloads
- Eases deployment and maintenance of multiple z/VM images
- ***Live Guest Relocation*** provides virtual server mobility
  - Dynamically move virtual servers (guests) from one member to another
  - Less disruptive workload balancing

z/VM SSI Cluster

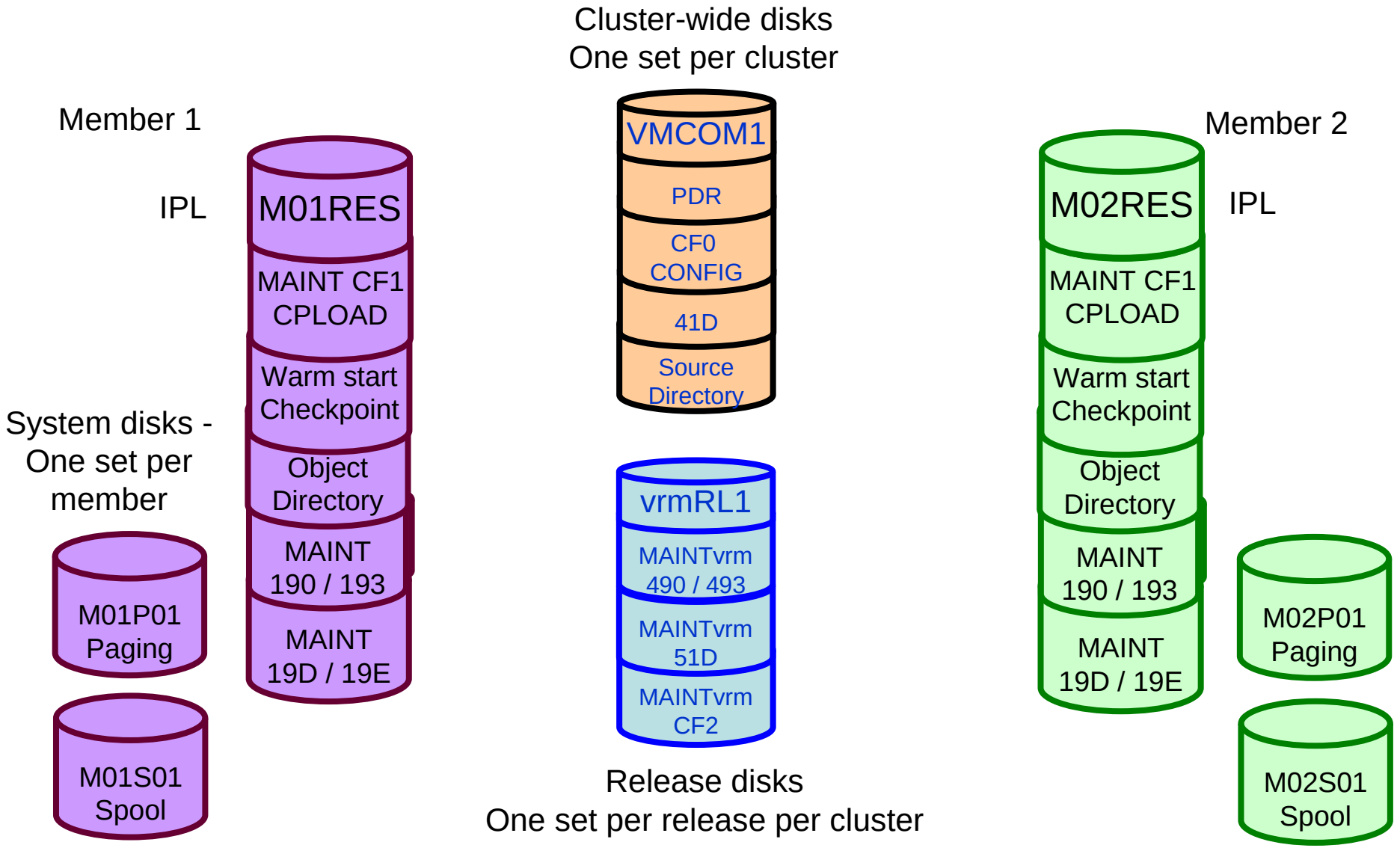


Major Attributes of a z/VM SSI Cluster

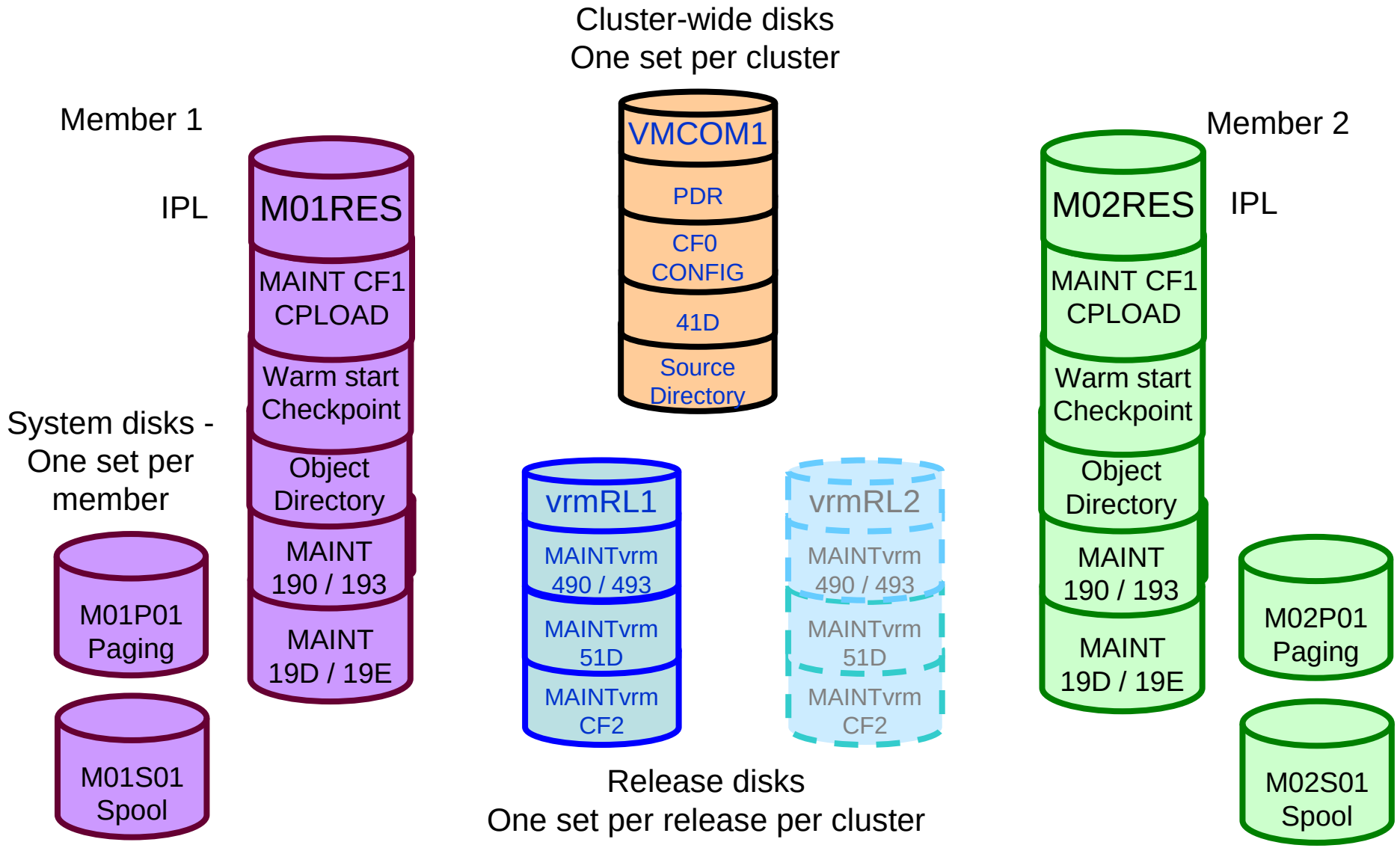
Multisystem Installation

- SSI cluster can be created with a single z/VM install
  - Customer provides information about the cluster on installation panels
    - DASD volumes
    - Channel-to-channel connections for ISFC
  - z/VM images 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



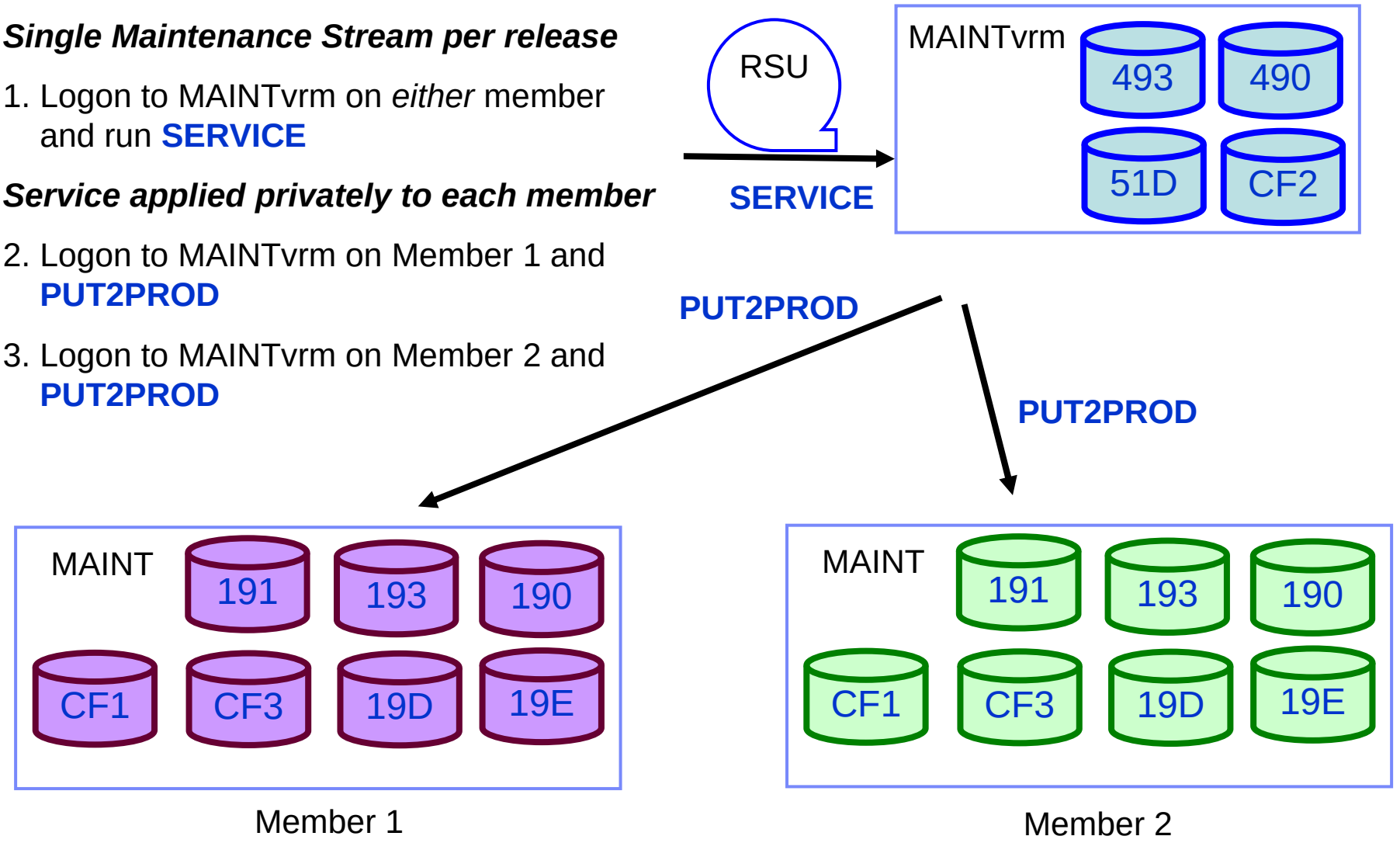
Applying Service

Single Maintenance Stream per release

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

Service applied privately to each member

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



Common System Configuration File

- Resides on new shared parm disk
- Define cluster name and each of its member systems
  - SYSTEM_IDENTIFIER enhanced
    - LPAR name can be matched to define system name
    - System name can be set to the LPAR name
- Identify direct ISFC links between members
- Define CP Owned volumes
  - Private
    - Paging
    - Tdisk
    - Sysres
  - Shared
    - Spool
    - SSI common volume
- Can include member-specific configuration

Common 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
```

Common 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
  - Created and viewed with a new 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 will be enhanced to display ownership information

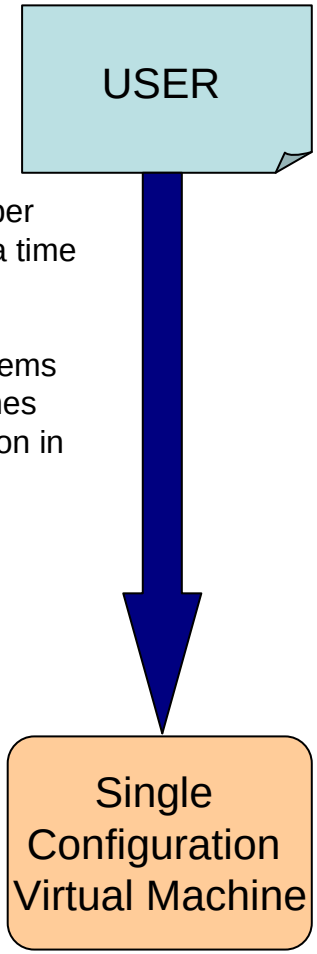
Defining Virtual Machines – Shared Source Directory

- All user definitions in a single shared source directory
- Run DIRECTXA on each member
- No system affinity (SYSAFFIN)
- Identical object directories on each member
- Single security context
  - Each user has same access rights and privileges on each member

Using a directory manager is strongly recommended!

Defining Virtual Machines – Shared Source Directory…

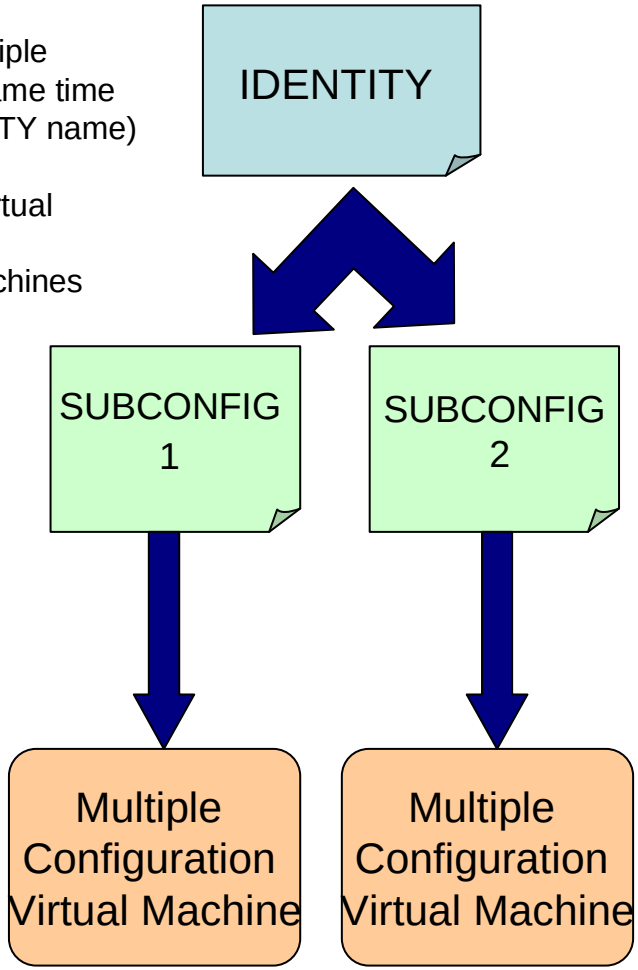
Traditional Definition



- 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



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

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

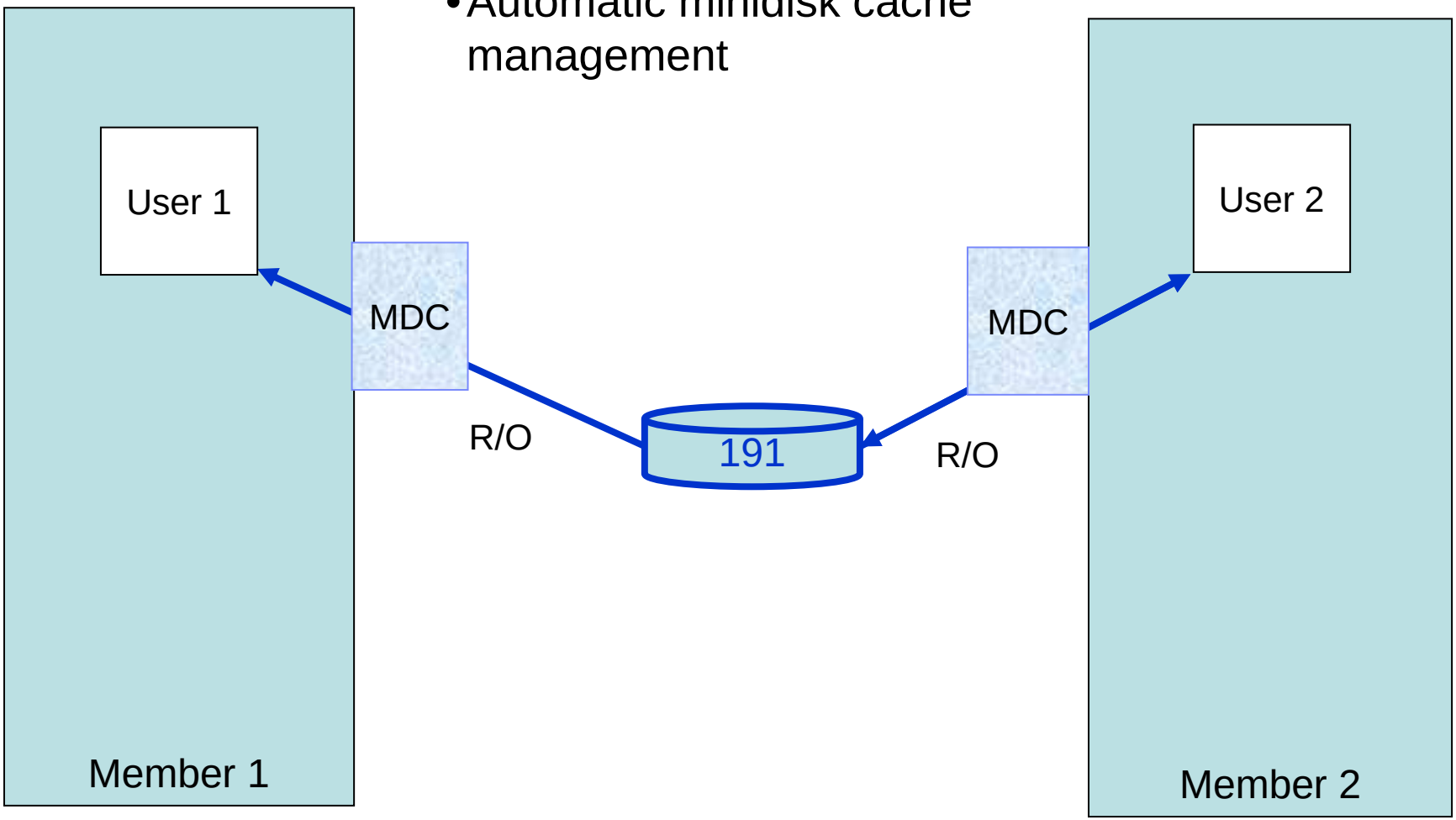
```
AT sysname CMD cmdname
```

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**

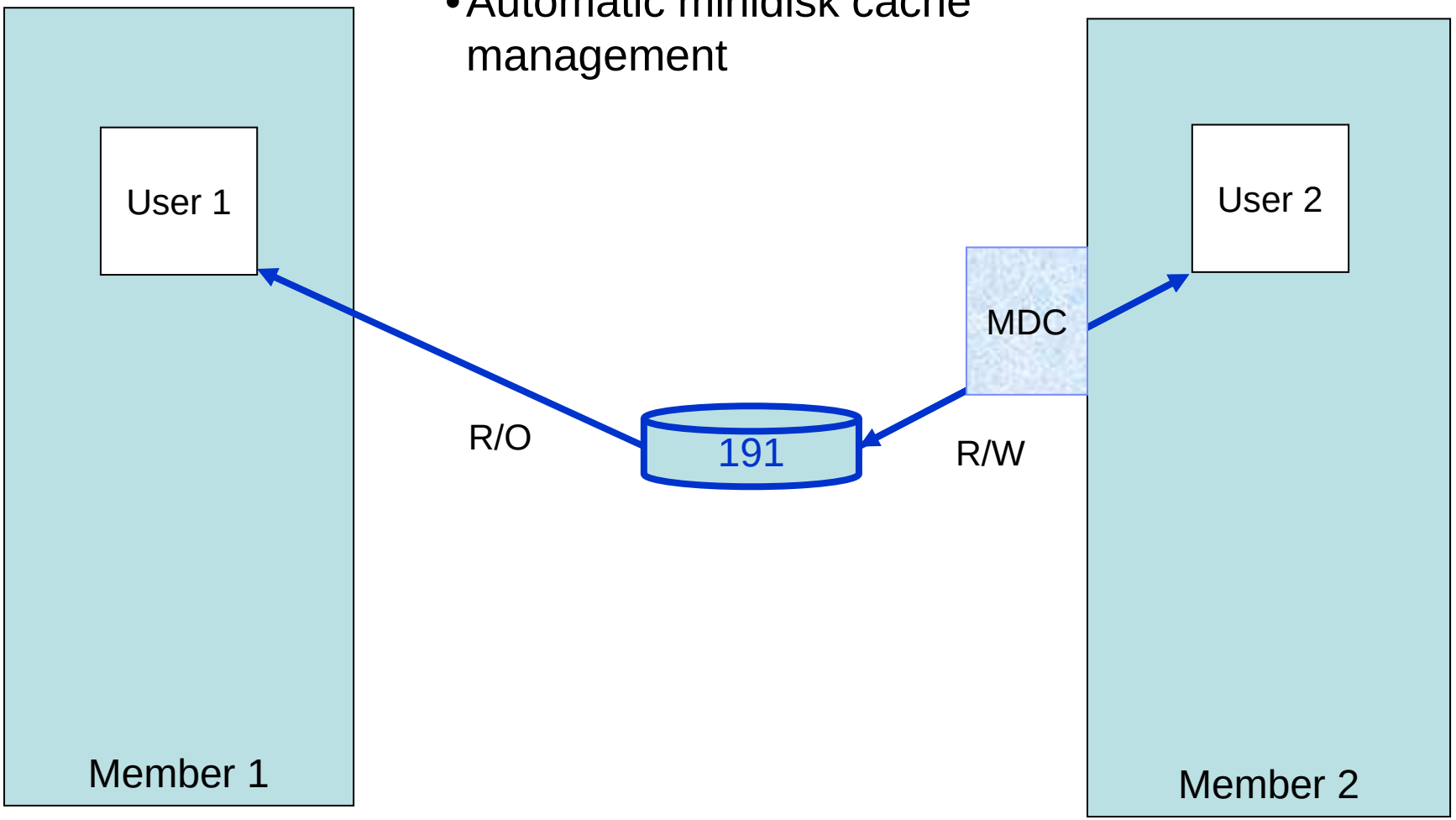
Cross-System Minidisk Management…

- Automatic minidisk cache management



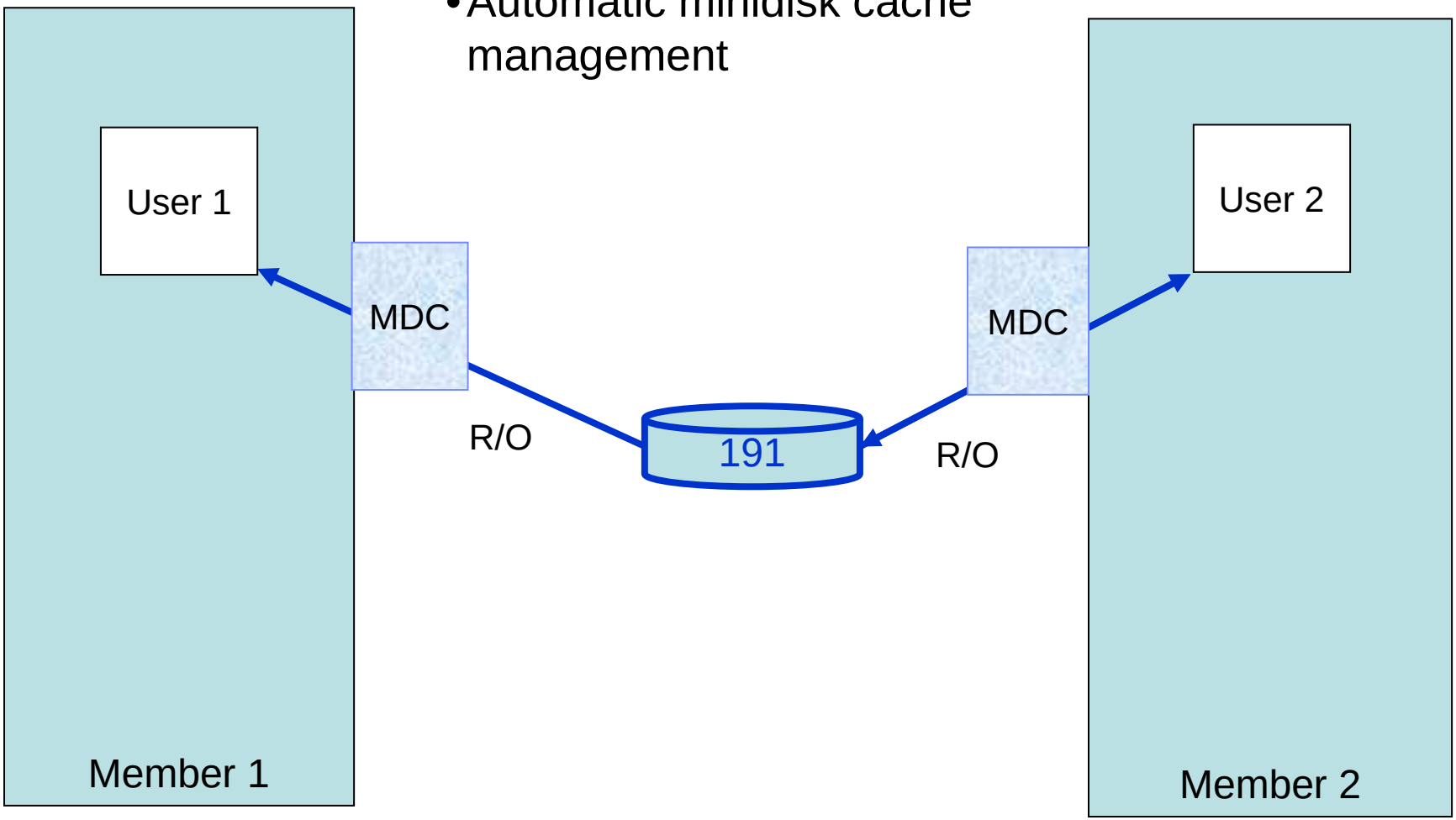
Cross-System Minidisk Management…

- Automatic minidisk cache management



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
  - 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 CP 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

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
- Members leave the SSI cluster when they shut down
- Status can be viewed with new commands

SSI Cluster Status – Example 1

```
SSI Name: CLUSTERA
SSI Level: 1
SSI Mode: Influx
Cross-System Timeouts: Enabled
SSI Persistent Data Record (PDR) device: VMCOM1 on EFE0

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

SSI Cluster Status – Example 2

```
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                       level: 1
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
```

Planning and Creating a z/VM SSI Cluster

SSI Cluster Requirements

- Servers must be IBM System z10 or z196
- Shared and non-shared DASD
  - 3390 volume required for the PDR
- 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

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 will be supported for only Linux on System z guests

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

Summary

- 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

- An SSI cluster gives you
  - Workload balancing (take the workload to the hardware)
  - Maintenance on your schedule (not the application owner)
  - Easier multi-system operation

Thanks!

Contact Information:

John Franciscovich
IBM
z/VM Development
Endicott, NY

francisj@us.ibm.com