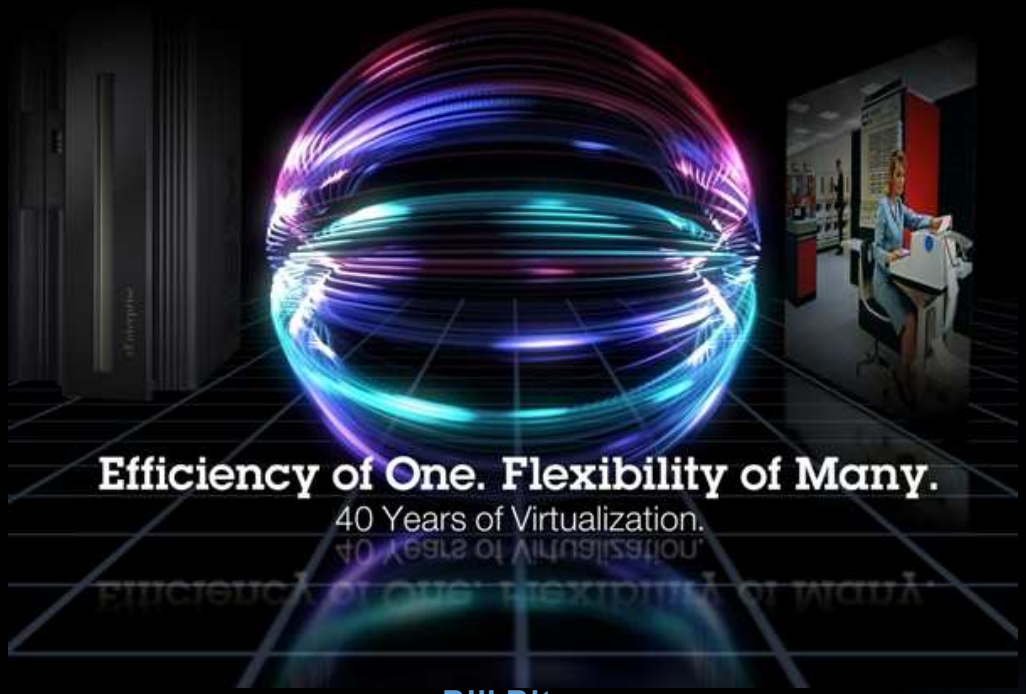
z/VM 6.2: Increasing the Endless Possibilities of Virtualization



Bill Bitner

IBM Endicott – The Birthplace of IBM

z/VM Customer Focus and Care Leader
bitnerb@*us.ibm.com*

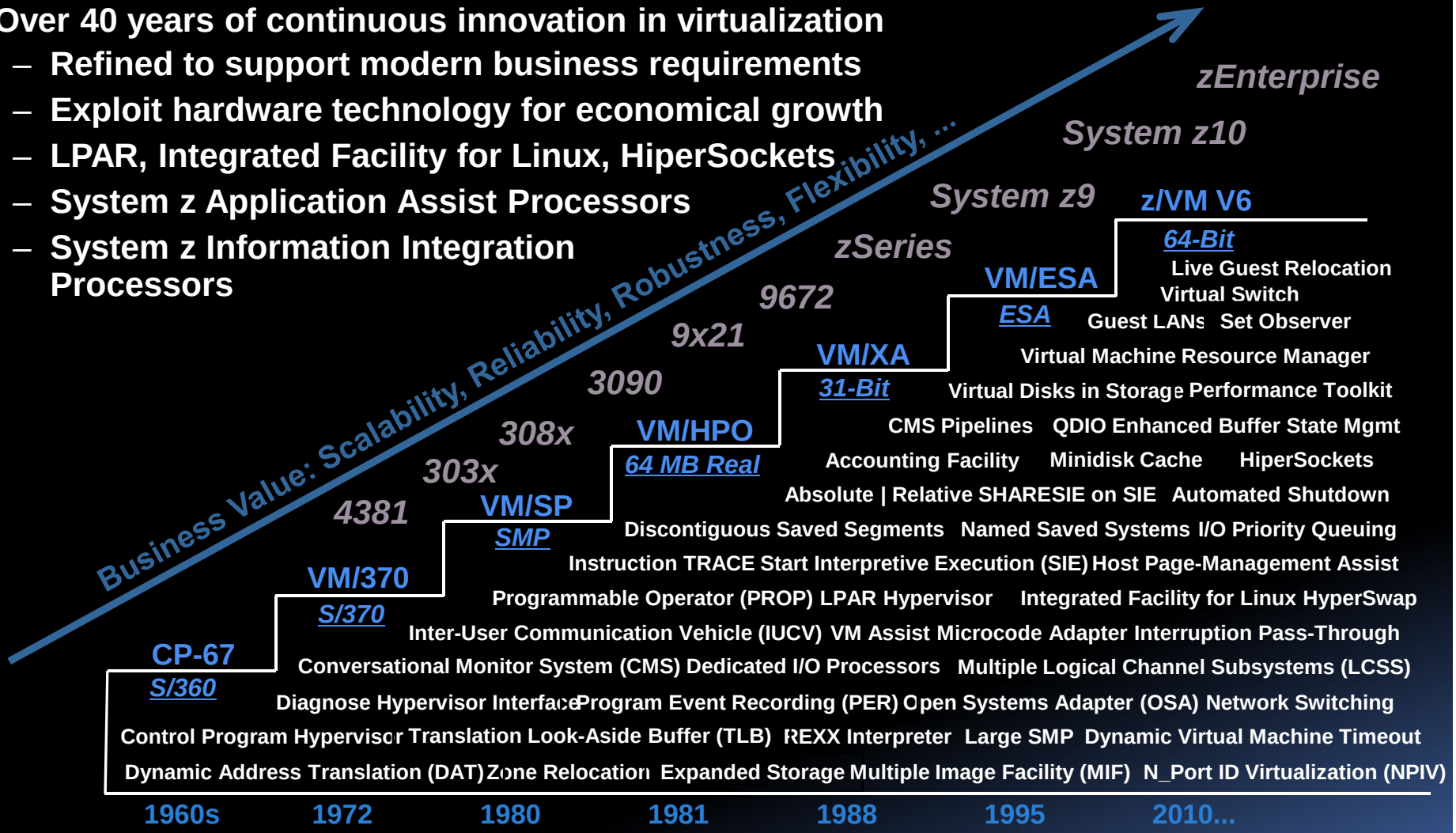
Agenda

z/VM Timeline	• Heritage and History • Constantly Expanding Function
Reflecting on the z/VM 6.2 Design	• Challenges • Speeds and Feeds
What Makes Live Guest Relocation Special?	• Making it safe • Making it manageable
New Possibilities	• Availability • Flexibility in Testing

IBM System z Virtualization Genetics

Over 40 years of continuous innovation in virtualization

- Refined to support modern business requirements
- Exploit hardware technology for economical growth
- LPAR, Integrated Facility for Linux, HiperSockets
- System z Application Assist Processors
- System z Information Integration Processors



IBM System z – a comprehensive and sophisticated suite of virtualization function

Reflecting on the z/VM 6.2 Design

z/VM Version 6 Release 2

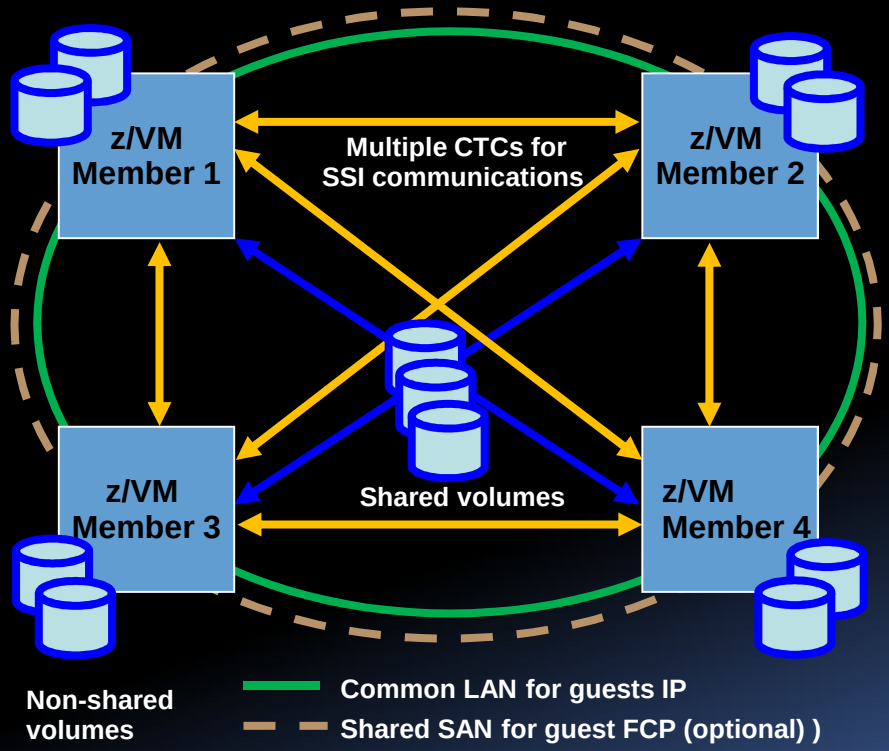
- Highlights
  - Announced **October 12, 2011**
  - Made generally available on **December 2, 2011**
  - Planned end of service is **April 30, 2015**

- New Priced Feature VM Single System Image:
  - Single System Image (SSI) Clustering
  - Live Guest Relocation (LGR)

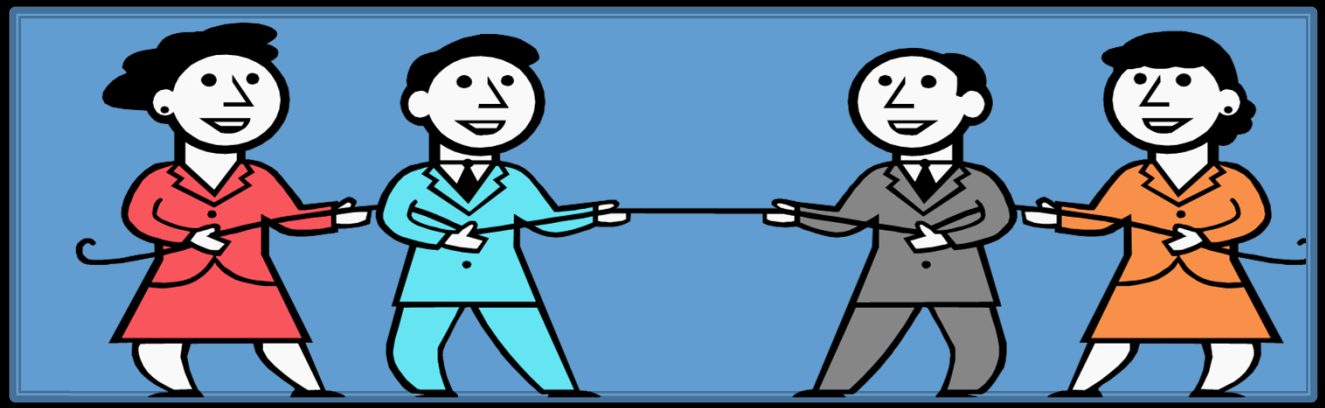
- References
  - z/VM Home Page: www.ibm.com/vm/
  - z/VM 6.2 Info: www.ibm.com/vm/zvm620/
  - z/VM SSI Info: www.ibm.com/vm/ssi/

SSI Feature: Clustered Hypervisor with LGR Support

- Connect up to four z/VM systems as members of a Single System Image (SSI) cluster
- Provides a set of shared resources for member systems and their hosted virtual machines
- Cluster members can be run on the same or different System z servers
- Simplifies systems 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



Key Early Design Struggle



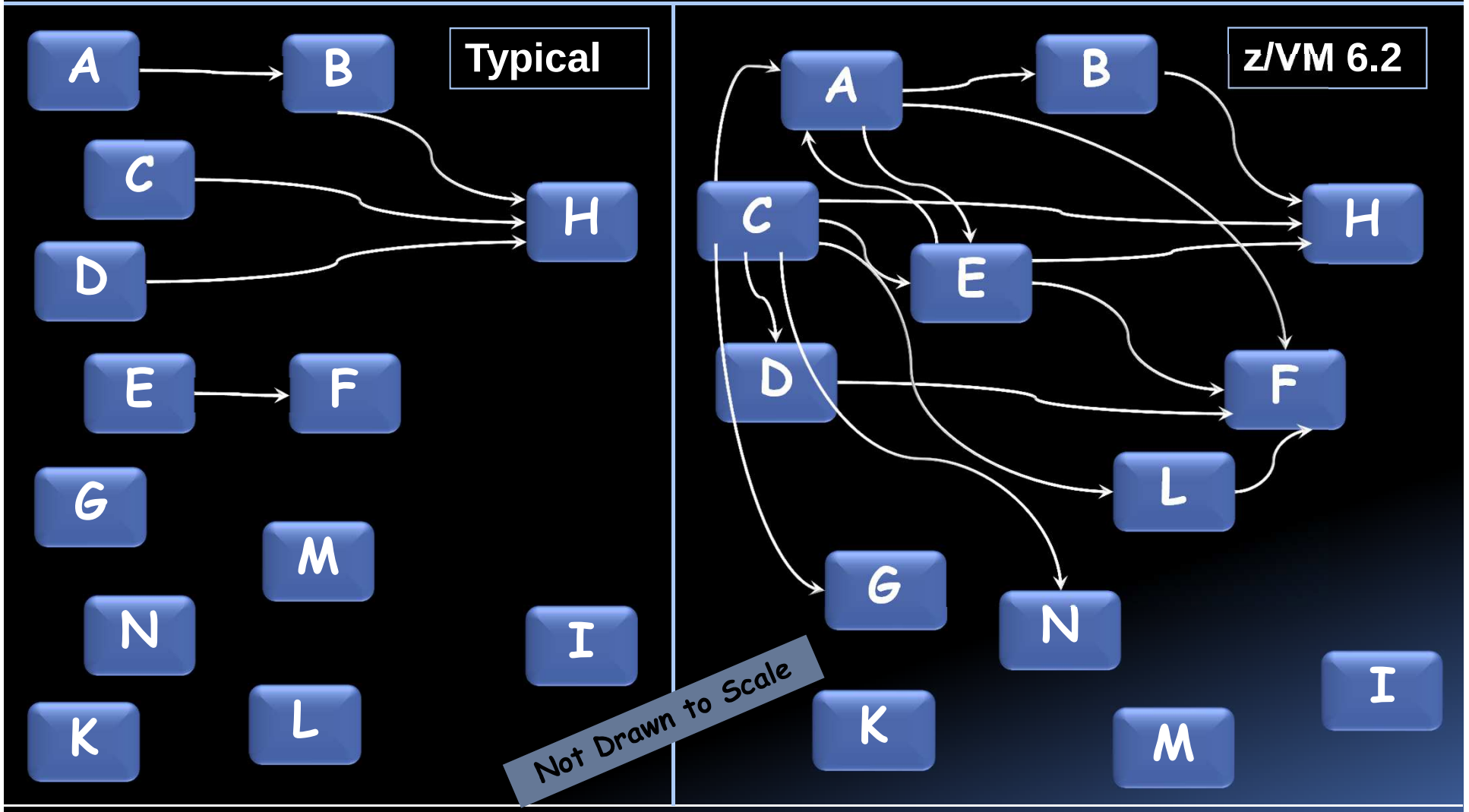
- Shorter Schedule
- Less Features
- Build off of early prototypes
  - Based on a Server Virtual Machine
- More Restrictions

- Longer Schedule
- Do it right
  - Various features
  - RAS
- Different design basics than other platforms
- Less Restrictions

Why So Difficult to do LGR?

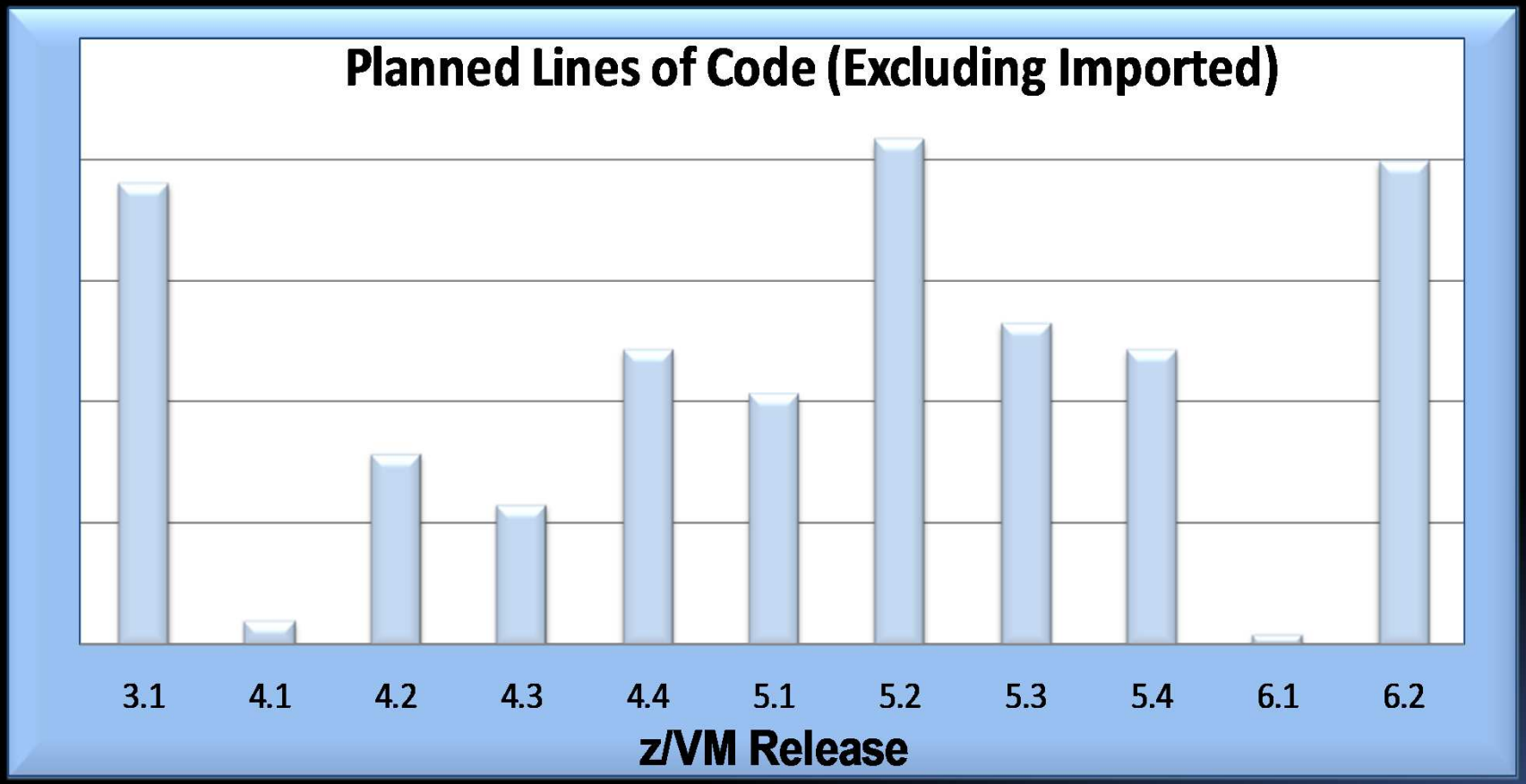
- Maintaining accurate representation of the architecture
  - A z/VM Strength
- Flexibility of the architecture
  - Different I/O devices
  - Crypto devices
  - Dynamic Resources
- Flexibility of z/VM
  - Tuning Options
  - CP System Services
  - Shared Memory
- Making it Worthy of the “System z” Brand

Typical Release vs. z/VM 6.2 Line Item Relations

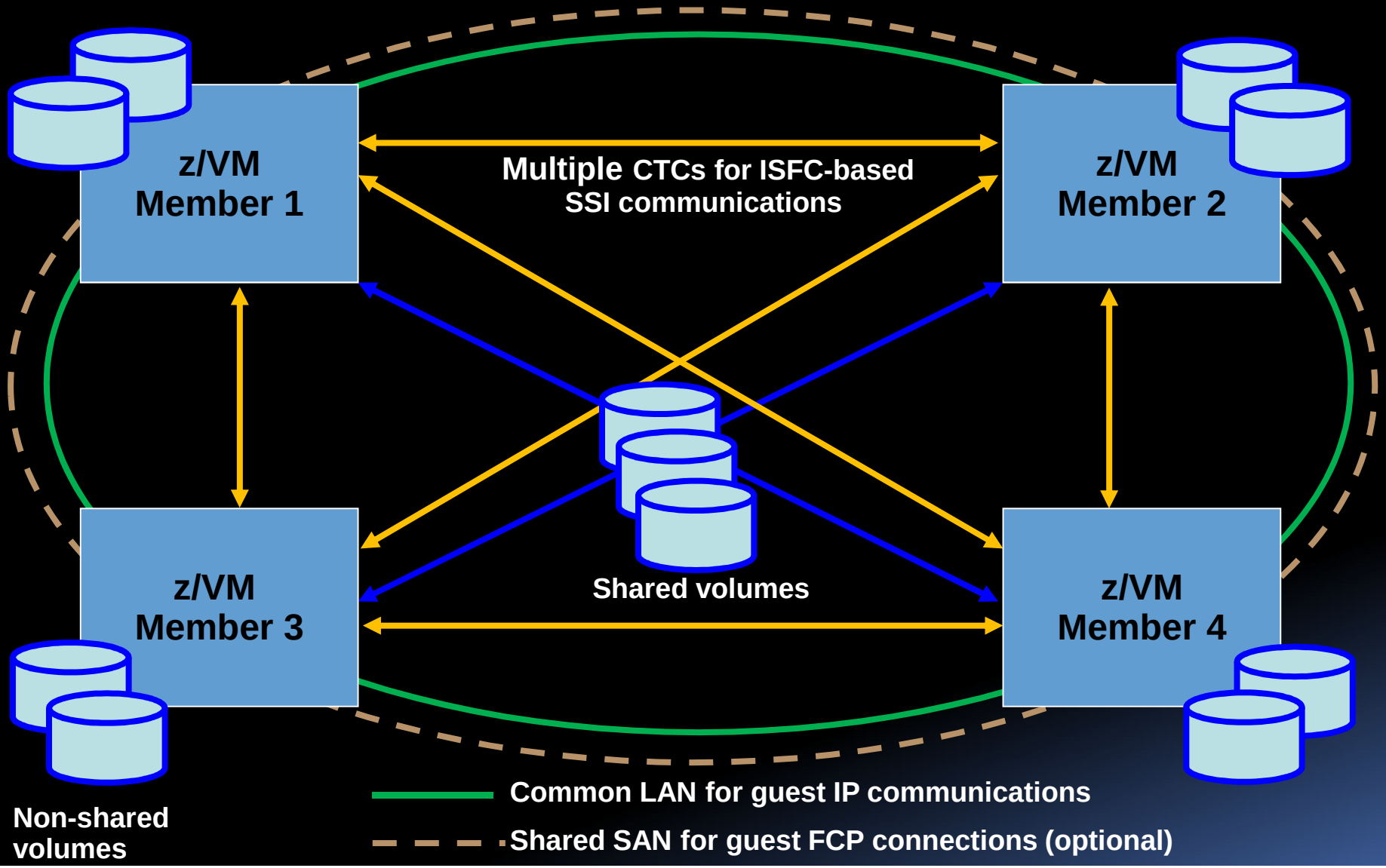


Speeds and Feeds

- Start of regular weekly team meetings: Feb 5, 2008



What Makes SSI & LGR Special?



SSI Cluster Management: Features for Greater Reliability

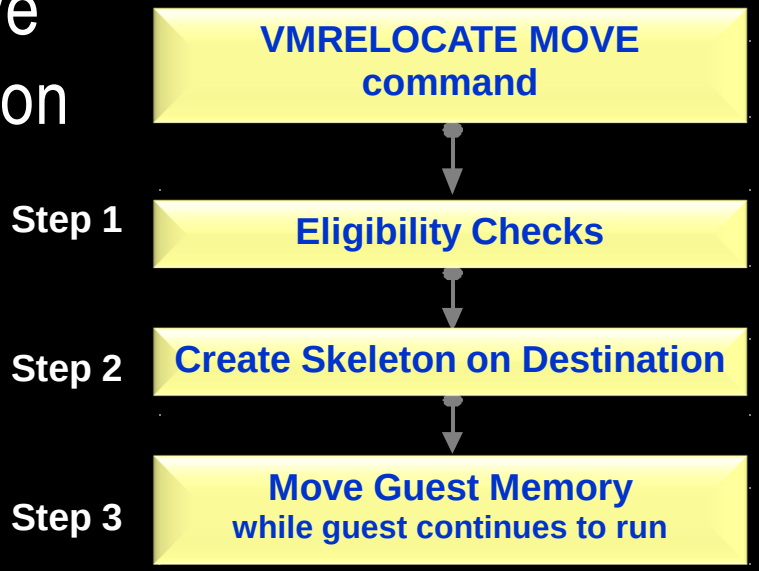
- Cross-checking of configuration details as members join cluster and as resources are used:
  - SSI membership definition and identity
  - Consistent definition of shared spool volumes
  - Compatible virtual network configurations (MAC address ranges, VSwitch definitions)
- Cluster-wide policing of resource access:
  - Volume ownership marking to prevent dual use
  - Coordinated minidisk link checking
  - Autonomic minidisk cache management
  - Single logon enforcement
- 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

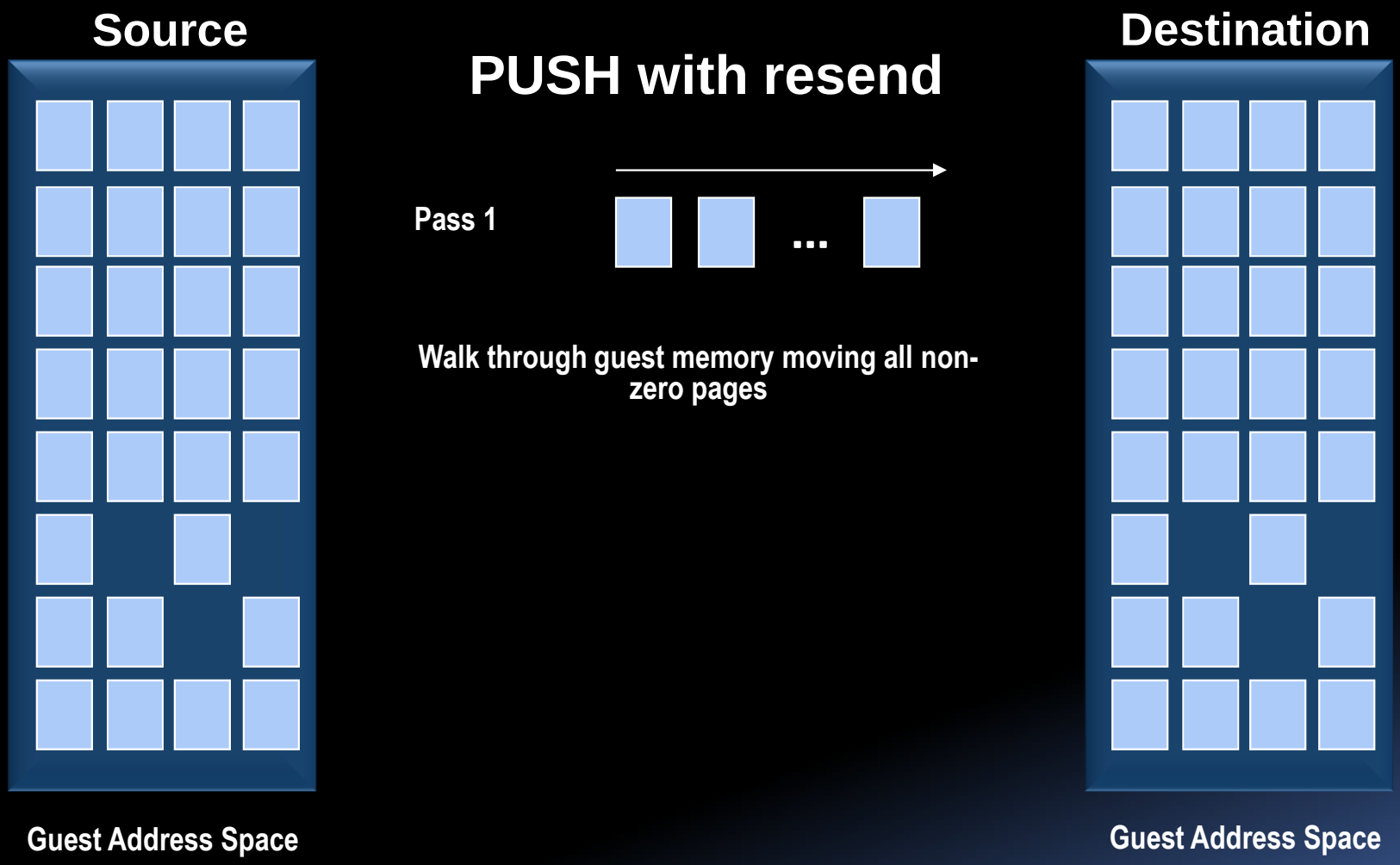
- Communications failure “locks down” future resource allocations until resolved
  - Existing running workloads continue to run
  - Prevents new accesses to resources
  - Cluster could temporarily be split and workloads continue to run

- 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.

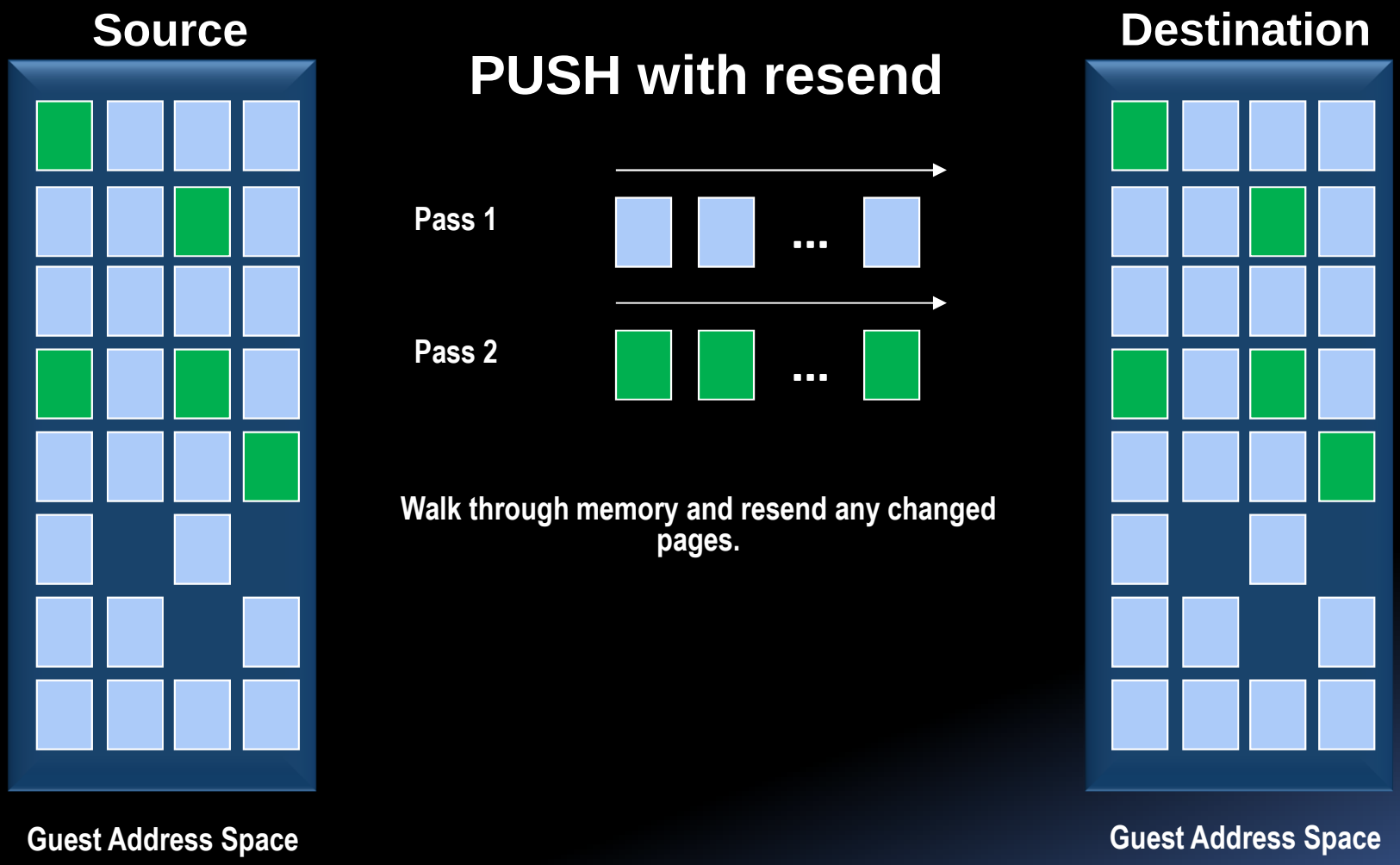
Stages of a Live Guest Relocation



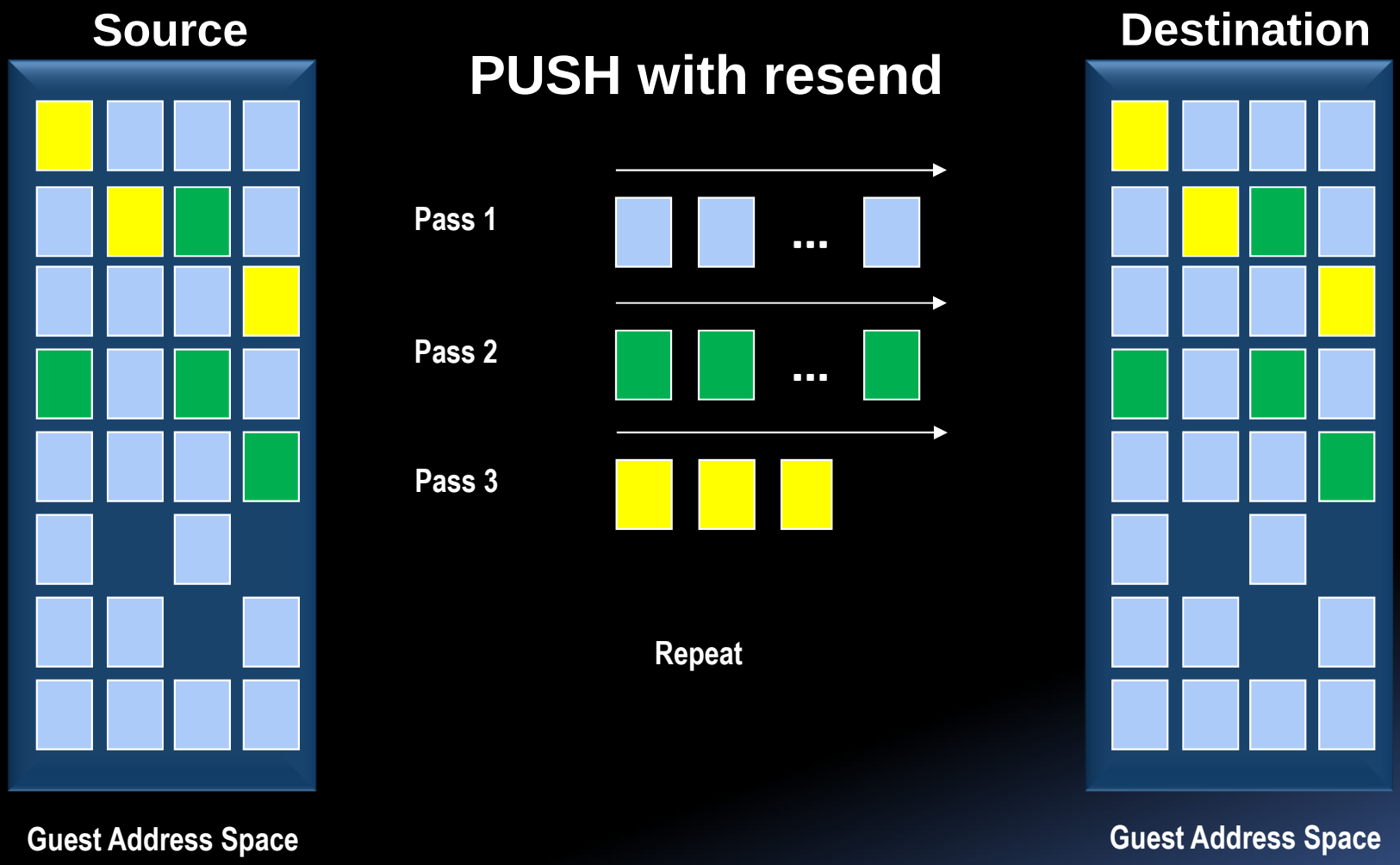
LGR, High-Level View of Memory Move



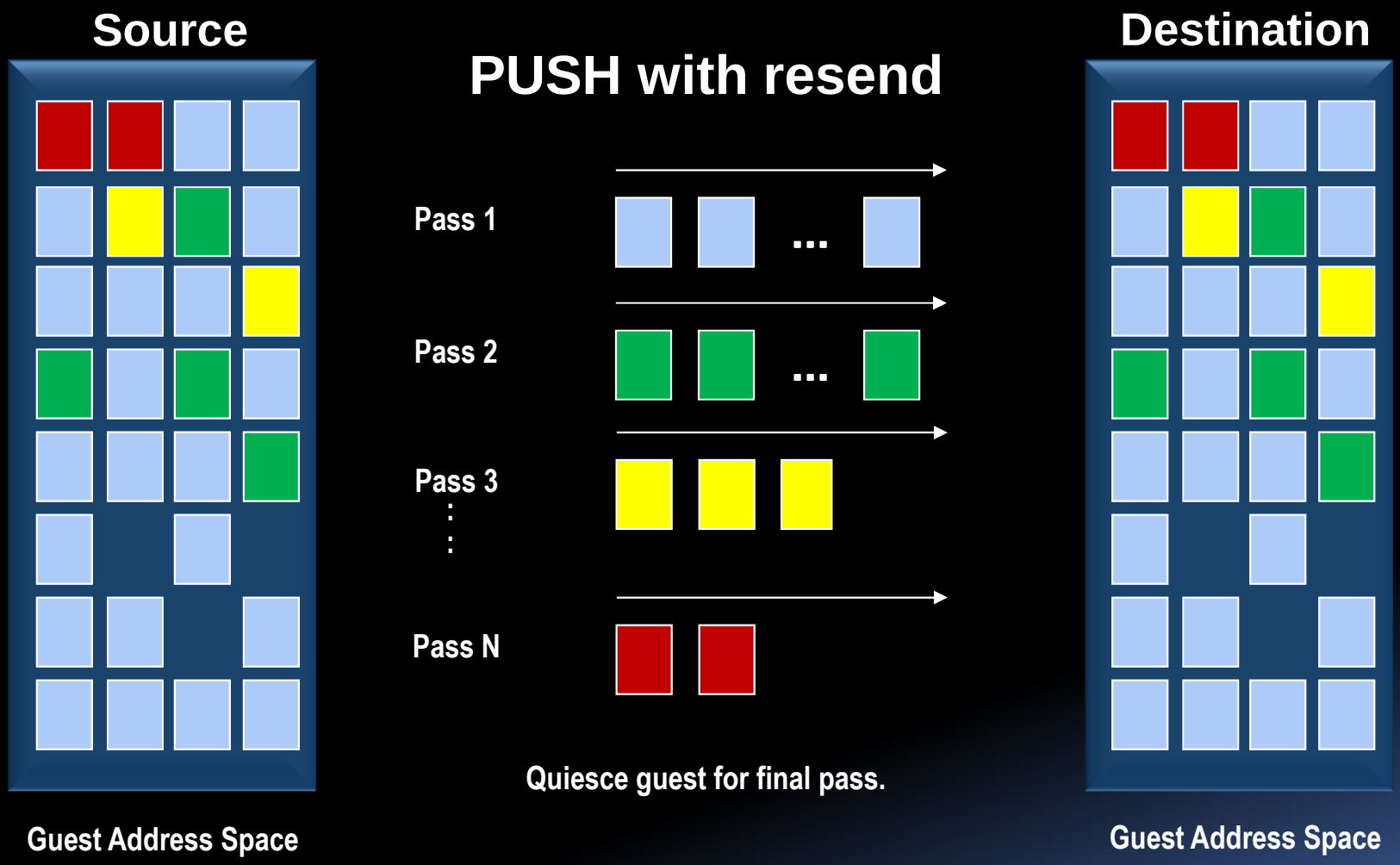
LGR, High-Level View of Memory Move



LGR, High-Level View of Memory Move



LGR, High-Level View of Memory Move



Stages of a Live Guest Relocation

Live Guest Relocation

- New CP Planning and Administration Chapter: Preparing for Live Guest Relocations in a z/VM SSI Cluster

- New CP VMRELOCATE command
  - VMRELOCATE command starts, cancels, or tests a Live Guest Relocation
  - CP Commands & Utilities Reference: 14 pages (6 pages messages)
  - Options to control behavior:
    - MAXQUIESCE – maximum quiesce time
    - MAXTOTAL – maximum total relocation time
    - TEST – test it first (what a concept!)
    - STATUS – find out status

Live Guest Relocation

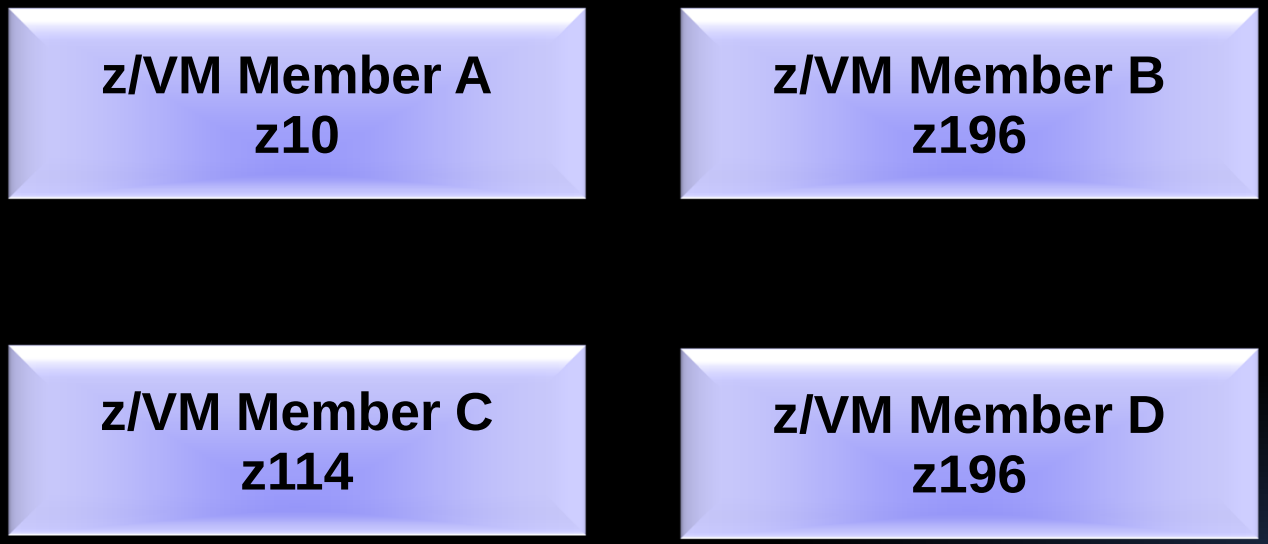
- New SMAPI interfaces
  - VMRELOCATE
  - VMRELOCATE_Image_Attributes
  - VMRELOCATE_Modify
- Other Interfaces of note:
  - *VMEVENT
  - *MONITOR
  - *ACCOUNT
- New CP Exit Points

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

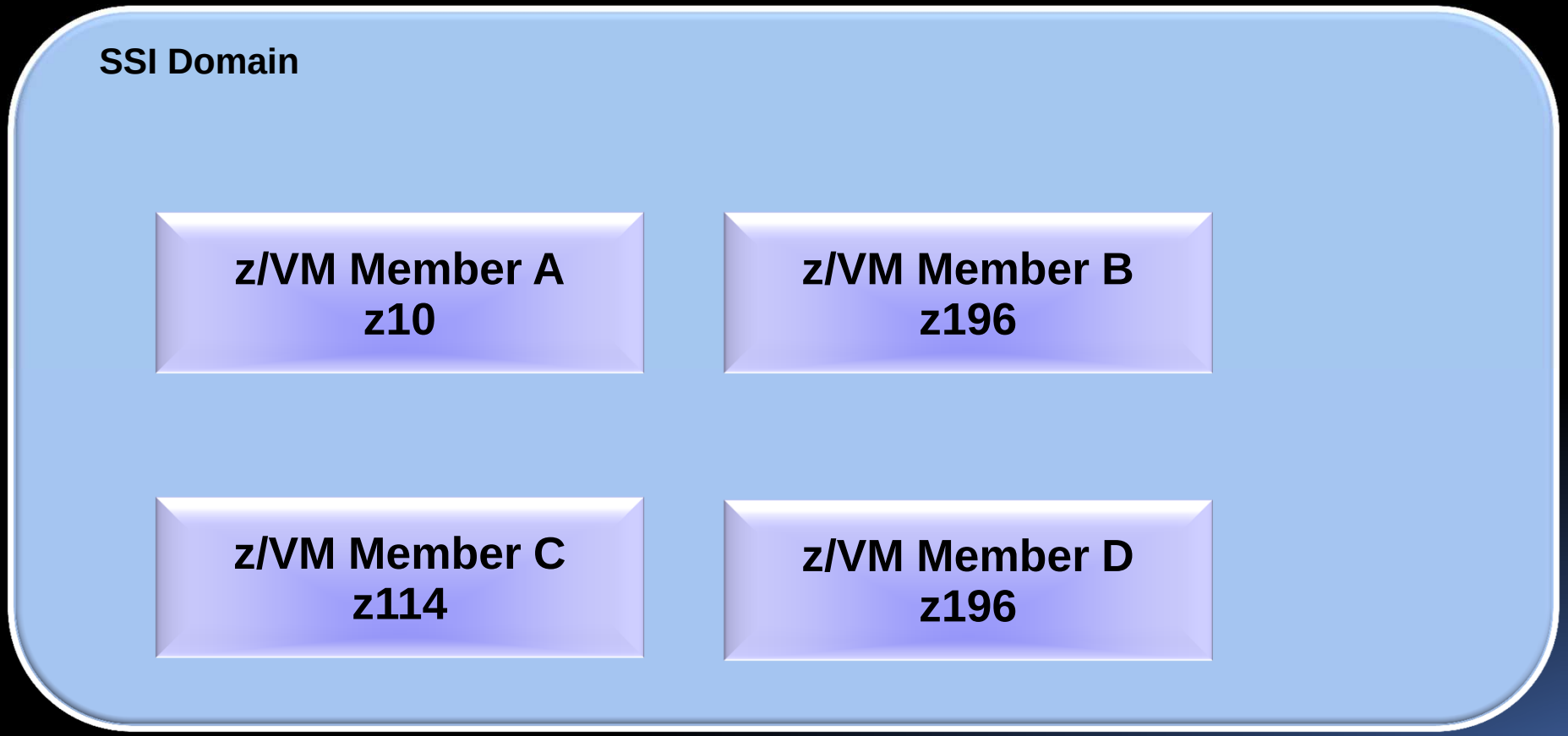
Relocation Domains

- Greater control over where virtual machines can relocate and what architecture features they will have.
- Architecture available to a virtual machine within a relocation domain is the maximal common subset.



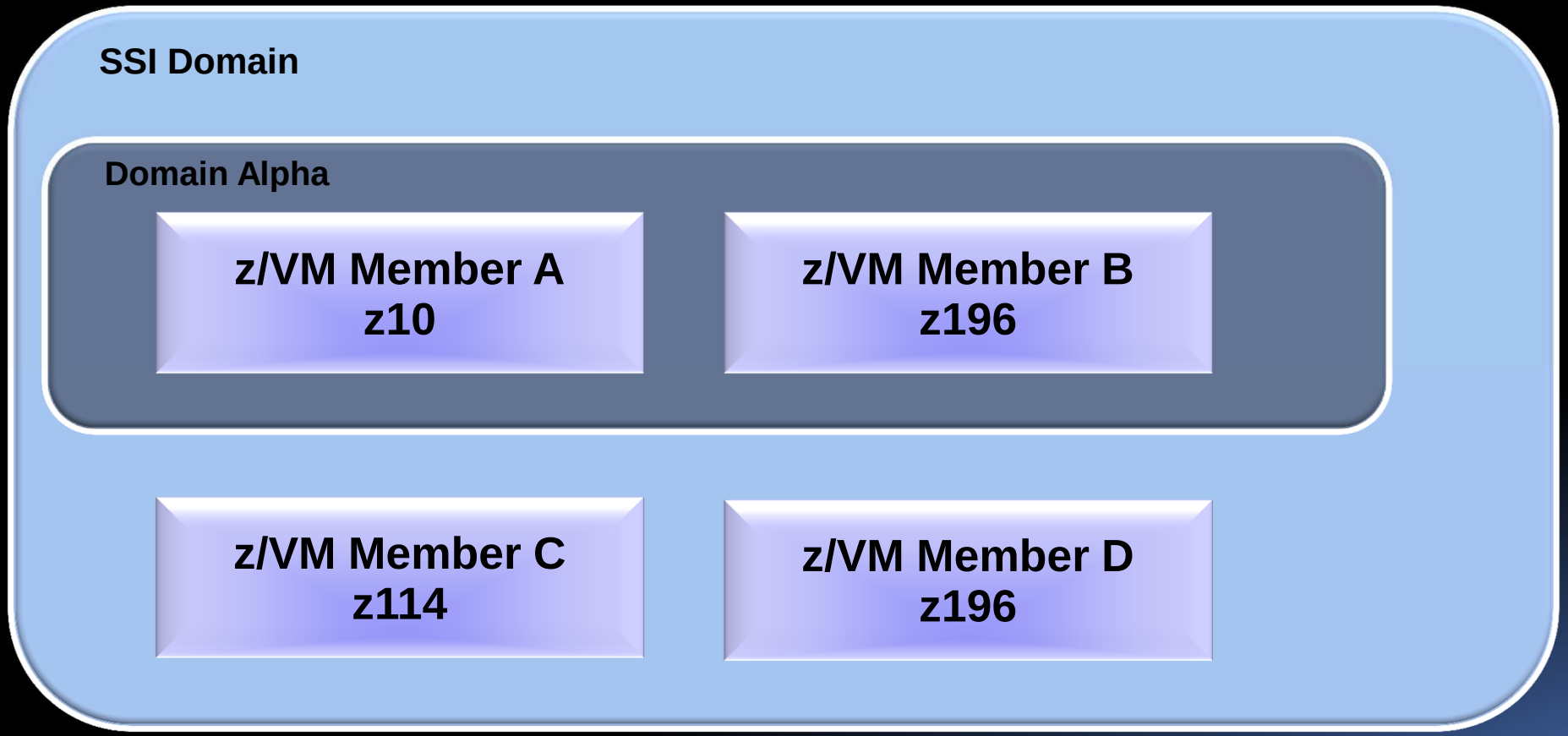
Relocation Domains

- By default, the SSI domain is a relocation domain that includes all members of an SSI Cluster.
- Additionally, there is a domain for each member which includes only that member.



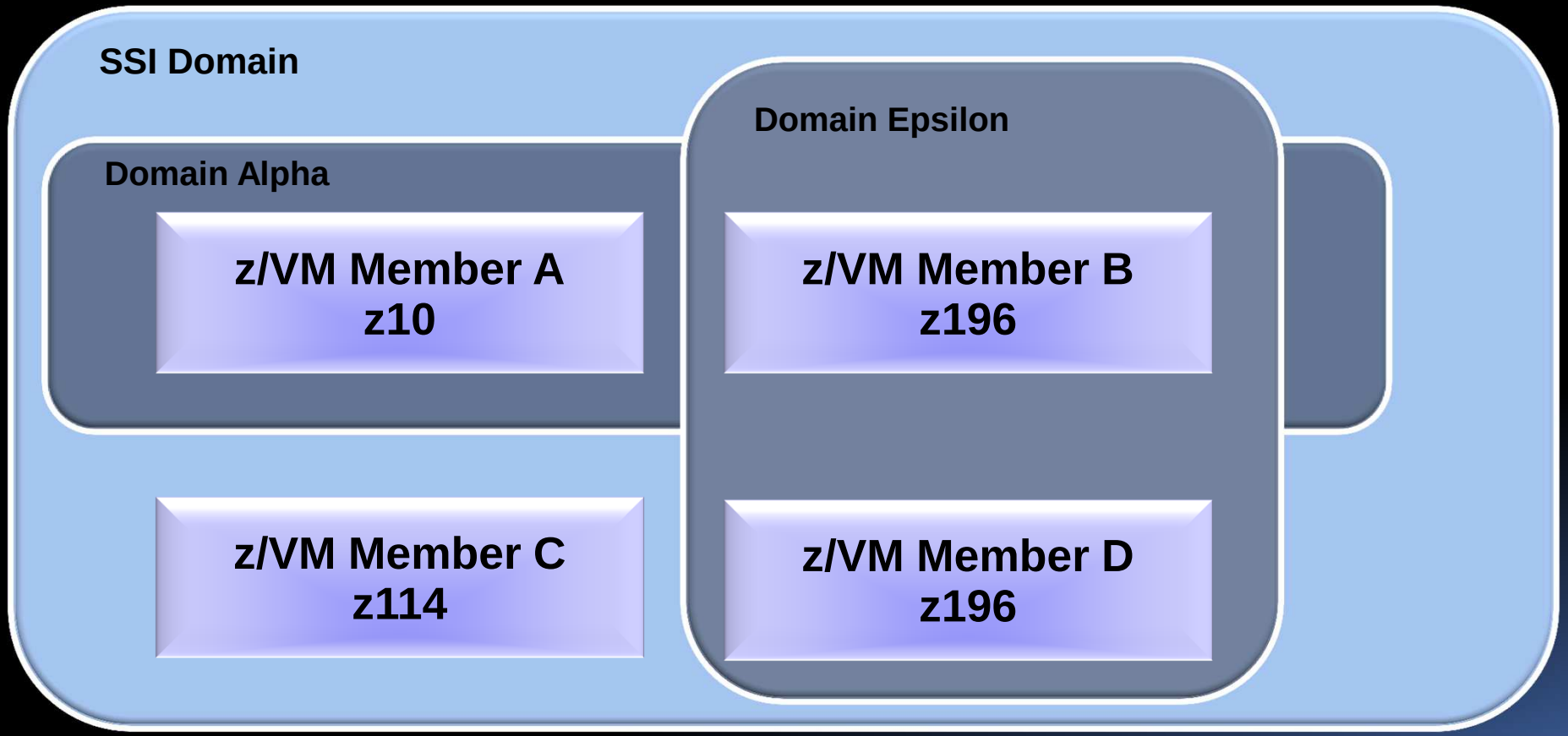
Relocation Domains

- Domain Alpha is created to span a z10 and a z196, this restricts the architecture exposed to the virtual machine assigned to Alpha to only the maximal common instructions and features. In this case, most likely a subset of the z196.



Relocation Domains

- Virtual machines in domain Epsilon are afforded the full z196 architecture.



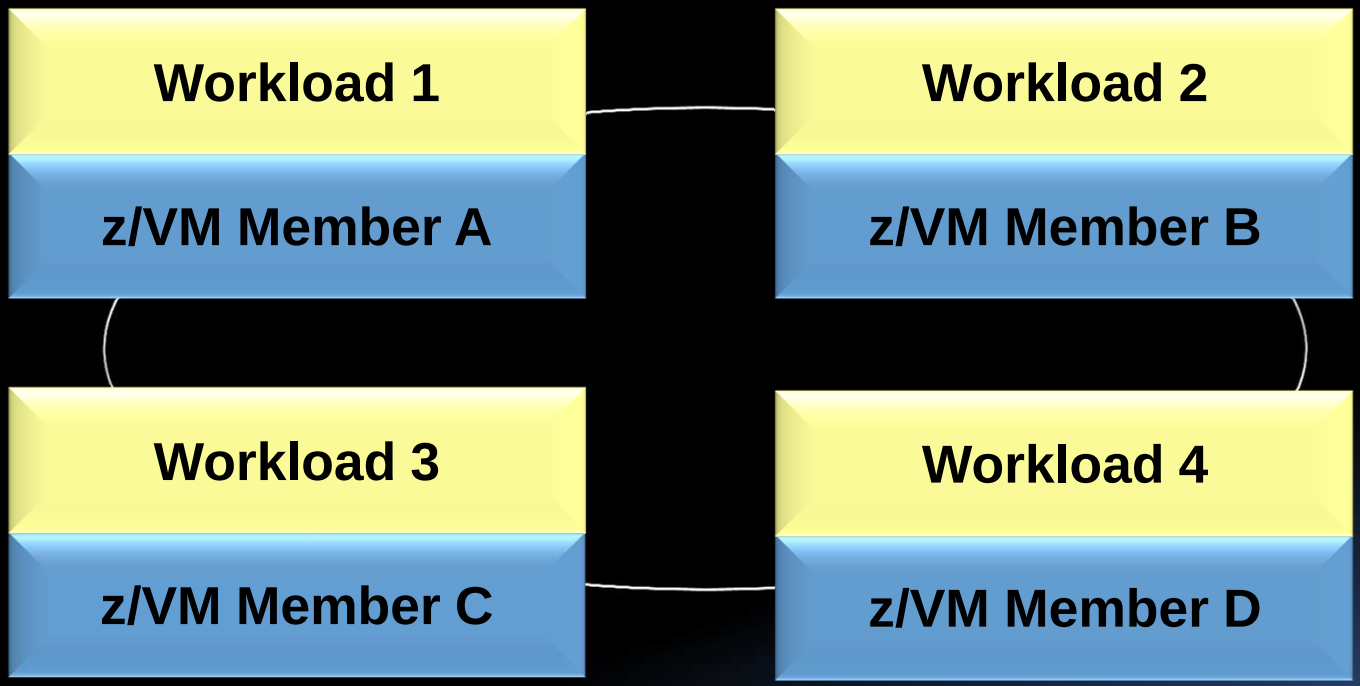
New Possibilities

What Can You Do with SSI Clusters and LGR?

1. Flexibility for Planned Outages
2. Methodically Testing at Current Levels
3. Increased Control Over Server Sprawl
4. Production with Protection
5. Managing Resource Distribution
6. Consistent Test Bed for Stress Tests
7. One From the Customers – Utility Migration LPAR
8. Local Disaster Recover (Business Continuity)
9. Migrate to New Processor

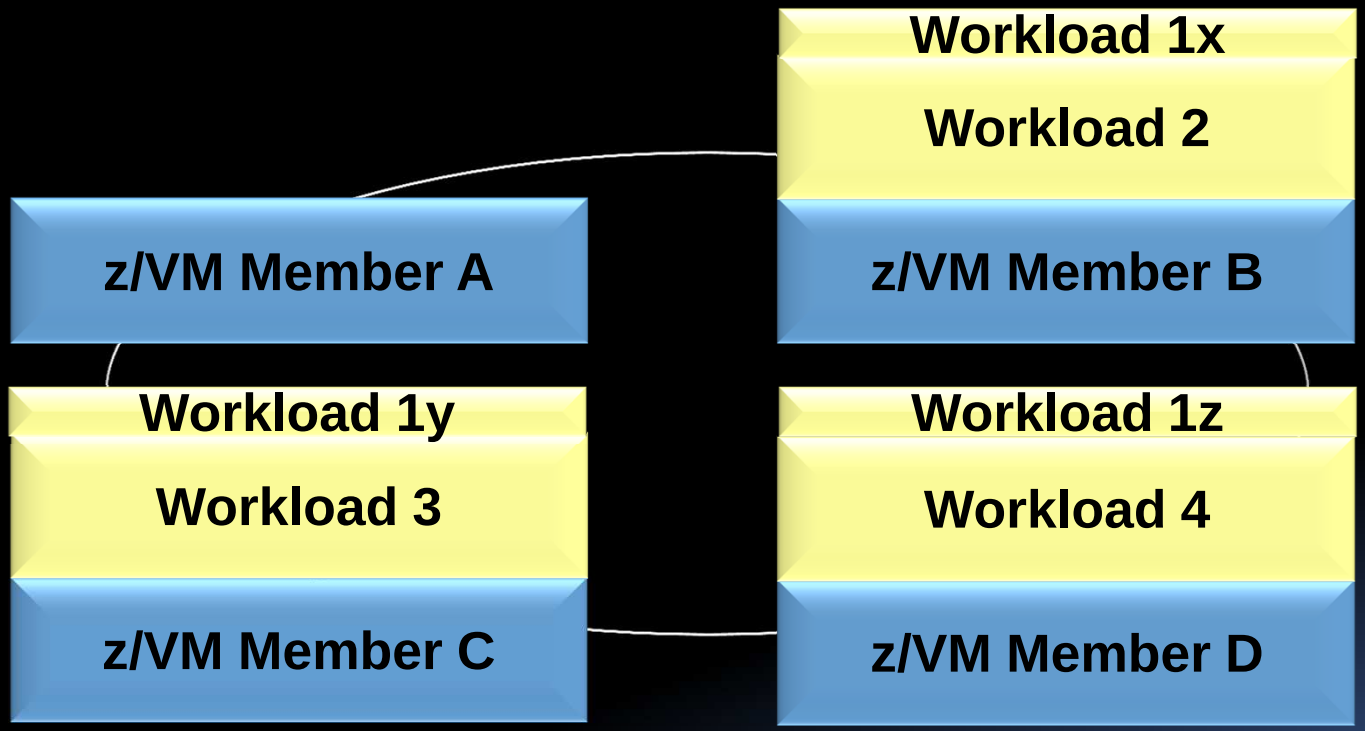
Flexibility for Planned Outages

- 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...



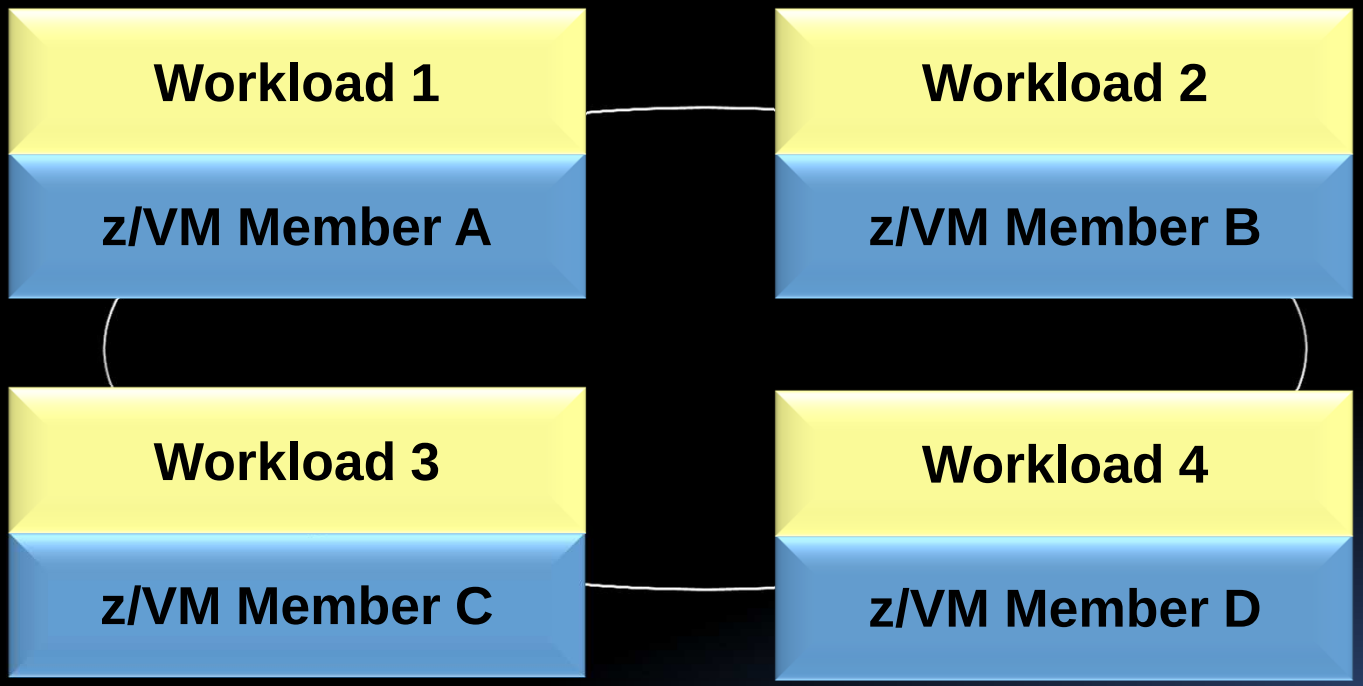
Flexibility for Planned Outages

1. Apply maintenance to Member A, having new CP load module ready for IPL.
2. Move critical work from Member A to the other 3 members in the cluster.
3. Shutdown Member A and bring back up with new CP load module.



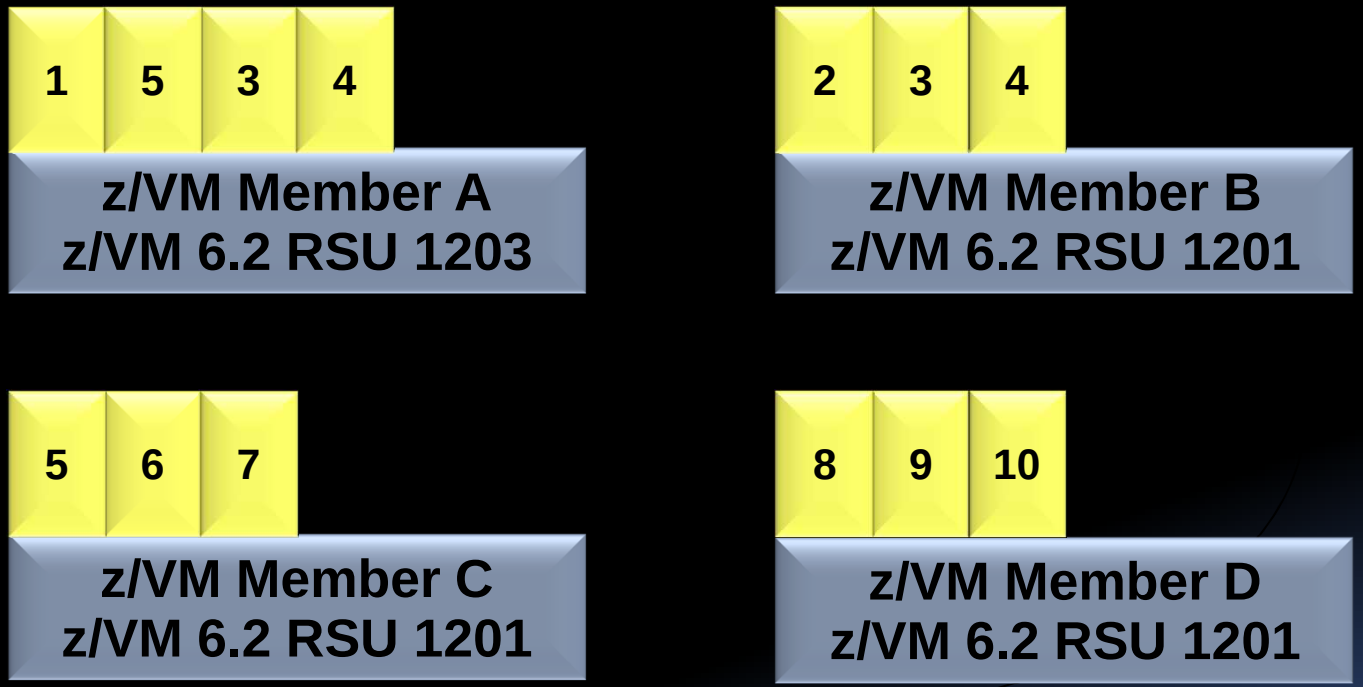
Flexibility for Planned Outages

1. Move workloads back to member A
2. Rejoice



Methodically Testing at Current Levels

- Testing for new levels of z/VM in the past often required use of second level systems and trade-offs between matching production environment.
- z/VM SSI clusters can be used to help test and migrate throughout the members.
- Perhaps start with System A at new service level and slowly move work there to test.

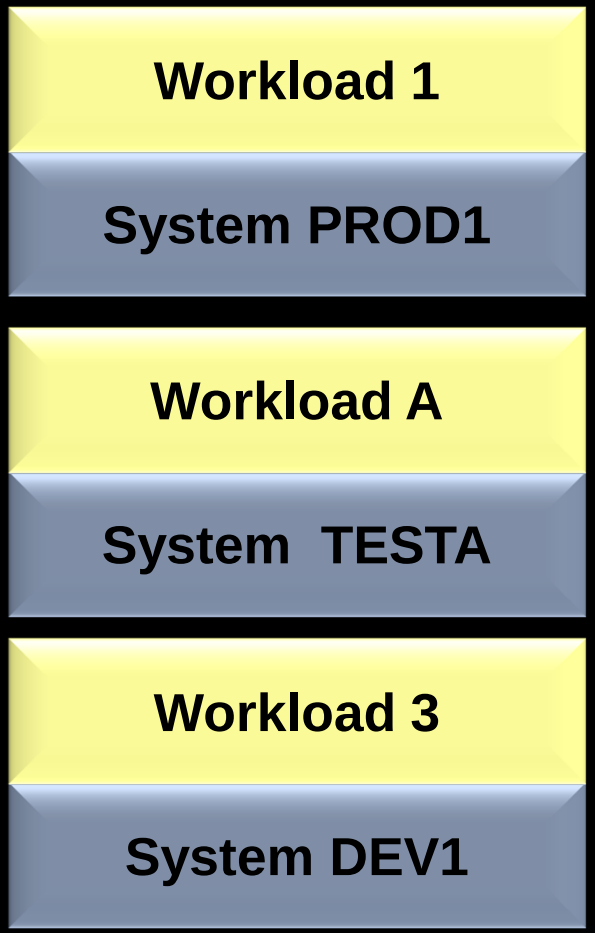


Increased Control Over Server Sprawl

- Server sprawl and the success of virtualization have led to virtual server sprawl, z/VM SSI Clusters improve the management characteristics for these environments.
- Consider customer with a single LPAR for production is sufficient today, but they are growing at a significant rate.
- Various reasons to expand past a single LPAR:
  - Out growing single LPAR capacity
  - Risk management: avoiding all eggs in one basket and diversification.
  - Flexibility for software licensing
- Move to z/VM 6.2 keeping your individual system, but prepare them to run as multi-member SSI in the future.
  - Bring in another LPAR and bring up an additional SSI member.

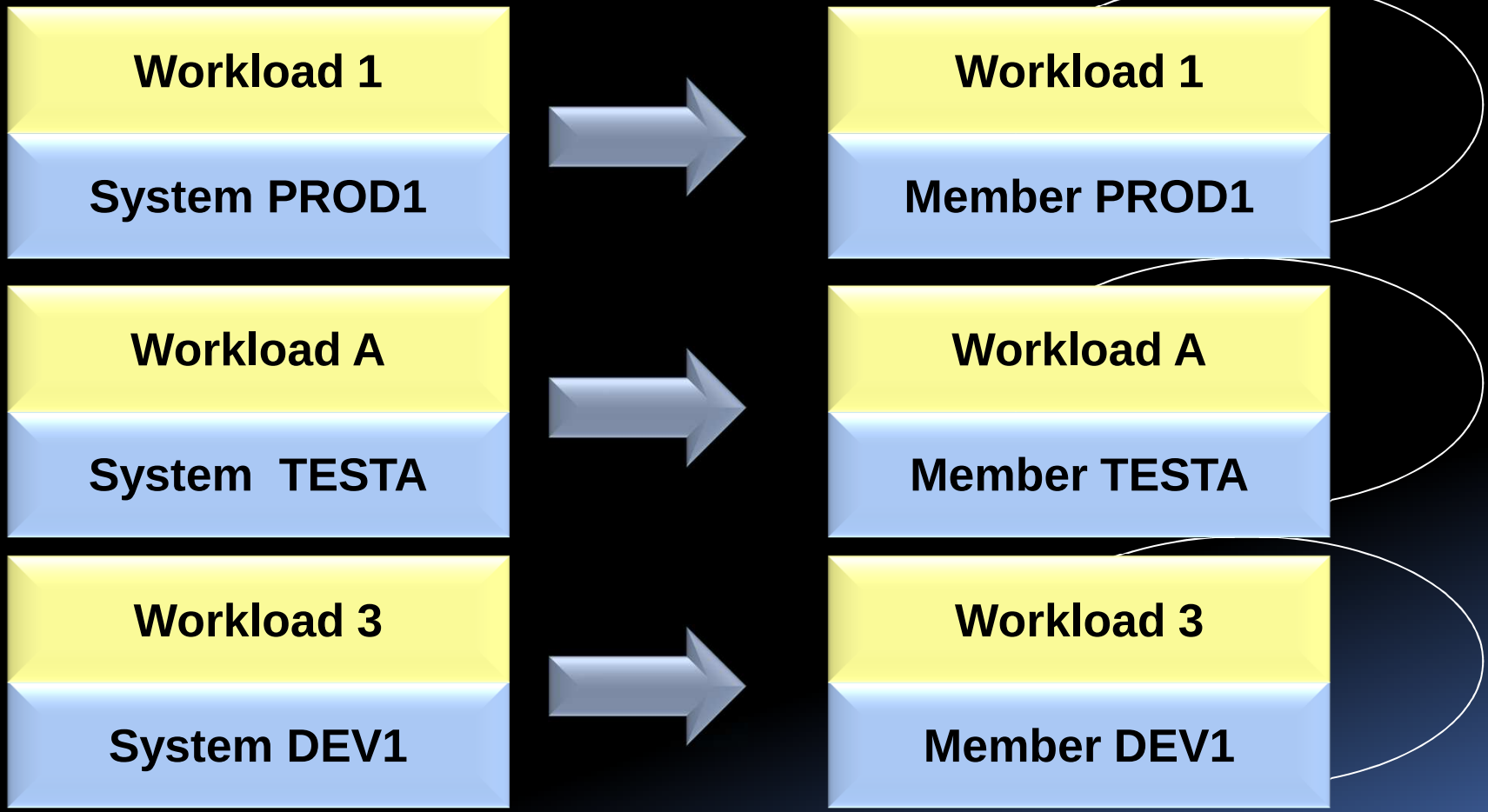
Increased Control Over Server Sprawl

Today, you may have 3 separate systems, but may not have compelling reason to combine them into a cluster.



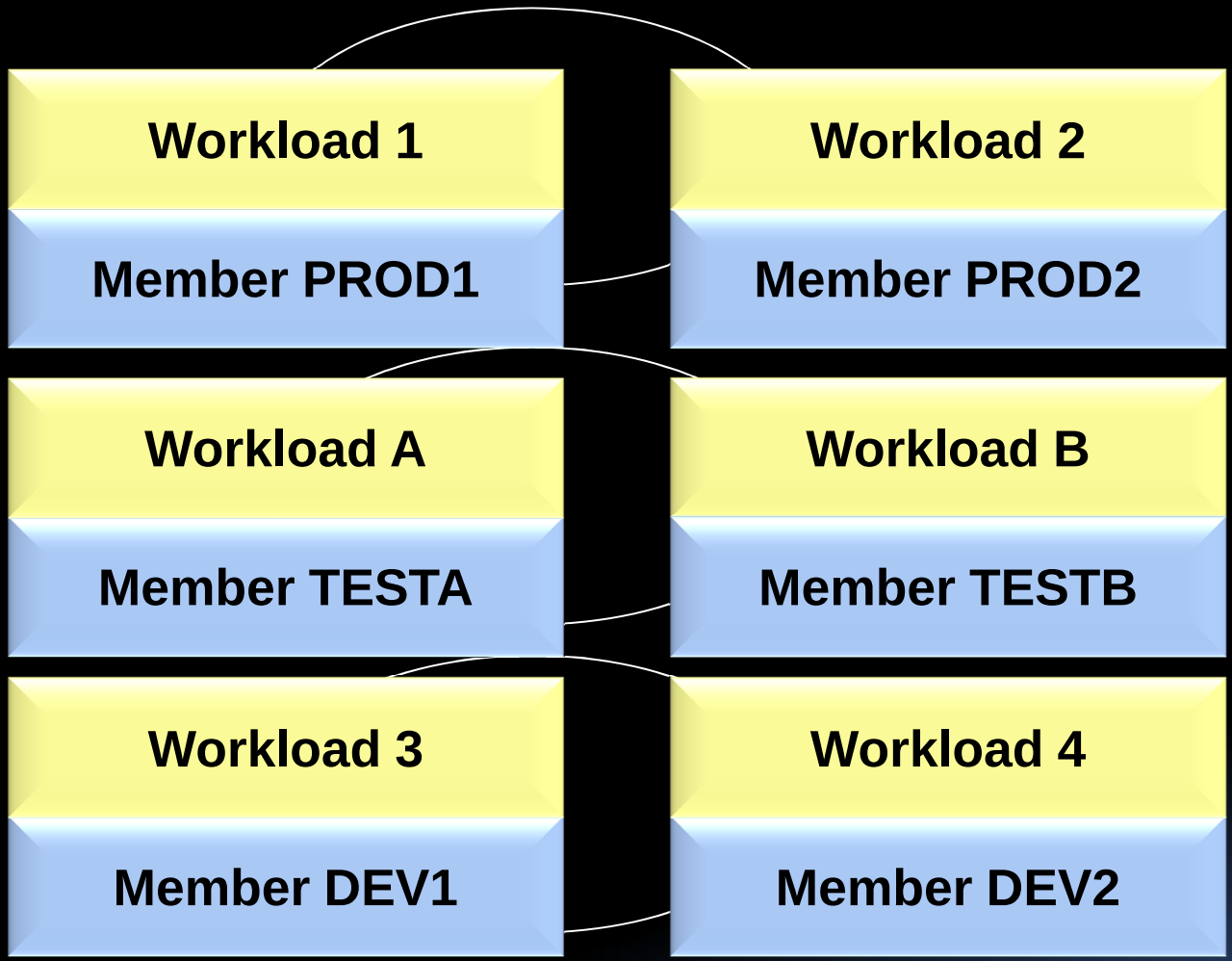
Increased Control Over Server Sprawl

Move to z/VM 6.2 and create clusters that just happen to be single member clusters for now.



Increased Control Over Server Sprawl

As workloads increase, create additional members in each cluster.

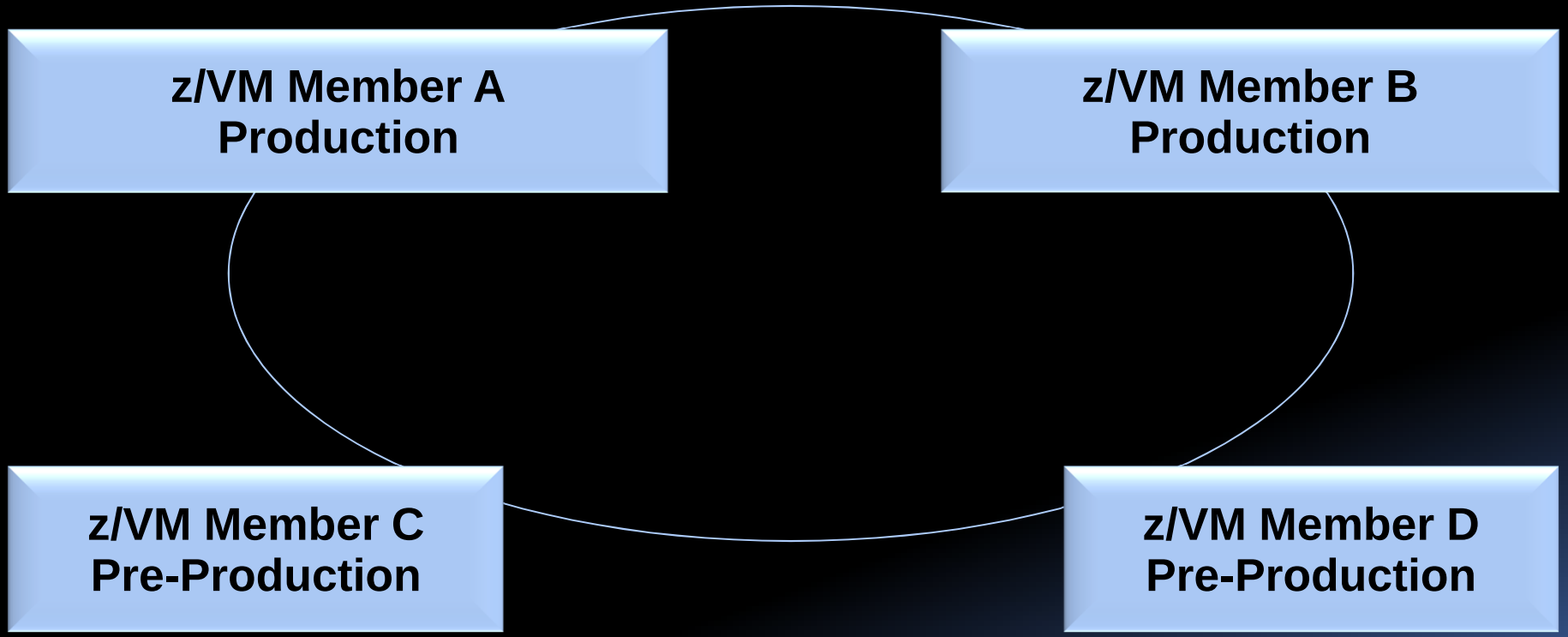


Production with Protection

- When adding a new application or upgrading an application in production, what is your confidence that you know how it will
  - Perform?
  - Impact other production workload?
  - Meet expectations?
- Single System Image provides a way to allow workload to be part of the production environment, and yet be isolated

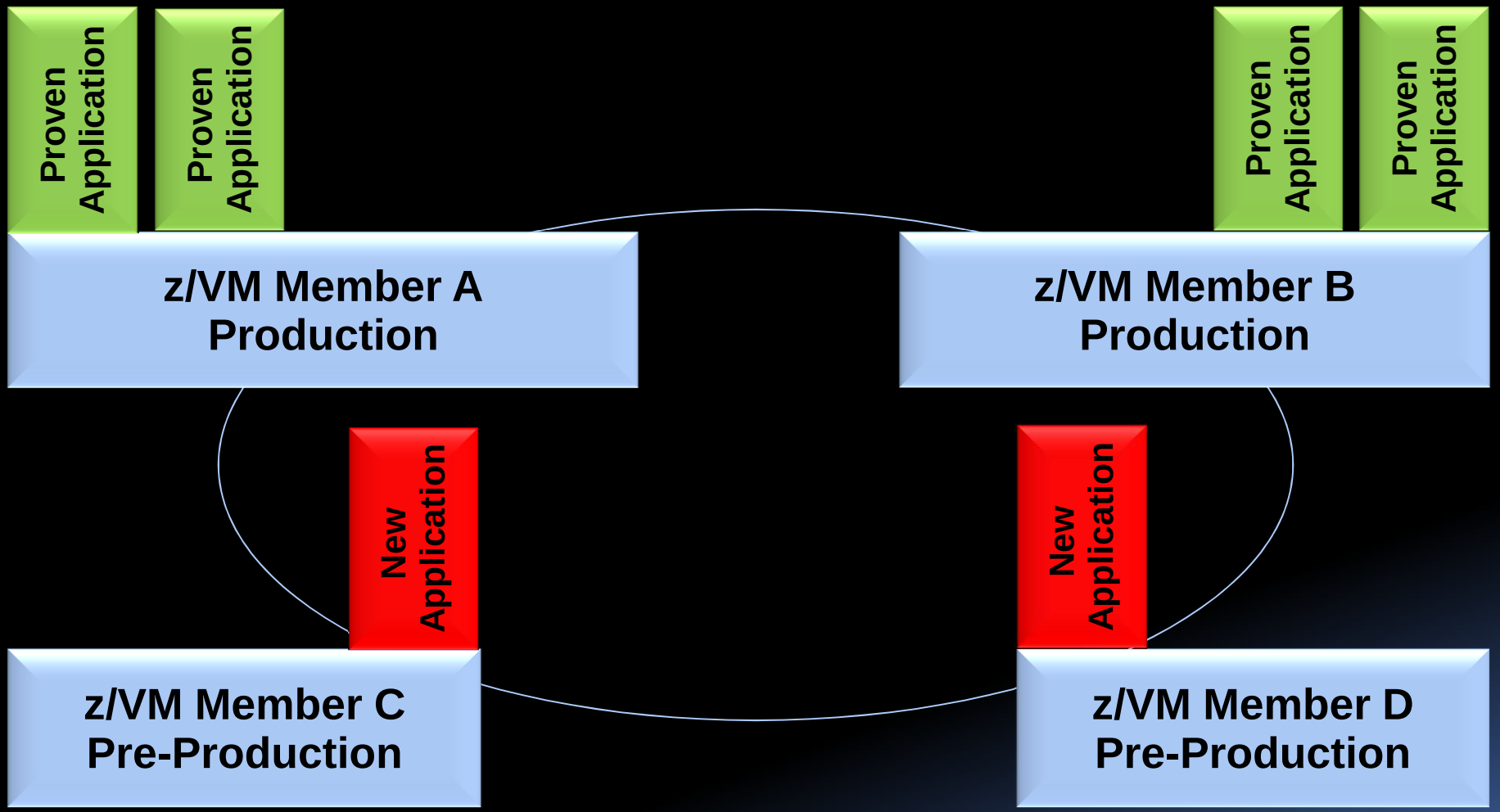
Production with Protection

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



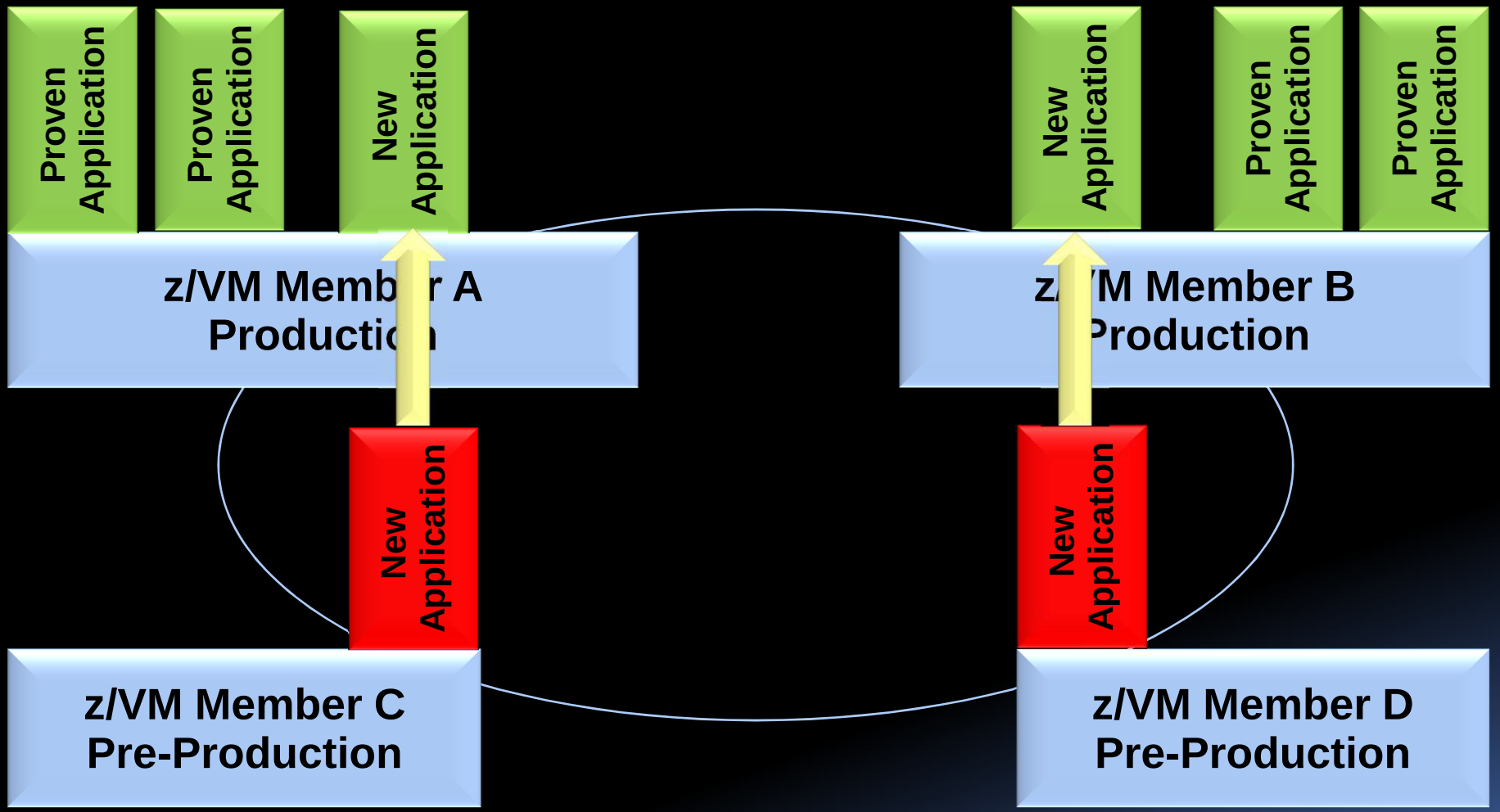
Production with Protection

- Allow new application to run in pre-production LPARs



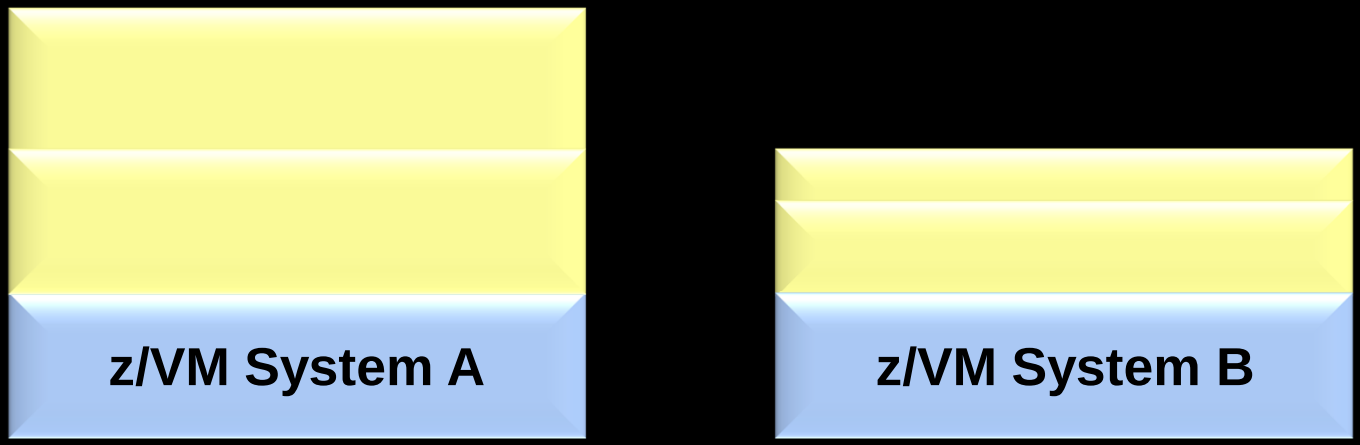
Production with Protection

- If all goes well, move into true production



Managing Resource Distribution

- Some customers have or are in processing of exceeding the capacity of a single z/VM system and split work across LPARs
- Determining how to divide the workloads across LPARs is a challenge, particularly in a dynamic world...



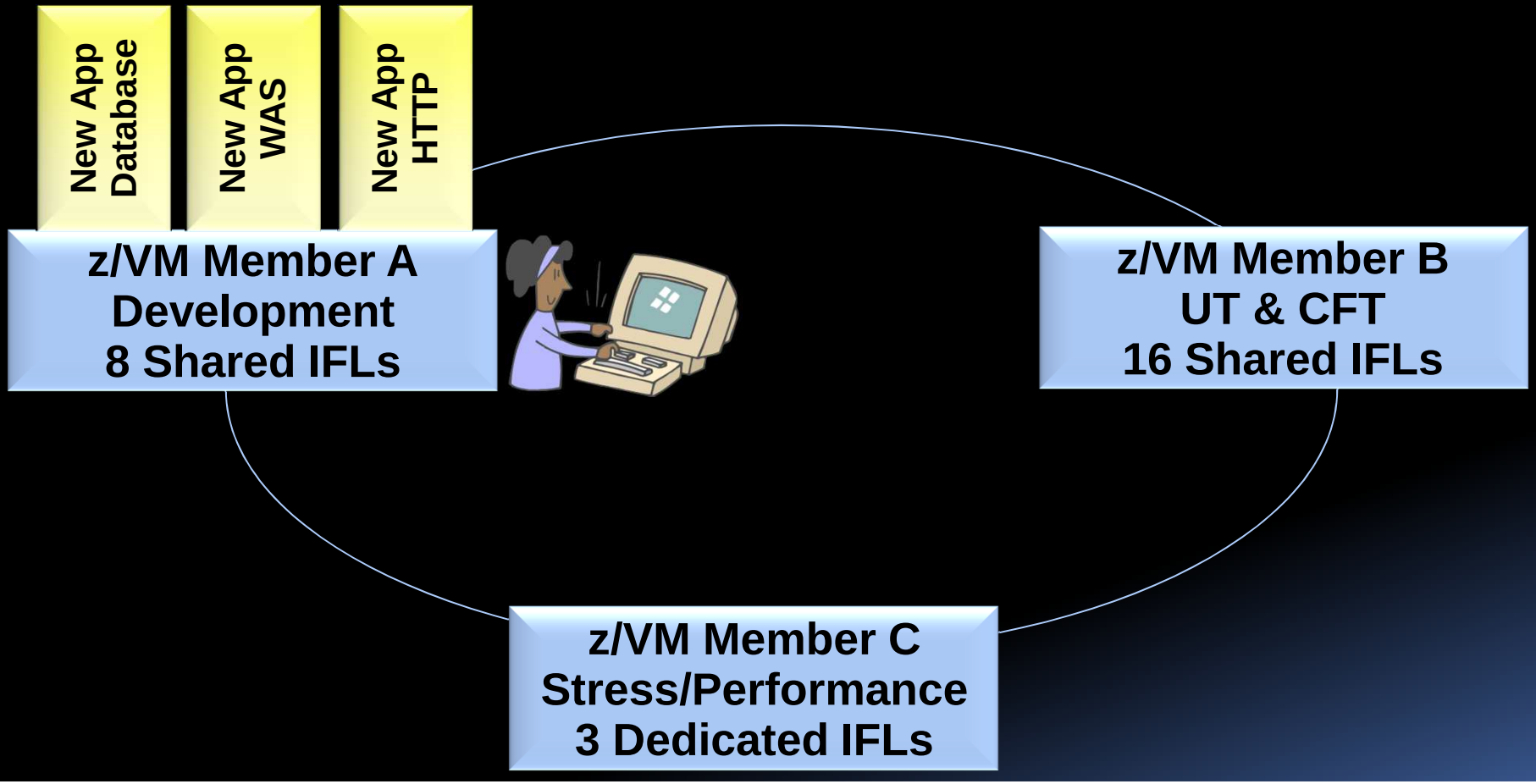
- With individual z/VM systems, one would need to define new virtual machines on B and remove the definitions on A
  - Responsibility of ensuring integrity during process is on shoulders of system programmer.
- With an SSI cluster, one can more easily redistribute the load through logoff/logon or in many cases with LGR.

Consistent Test Bed for Stress Tests

- Testing Challenges:
  - Controlling test environments, testing in **consistent** manner
  - Functional and QA testing of various test programs
  - Stress testing in a controlled environment
- Having an SSI cluster environment allows:
  - Virtual server with same resources, run in different members of cluster based on needs
  - Load in development probably not as heavy, run that in a smaller shared environment
  - Various testing in UT & CFT could create a heavier load for various testing
  - An isolated LPAR (member) for stress testing or establishing performance characteristics of workload.

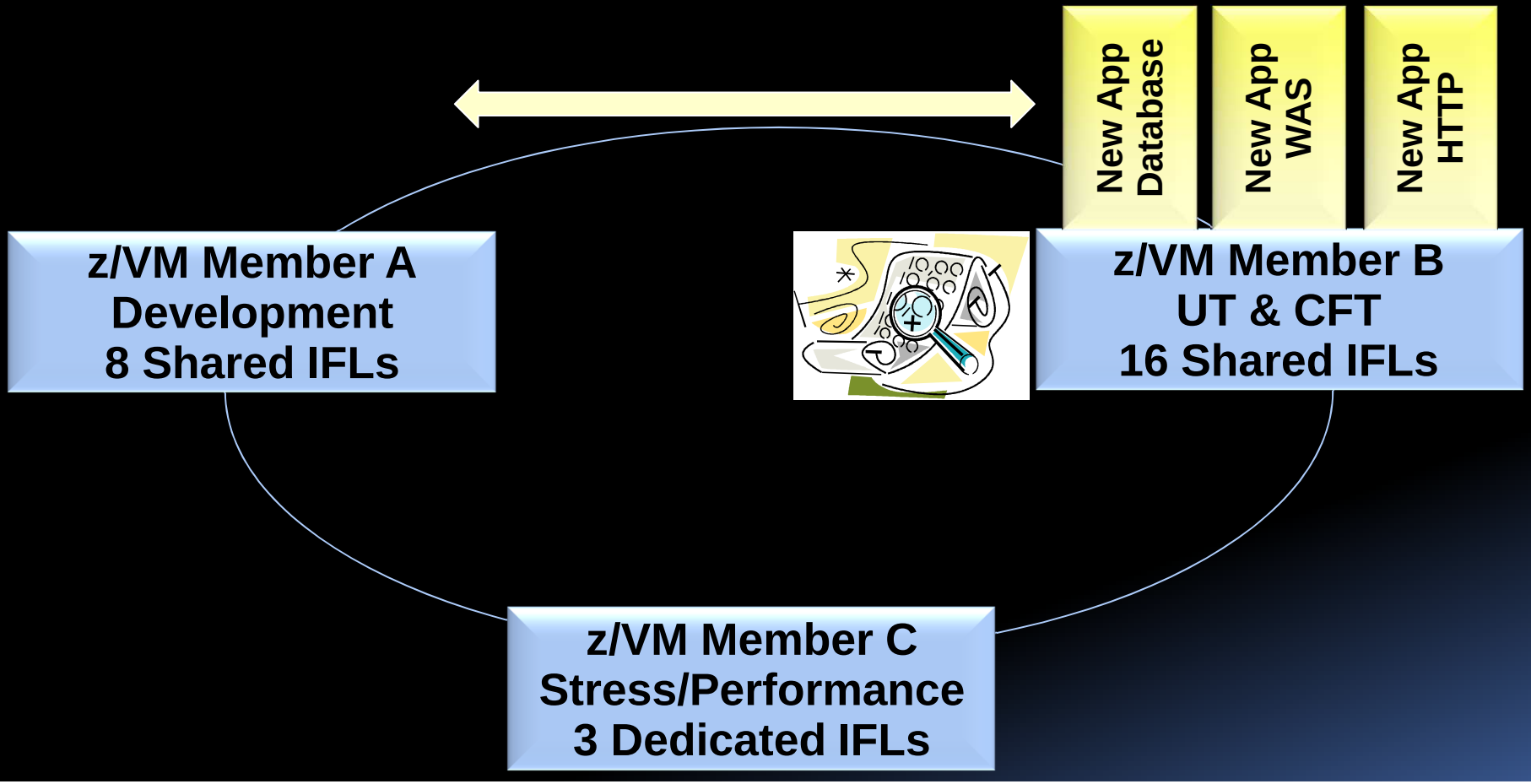
Consistent Test Bed for Stress Tests

- Consider this example with development, unit test, component function test, performance test, and stress tests.
- Build it all in the development member.



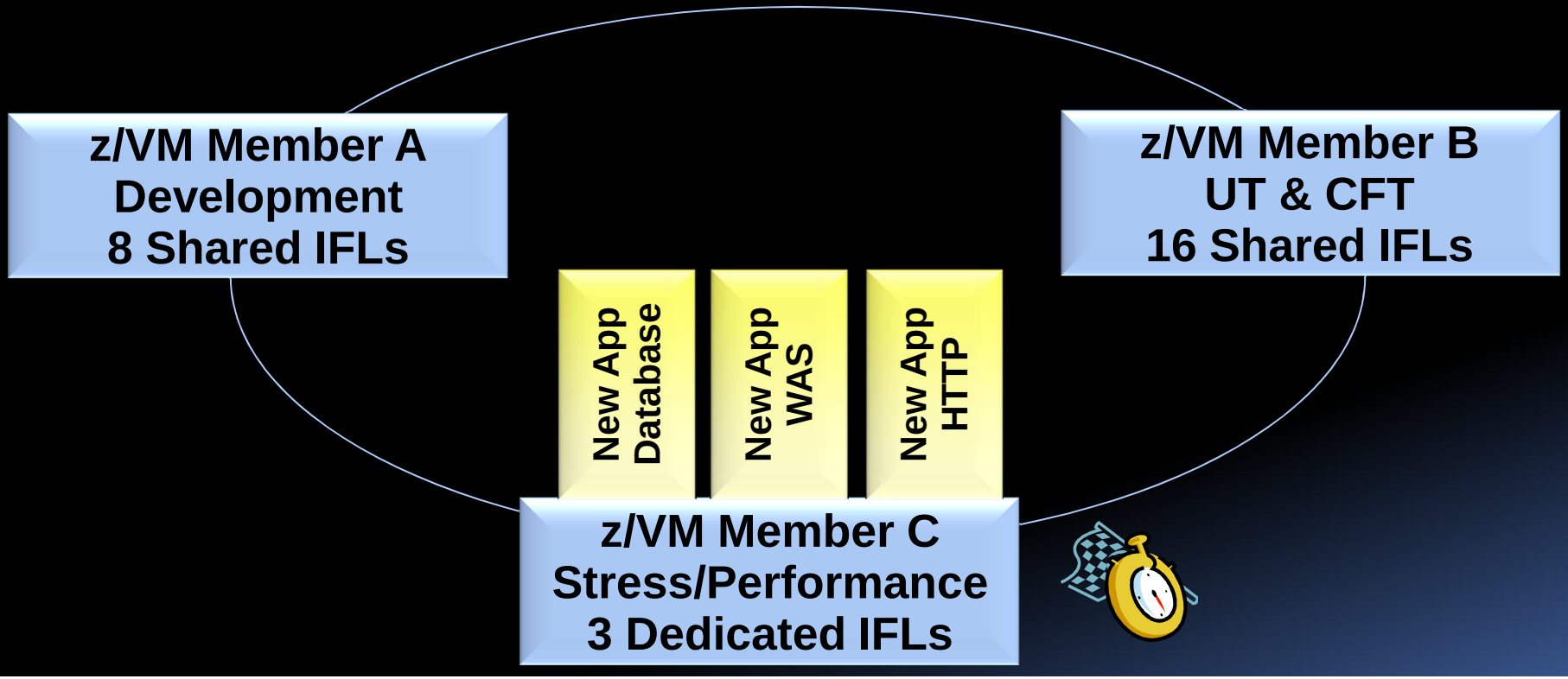
Consistent Test Bed for Stress Tests

- Development and Test could share the virtual machines involved, passing them back and forth between the systems as needed.

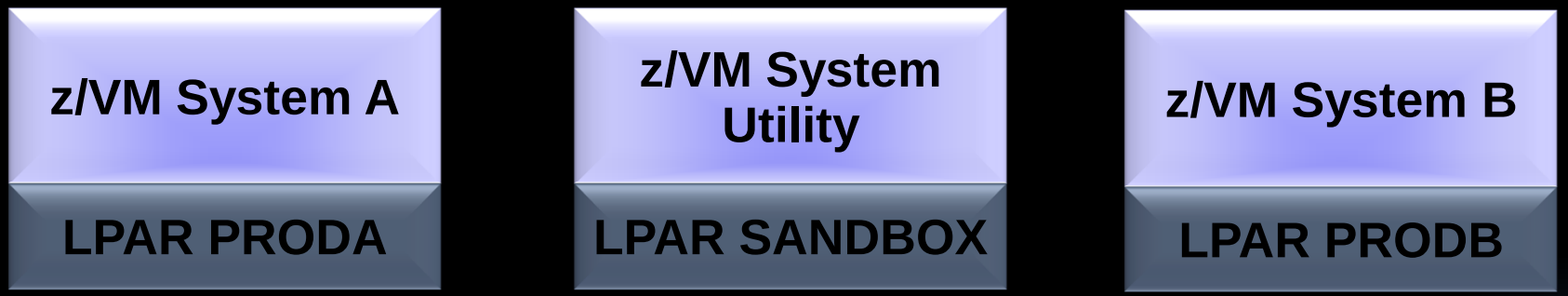


Consistent Test Bed for Stress Tests

- When ready for performance or stress test, move to Member C with Dedicated resources
- More control over what has changed

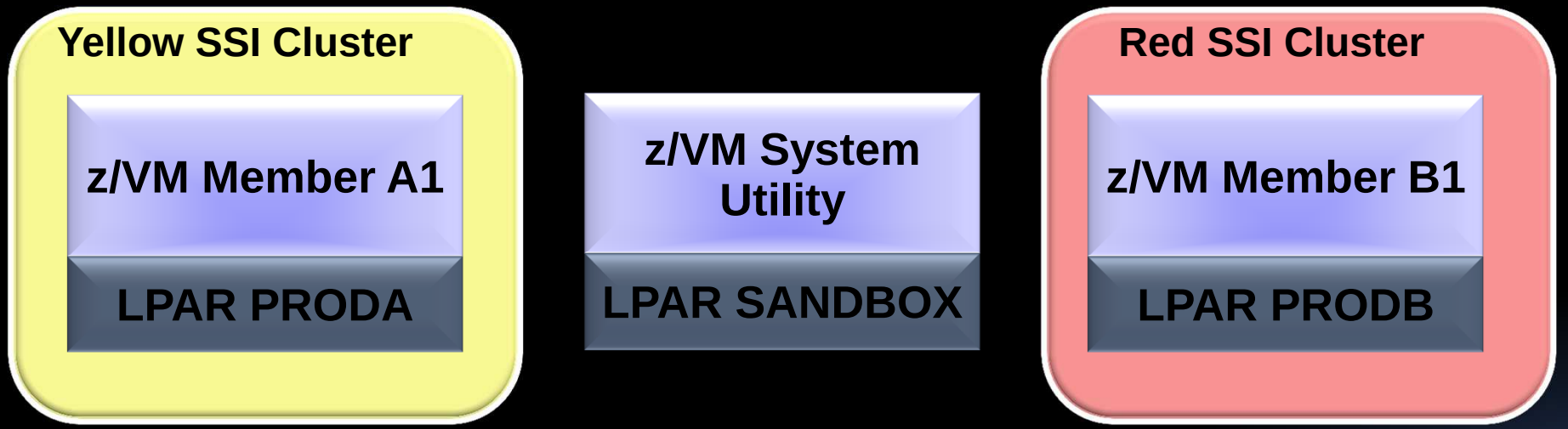


One From the Customers – Utility Migration LPAR



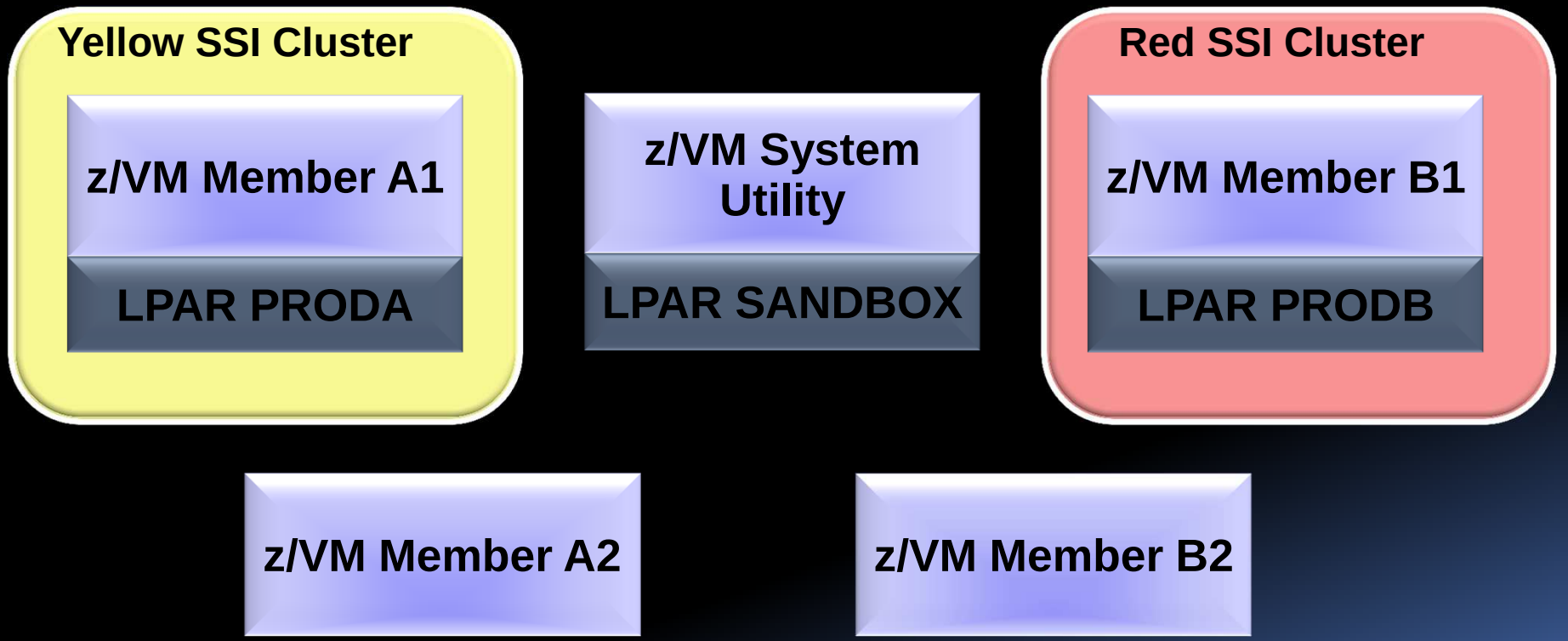
One From the Customers – Utility Migration LPAR

- Create SSI Cluster for each production System
  - Two Two-Member Clusters
  - But only include one of the production LPARs in each
- Utility System can stay a singleton or even a non-SSI system



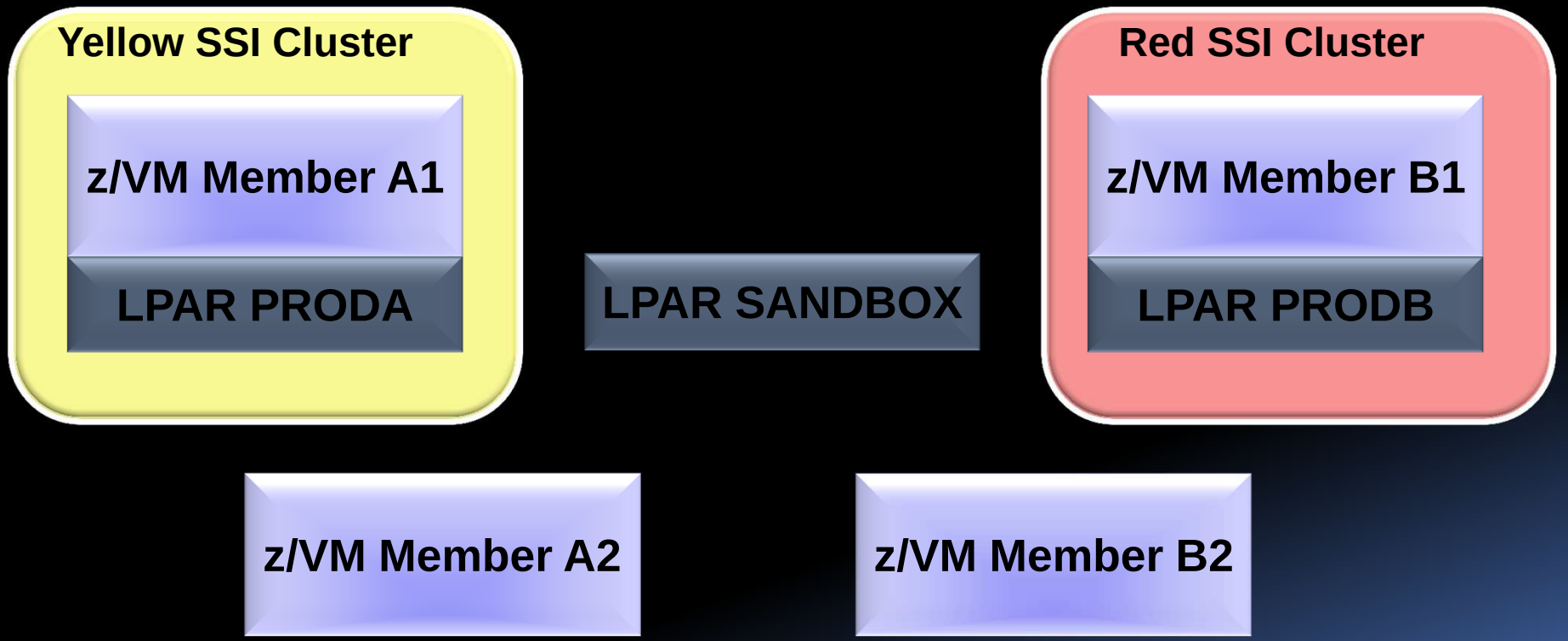
One From the Customers – Utility Migration LPAR

- Clone the production members so there is a second system (member) for each of the production LPARs.



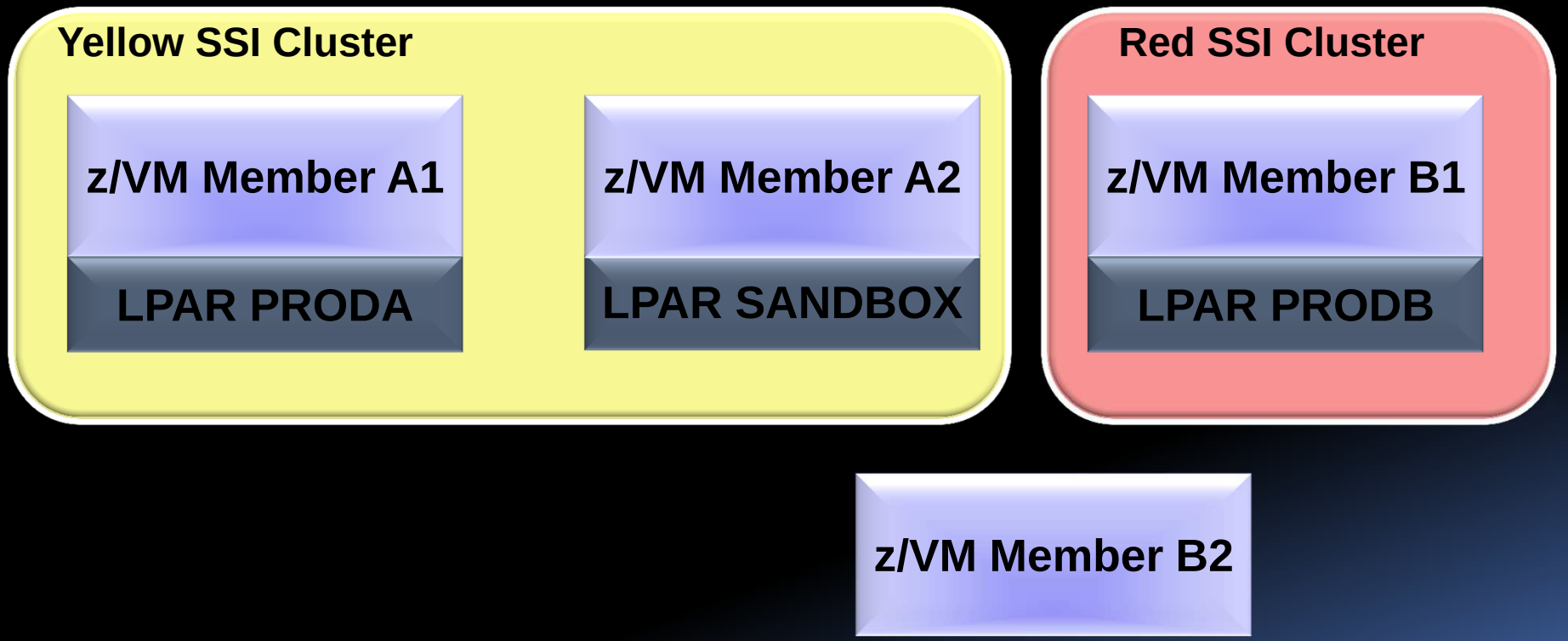
One From the Customers – Utility Migration LPAR

- To update CP on production LPAR PRODA
  1. Shutdown Utility System



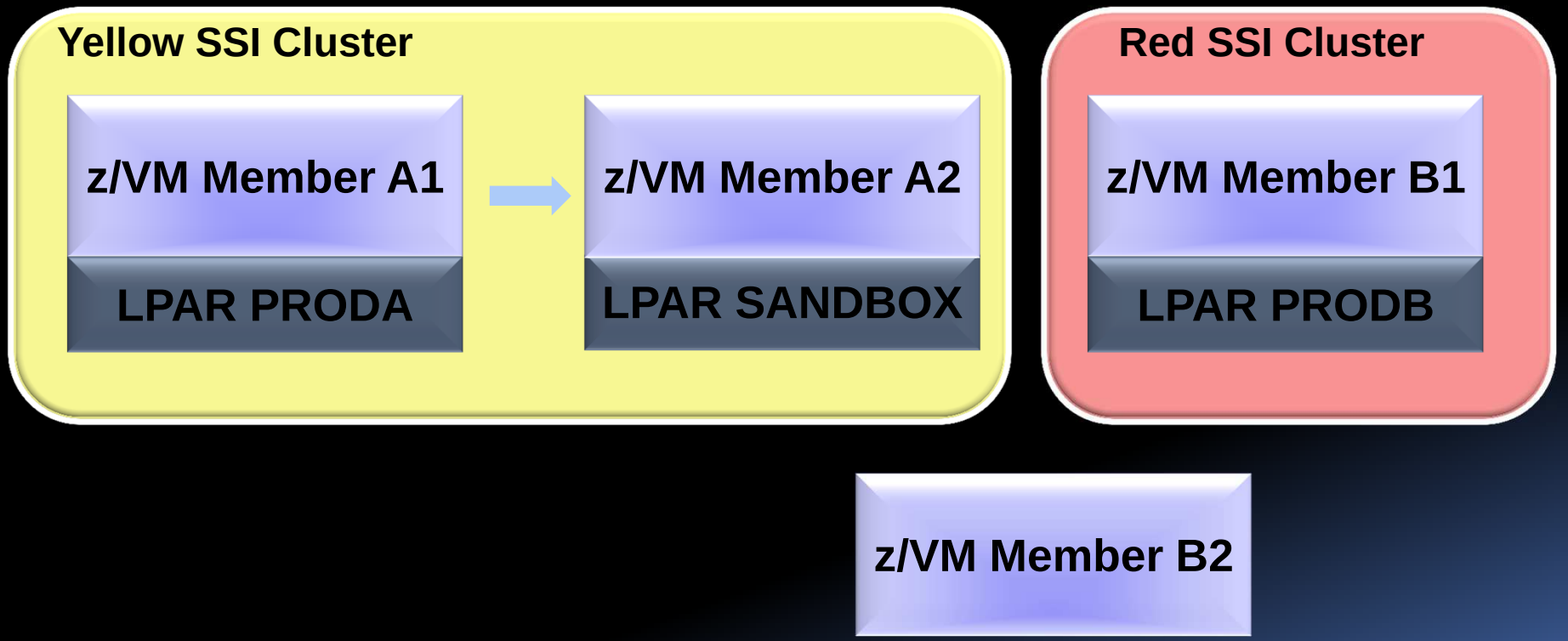
One From the Customers – Utility Migration LPAR

- To update CP on production LPAR PRODA
  1. Shutdown Utility System
  2. Bring up the other Member in SANDBOX LPAR



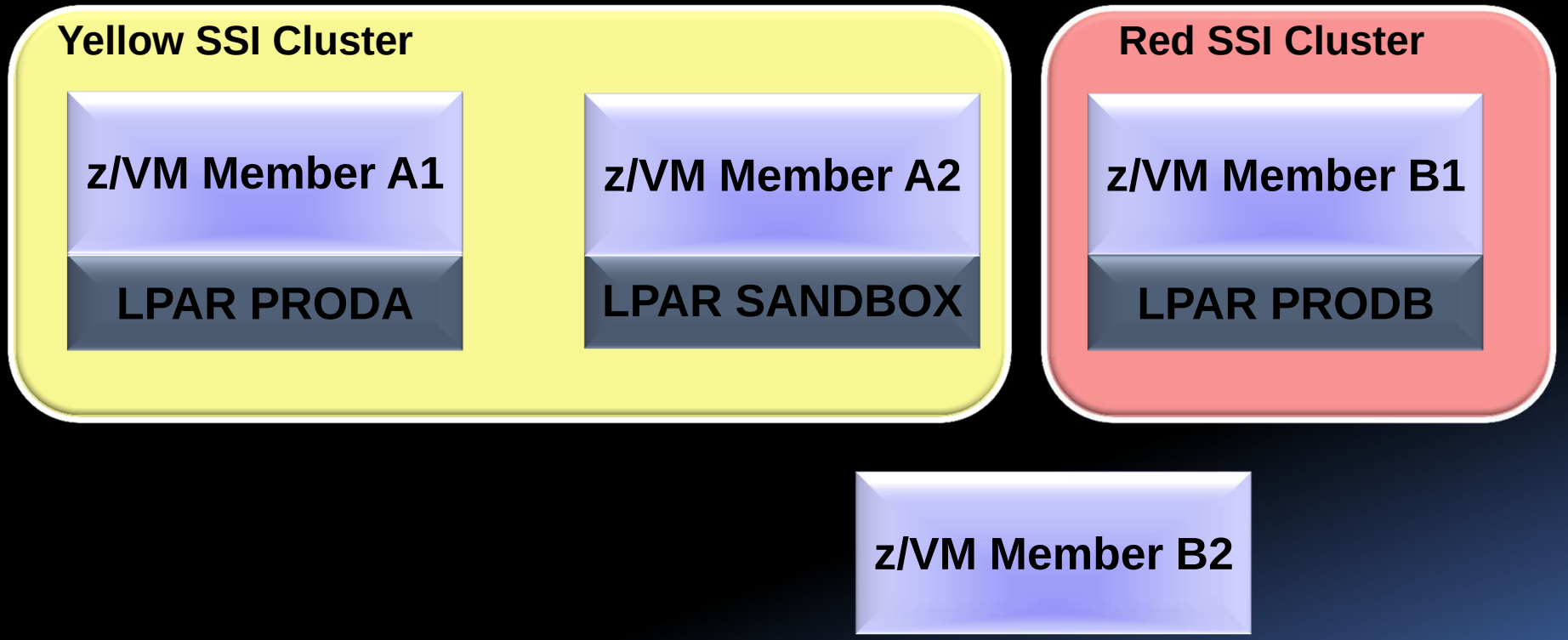
One From the Customers – Utility Migration LPAR

- To update CP on production LPAR PRODA
  1. Shutdown Utility System
  2. Bring up the other Member in SANDBOX LPAR
  3. Move work from A1 to A2



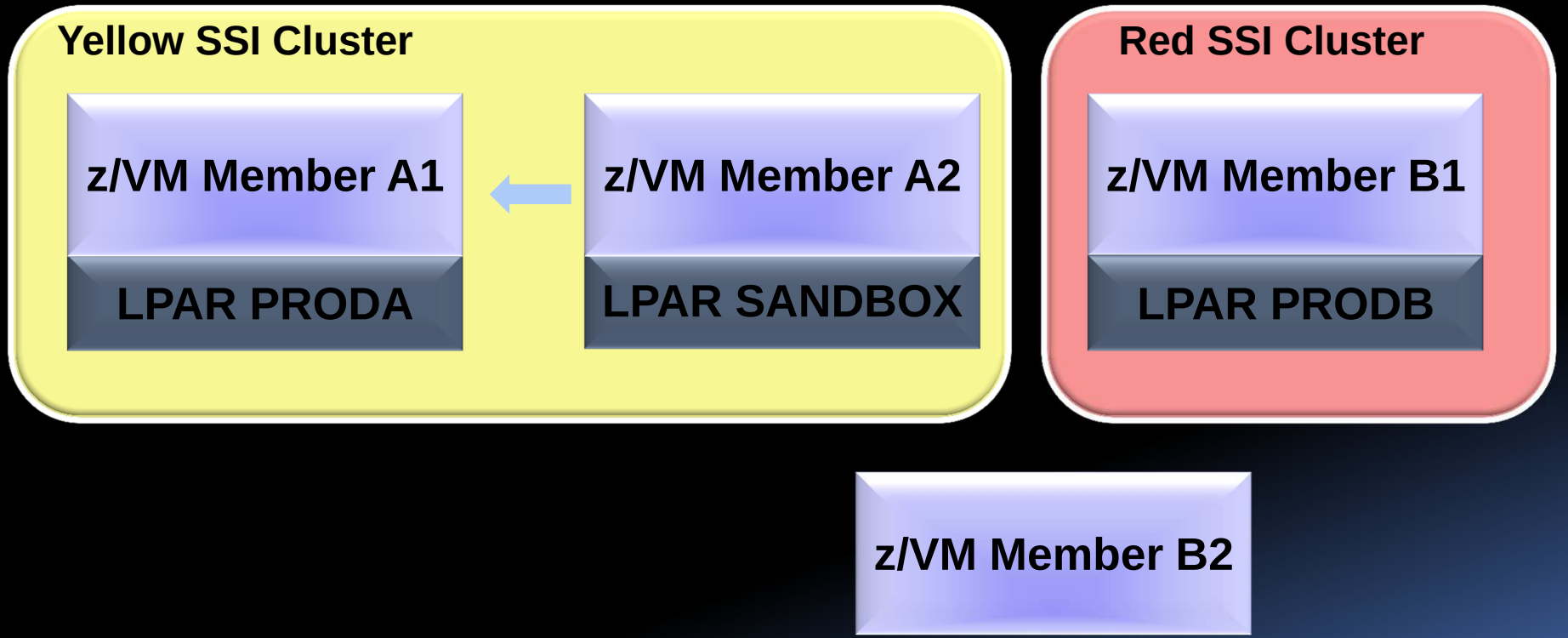
One From the Customers – Utility Migration LPAR

- To update CP on production LPAR PRODA
  1. Shutdown Utility System
  2. Bring up the other Member in SANDBOX LPAR
  3. Move work from A1 to A2
  4. Bounce A1 to pick up service



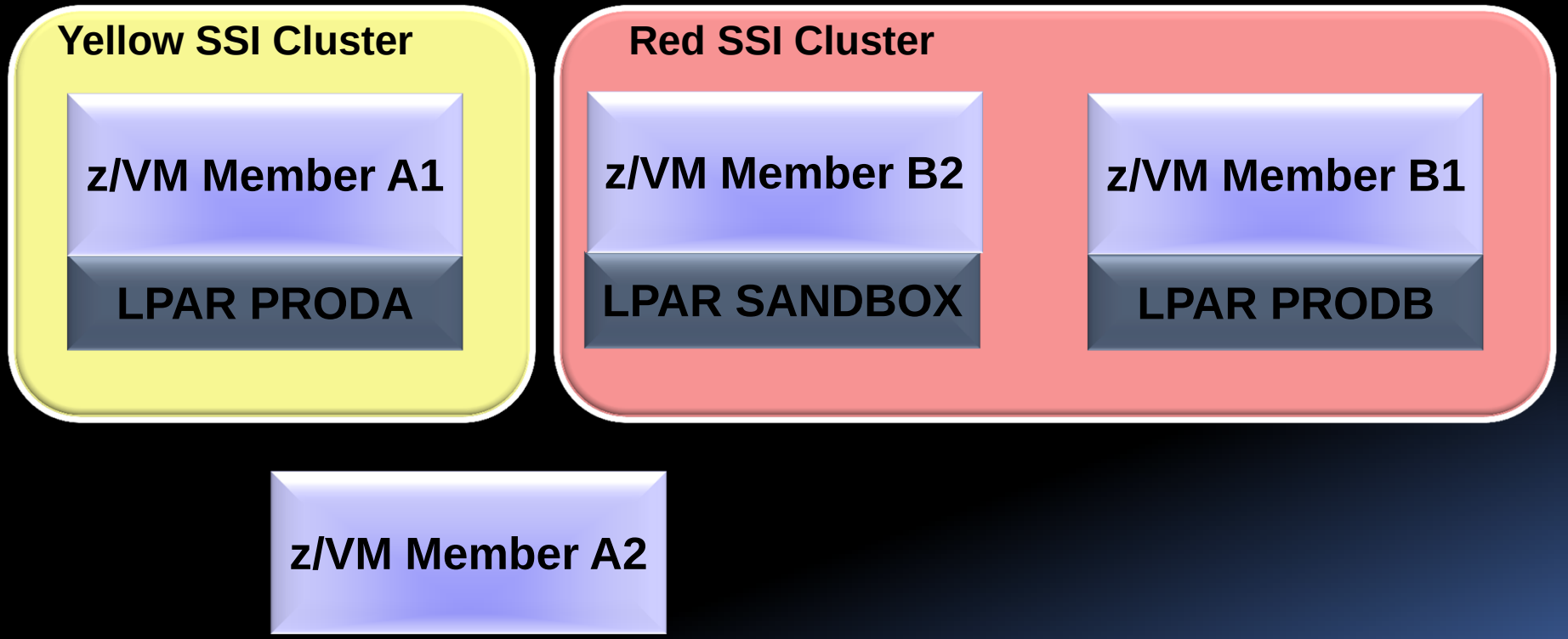
One From the Customers – Utility Migration LPAR

- To update CP on production LPAR PRODA
  1. Shutdown Utility System
  2. Bring up the other Member in SANDBOX LPAR
  3. Move work from A1 to A2
  4. Bounce A1 to pick up service
  5. Move work back to A1 from A2



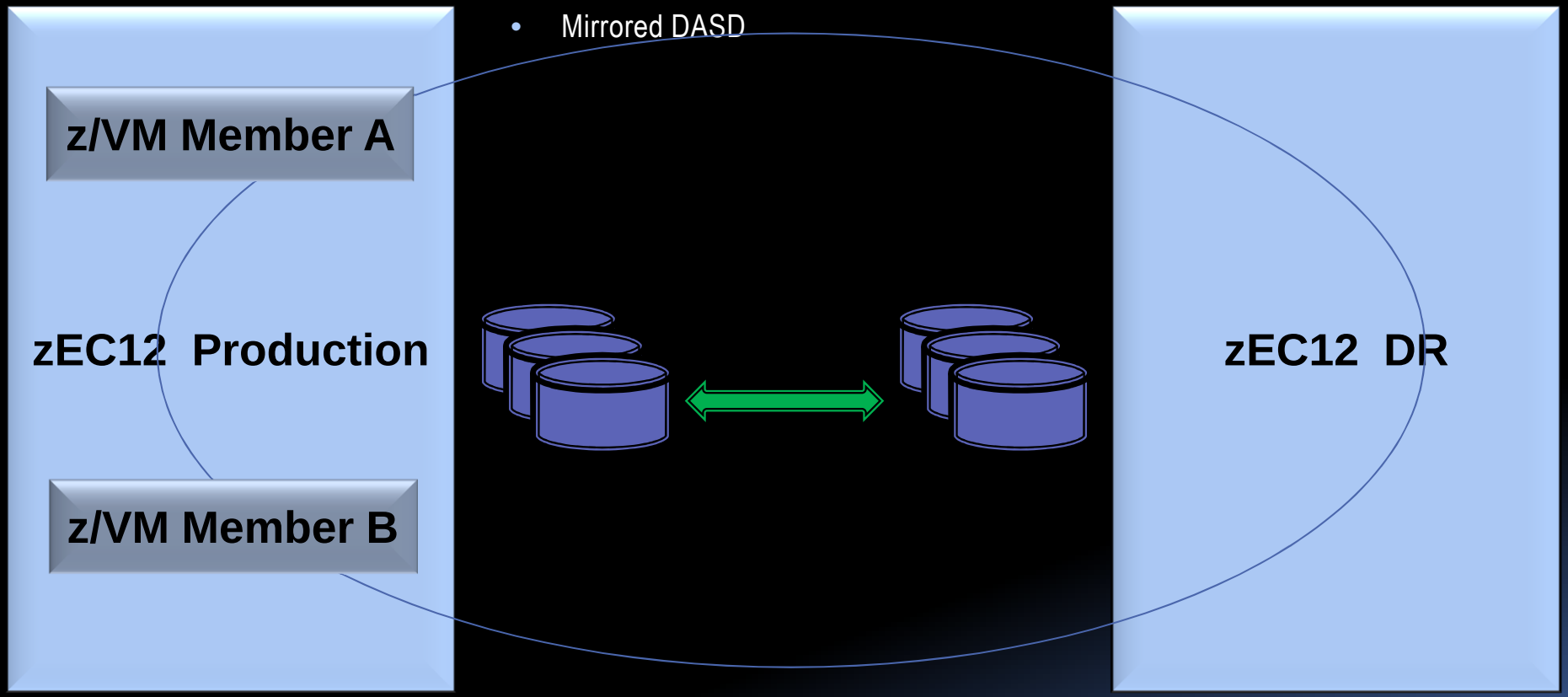
One From the Customers – Utility Migration LPAR

- Repeat on Red SSI Cluster



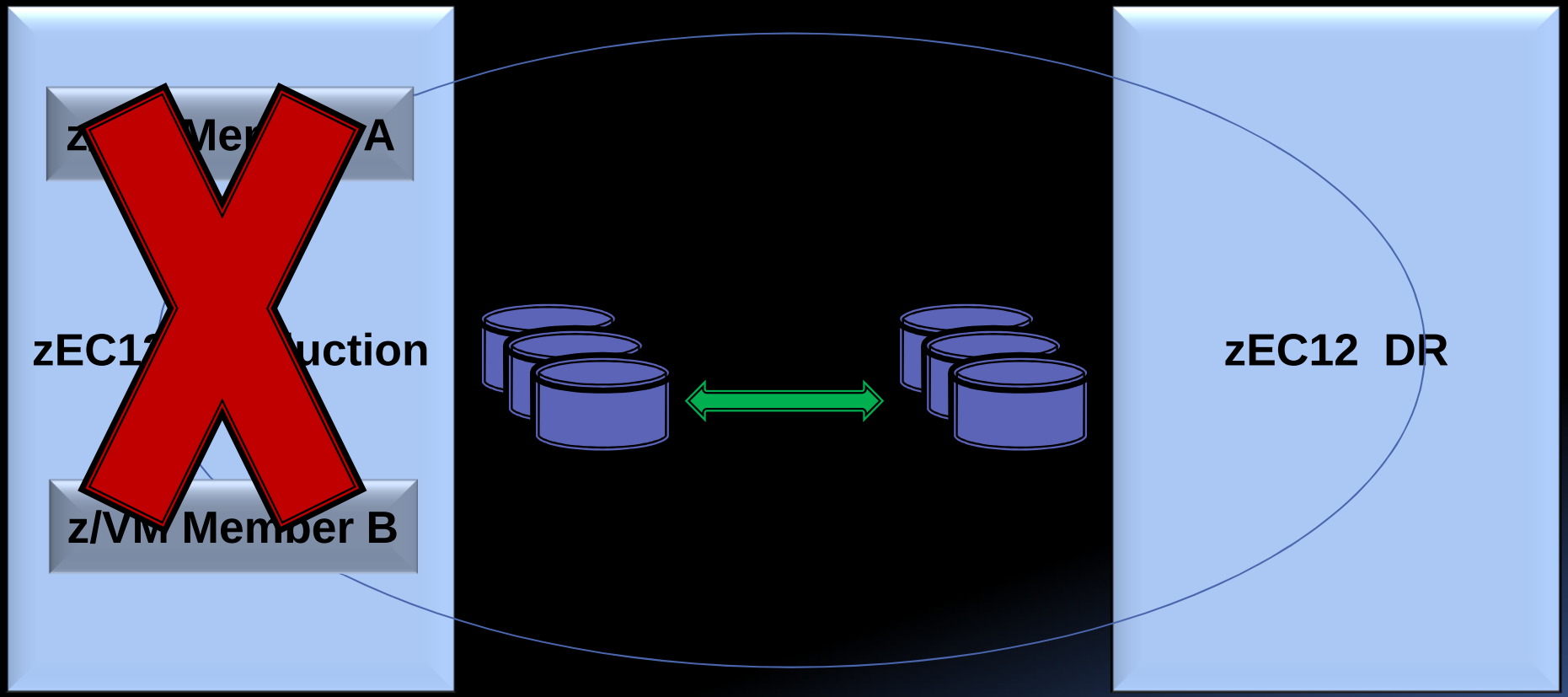
Local Disaster Recover (Business Continuity)

- Four Members Defined:
  - 2 Members active in production (A & B)
  - 2 Members standby in DR (C & D)
  - Mirrored DASD



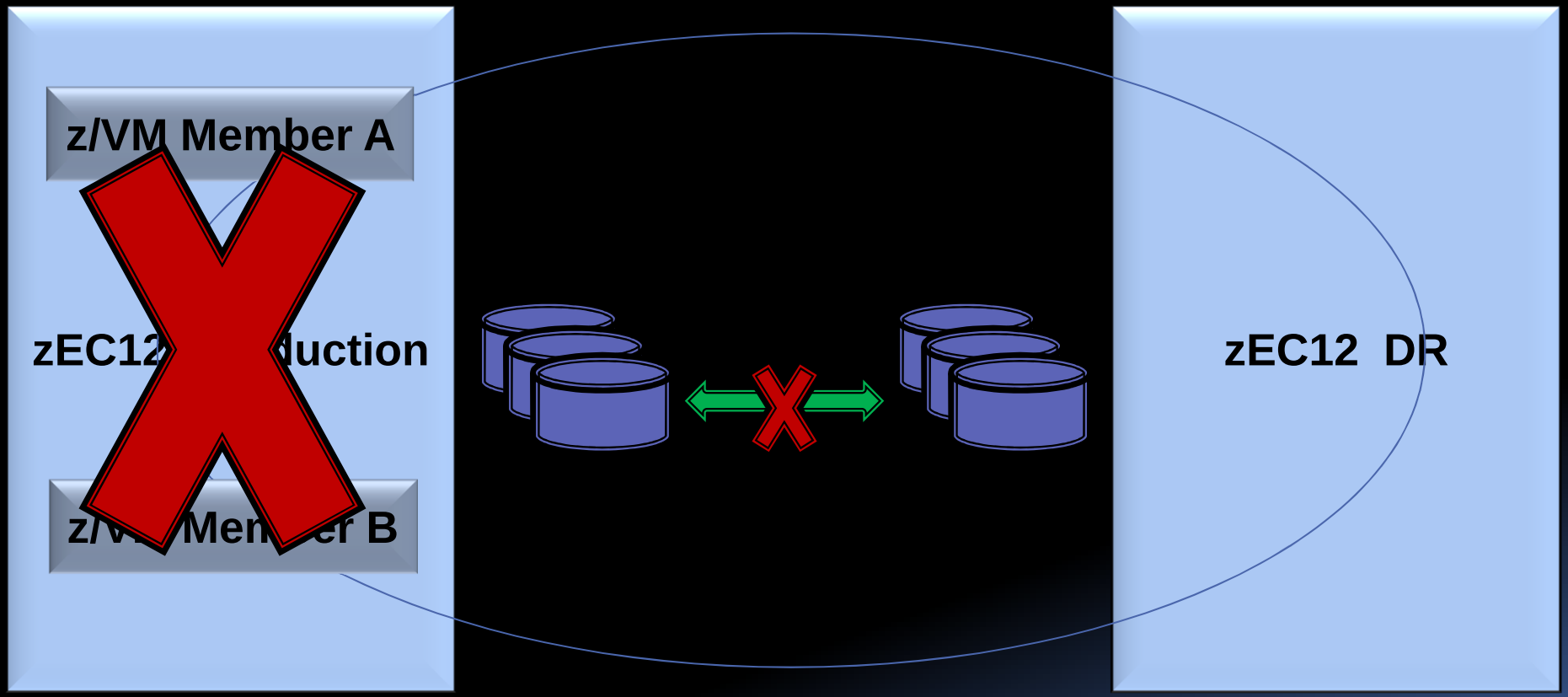
Local Disaster Recover (Business Continuity)

- Assume Production Side goes down



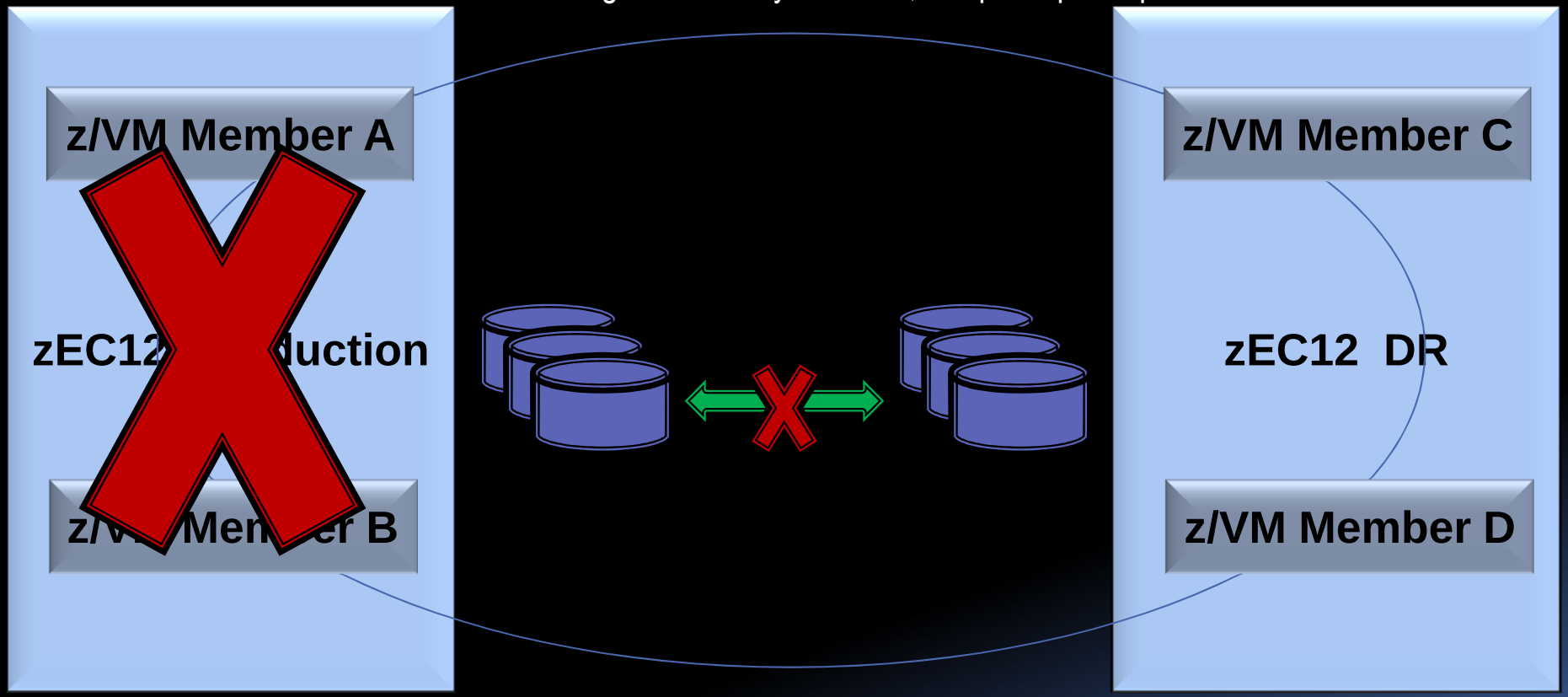
Local Disaster Recover (Business Continuity)

- Assume Production Side goes down
- Sever mirroring of DASD



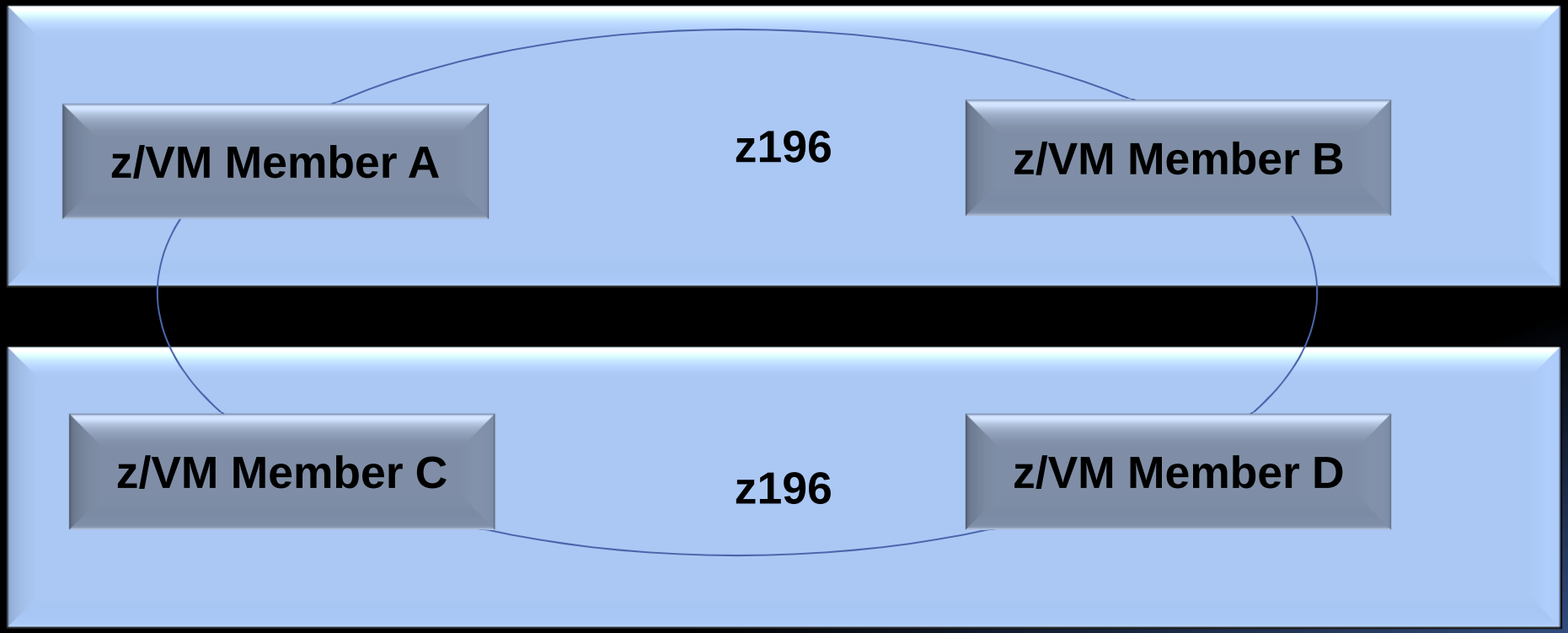
Local Disaster Recover (Business Continuity)

- Bring up Member C & D
- Logon virtual machines (shared directory)
- Not a High Availability Solution, but perhaps helpful.



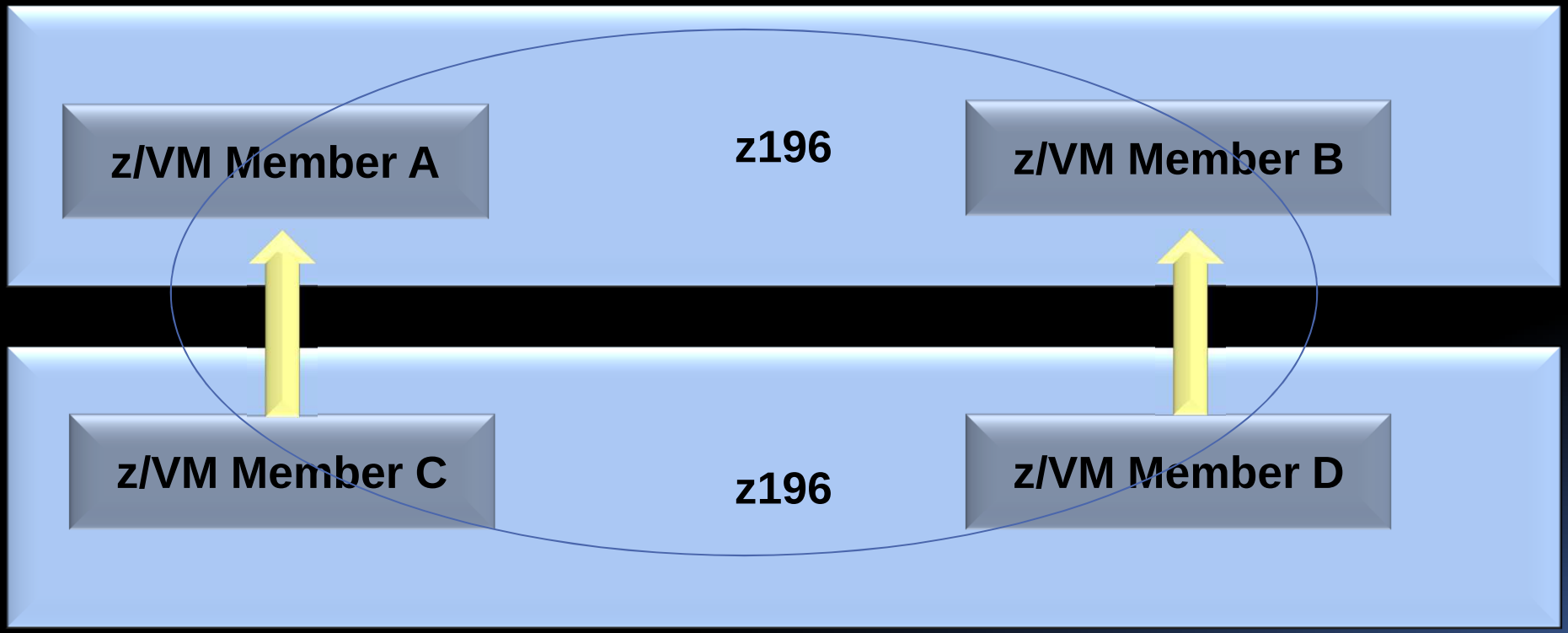
Migrate to New Processors

- Four Members Defined:
  - 2 Members on each of 2 CECs



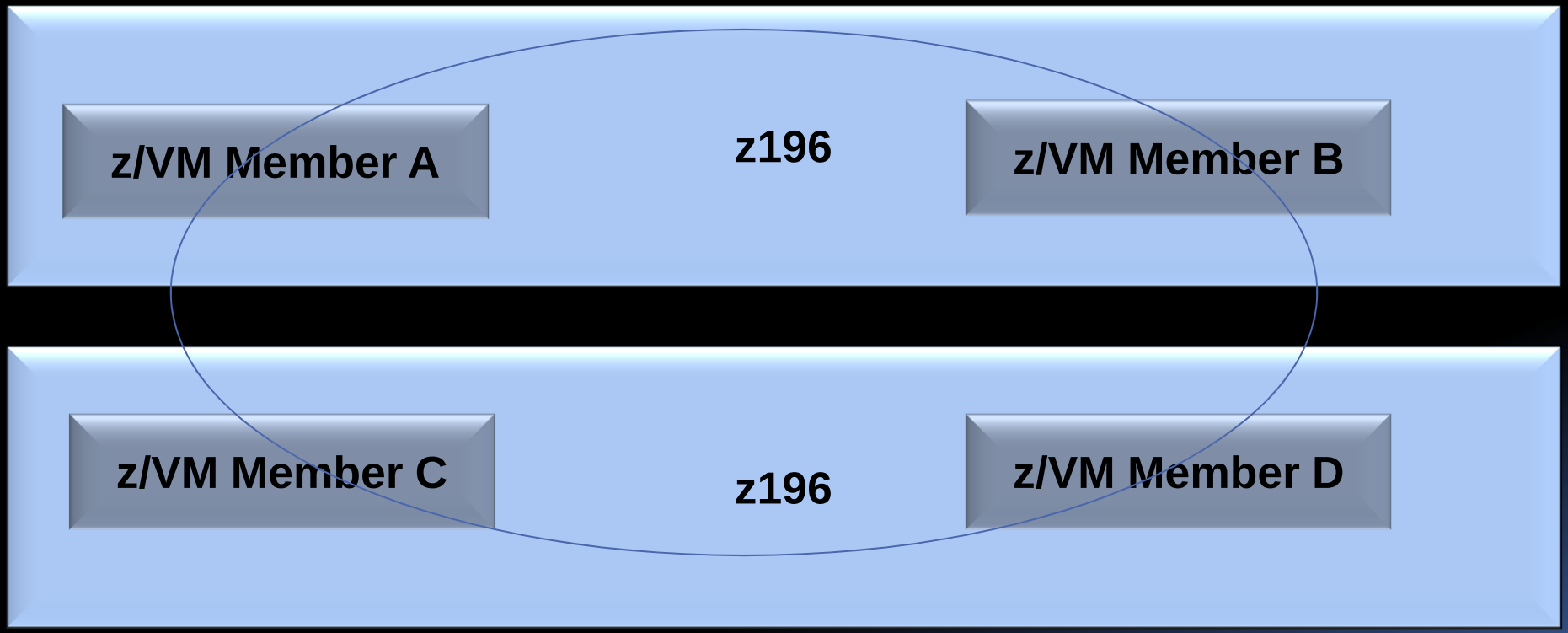
Migrate to New Processors

- Move work off of second z196 to first z196, unto just Members A & B



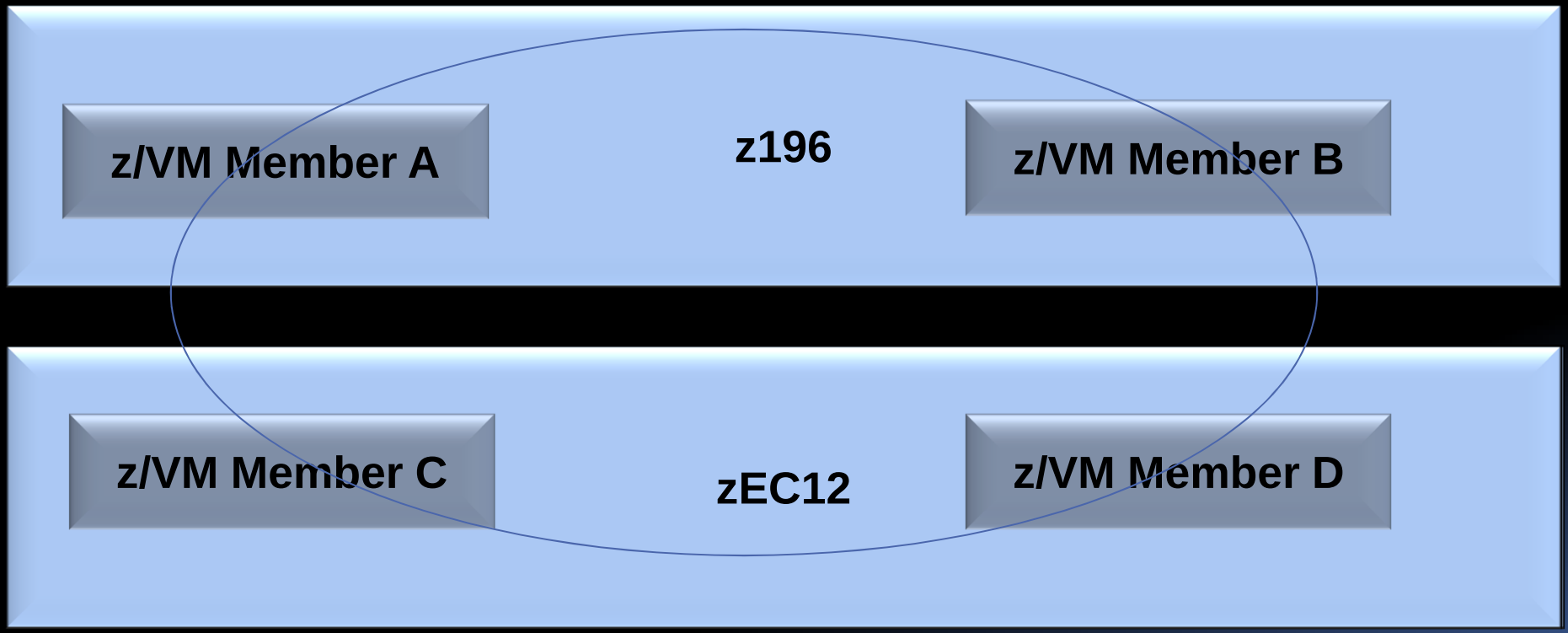
Migrate to New Processors

- Move work off of second z10 to first z196, unto just Members A & B
- Shutdown Members C & D



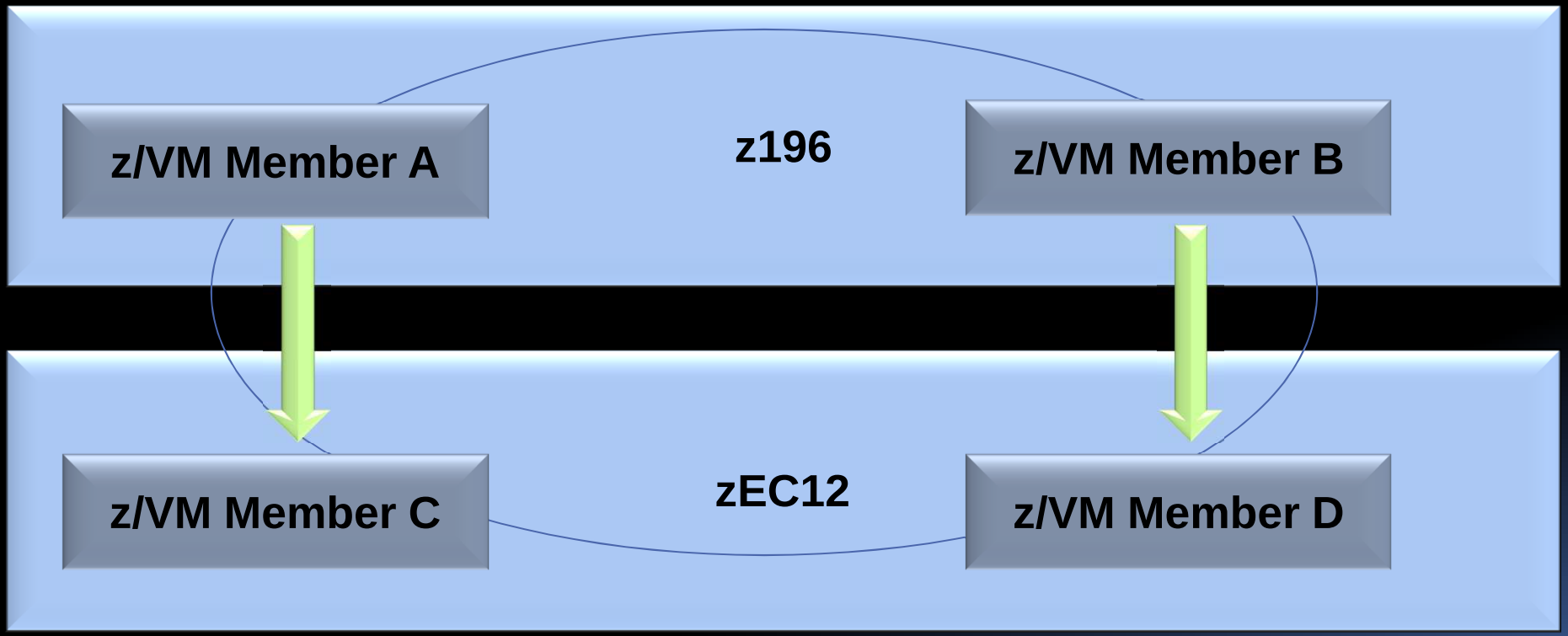
Migrate to New Processors

- Push out z196 and pull in the new zEC12
- Start up Members C & D on the new zEC12



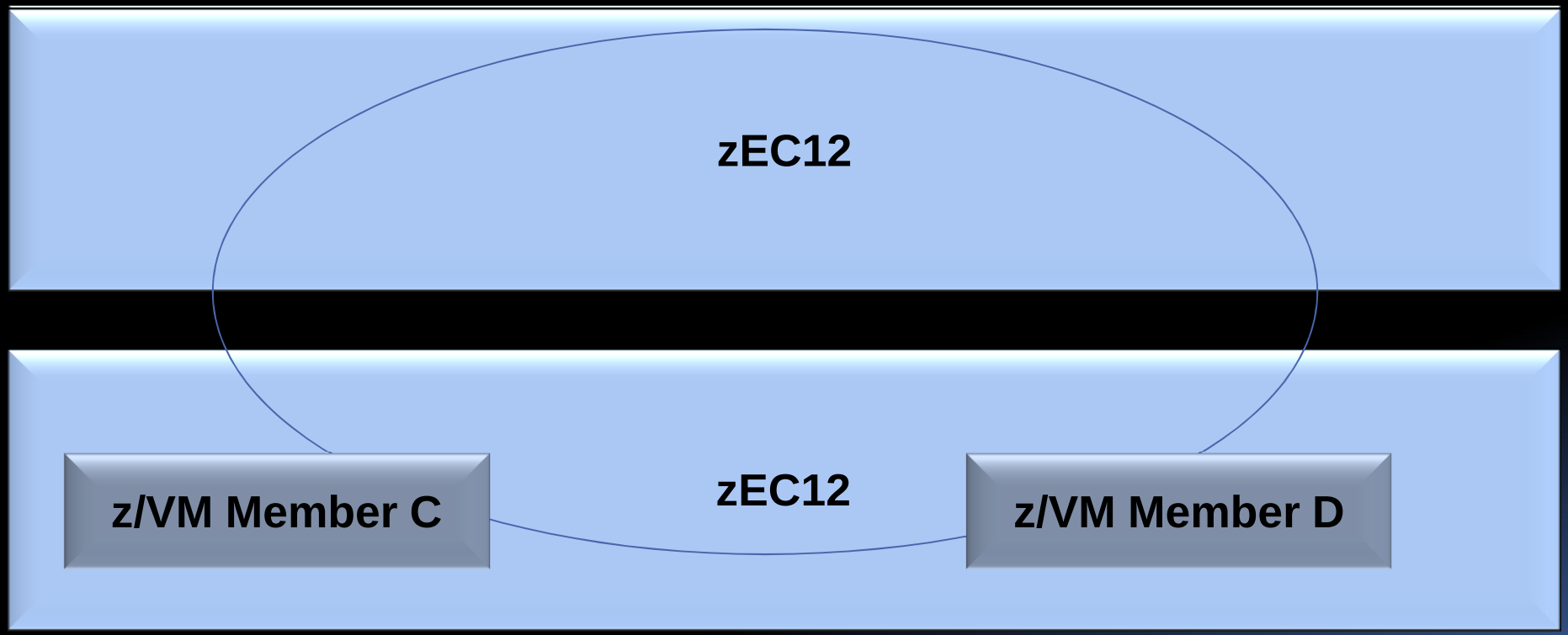
Migrate to New Processor

- Now, move Member A and B workloads to the Members C and D.



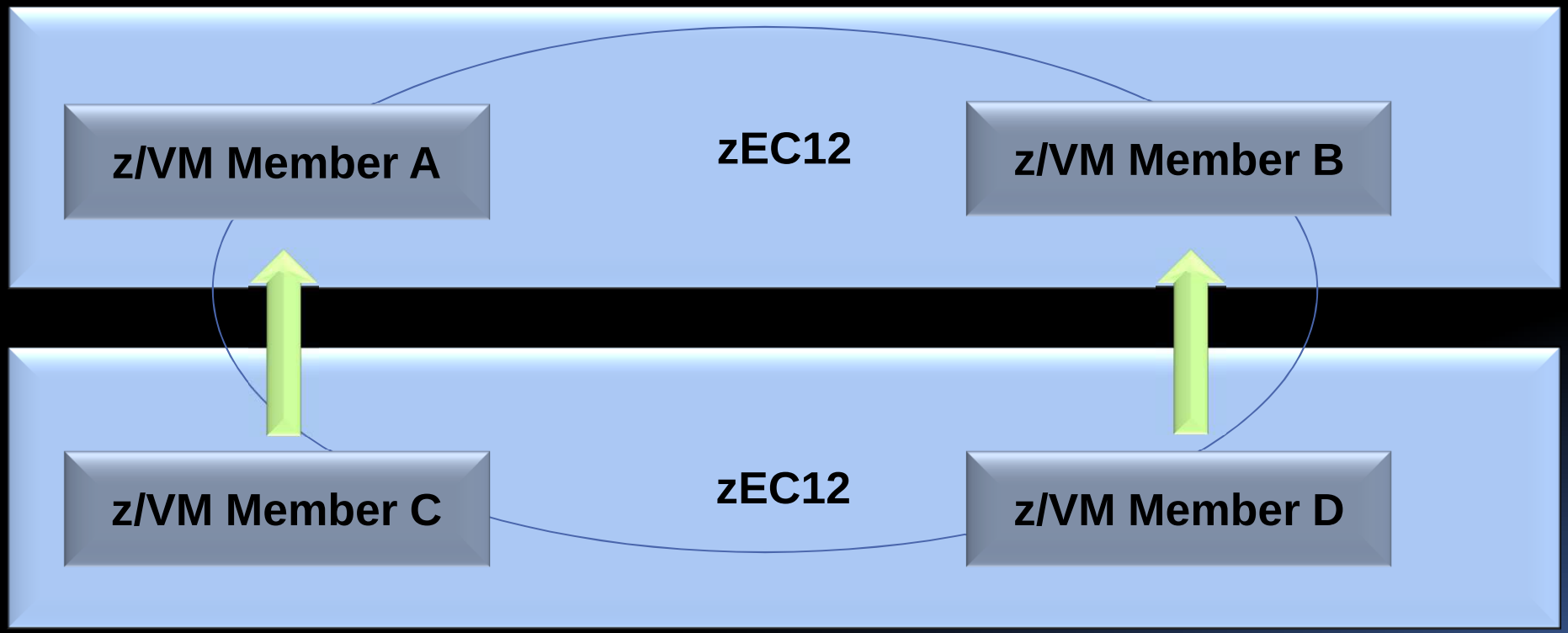
Migrate to New Processor

- Shutdown Members A and B
- Pull out old z196
- Push in new zEC12



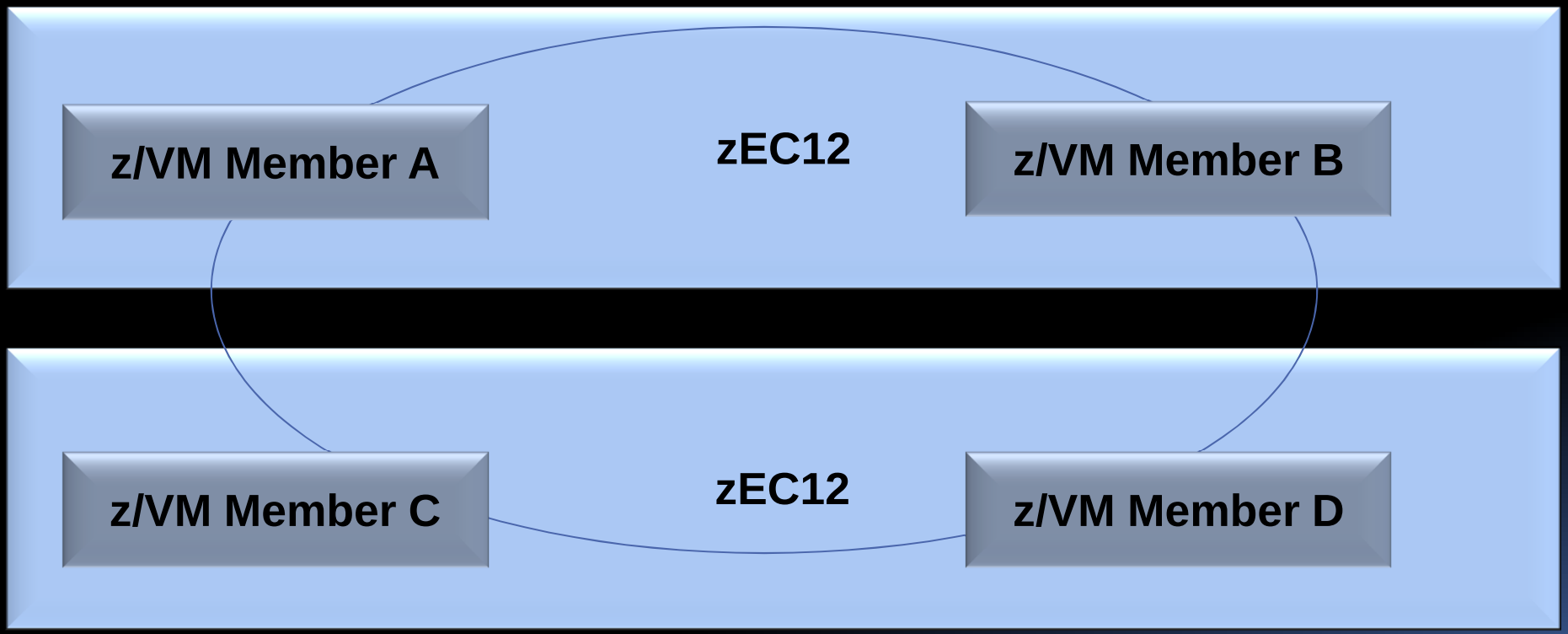
Migrate to New Processor

- Bring back up Members A and B
- Move workloads back to Members A & B



Migrate to New Processor

- Running on new processors without shutting down servers!!
- Would need to re-boot Linux to pick up new zEC12 hardware facilities.



Summary

Summary: z/VM 6.2 – Another Milestone for Virtualization

Manage Resources & Workloads

- For decades, System z has shown the strength of moving resources to the work that needed it. SSI and LGR add more value by allowing work to move to the resources in a non-disruptive manner.

Optimize Success

- The SSI clustering takes advantage of hardware and software technology to optimize success by minimizing the complex system programmer steps required for clustering technology, with low overhead and without specialized hardware.

Protect the Advantage

- Guest mobility in general is remarkable technology. z/VM Live Guest Relocation takes it to the next level. Exploiting LGR doesn't mean giving up the rich resource control and management features customers have come to love with z/VM.

Contact Info:

Bill Bitner
z/VM Customer Focus and Care
z/VM Development Lab – Endicott, NY
bitnerb@us.ibm.com
+1 607-429-3286

Trademarks

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

IBM*	System z10*	System z196
IBM Logo*	Tivoli*	System z114
DB2*	z10 BC	
Dynamic Infrastructure*	z9*	
GDPS*	z/OS*	
HiperSockets	z/VM*	
Parallel Sysplex*	z/VSE	
RACF*	zEnterprise*	
System z*		

* Registered trademarks of IBM Corporation

The following are trademarks or registered trademarks of other companies.

OpenSolaris, 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.
INFINIBAND, InfiniBand Trade Association and the INFINIBAND design marks are trademarks and/or service marks of the INFINIBAND Trade Association.
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.

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.

Notice Regarding Specialty Engines (e.g., zIIPs, zAAPs and IFLs):

Any information contained in this document regarding Specialty Engines ("SEs") and SE eligible workloads 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). 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 SEs 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.