Understanding z/VM's LGR and Relocation Domains

Version 1.3

Bill Bitner
z/VM Development Lab Client Focus & Care
bitnerb@us.ibm.com

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

Db2*	FlashCopy*	IBM (logo)*	OMEGAMON*	z13*	z/Architecture*	zSeries*
DirMaint	FlashSystem	IBM Z*	PR/SM	z13s	zEnterprise*	z/VM*
DS8000*	GDPS*	LinuxONE*	RACF*	z14	z/OS*	z Systems*
ECKD	ibm.com	LinuxONE Emperor	System z10*	z10 BC	zSecure	
FICON*	IBM eServer	LinuxONE Rockhopper	XIV*	z10EC		

* Registered trademarks of IBM Corporation

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.
IT Infrastructure Library is a Registered Trade Mark of AXELOS Limited.
ITIL is a Registered Trade Mark of AXELOS Limited.
Linear Tape-Open, LTO, the LTO Logo, Ultrium, and the Ultrium logo are trademarks of HP, IBM Corp. and Quantum in the U.S. and other countries.
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.
Linux is a registered trademark of Linus Torvalds 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.
OpenStack is a trademark of OpenStack LLC. The OpenStack trademark policy is available on the OpenStack website.
Java and all Java-based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.
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.
UNIX is a registered trademark of The Open Group in the United States and other countries.
VMware, the VMware logo, VMware Cloud Foundation, VMware Cloud Foundation Service, VMware vCenter Server, and VMware vSphere are registered trademarks or trademarks of VMware, Inc. or its subsidiaries in the United States and/or other jurisdictions.
Other product and service names might be trademarks of IBM or other 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.
This information provides only general descriptions of the types and portions of workloads that are eligible for execution on Specialty Engines (e.g, zIIPs, zAAPs, and IFLs) ("SEs"). IBM authorizes customers to use IBM SE only to execute the processing of Eligible Workloads of specific Programs expressly authorized by IBM as specified in the “Authorized Use Table for IBM Machines” provided at www.ibm.com/systems/support/machine_warranties/machine_code/aut.html (“AUT”). No other workload processing is authorized for execution on an SE. IBM offers SE at a lower price than General Processors/Central Processors because customers are authorized to use SEs only to process certain types and/or amounts of workloads as specified by IBM in the AUT.

Abstract

It sounds like an old real estate joke, but the three most important things to understand about Live Guest Relocation: Domains, Domains, Domains. Live Guest Relocation was introduced in z/VM 6.2 and many customers have grown to love it and depend on it. One of the key concepts associated with LGR is relocation domains. If you've ever wanted to really understand this part of the whole relocation construct, then come to this session. We'll describe what relocation domains are and how they affect LGR. You'll leave being able to understand how virtual architectures are influenced by relocation domains. In other words, you'll leave being able to amaze your friends and coworkers.

Agenda

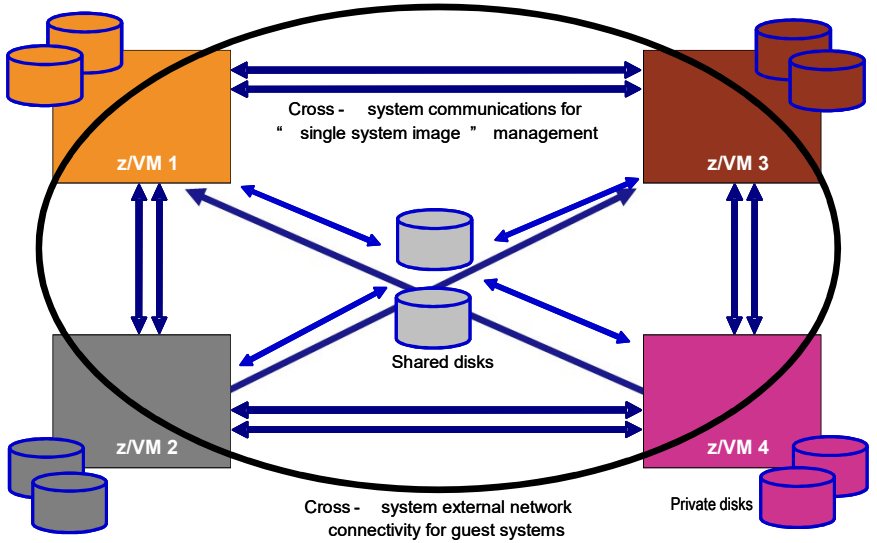
- Background
  - Single System Image
  - Live Guest Relocation
- Relocation Domains
  - Why?
  - What?
  - How?
- Uses of Relocation Domains

Single System Image and Live Guest Relocation

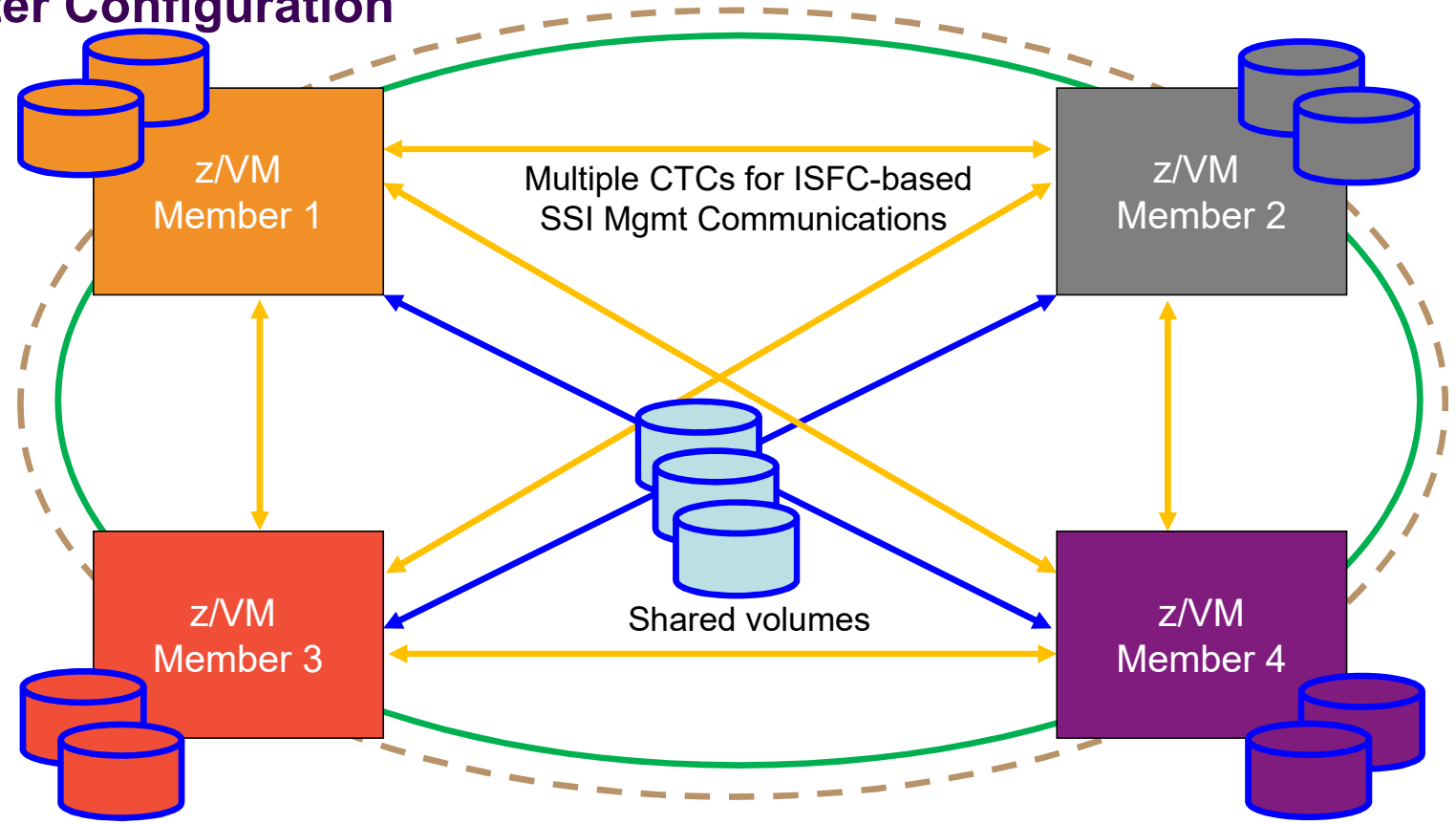
Single System Image (SSI) Feature

Clustered Hypervisor with Live Guest Relocation

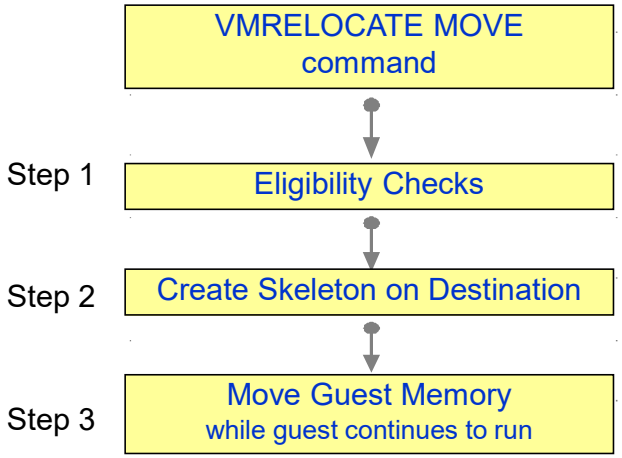
- Optional feature, available starting with z/VM 6.2 (No cost starting in z/VM 7.1)
- Connect up to four z/VM systems as members of a Single System Image cluster
- Cluster members can be run on the same or different IBM Z or LinuxONE servers
- Simplifies management of a multi-z/VM environment
  - Single user directory
  - Cluster management from any member
    - Apply maintenance to all members in the cluster from one location
    - Issue commands from one member to operate on another
  - Built-in cross-member capabilities
  - Resource coordination and protection of network and disks
- Allows Live Guest Relocation of running Linux guests



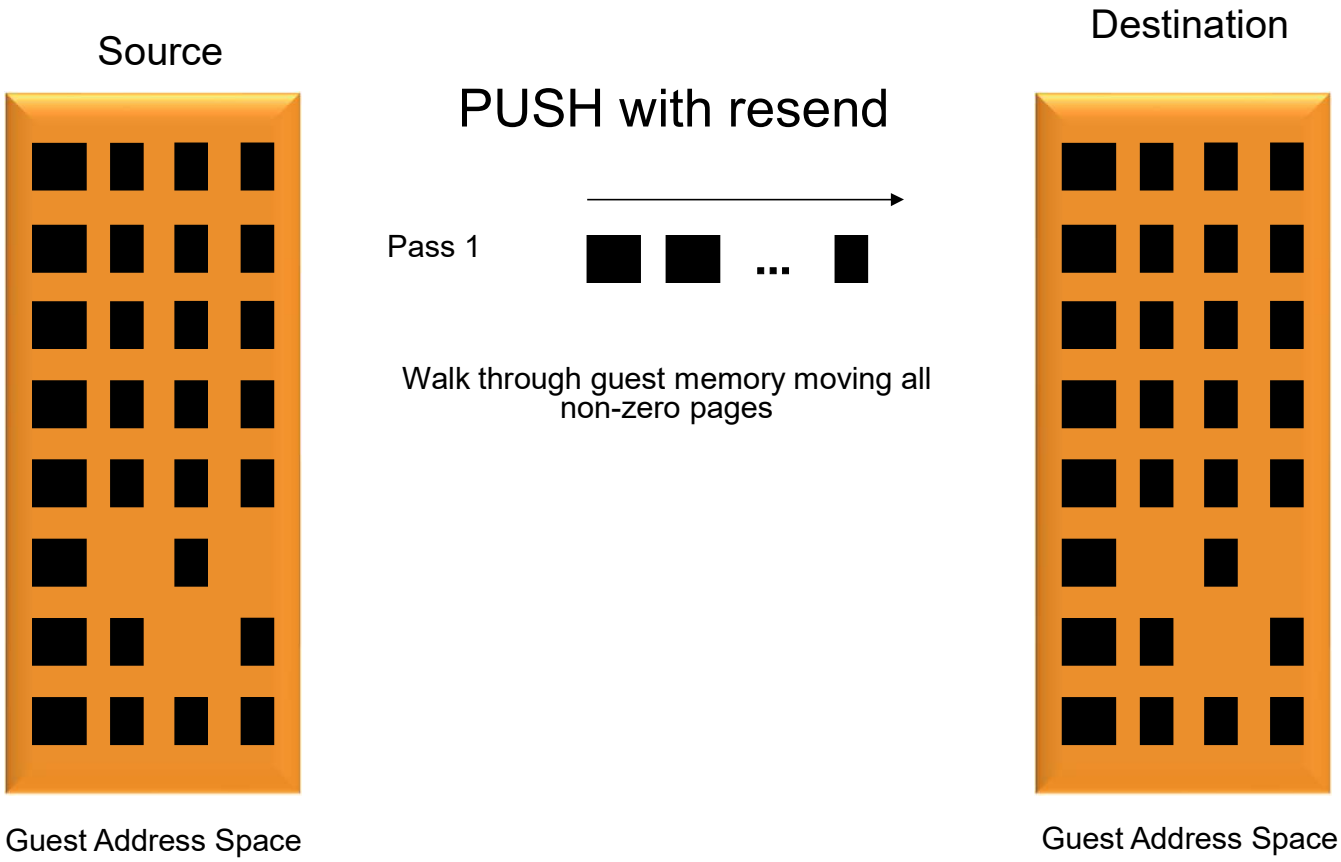
SSI Cluster Configuration



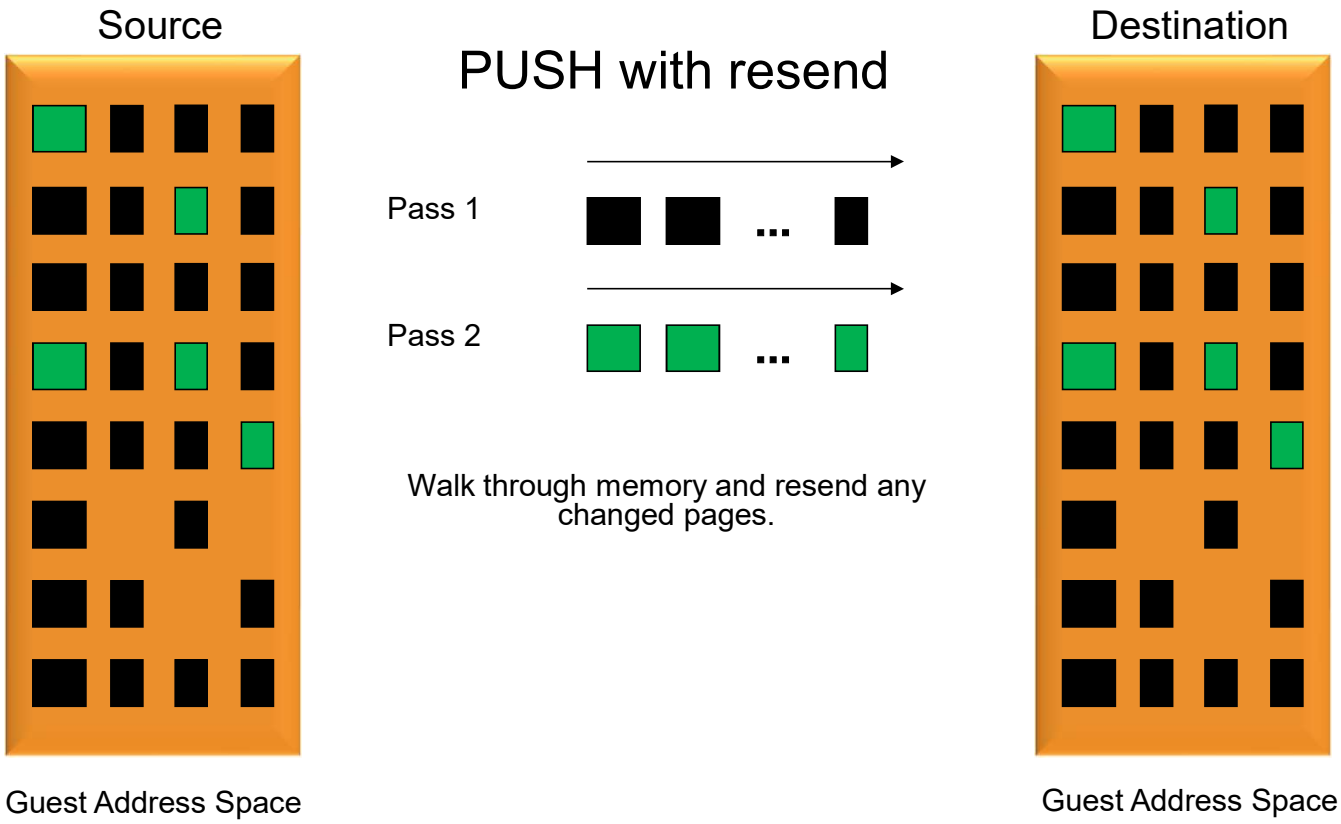
Stages of a Live Guest Relocation



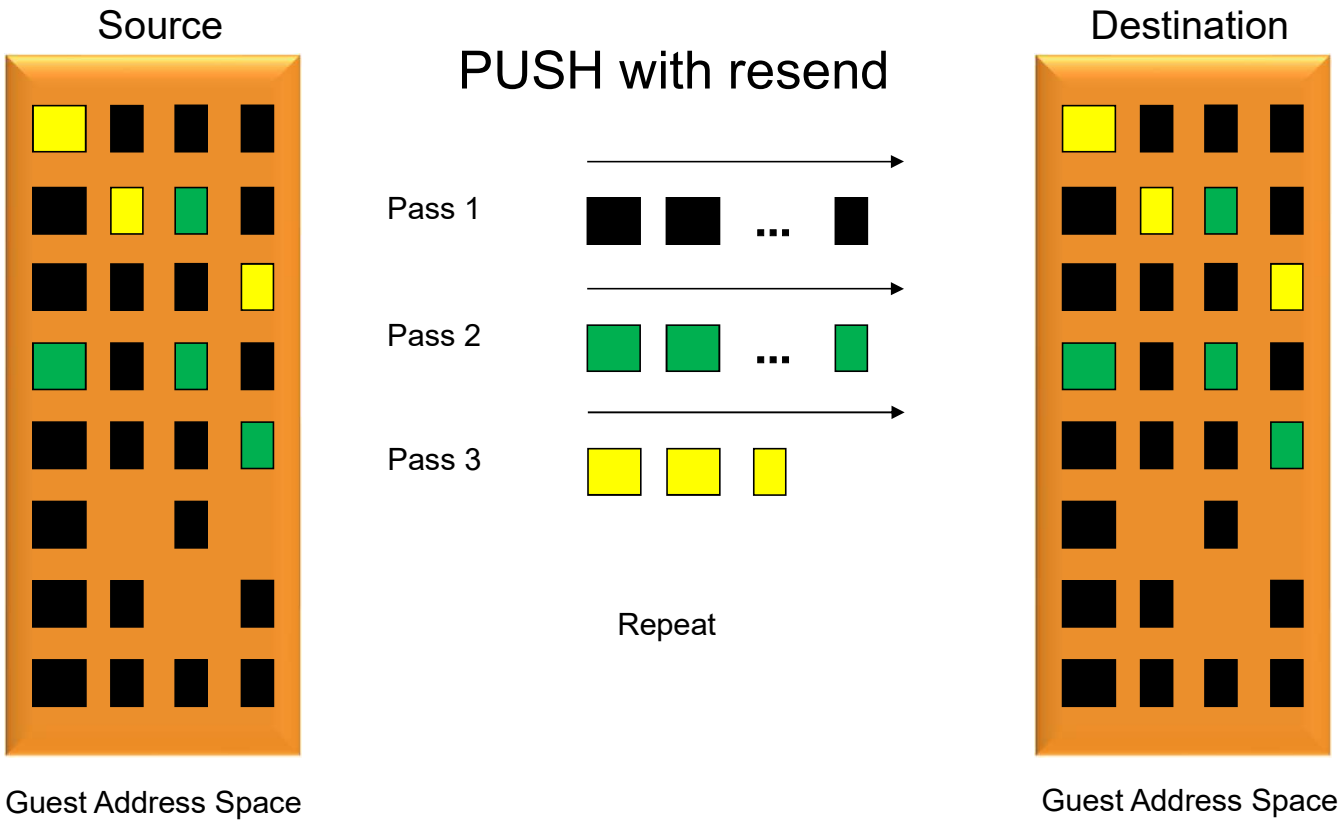
LGR, High-Level View of Memory Move



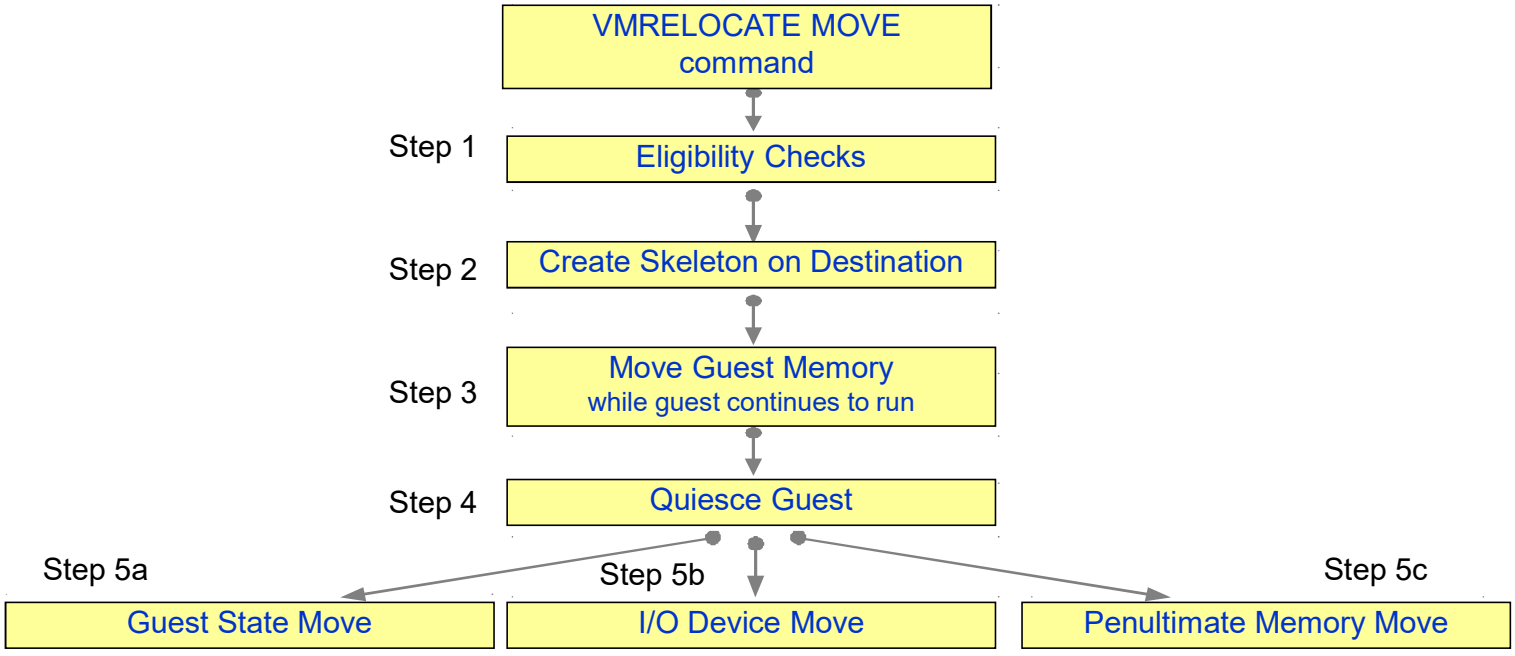
LGR, High-Level View of Memory Move



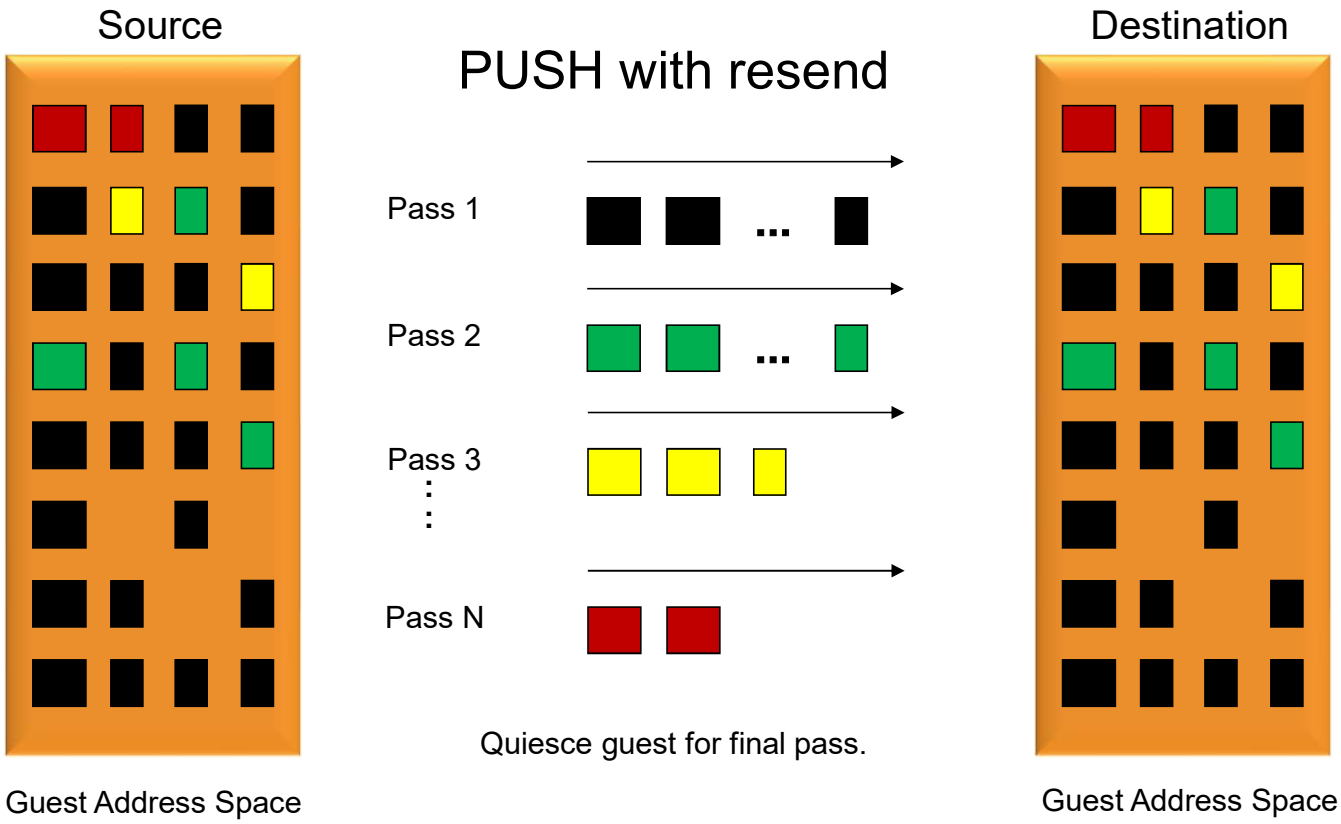
LGR, High-Level View of Memory Move



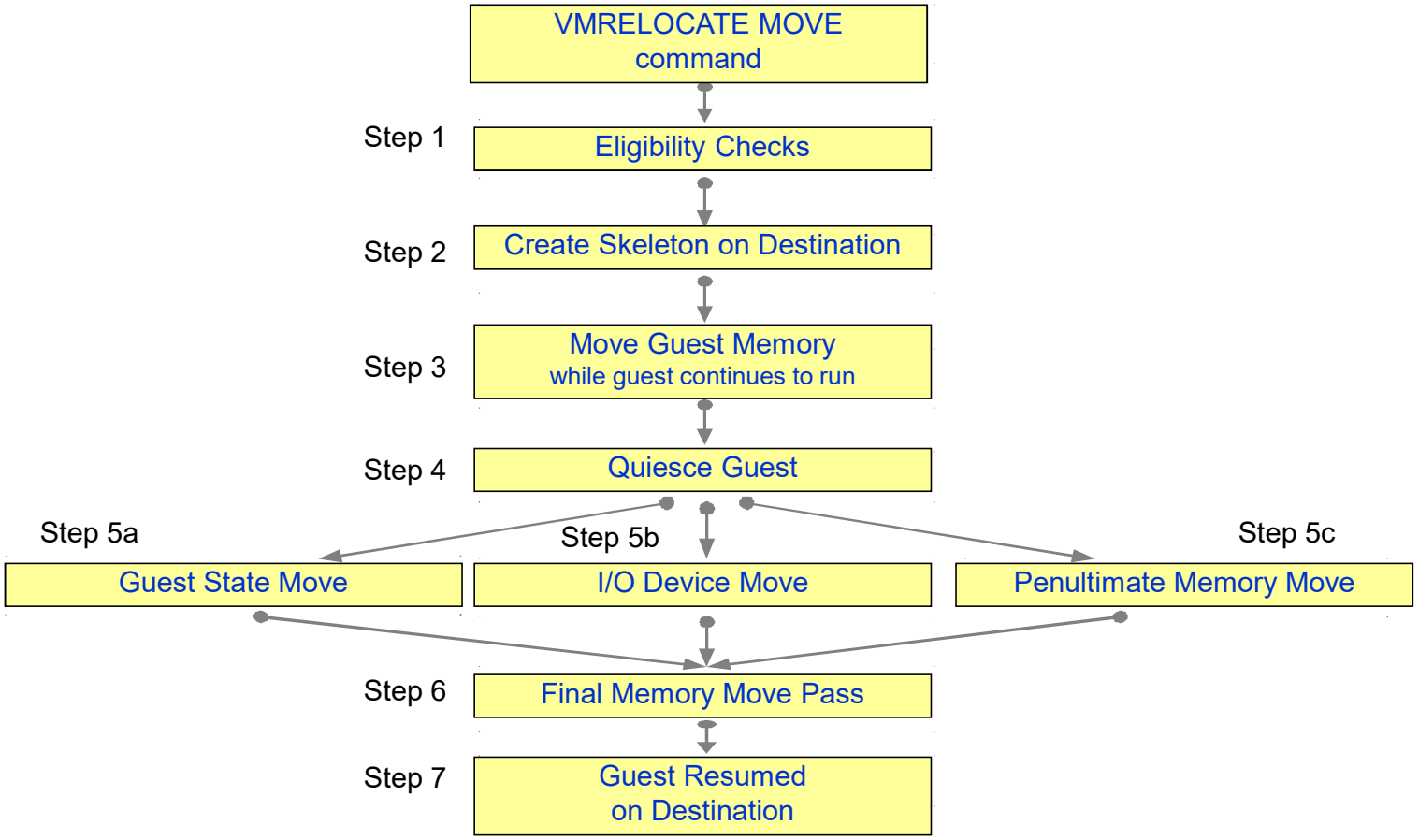
Stages of a Live Guest Relocation



LGR, High-Level View of Memory Move



Stages of a Live Guest Relocation



SSI Cluster Management: 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

- DirMaint
  - Main DirMaint virtual machine which can run on any of the members
  - Main DirMaint coordinates with satellite virtual machines on other members
  - A member that is down will be brought “up to speed” when re-started.

SSI Cluster Management: Addressing Problems

- Communications failure “locks down” future resource allocations until resolved
  - Existing running workloads continue to run
  - Prevents new accesses to resources
  - Cluster could temporarily be split and workloads continue to run

- Added the new “REPAIR” option to IPL for severe problem resolution
  - Meant for use with a single member cluster to repair
  - Allows correcting various problems that aren’t addressable in standard cluster.

Safe Guest Relocation

- Eligibility checks done multiple times throughout the relocation process.

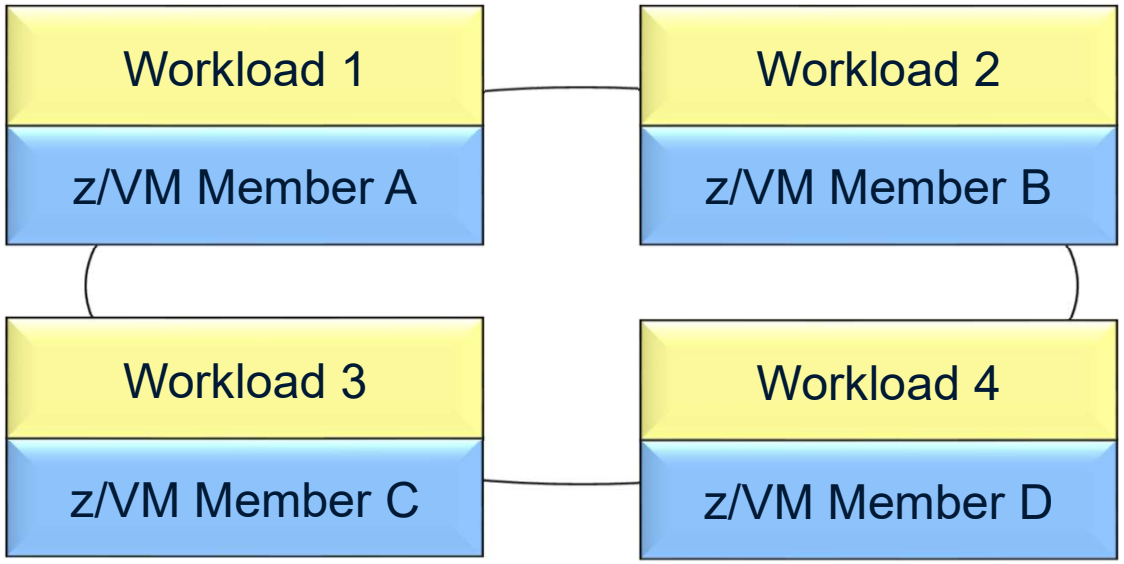
- Check more than just eligibility to move the virtual machine, but also check is it “safe” to move.
  - Overrides are available via *force* options

- Checks for:
  - Does virtual machine really have access to all the same resources and functions?
  - Will moving the virtual machine over commit resources to the point of jeopardizing other workload on the destination system?

- Pacing logic to minimize impact to other work in more memory constrained environments

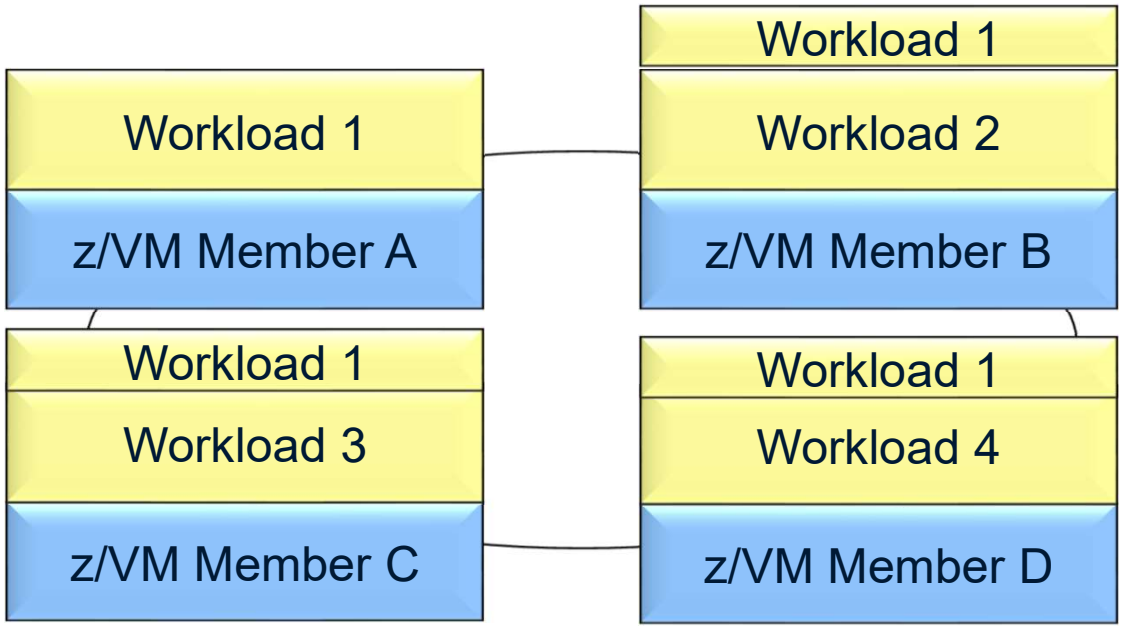
Flexibility for Planned Outages

- The good news is workload running on z/VM is becoming more and more critical; the bad news is that brings greater availability challenges.
  - Maintenance windows for down time get smaller
- SSI and LGR allow moving work and rolling out service...



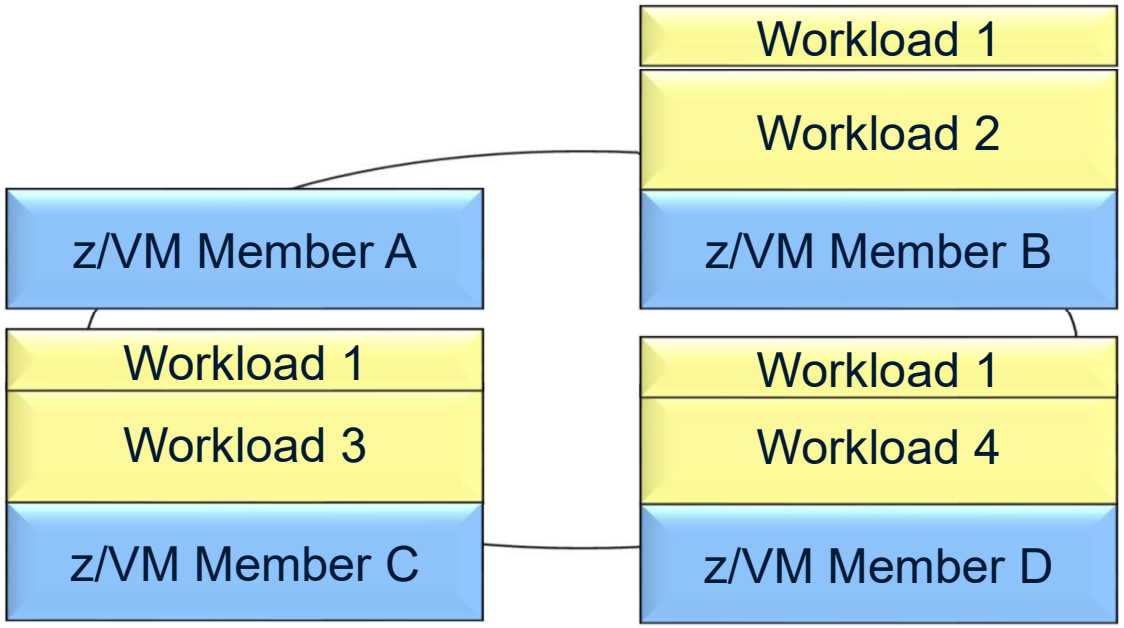
Flexibility for Planned Outages

- Move Workload 1 to other members in the SSI cluster



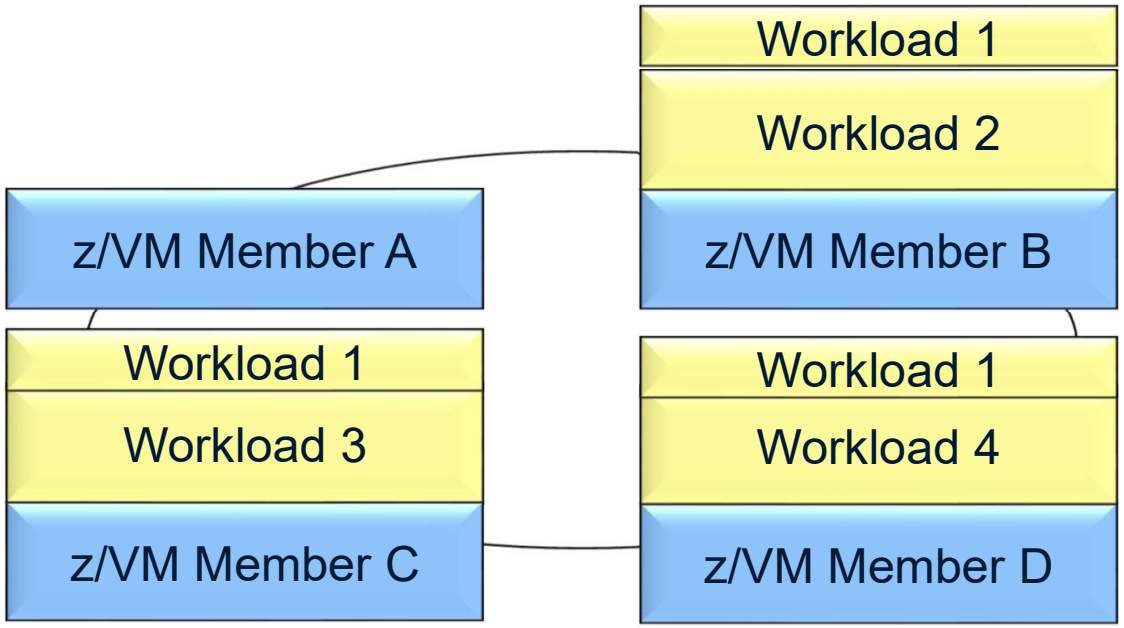
Flexibility for Planned Outages

- Shutdown Member A



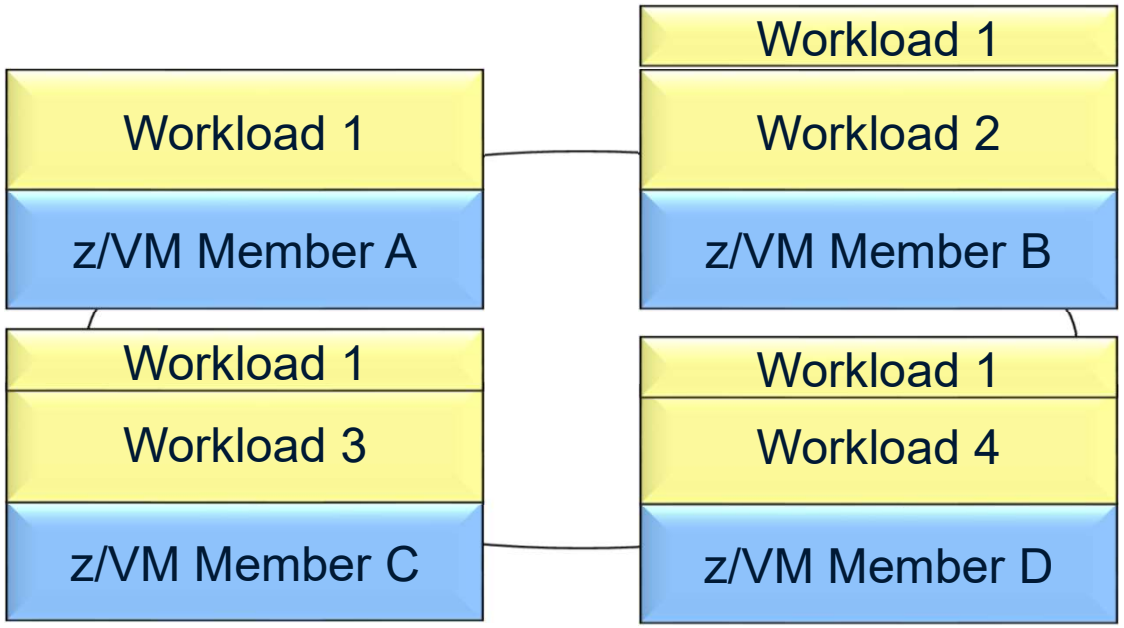
Flexibility for Planned Outages

- Bring up Member A with new service



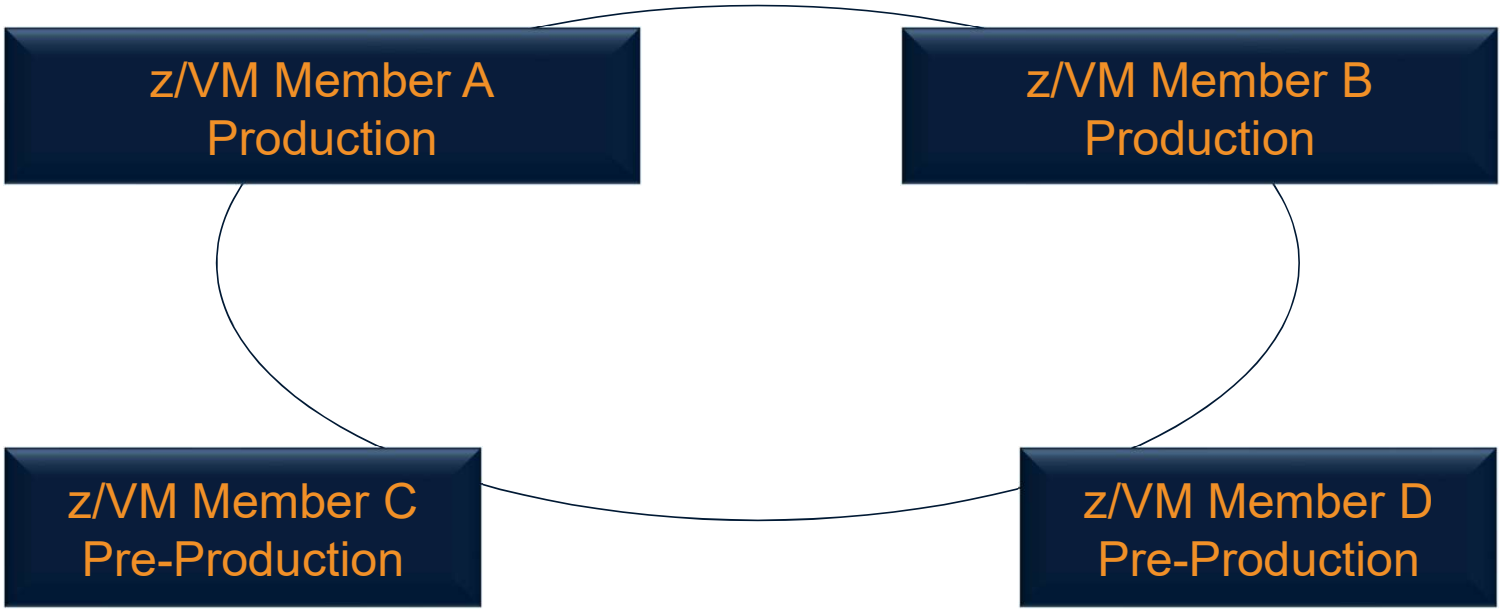
Flexibility for Planned Outages

- Move workload back to Member A



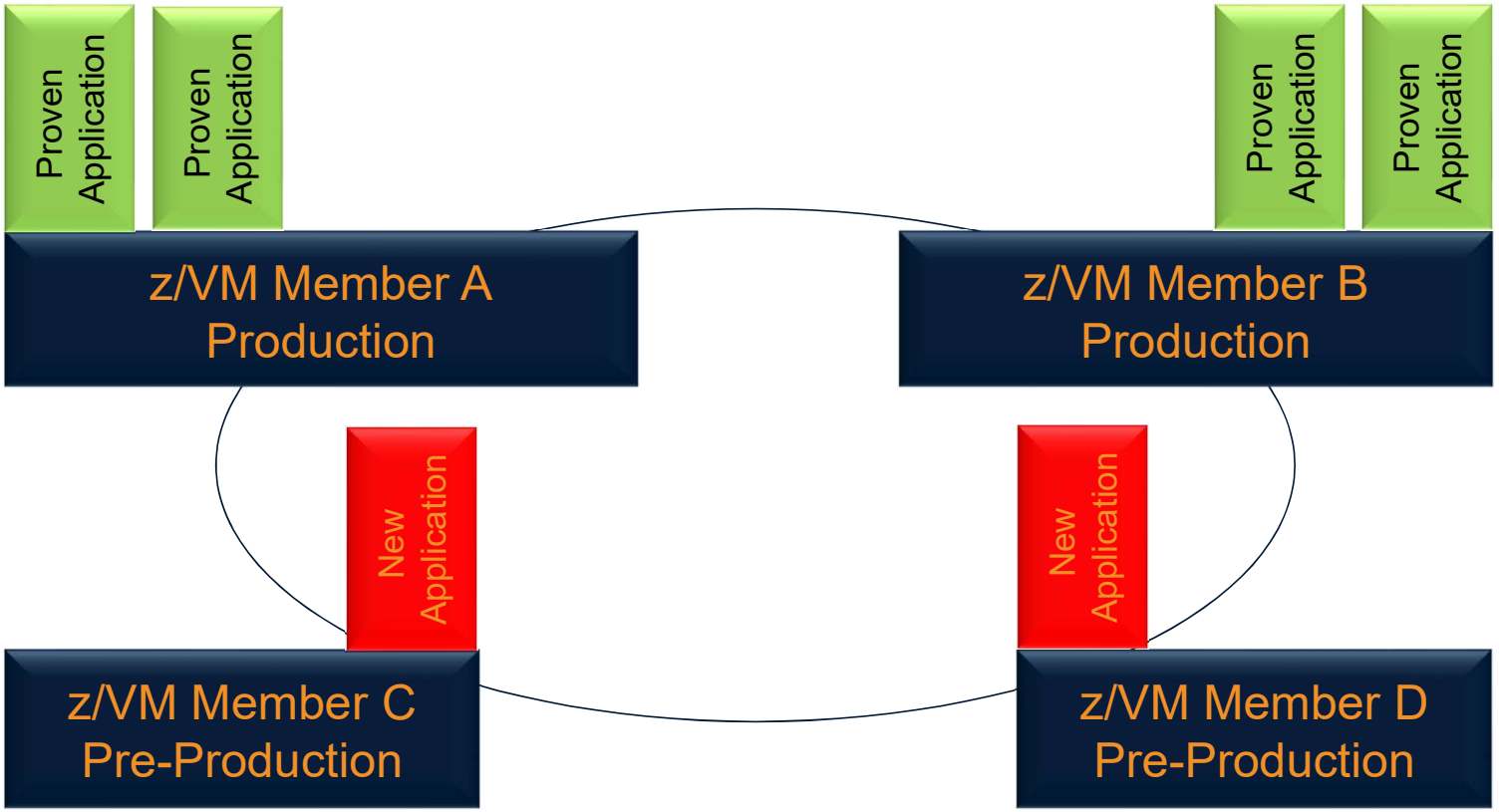
Production with Protection

- Four Members
  - True Production – two for redundancy
    - Full amount of resources.
  - Pre-Production: proving grounds
    - Limited resources.



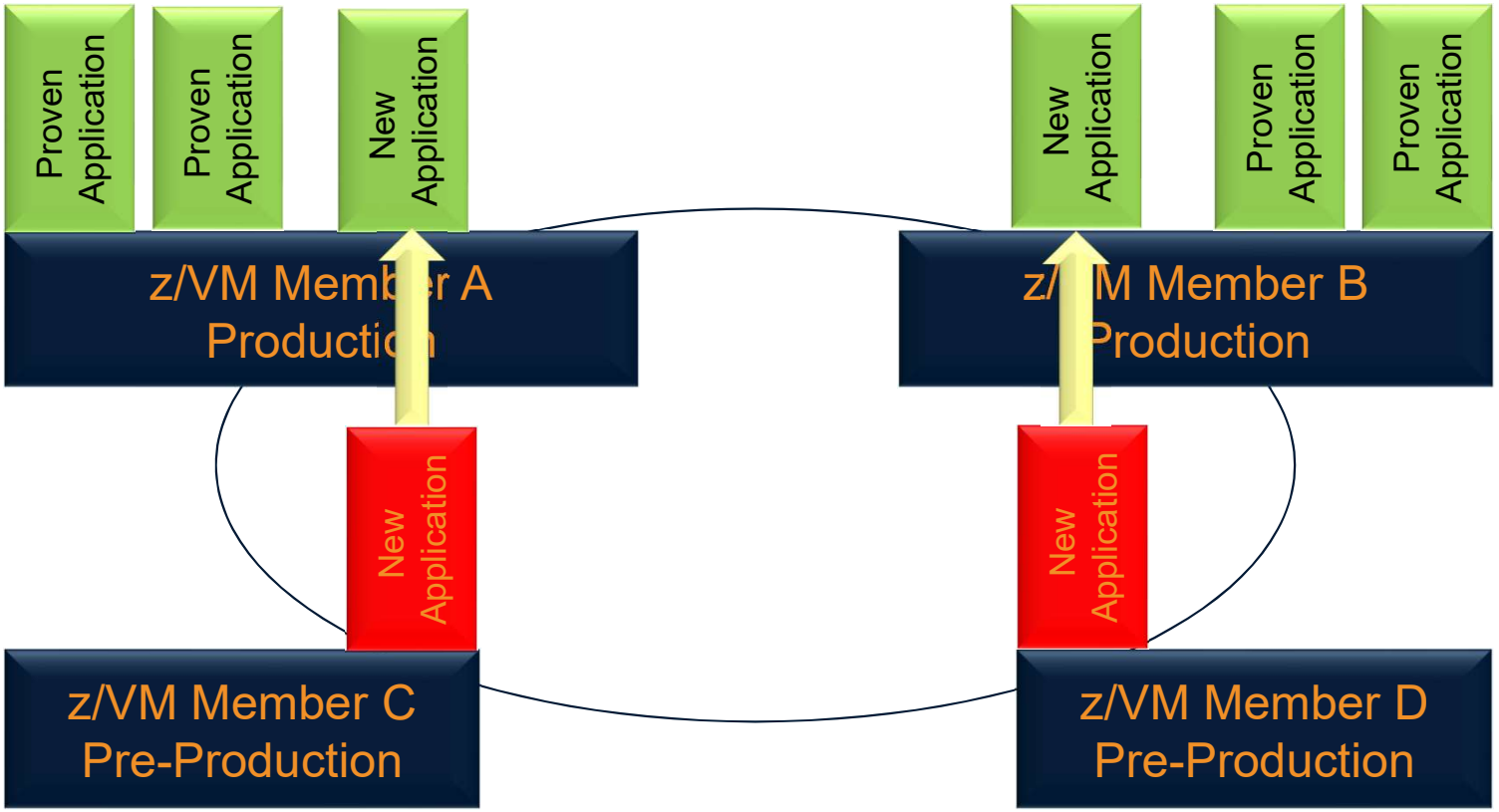
Production with Protection

- Allow new application to run in pre-production LPARs



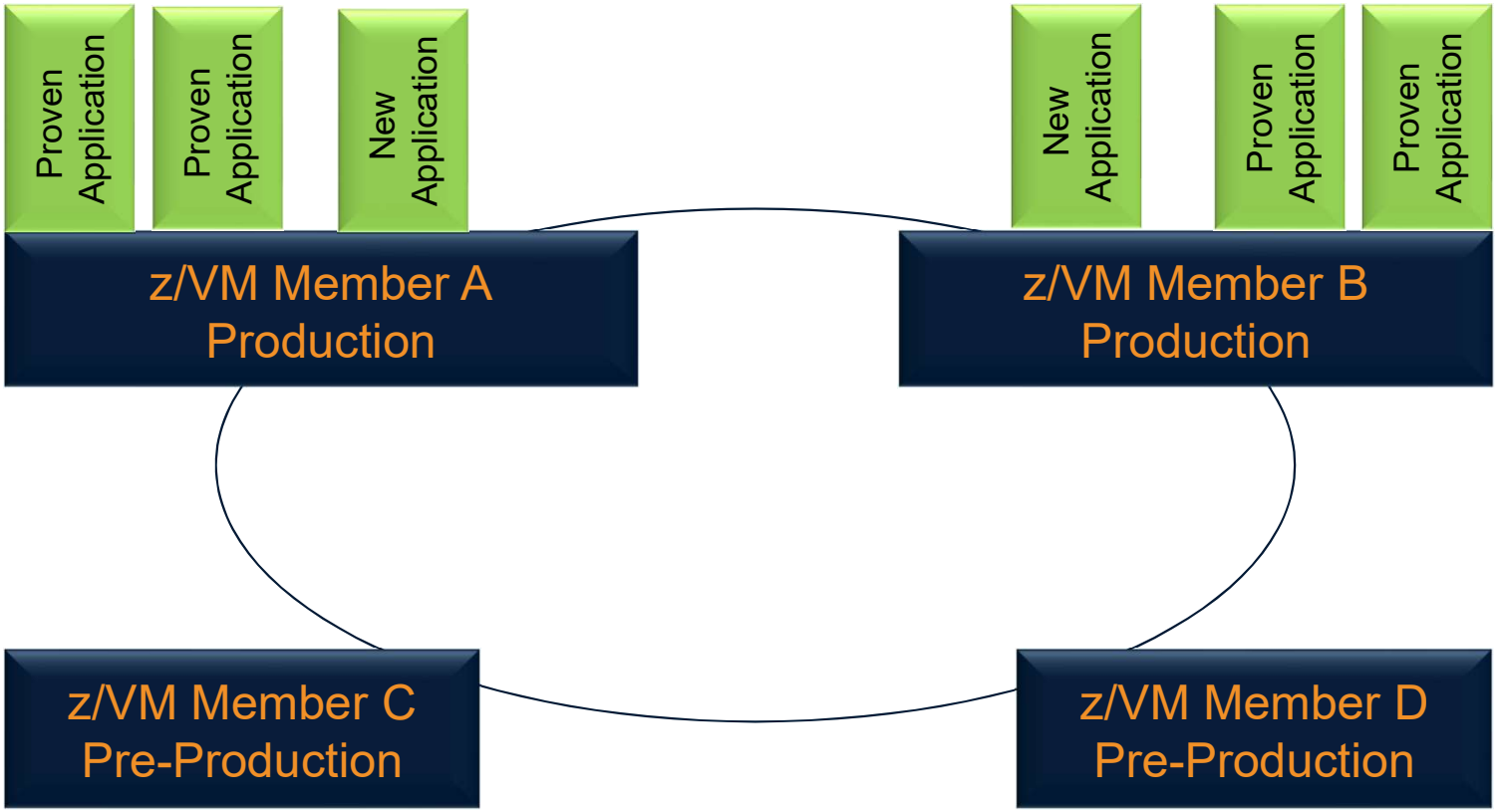
Production with Protection

- If all goes well, move into true production



Production with Protection

- If all goes well, move into true production



Relocation Domains

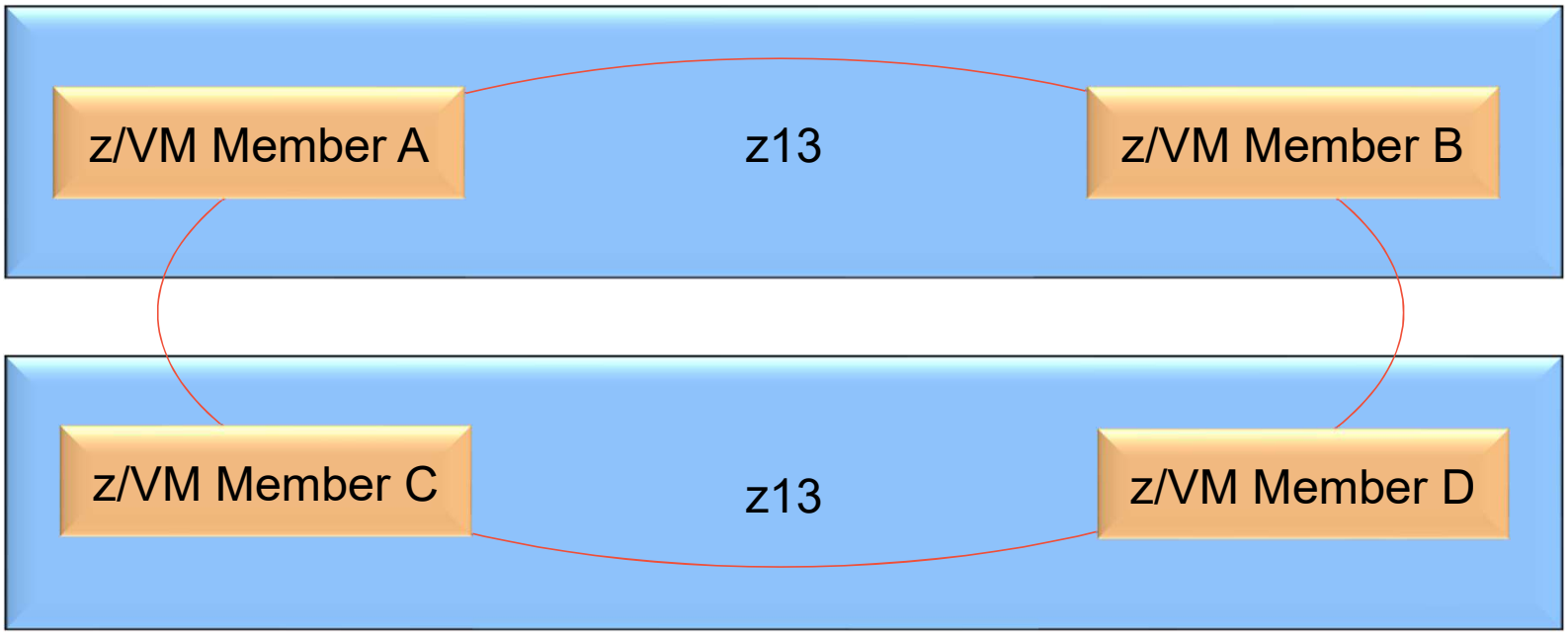
Why Relocation Domains?

- IBM Z and LinuxONE systems are highly flexible and dynamic

- You want to be able to move virtual machines around without requiring exact duplicates of:
  - Hardware features
  - Hardware facilities
  - Hardware configurations
  - Hypervisor software levels
  - Firmware levels

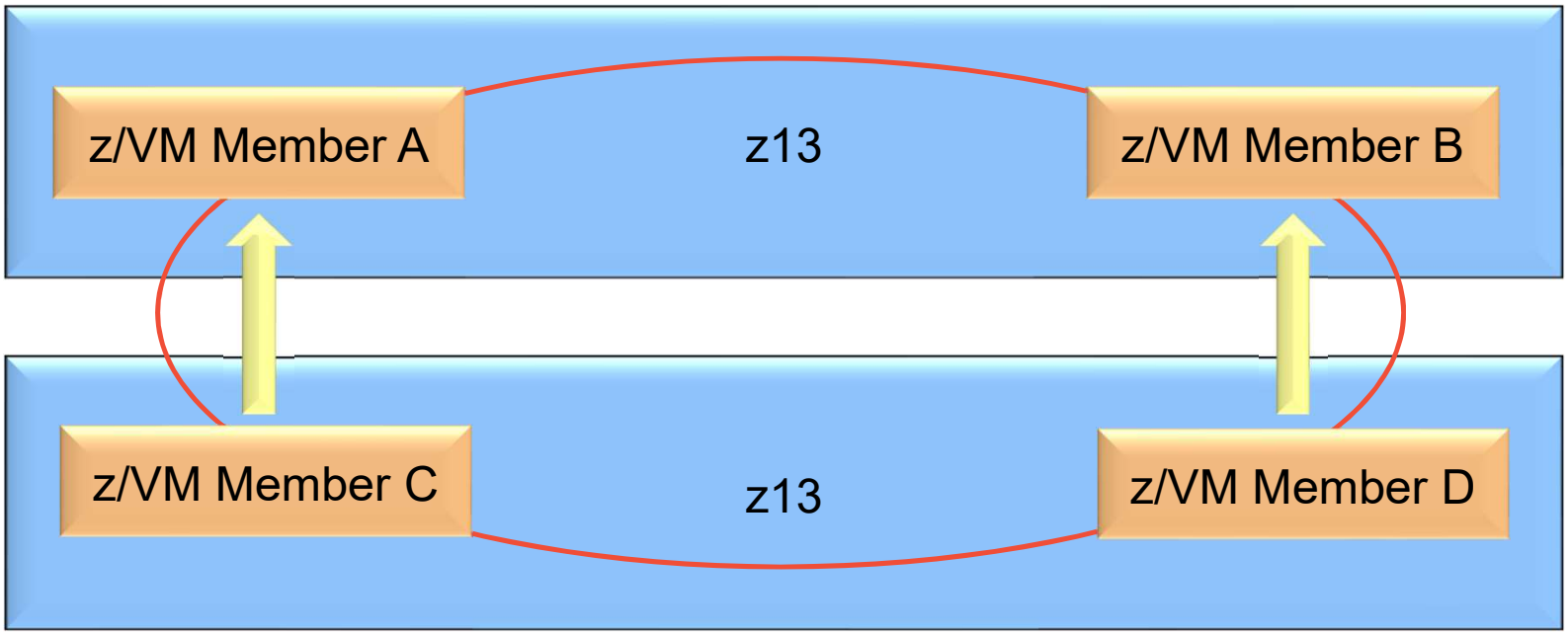
Migrate to New Processors

- Four members defined:
  - 2 members on each of 2 z13 servers



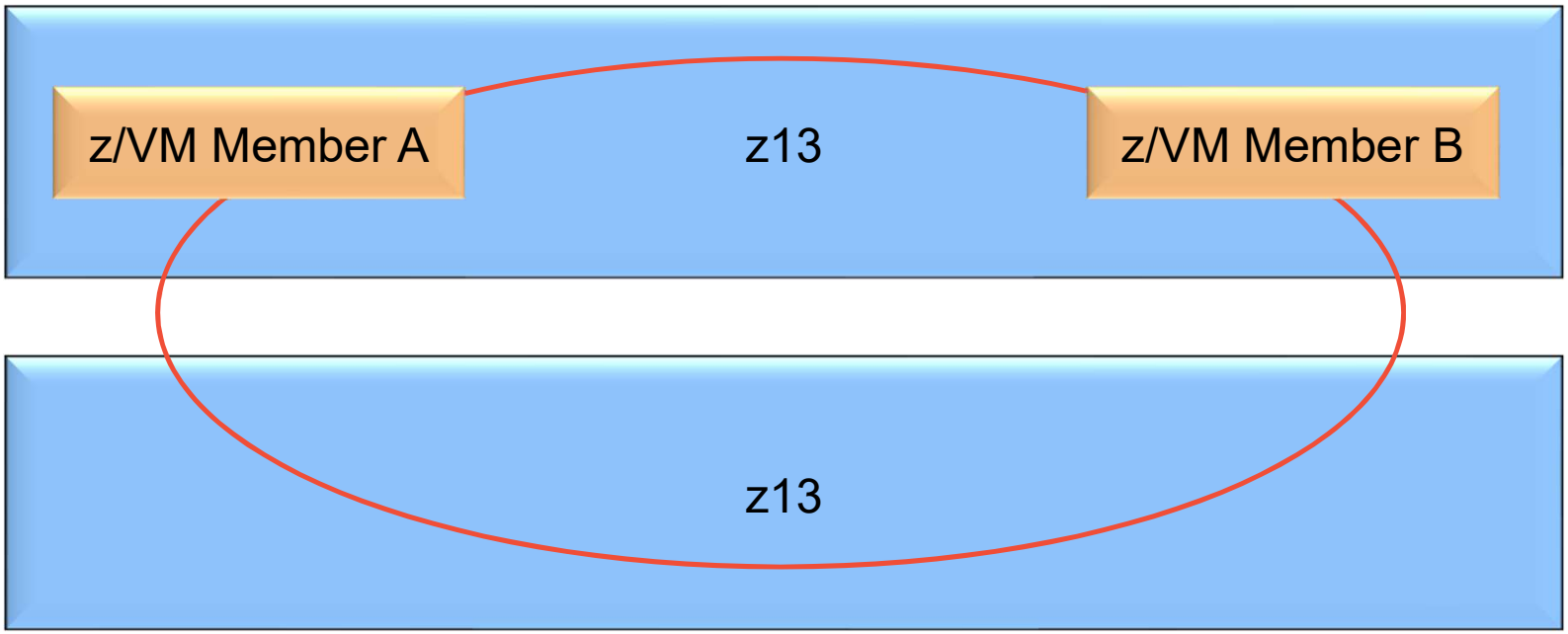
Migrate to New Processors

- Move work from members C & D (2nd z13) to members A & B (1st z13)



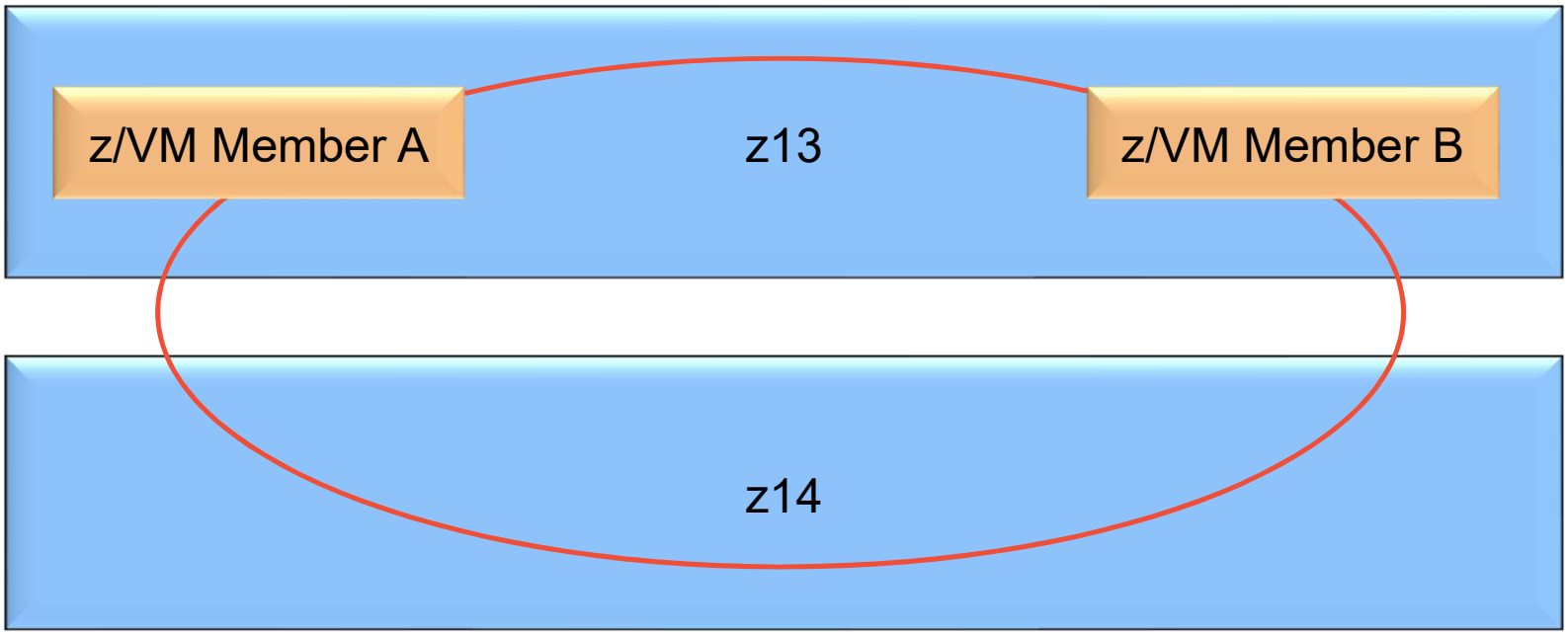
Migrate to New Processors

- Remove the z13



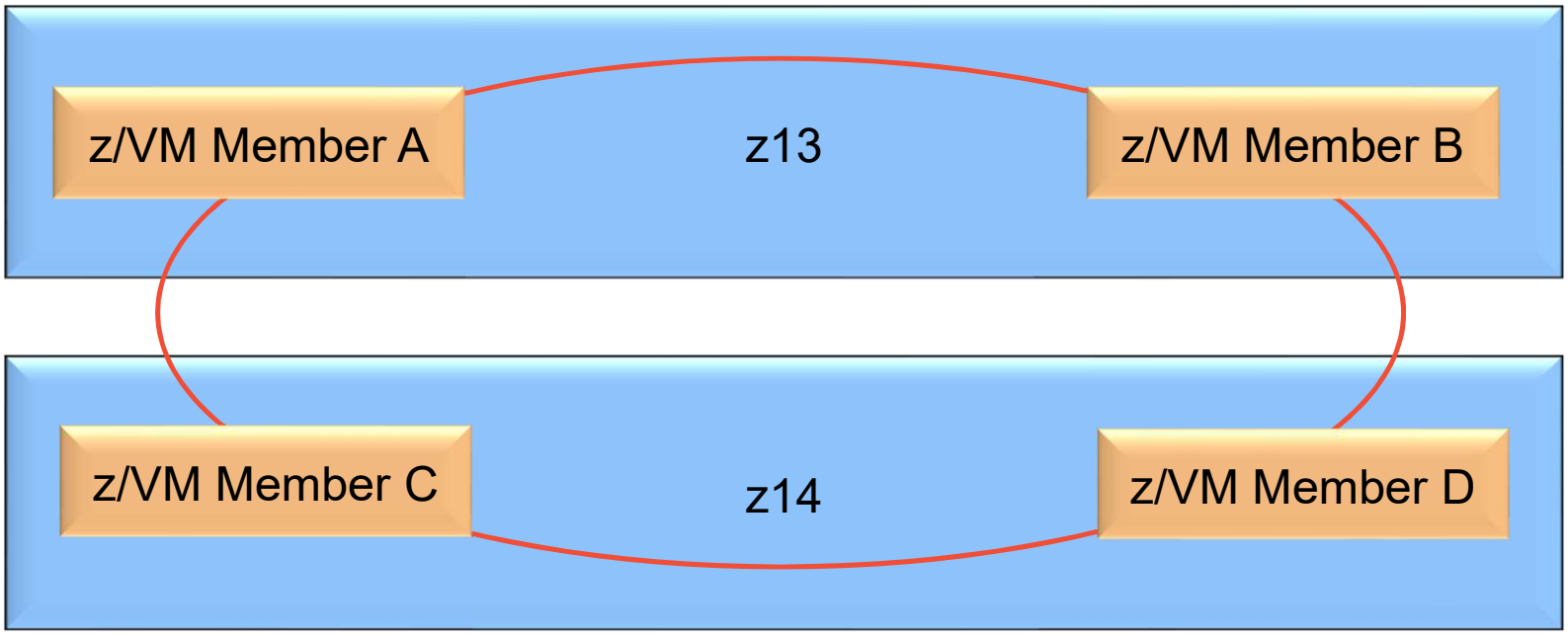
Migrate to New Processors

- Bring up new z14



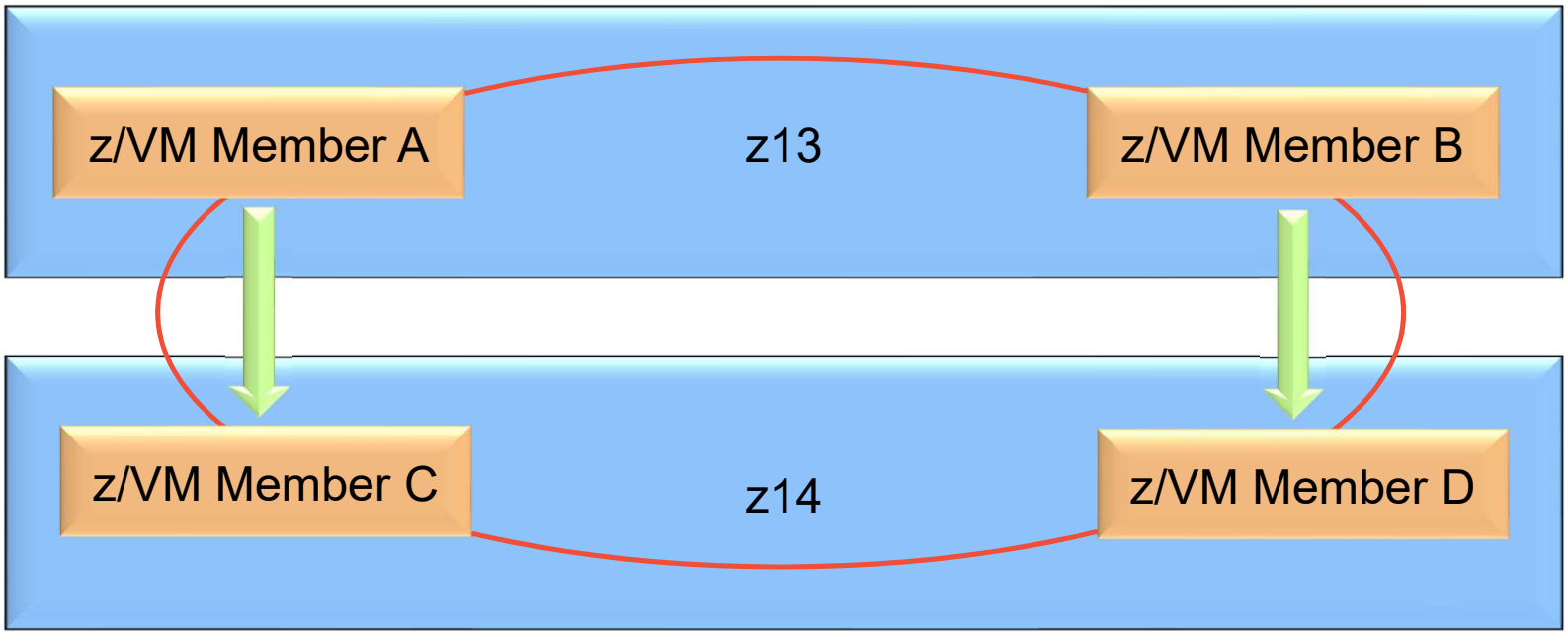
Migrate to New Processors

- Bring back up members C and D



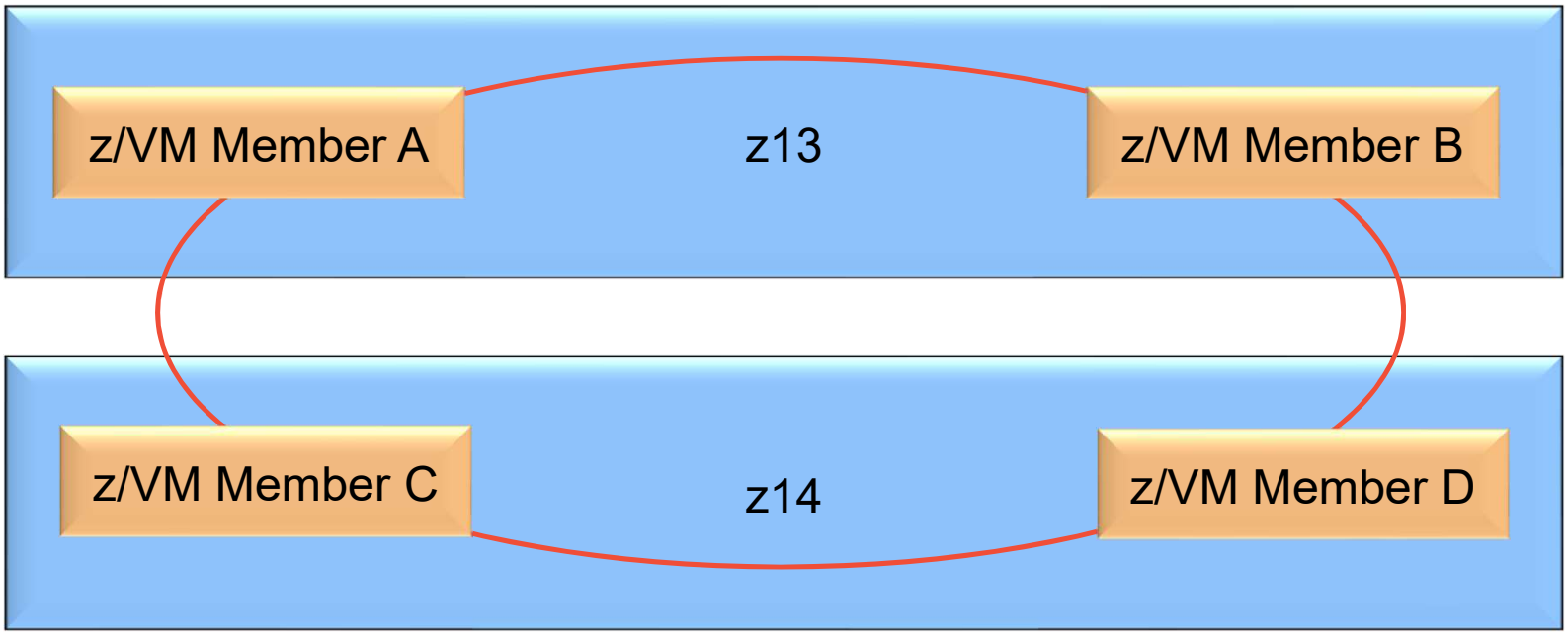
Migrate to New Processors

- Move workloads back to members C and D



Migrate to New Processors

- Hardware on 100s of Linux servers upgraded without an outage
- Duplicate the process for the other z13
- You do need to restart virtual machines to pick up new features from z14 after both machines are z14.



What is a Relocation Domain?

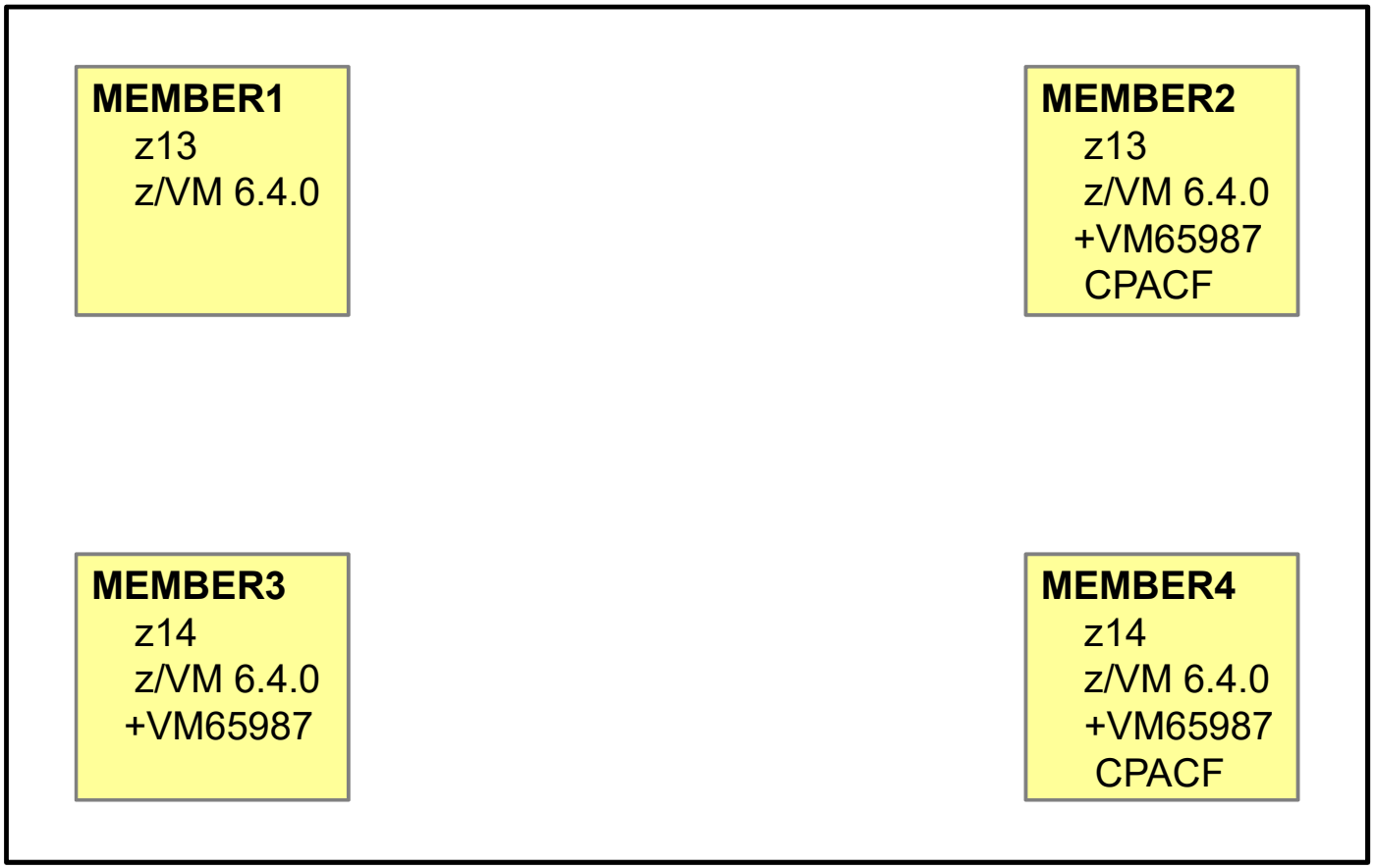
- A relocation domain defines a set of members of an SSI cluster among which virtual machines can relocate freely

- Regardless of differences in the facilities of the individual members, a domain has a common architectural level
  - This is the **maximal common subset** of all the members' facilities

- Several default domains are automatically defined by CP
  - Single member domains for each member in the SSI – domain name is member name
  - An SSI domain that will have the features and facilities common to all members, named “SSI”

- Defining your own domains is useful in a 3+ member cluster
  - In a 1 or 2 member cluster, all possible domains are defined by default
  - Defined via a SYSTEM CONFIG statement or dynamically by command

Relocation Domains – Default Domains



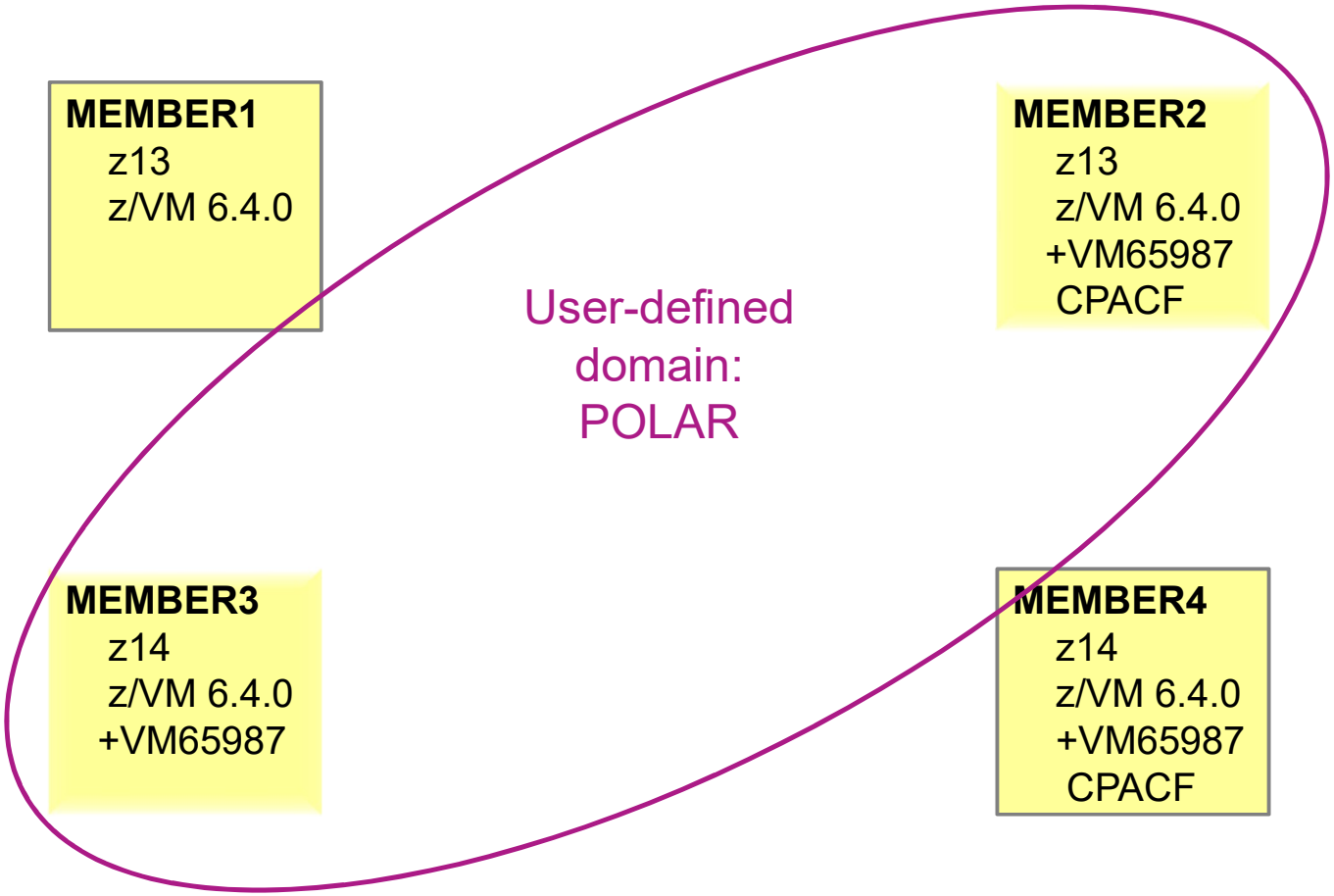
SSI Domain

z13 instruction set
z/VM 6.4.0

VM65987 – APAR for Guarded Storage Facility virtualization.

CPACF – CP Assist for Cryptographic Function – a no charge processor feature

Relocation Domains – User Defined MEMBER2 and MEMBER3



SSI Domain
z13 instruction set
z/VM 6.4.0

POLAR Domain
z13 instruction set
z/VM 6.4.0
+VM65987

Relocation Domains

MEMBER1
z13
z/VM 6.4.0

MEMBER3
z14
z/VM 6.4.0
+VM65987

MEMBER2
z13
z/VM 6.4.0
+VM65987
CPACF

User-defined domain: PANDA

MEMBER4
z14
z/VM 6.4.0
+VM65987
CPACF

SSI Domain (1,2,3,4)
z13 instruction set
z/VM 6.4.0

POLAR Domain (2,3)
z13 instruction set
z/VM 6.4.0
+VM65987

PANDA Domain (2,4)
z13 instruction set
z/VM 6.4.0
+VM65987
CPACF

Relocation Domains

MEMBER1
z13
z/VM 6.4.0

MEMBER2
z13
z/VM 6.4.0
+VM65987
CPACF

MEMBER3
z14
z/VM 6.4.0
+VM65987

User-defined domain: GRIZZLY

MEMBER4
z14
z/VM 6.4.0
+VM65987
CPACF

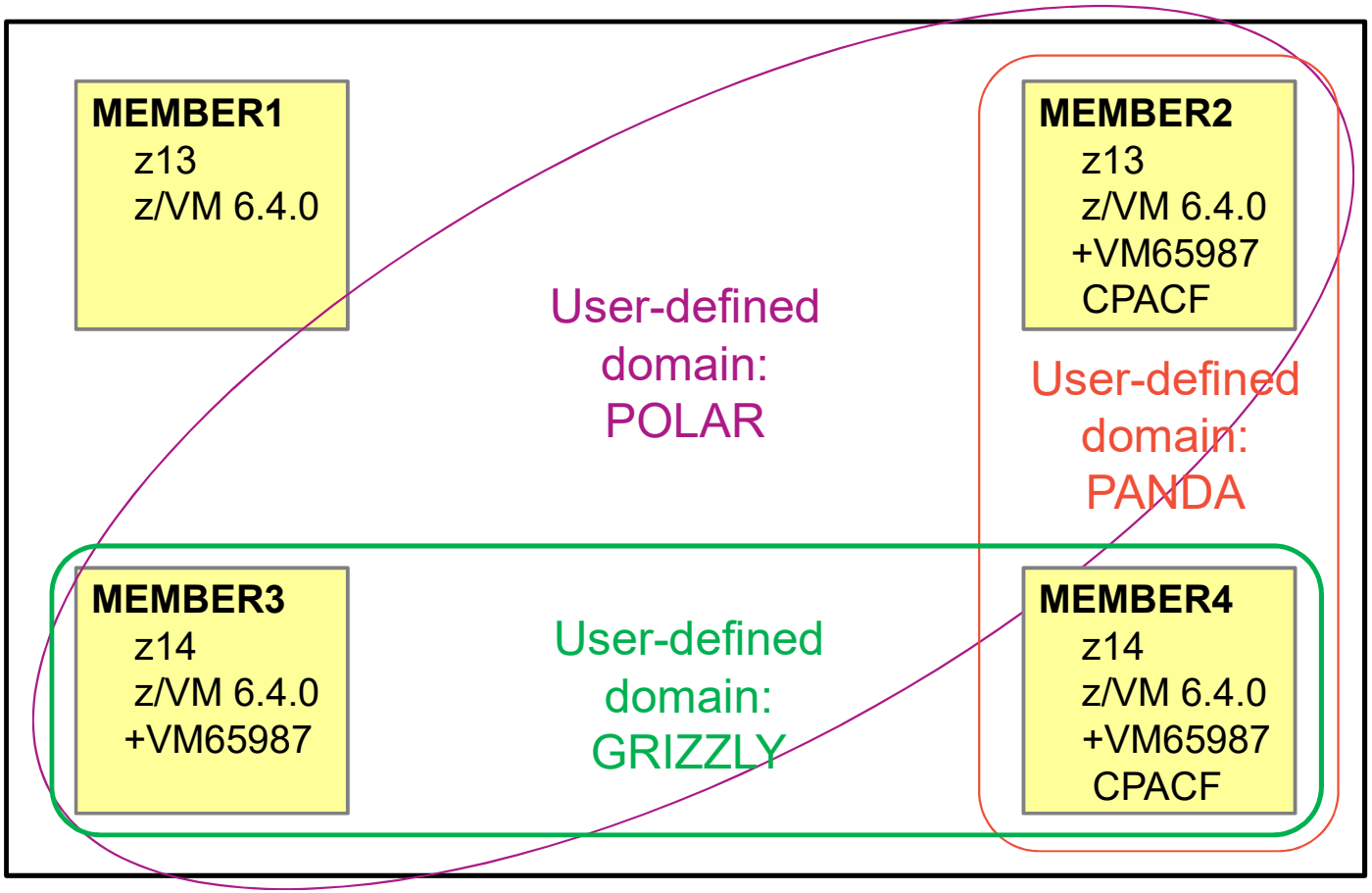
SSI Domain (1,2,3,4)
z13 instruction set
z/VM 6.4.0

POLAR Domain (2,3)
z13 instruction set
z/VM 6.4.0
+VM65987

PANDA Domain (2,4)
z13 instruction set
z/VM 6.4.0
+VM65987
CPACF

GRIZZLY Domain (3,4)
z14 instruction set
z/VM 6.4.0
+VM65987

Relocation Domains



SSI Domain (1,2,3,4)
- z13 instruction set
- z/VM 6.4.0

POLAR Domain (2,3)
- z13 instruction set
- z/VM 6.4.0
- +VM65987

PANDA Domain (2,4)
- z13 instruction set
- z/VM 6.4.0
- +VM65987
- CPACF

GRIZZLY Domain (3,4)
- z14 instruction set
- z/VM 6.4.0
- +VM65987

Defining Relocation Domains

- CP Command
  - **DEFINE RELODOMAIN** *domain_name member_names*
- System Configuration File
  - **RELOCATION_DOMAIN** *domain_name member_names*
- Note there is no way to delete a defined domain
- These identify the members that make up Relocation domains, but the commands alone do not determine the virtual architecture level associated with virtual machines

Assigning Virtual Machines to Relocation Domains

- All virtual machines have a relocation domain associated with them
  - Multiconfiguration virtual machines default to the member specific domain and cannot be changed
  - Single configuration virtual machines default to the SSI relocation domain
- **CP SET VMRELOcate USER *userid* ON DOMIAN *domain_name***
  - **SET VMRELOCATE** OFF without an explicit domain assignment will leave the virtual machine in the SSI relocation domain, even though it can not be relocated.
- Directory Entry

```
USER LGRRH56 E 2G 3G ABCDEFG
   INCLUDE LGRDFLT
   IPL 150
   VMRELOCATE ON DOMAIN WINNIE
   LINK PMAINT 0193 0F93 RR
   MDISK 0150 3390 1 END FL4BC8 MR ALL WRITE MULTI
   MDISK 0151 3390 1 END FL4BC9 MR ALL WRITE MULTI
   MDISK 0152 3390 1 END FL4BCA MR ALL WRITE MULTI
```

Architecture Fencing in Domains

- To allow virtual machines to move between members that have different facilities or features, they cannot use facilities that are not available on all the members in their relocation domain.

- *Fencing* is the process of preventing virtual machines from using facilities or features not included in their domain even if the SSI member they are on has access to those features

- Examples of commands/instructions with “fenced” responses:
  - **Q CPUID** -the model number will always reflect the virtual architecture level, the processor number is set at logon and not affected by relocation or relocation domain changes
  - **Diagnose x'00'** – will reflect the virtual CPLEVEL
  - **STFL –** Store Facility List
  - **STFLE –** Store Facility List Extended instruction
  - Certain other ‘query’ like instructions

- In **most** cases, there is minimal overhead due to IBM Hardware and z/VM Development collaboration.

What could influence the Architecture Description?

- Processor Facilities
  - Instructions (e.g. new Vector Decimal instructions added with z14 processor family)
    - Instructions are added as architected groups. While they are typically known as ‘new ops introduced with processor xyz, there is an architected way to determine their availability.
  - Architecture (e.g. Enhanced DAT 1 aka Large Page)
  - Whether these are enabled for the logical partition profile (e.g. I/O Priority Queueing)

- Processor Features
  - Examples:
    - Crypto Cards & model level
    - CPACF

- z/VM Level – Both release and service
  - New z/VM capabilities
  - Virtualization of new hardware facilities and features

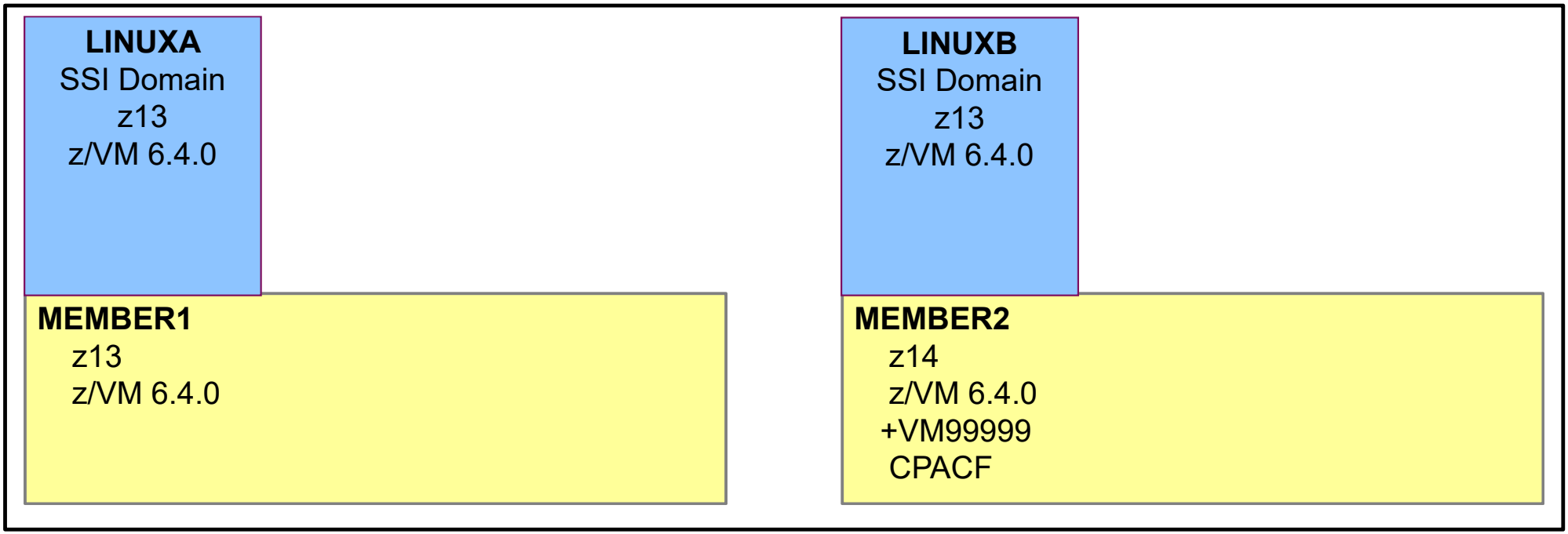
Relocation Eligibility

- Relocation domains allow management of which members a guest can freely relocate between and the virtual architecture level presented to them.
- Additionally, virtual machines can be ineligible for live guest relocation for other reasons.
  - See Chapter 29 of the CP Planning and Administration Book for more details.
  - Guest requirement examples:
    - One with all virtual CP or all virtual IFL processors
    - IPLed from a device or an NSS that meets requirements
    - Etc.
  - Guest ineligibility examples:
    - Virtual machine is an XC mode virtual machine
    - Virtual machine has a temporary disk attached
    - Virtual machine has an open spool file other than a console file
    - Virtual machine has used diagnose x'214' to establish a pending page release
    - Virtual machine is using a DCSS that does not exist on the destination system

Dynamic Nature of Virtual Architecture Levels

- Many of the capabilities that make up an architecture are dynamic.
- Members, and their capabilities, can shutdown and restart.
- z/VM relocation processing has to be ready to handle that.

Virtual Architecture Levels

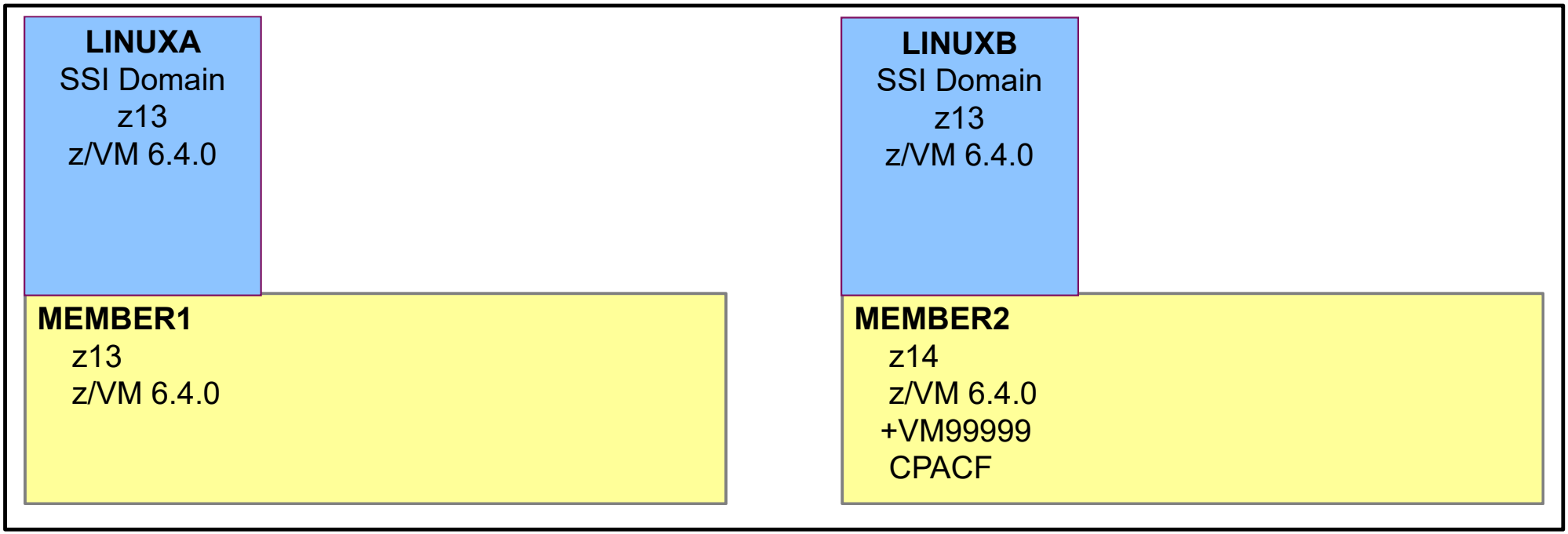


SSI Domain
- z13 instruction set
- z/VM 6.4.0

VM99999 – fictitious service that adds a feature to z/VM that guests could use.

CPACF – CP Assist for Cryptographic Function – a no charge processor feature

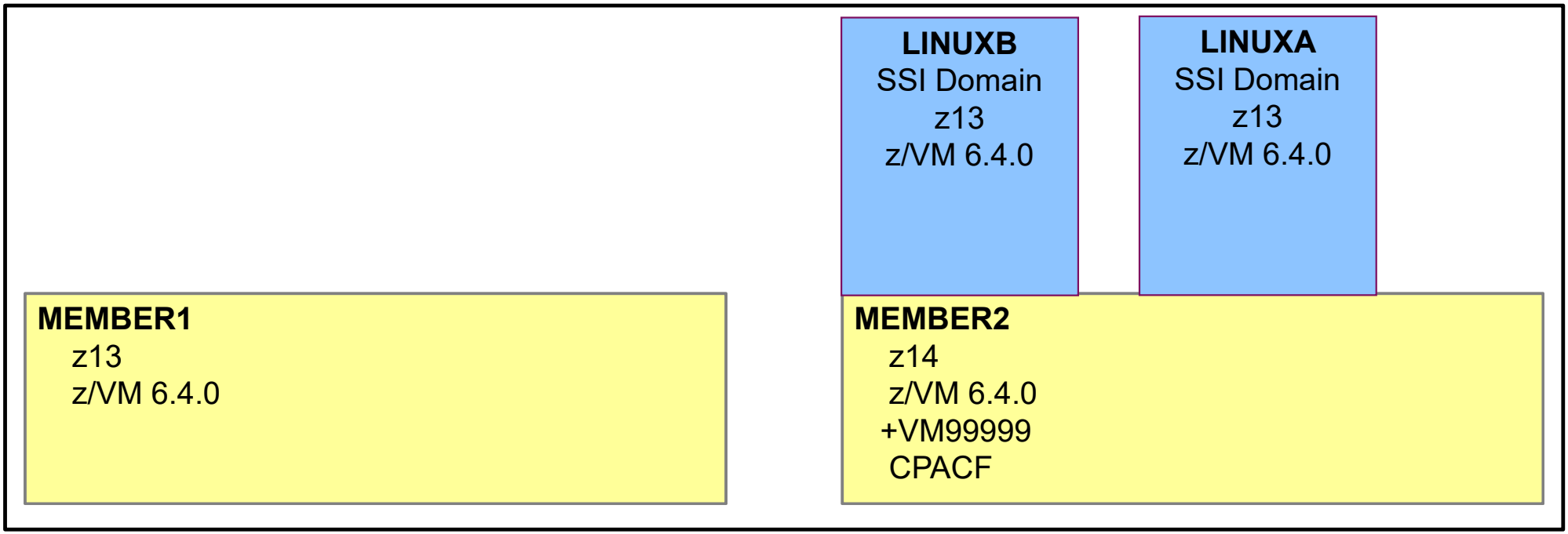
Virtual Architecture Levels



SSI Domain
- z13 instruction set
- z/VM 6.4.0

Relocate LINUXA from Member1 to Member2

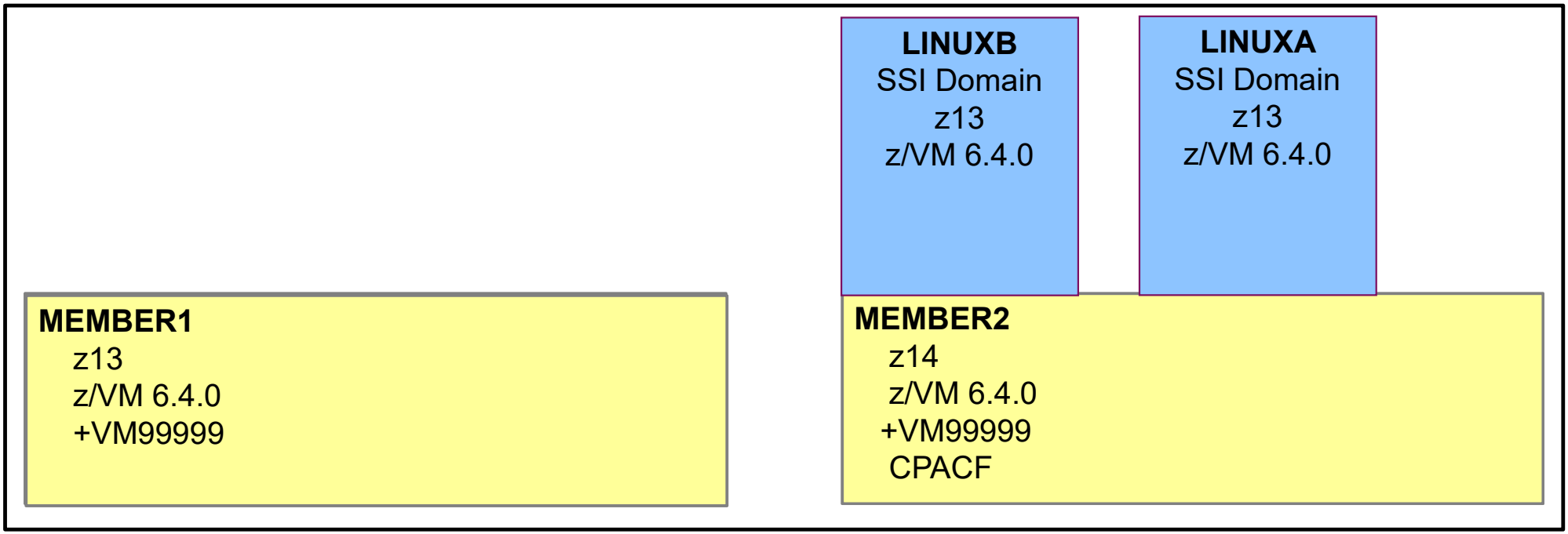
Virtual Architecture Levels



SSI Domain
z13 instruction set
z/VM 6.4.0

Shutdown MEMBER1, and bring back up with VM99999 applied

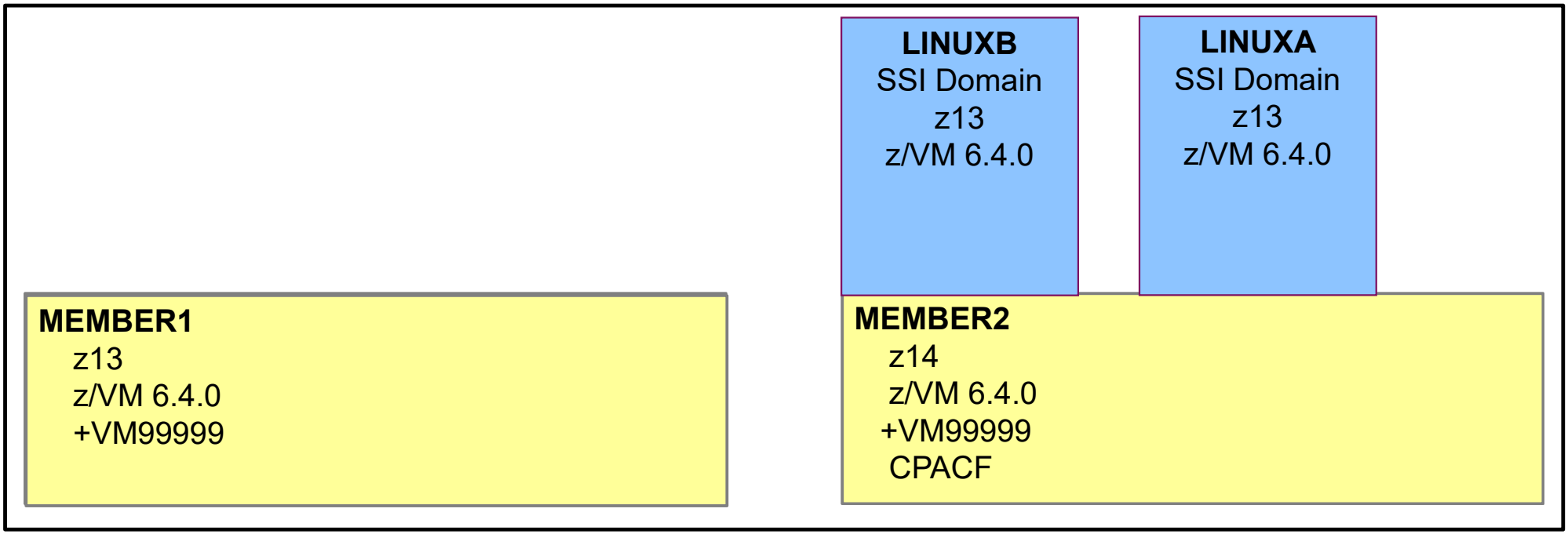
Virtual Architecture Levels



SSI Domain (VAL2)
z13 instruction set
z/VM 6.4.0
+VM99999

Shutdown MEMBER1, and bring back up with VM99999 applied

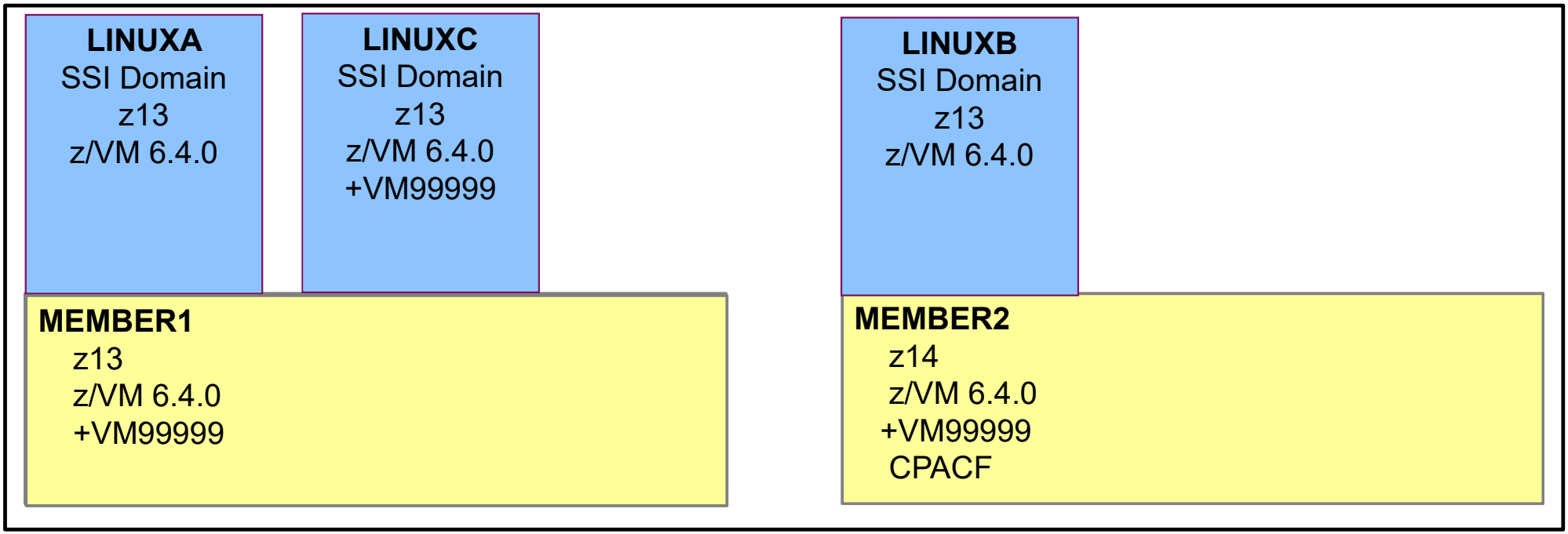
Virtual Architecture Levels



SSI Domain (VAL2)
z13 instruction set
z/VM 6.4.0
+VM99999

Relocate LINUXA back to MEMBER1. It keeps it's Virtual Architecture Level

Virtual Architecture Levels



SSI Domain (VAL2)
z13 instruction set
z/VM 6.4.0
+VM99999

Log on and start LINUXC in the SSI Domain. It gets new feature (VM99999).

Relocation Domain – Life Cycle

- Relocation Domain Information is included on the Persistent Data Record (PDR) section on disk.
- It survives members being down.
- If you are copying the PDR volume in some way, perhaps a DR solution, that needs to be considered as well.
- Using **CLEARPDR** option at IPL, not only clears out member status from the PDR but also relocation architecture information.

Other Uses of Relocation Domains

- Separate work based on Security
  - Domain ENCRYPT on members where z/VM encrypted paging is configured so that virtual machines cannot be relocated to a member without encrypted paging

- Separate for Performance
  - Domains for SMT1 and SMT2 keeping workloads that may run better in SMT1 from being relocated onto an SMT-2 system

- To protect redundancy
  - Members 1,2, 3, 4
  - Have domains ODD and EVEN
  - Virtual machines paired for redundancy LINUX01 (ODD) and LINUX02 (EVEN)

Problem Determination

- Useful Commands
  - **CP QUERY RELODOMAIN** – gives a list of the defined domains and which members are in the domain
  - **CP QUERY VMRELOCATE** – indicates which domain the virtual machine is in
  - **CP VMRELOCATE** with **TEST** option

- **RELODOM Package on z/VM download page**
  - http://www.vm.ibm.com/Download/packages/descript.cgi?RELODOM
  - Provides information typically for IBM Level 2 for analysis of:
    - Defined relocation domains
    - The virtual architecture level variations being used in the z/VM SSI Cluster
    - Virtual machines associated with each virtual architecture level variation
    - Output mostly meant for IBM analysis
  - Recommend run and keep data if you're doing LGR for first time or first time in a new/unique environment

Summary

Summary

✓Relocation domains add flexibility for SSI clusters

✓The virtual architecture level is the maximal common subset of facilities/features of the members of the relocation domain

✓Virtual Architecture Levels are dependent on the relocation domain a virtual machine is assigned

✓Setting relocation off does not remove the virtual machine from a relocation domain. It will continue to be in its relocation domain or the SSI domain by default

✓There can be multiple virtual architecture levels for a given relocation domain