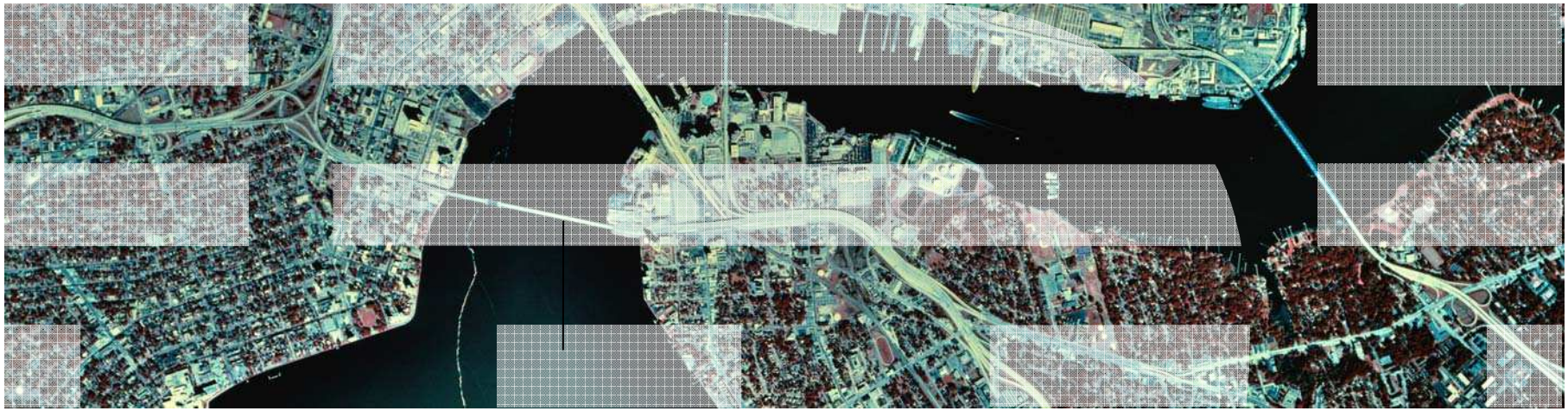
z/VM Live Guest Relocation - Planning and Use

John Franciscovich
francisj@us.ibm.com

Emily Kate Hugenbruch
ekhugen@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™ System z196 System z114

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.

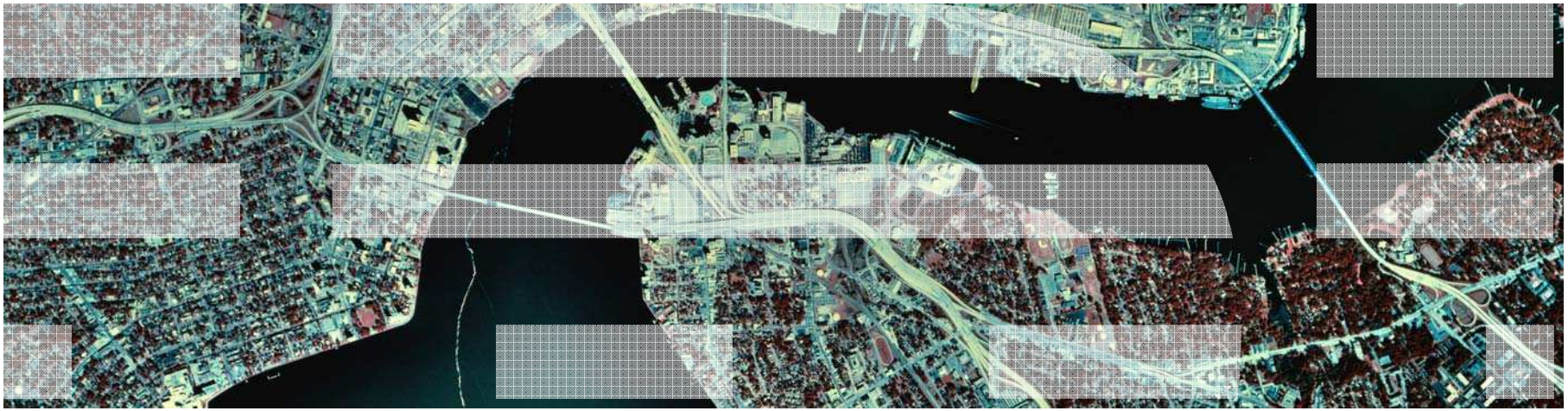
In this document, any references made to an IBM licensed program are not intended to state or imply that only IBM's licensed program may be used; any functionally equivalent program may be used instead.

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.

It is possible that this material may contain reference to, or information about, IBM products (machines and programs), programming, or services that are not announced in your country. Such references or information must not be construed to mean that IBM intends to announce such IBM products, programming or services in your country.

Agenda

- Planning for Live Guest Relocation (LGR)
- Relocation Domains
- Performing Live Guest Relocations
- Helpful Hints



Planning
for
Live Guest Relocation

General Guidelines for Relocating a Guest

Make sure all resources used by the virtual machine are available on the destination member

- Devices
- Facilities (will be handled automatically if you are relocating within a domain)
- Crypto cards
- Capacity for the virtual machine's memory and processor requirements
- Equivalency ids (**EQID**s) are defined for devices that need them
  - OSAs and FCPs
- Make sure that the devices really are equivalent
    - OSAs should be connected to the same LAN segment
    - FCPs should have access to the same SAN fabric
    - WWPNs and LUNs
  - If possible, use the same device numbers to refer to equivalent devices
- If connected to a VSWITCH, make sure the same VSWITCH is defined on the destination and the OSAs have been assigned EQIDs.
- If the virtual machine has an FCP, make sure the “queue_if_no_path” option is specified in Linux
- **OPTION CHPIDVIRTUALIZATION ONE** should be specified in guest's directory entry

Guest Configuration for Live Guest Relocation

- In order to be eligible to relocate, a guest must be:
  - Defined as a single configuration virtual machine
  - Running in an ESA or XA virtual machine in ESA/390 or z/Architecture mode
  - Logged on and 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 virtual disks in storage (created with DEFINE VFB-512 command)
  - 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 (APPLDATA) collection
    - If guest buffer is not in a shared DCSS
  - Single Console Image Facility
  - Collaborative Memory Management Assist (CMMA)

Memory Requirements for Live Guest Relocation

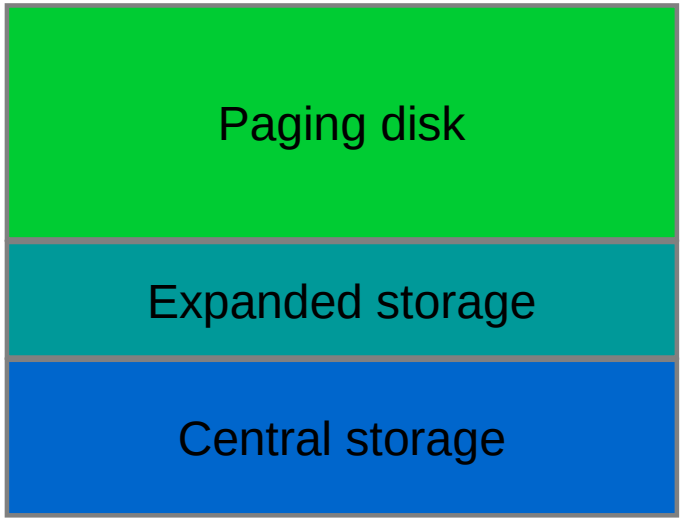
- A relocating guest's current memory size ***must*** fit in available space on the destination member

Guest's Current Memory Size

Virtual memory fully populated, including
- Private Vdisks
- Estimated size of supporting CP structures

<

Available space
- sum of available memory



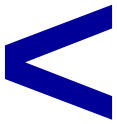
Memory Requirements for Live Guest Relocation…

- Additional checks

1. Does the guest's current memory size exceed paging capacity on the destination?

Guest's Current Memory Size

Virtual memory fully populated, including
- Private Vdisks
- Estimated size of supporting CP structures



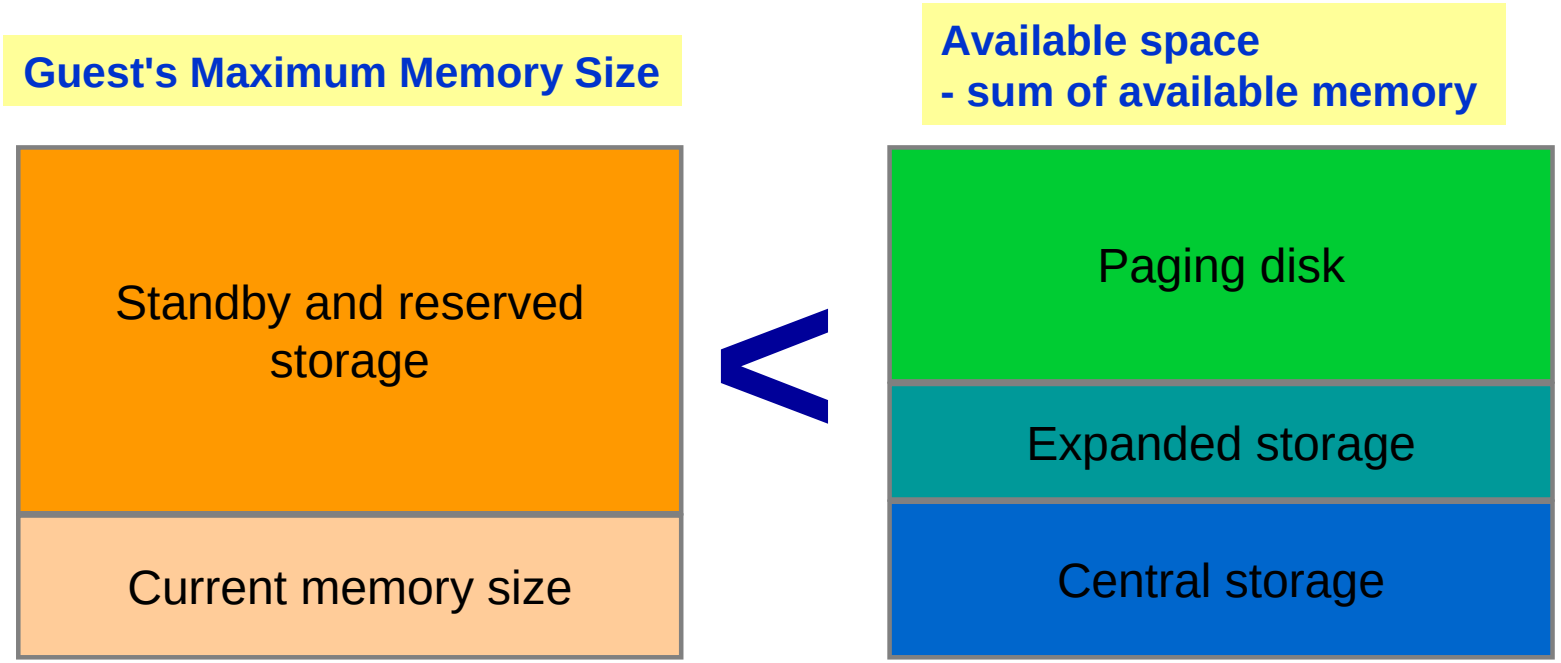
Paging disk capacity

May be overridden if you are certain that this is not applicable to your environment

Memory Requirements for Live Guest Relocation…

- Additional checks

2. Does the guest's maximum memory size exceed available space on the destination?

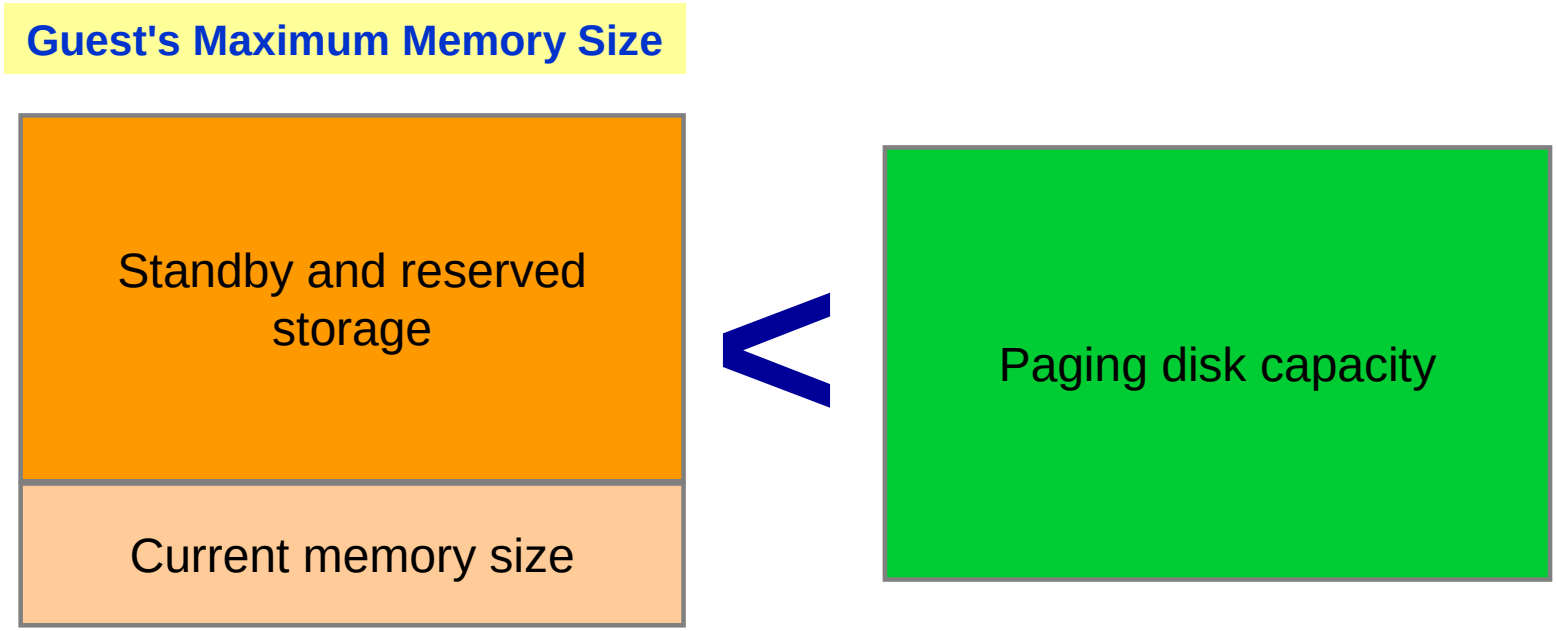


May be overridden if you are certain that this is not applicable to your environment

Memory Requirements for Live Guest Relocation…

- Additional checks

3. Does the guest's maximum memory size exceed paging capacity on the destination?



May be overridden if you are certain that this is 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

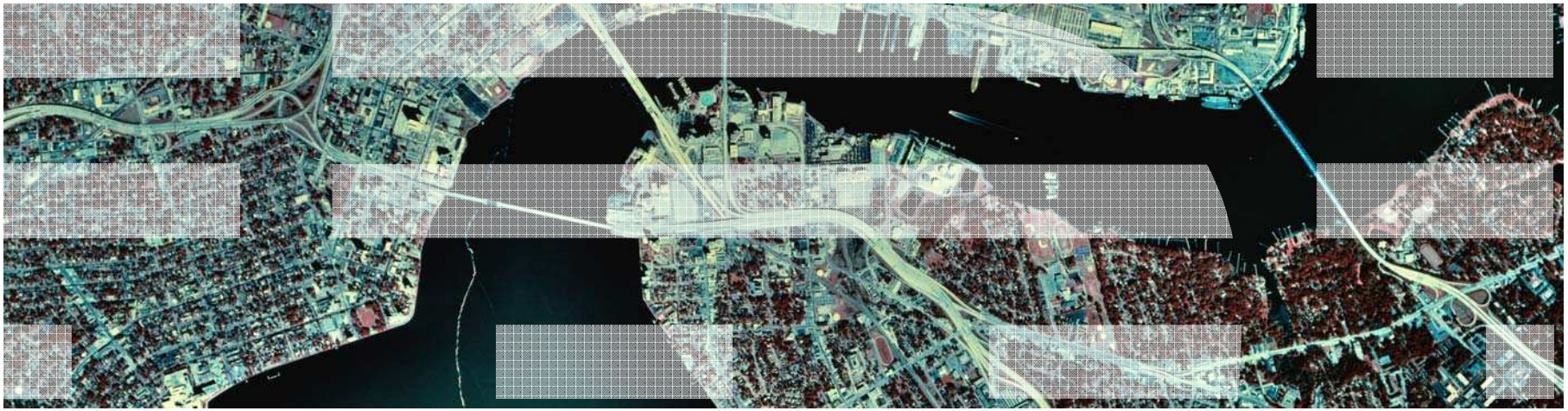
- Relocations may increase paging demand
  - Available paging space should be at least 2x total virtual memory of all guests
    - Including guests to be relocated to this member
  - Avoid allocating more than 50% of available paging space
    - If size of guests to be relocated increase in-use amount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

```
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%
```

Conditions That Prevent a Relocation

- Conditions in the following categories could prevent a relocation from completing:
  - Guest State Conditions
  - Device Conditions
  - Device State Conditions
  - Virtual Facility Conditions
  - Configuration Conditions
  - Resource Limit Conditions
  - Other…
- Entire list of conditions documented in CP Planning and Administration
  - "Preparing for Live Guest Relocation in a z/VM SSI Cluster"

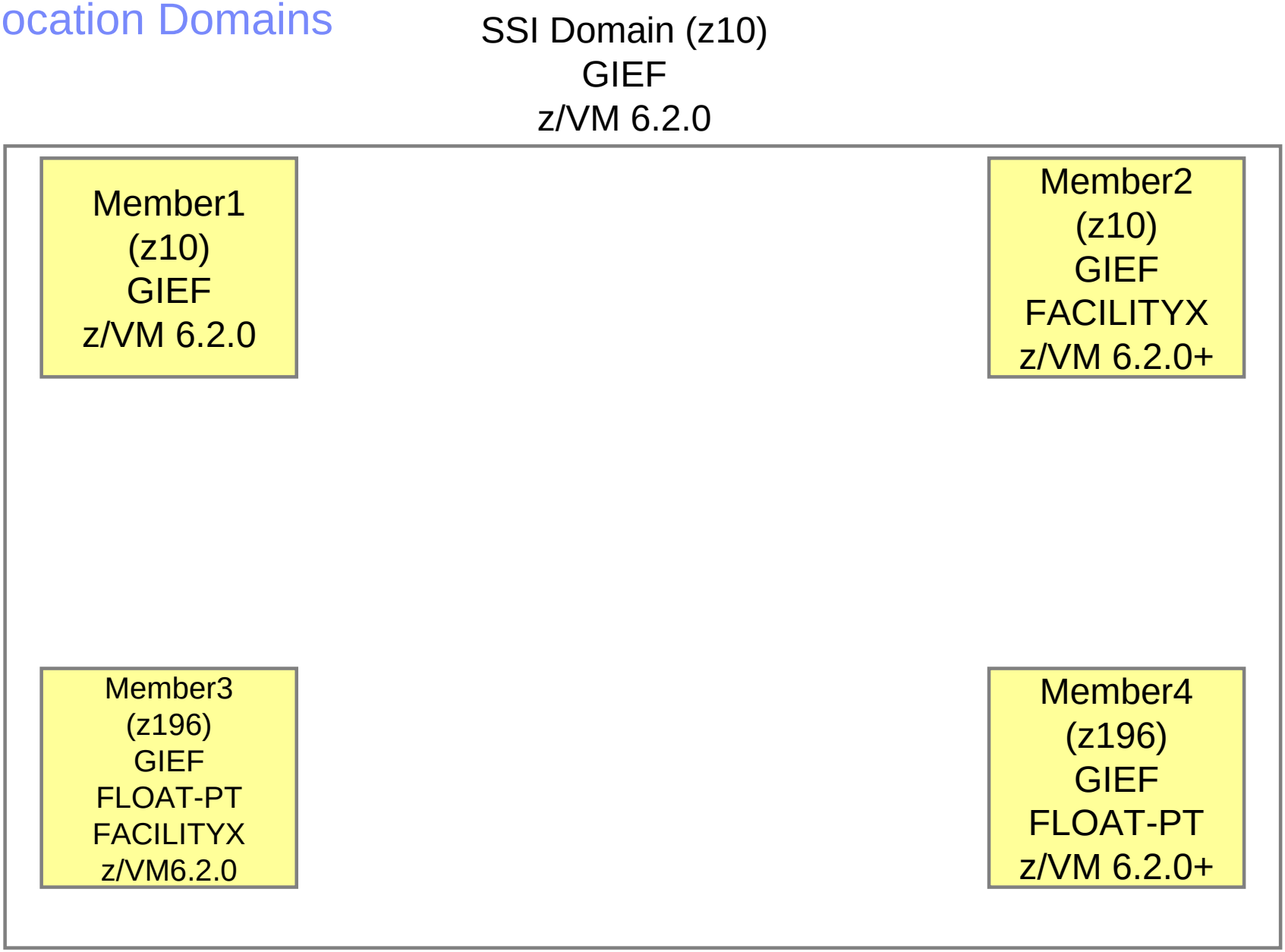


Relocation Domains

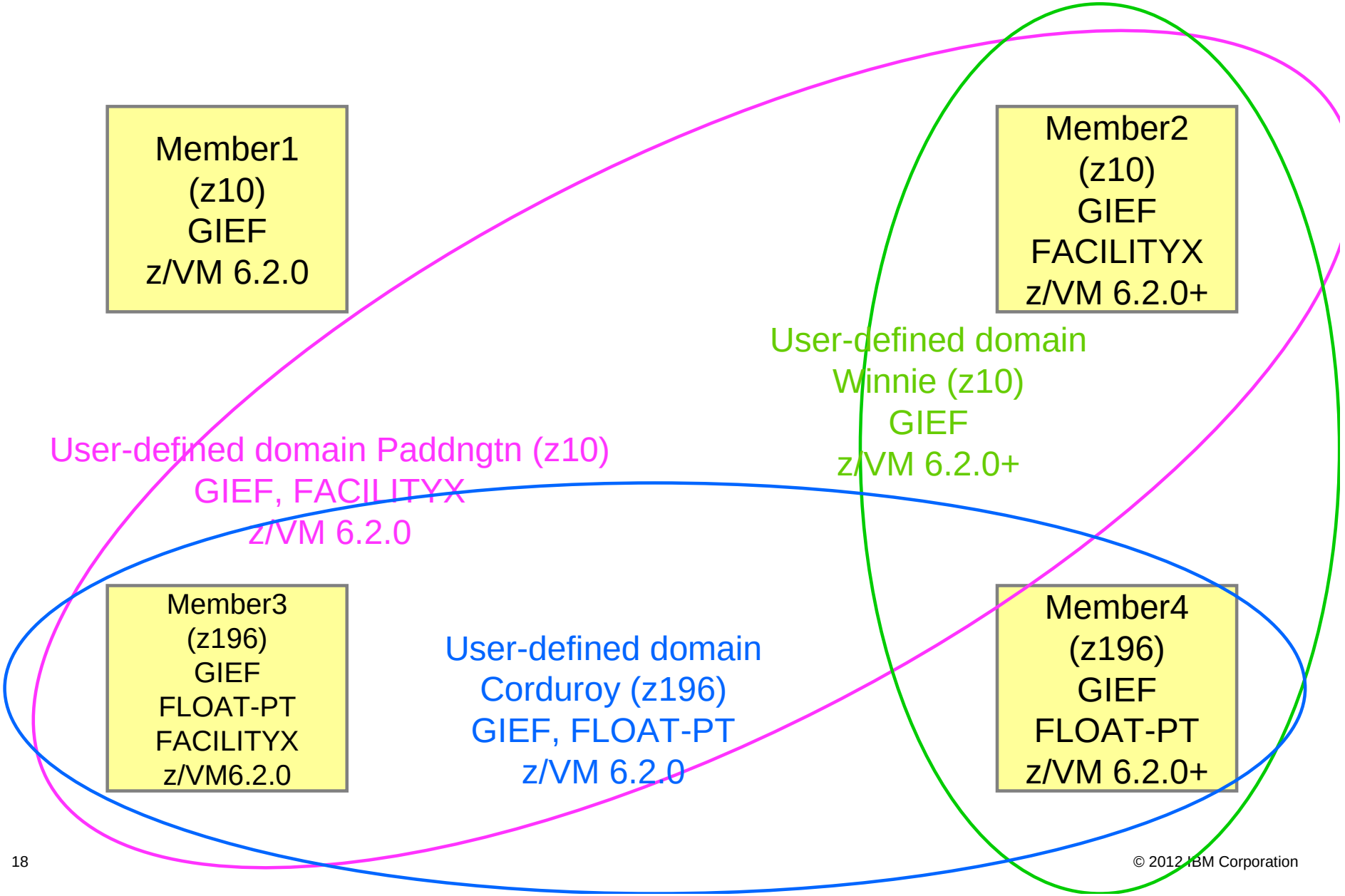
What is a Relocation Domain?

- A relocation domain defines a set of members of an SSI cluster among which virtual machines can relocate freely
- Relocation domains can be defined for business or technical reasons
- 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
  - An SSI domain that will have the features and facilities common to all members
- 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

Relocation Domains



Relocation Domains



Defining Relocation Domains

- In system configuration file:

```
88
89 RELOCATION_DOMAIN PADDNGTN MEMBER2 MEMBER3
90 RELOCATION_DOMAIN WINNIE MEMBER2 MEMBER4
91 RELOCATION_DOMAIN CORDUROY MEMBER3 MEMBER4
92
```

- Dynamically via a **DEFINE** command:

```
define relodomain paddngtn members member2 member3

define relodomain winnie members member2 member4

define relodomain corduroy members member3 member4
```

Assigning Relocation Domains

- Virtual machines may be assigned to a domain in their directory entry
  - Default for single configuration virtual machines is the SSI domain
  - Default for multiconfiguration virtual machines is their single member domain, which cannot be changed

- Virtual machines are assigned a virtual architecture level when they log on, according to what domain they are in

- They cannot use facilities or features not included in the domain even if the member they are on has access to those features
  - We call this “fencing”

- 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
  - **STFLE**

Assigning Relocation Domains - Directory

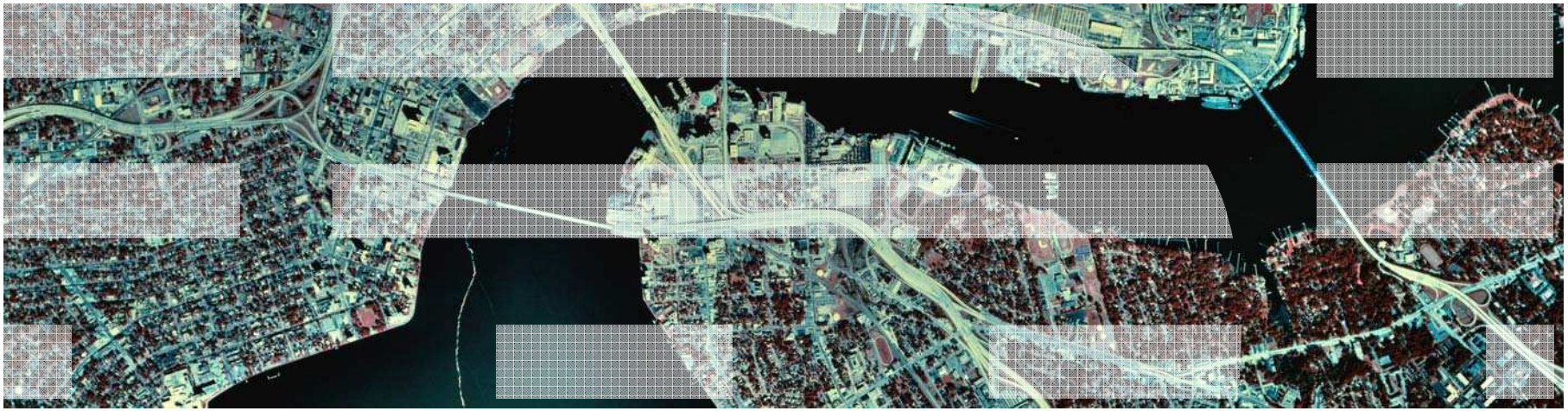
```
dirm for lgrrh56 vmrelocate on domain winnie
DVHXMT1191I Your VMRELOCATE request has been sent for processing to
DVHXMT1191I DIRMAINT at MEMBER1 via DIRMSAT2.
Ready; T=0.01/0.02 11:32:46
DVHREQ2288I Your VMRELOCATE request for LGRRH56
DVHREQ2288I at * has been accepted.
DVHBIU3450I The source for directory e
DVHBIU3450I LGRRH56 has been updated.
DVHBIU3424I The next ONLINE will take
DVHBIU3424I immediately.
DVHRLA3891I Your DSATCTL request has b
DVHRLA3891I for processing.
DVHRLA3891I Your DSATCTL request has b
DVHRLA3891I for processing.
DVHRLA3891I Your DSATCTL request has been relayed
DVHRLA3891I for processing.
DVHRLA3891I Your DMVCTL request has been relayed
DVHRLA3891I for processing.
DVHRLA3891I Your DMVCTL request has been relayed
DVHRLA3891I for processing.
DVHRLA3891I Your DMVCTL request has been relayed
DVHRLA3891I for processing.
DVHBIU3428I Changes made to directory entry LGRRH56
DVHBIU3428I have been placed online.
DVHREQ2289I Your VMRELOCATE request for LGRRH56
DVHREQ2289I at * has completed; with RC = 0.
```

```
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
```

Assigning Relocation Domains - Dynamic

- A running virtual machine may be dynamically reassigned to a domain with the same or greater facilities, so long as the member he is currently on has access to those facilities
- For example, a guest may be in the SSI domain, but relocate to a member with access to more facilities, so you may want to reassign him to a domain with higher facilities

```
set vmrelocate * domain ssi
Running on member GDLRCTS2
Relocation enabled in Domain SSI
Ready;
q cpuid
CPUID = FF3B6D8520978000
Ready;
define relodomain winnie gdlrcts1 gdlrcts2
Ready;
set vmrelocate * domain winnie
Running on member GDLRCTS2
Relocation enabled in Domain WINNIE
Ready;
q cpuid
CPUID = FF3B6D8528178000
Ready;
```



Live Guest Relocation

What to Know Before Starting Relocations

- CP moves the virtual machine in several different stages
- While the majority of the virtual machines' memory is being moved, the virtual machine is running
- Then the virtual machine is stopped (quiesced), checked again for eligibility and the final state is moved
- If there are any eligibility failures at any point until after the final state is moved, the relocation cancels and the virtual machine is resumed on the source member
- At any point until after the final state is moved, the relocation can be canceled
  - By the **VMRELOCATE CANCEL** command, from the source or destination
  - **CPHX** will cancel a **VMRELOCATE SYNC** command

What to Know Before Starting Relocations…

- Use the **VMRELOCATE TEST** command before you try a **VMRELOCATE MOVE**

- Choose one class A user to always issue your **VMRELOCATE** commands from and only issue one **VMRELOCATE** command at a time
  - Use the default option, **SYNCHRONOUS** to enforce one-at-a-time relocations

- Use the **AT** command to issue **VMRELOCATE**s on another member in your SSI cluster

- Know how long your Linux machine can be quiesced, look at applications and when they will timeout (30 seconds? 5 seconds?)
  - Use the **MAXQUIESCE** option to tell CP how long quiesce time can be
  - If this is exceeded, the relocation will be canceled and the virtual machine resumed on the source member

Live Guest Relocation – Example

```
q ssi
SSI Name: SSITEST
SSI Mode: Stable
Cross-System Timeouts: Enabled
SSI Persistent Data Record (PDR) device: FL4B84 on 4B84
SLOT SYSTEMID STATE      PDR HEARTBEAT        RECEIVED HEARTBEAT
   1 GDLLCPX1 Joined     2011-10-13 15:10:18 2011-10-13 15:10:18
   2 GDLLCPX2 Joined     2011-10-13 15:10:12 2011-10-13 15:10:12
   3 GDLLCPX3 Joined     2011-10-13 15:10:26 2011-10-13 15:10:26
   4 GDLMCPX4 Joined     2011-10-13 15:10:35 2011-10-13 15:10:35
Ready; T=0.01/0.01 15:10:41
```

Live Guest Relocation – Example

```
formssi display 141
HCPPDF6618I Persistent Data Record on device 0141 (label FL4B84) is for
HCPPDF6619I PDR                    state: Unlocked
HCPPDF6619I                   time stamp: 10/13/11 15:10:42
HCPPDF6619I      cross-system timeouts: Enabled
HCPPDF6619I PDR slot  1           system: GDLLCPX1
HCPPDF6619I                        state: Joined
HCPPDF6619I                   time stamp: 10/13/11 15:10:18
HCPPDF6619I                  last change: GDLLCPX1
HCPPDF6619I PDR slot  2           system: GDLLCPX2
HCPPDF6619I                        state: Joined
HCPPDF6619I                   time stamp: 10/13/11 15:10:42
HCPPDF6619I                  last change: GDLLCPX2
HCPPDF6619I PDR slot  3           system: GDLLCPX3
HCPPDF6619I                        state: Joined
HCPPDF6619I                   time stamp: 10/13/11 15:10:26
HCPPDF6619I                  last change: GDLLCPX3
HCPPDF6619I PDR slot  4           system: GDLMCPX4
HCPPDF6619I                        state: Joined
HCPPDF6619I                   time stamp: 10/13/11 15:10:35
HCPPDF6619I                  last change: GDLMCPX4
Ready; T=0.01/0.01 15:10:48
```

Live Guest Relocation – Example

```
xautolog lgrlin21
Command accepted
Ready; T=0.01/0.01 15:11:44
AUTO LOGON  ***       LGRLIN21 USERS = 21
HCPCLS6056I XAUTOLOG information for LGRLIN21: The IPL command is verifi
set secuser lgrlin21 *
HCPCFX6768I SECUSER of LGRLIN21 initiated.
Ready; T=0.01/0.01 15:11:50
LGRLIN21: Booting default (ipl)...
LGRLIN21: Linux version 2.6.16.60-0.21-default (geeko@buildhost) (gcc ve
 UTC 2008
```

▪ ▪ ▪

```
Welcome to SUSE Linux Enterprise Server 10 SP2 (s390x) - Kernel 2.6.16.6
"
"
linux-nxpt login:
```

Live Guest Relocation – Example

```
q lgrlin21 at all
GDLLCPX2 : LGRLIN21 - DSC
Ready; T=0.01/0.01 15:44:52
```

```
vmrelocate test lgrlin21 to gdllcpx1
User LGRLIN21 is eligible for relocation to GDLLCPX1
Ready; T=0.01/0.01 15:45:21
```

```
VMRELOCATE MOVE LGRLIN21 TO GDLLCPX1 MAXQ 5 SEC
Relocation of LGRLIN21 from GDLLCPX2 to GDLLCPX1 started
User LGRLIN21 has been relocated from GDLLCPX2 to GDLLCPX1
LGRLIN21: User LGRLIN21 has been relocated from GDLLCPX2 to GDLLCPX1
```

Live Guest Relocation – Example

```
LGRLIN21: qeth: check on device 0.0.0700, dstat=x0, cstat=x2 <4>qeth: ir
qeth: irb: 00 00 00 00  00 00 00 00  00 00 00 00  00 00 00 00
qeth: irb: 00 00 00 00  00 00 00 00  00 00 00 00  00 00 00 00
qeth: irb: 00 00 00 00  00 00 00 00  00 00 00 00  00 00 00 00
qdio : received check condition on activate queues on device 0.0.0702 (c
qeth: Recovery of device 0.0.0700 started ...
qeth: Device 0.0.0700/0.0.0701/0.0.0702 is a OSD Express card (level: 03
with link type OSD_100 (portname: whatever)
qeth: Hardware IP fragmentation not supported on eth0
qeth: VLAN enabled
qeth: Multicast enabled
qeth: IPV6 enabled
qeth: Broadcast enabled
qeth: Using SW checksumming on eth0.
qeth: Outbound TSO enabled
USER DSC    LOGOFF AS   LGRLIN21 USERS = 20     FORCED BY SYSTEM
Ready; T=0.01/0.01 15:45:52
LGRLIN21: qeth: Device 0.0.0700 successfully recovered!
Oct 13 15:45:51 linux-nxpt kernel: qeth: check on device 0.0.0700, dstat
00 00  00 80 e0 80"
Oct 13 15:45:51 linux-nxpt kernel: qeth: irb: 00 00 00 00  00 00 00 00
Oct 13 15:45:51 linux-nxpt kernel: qeth: irb: 00 00 00 00  00 00 00 00
Oct 13 15:45:51 linux-nxpt kernel: qeth: irb: 00 00 00 00  00 00 00 00
LGRLIN21: Oct 13 15:45:51 linux-nxpt kernel: qdio : received check condi
Oct 13 15:45:51 linux-nxpt kernel: qeth: Recovery of device 0.0.0700 sta
Oct 13 15:45:56 linux-nxpt kernel: qeth: Device 0.0.0700/0.0.0701/0.0.07
Oct 13 15:45:56 linux-nxpt kernel: with link type OSD_100 (portname: wha
Oct 13 15:45:56 linux-nxpt kernel: qeth: Using SW checksumming on eth0."
```

Live Guest Relocation – Example

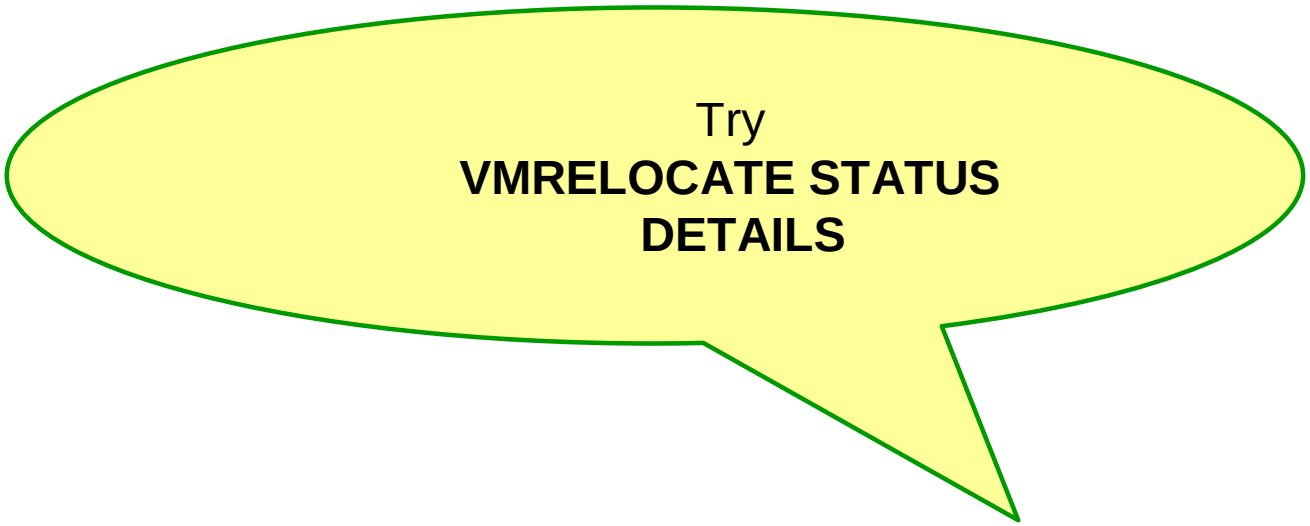
```
q lgrlin21 at all
GDLLCPX1 : LGRLIN21 - DSC
Ready; T=0.01/0.01 15:46:35
```

```
AT GDLLCPX1 CMD VMRELOCATE MOVE LGRLIN21 TO GDLLCPX2 MAXQ 5 SEC
Relocation of LGRLIN21 from GDLLCPX1 to GDLLCPX2 started
LGRLIN21: User LGRLIN21 has been relocated from GDLLCPX1 to GDLLCPX2
User LGRLIN21 has been relocated from GDLLCPX1 to GDLLCPX2
LGRLIN21: qeth: check on device 0.0.0700, dstat=x0, cstat=x2 <4>qeth: ir
qeth: irb: 00 00 00 00  00 00 00 00  00 00 00 00  00 00 00 00
qeth: irb: 00 00 00 00  00 00 00 00  00 00 00 00  00 00 00 00
qeth: irb: 00 00 00 00  00 00 00 00  00 00 00 00  00 00 00 00
qdio : received check condition on activate queues on device 0.0.0702 (c
qeth: Recovery of device 0.0.0700 started ...
qeth: Device 0.0.0700/0.0.0701/0.0.0702 is a OSD Express card (level: 03
with link type OSD_100 (portname: whatever)
qeth: Hardware IP fragmentation not supported on eth0
qeth: VLAN enabled
qeth: Multicast enabled
qeth: IPV6 enabled
qeth: Broadcast enabled
qeth: Using SW checksumming on eth0.
qeth: Outbound TSO enabled
Ready; T=0.01/0.01 15:47:10
LGRLIN21: qeth: Device 0.0.0700 successfully recovered!
```

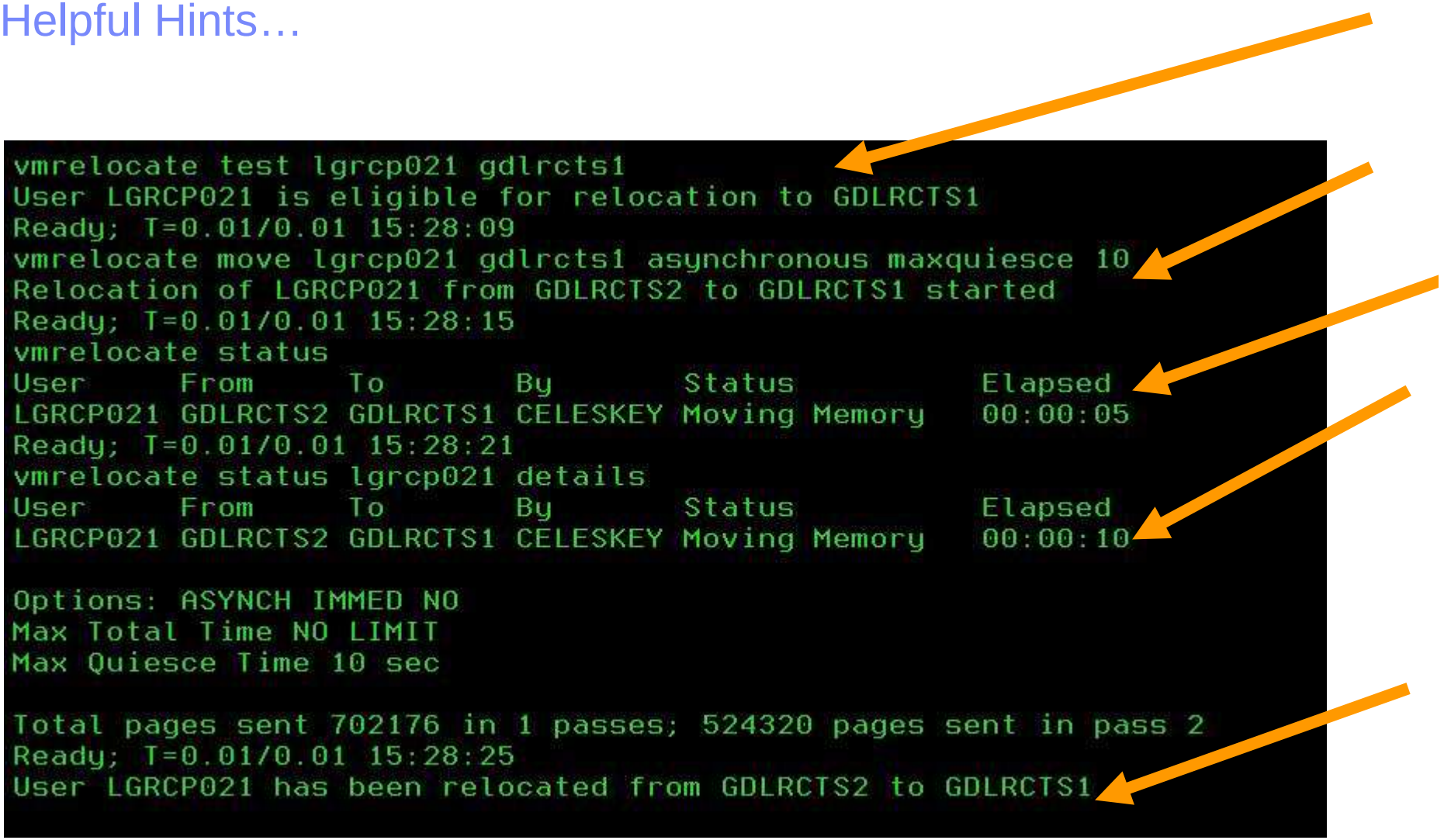
Live Guest Relocation – Example

```
q LGRLIN21 AT ALL
GDLLCPX2 : LGRLIN21 - DSC
Ready; T=0.01/0.01 15:47:41
```

Helpful Hints



Helpful Hints…



Helpful Hints…

Help! My relocation always exceeds my quiesce time!

Are you issuing relocations one at a time?

Check system constraints, are you trying to relocate to a member that really can't handle another user?

How many CTCs do you have between your members? How fast are they? How many devices do you have on each CHPID?

Helpful Hints…

I don't trust that you're really leaving the guest running, I want to see what my guest is doing as he relocates!

Use SCIF from another single configuration virtual machine -
SET OBSERVER LINUX01 *

Have the virtual machine spool his console
SPOOL CONS * START

Connect to Linux via SSH or VNC

Helpful Hints…

What are all these messages I see on my Linux console after relocation?

You may see recovery messages for OSA or FCP devices after relocation, this is normal

More Information

z/VM 6.2 resources

http://www.vm.ibm.com/zvm620/

z/VM Single System Image Overview

http://www.vm.ibm.com/ssi/

Redbook – An Introduction to z/VM SSI and LGR

http://publib-b.boulder.ibm.com/redpieces/abstracts/sg248006.html?Open

Thanks!

Contact Information:

John Franciscovich
IBM
z/VM Development
Endicott, NY

francisj@us.ibm.com