z/VM SSI and LGR Performance

Version 1.0

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:

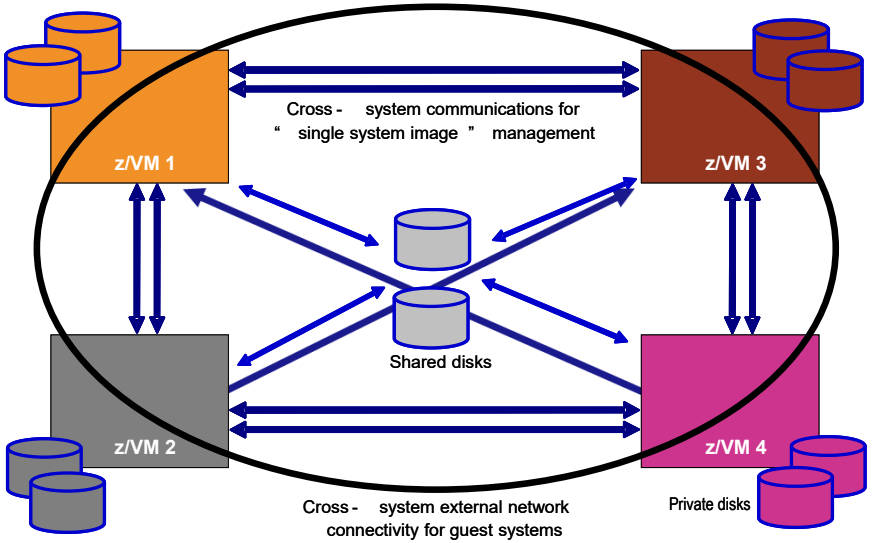
Performance is in Internal Throughput Rate (ITR) ratio based on measurements and projections using standard IBM benchmarks in a controlled environment. The actual throughput that any user will experience will vary depending upon considerations such as the amount of multiprogramming in the user's job 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.

Background

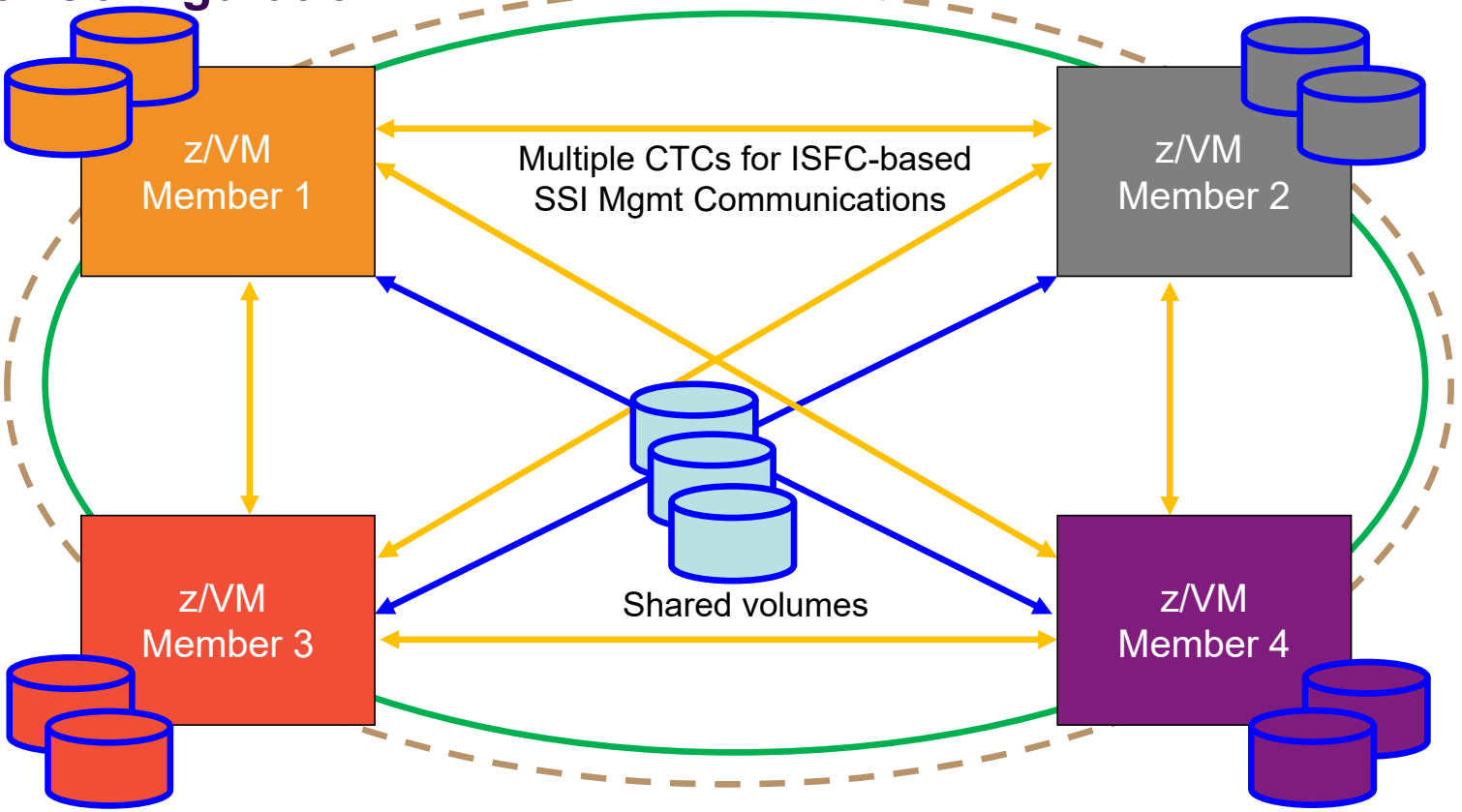
Single System Image (SSI) Feature

Clustered Hypervisor with Live Guest Relocation

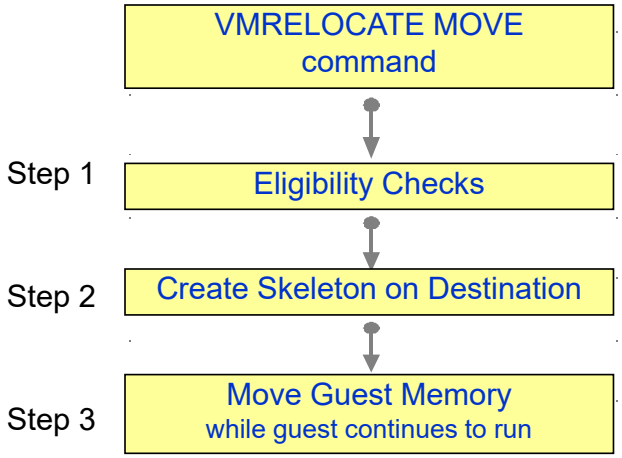
- Optional feature, available starting with z/VM 6.2 (No cost starting in z/VM 7.1)
- Connect up to four z/VM systems as members of a Single System Image cluster
- Cluster members can be run on the same or different IBM Z or LinuxONE servers
- Simplifies management of a multi-z/VM environment
  - Single 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



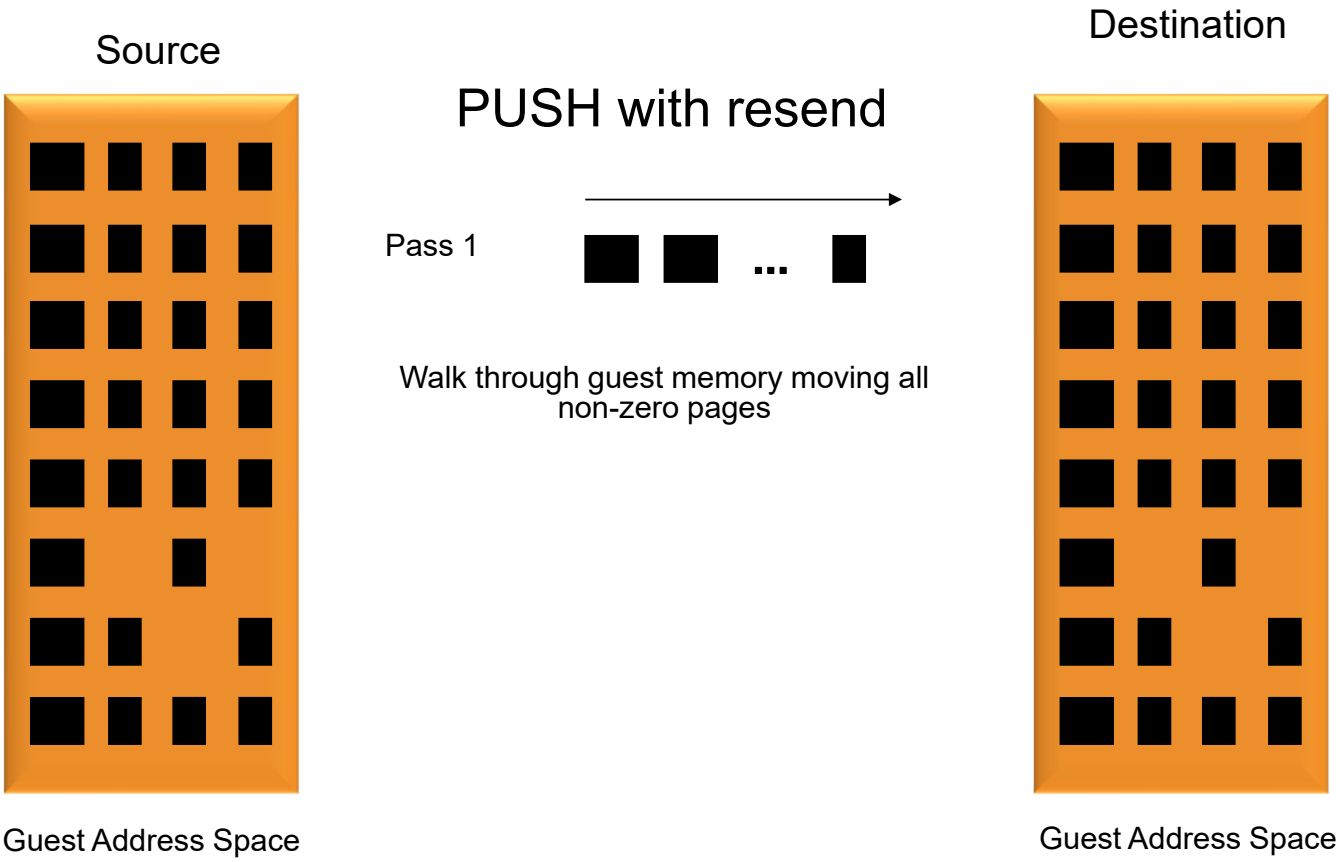
SSI Cluster Configuration



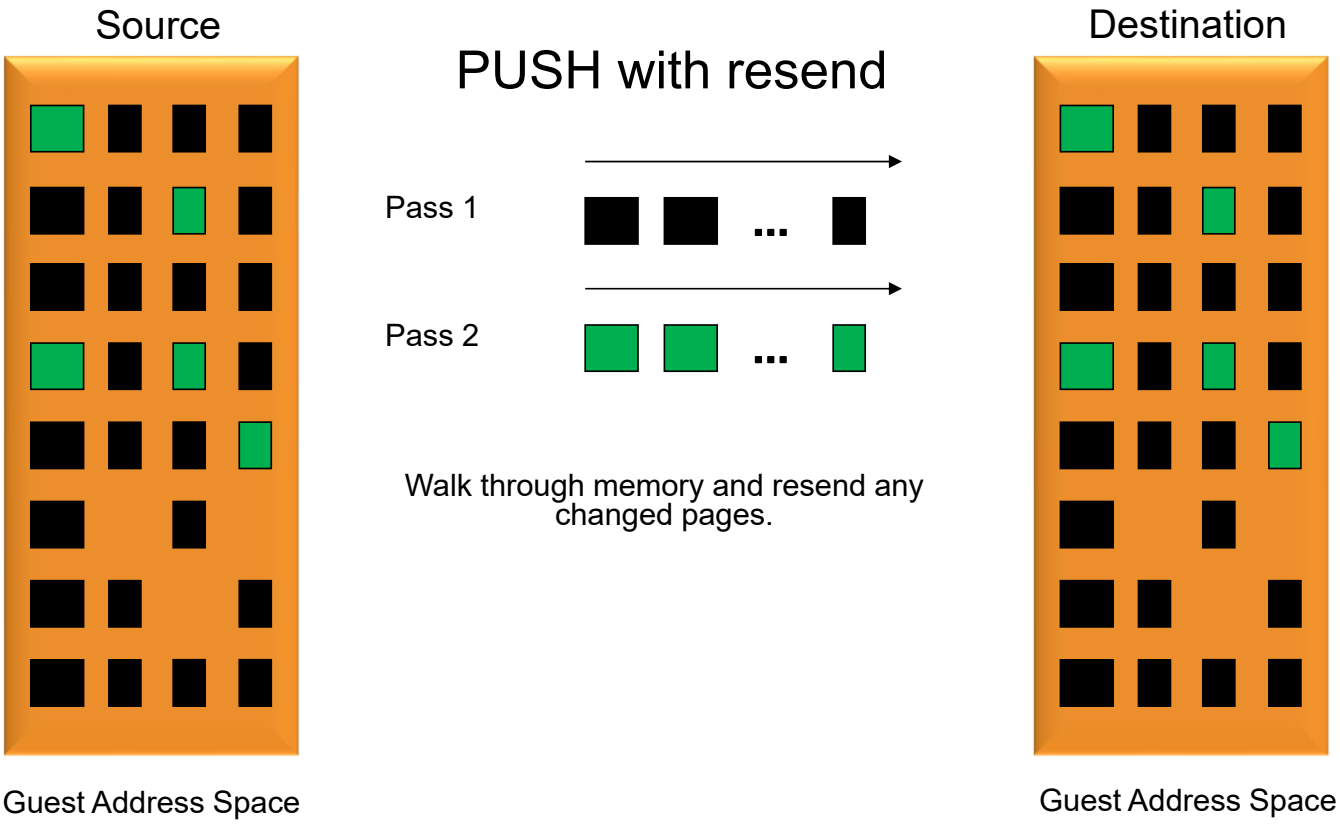
Stages of a Live Guest Relocation



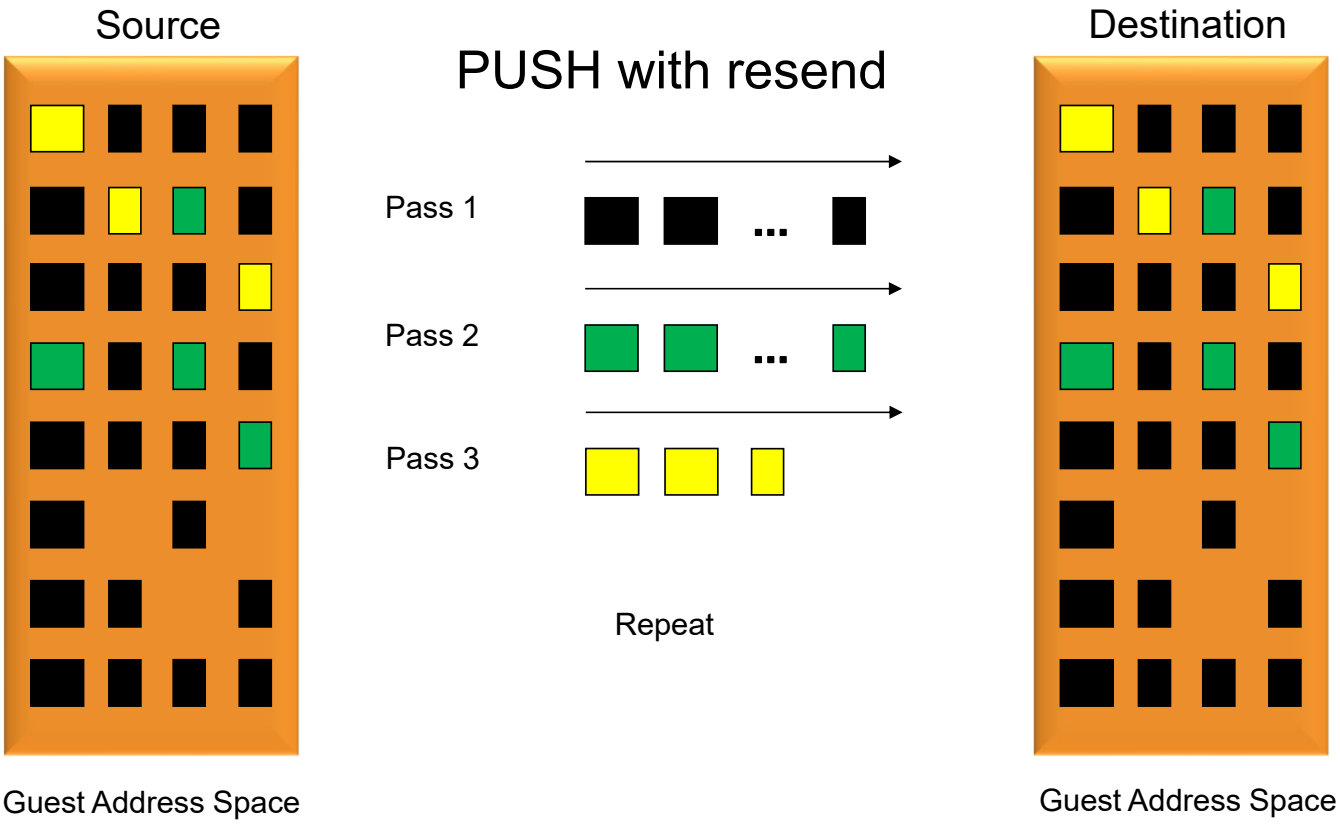
LGR, High-Level View of Memory Move



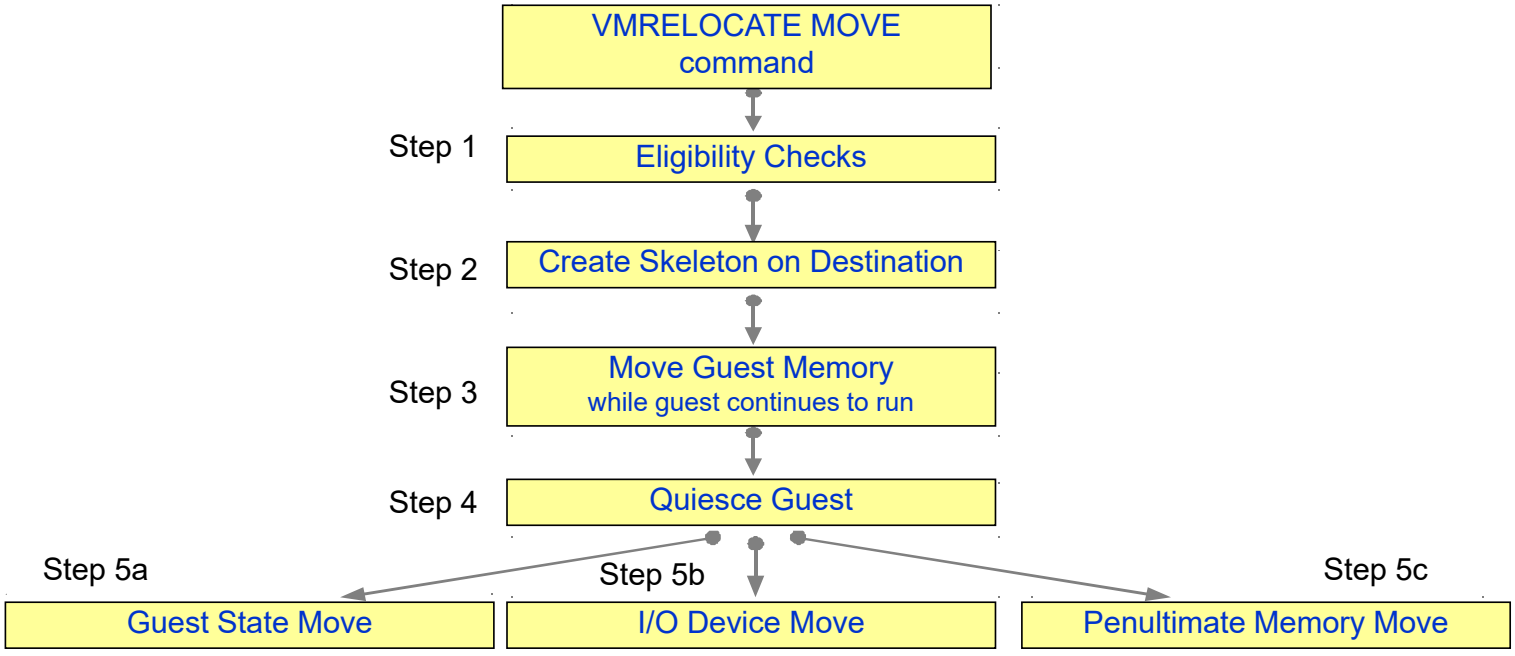
LGR, High-Level View of Memory Move



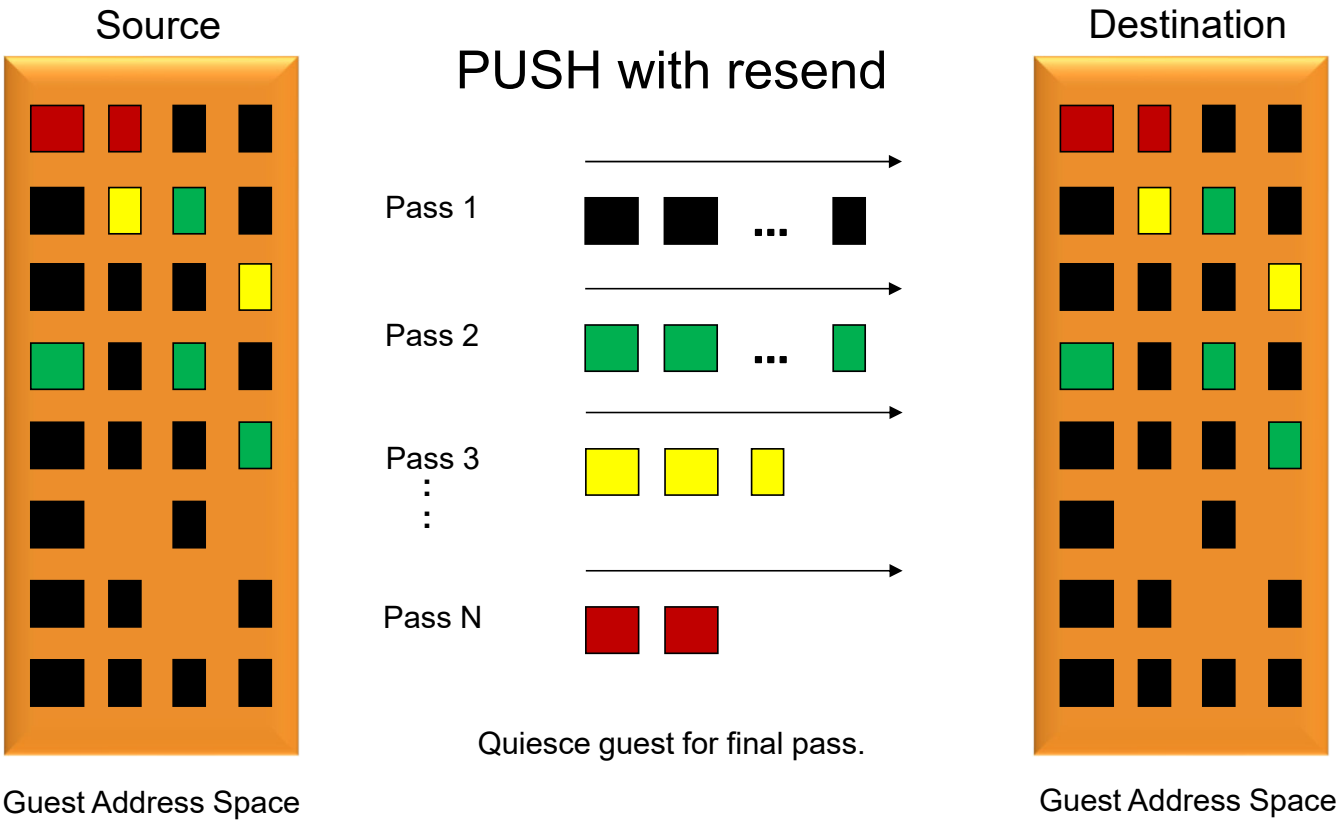
LGR, High-Level View of Memory Move



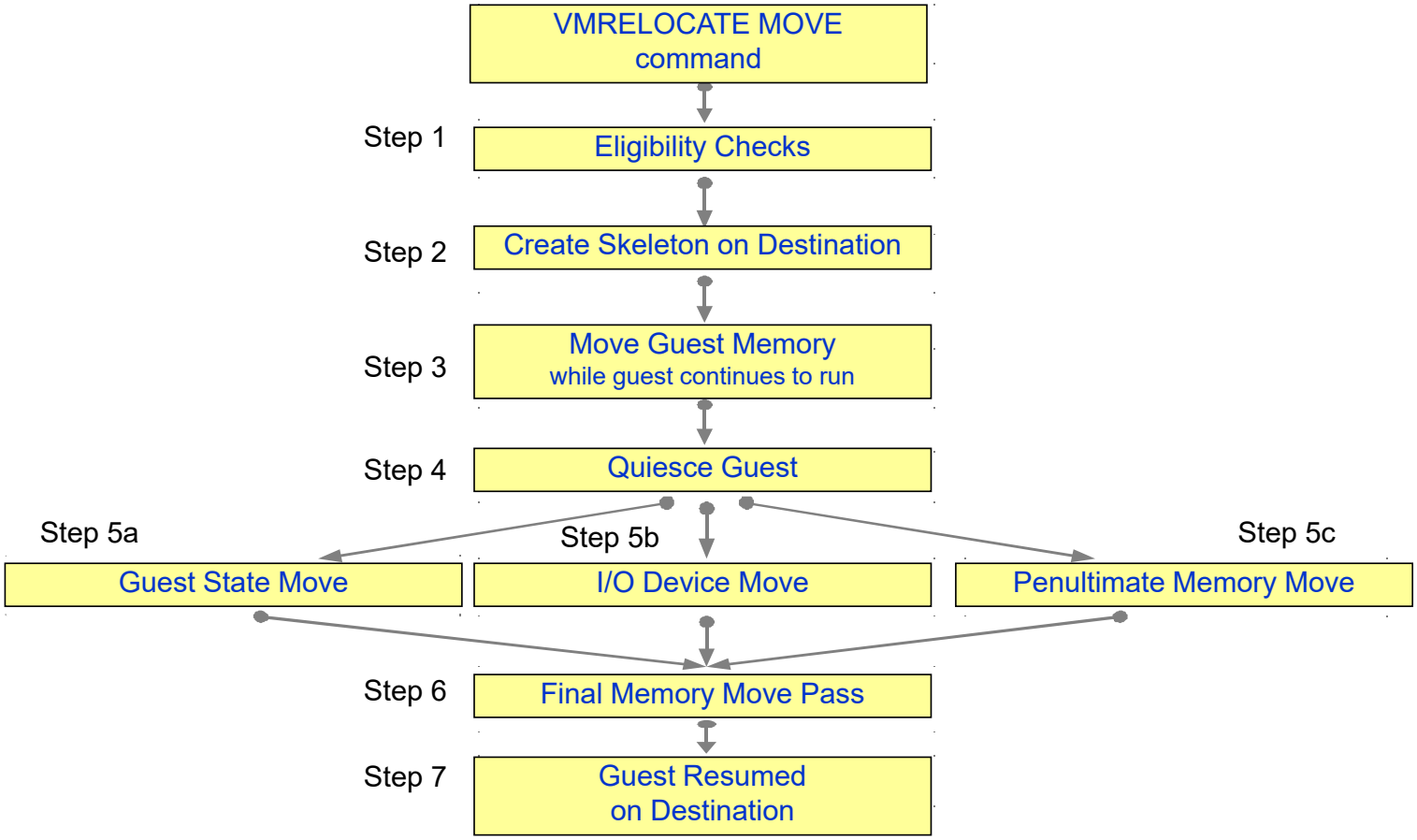
Stages of a Live Guest Relocation



LGR, High-Level View of Memory Move



Stages of a Live Guest Relocation



Performance of LGR and SSI

Live Guest Relocation – Key Performance Metrics

- Quiesce Time (QT)
  - Elapsed time that the guest is stopped (stunned) so z/VM can move the guest's last set of storage pages – probably the frequently-changed ones
  - To tolerate relocation, the guest and its applications must tolerate the quiesce time
  - VMRELOCATE can be invoked with a specified maximum quiesce time
    - If the quiesce would run past the maximum, z/VM cancels the relocation

- Relocation Time (RT)
  - Elapsed time from when the VMRELOCATE command is issued to when the guest is successfully restarted on the destination system.
  - Elapsed time must fit within the customer's window of time for planned outages for system maintenance, etc.

Bottom line: there are some scenarios where LGR is not feasible as a result of the requirements for relocation time and quiesce time

LGR: Factors Affecting QT and RT

- Size of the guest
  - Amount of memory to move, time required to walk its DAT tables
- How broadly or frequently the guest changes its pages
  - It's an iterative memory push from source to destination
- Time needed to relocate the guest's I/O configuration
  - I/O device count, I/Os to quiesce, OSA recovery on target side
- Capacity of the ISFC logical link
  - Number of chpids, their speeds, number of RDEVs
- Storage constraints on source and target systems
- Performance of paging subsystem
- Other work the systems are doing
- Other relocations happening concurrently with the one of interest
- Delays injected when LGR throttles itself back to prevent abends and other problems.
  - End-to-end LGR throttling – triggered by paging intensities
  - Memory-move endpoint throttling – triggered by memory consumption
  - ISFC logical link throttling – triggered by ISFC running out of queued traffic buffers

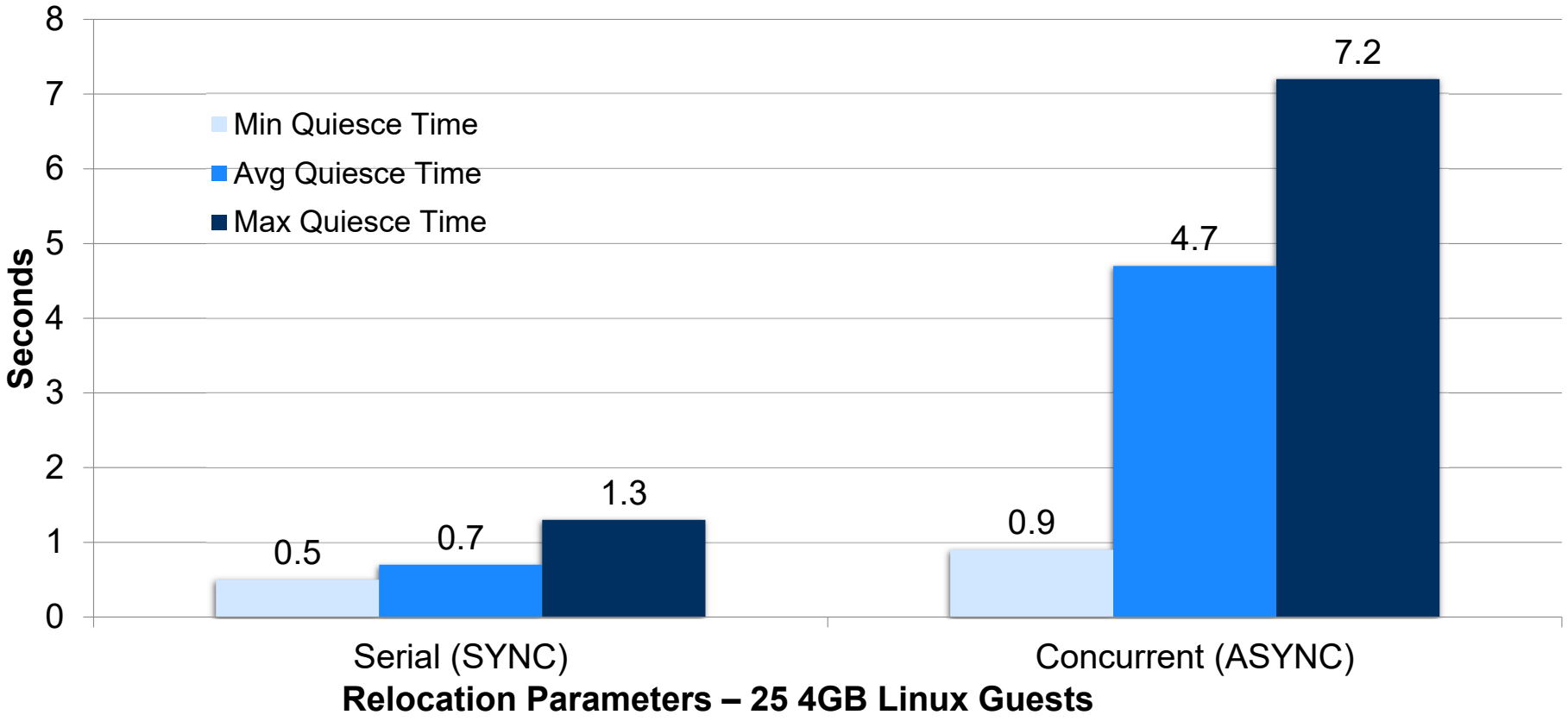
LGR: Serial vs. Concurrent Relocations

- By default, the VMRELOCATE command operates synchronously.
- There is a command option (ASYNCH) to run it asynchronously (similar to SPXTAPE)
- You could also achieve concurrent relocations by:
  - Use the asynchronous version of VMRELOCATE multiple times.
  - Run VMRELOCATE commands in multiple users concurrently.

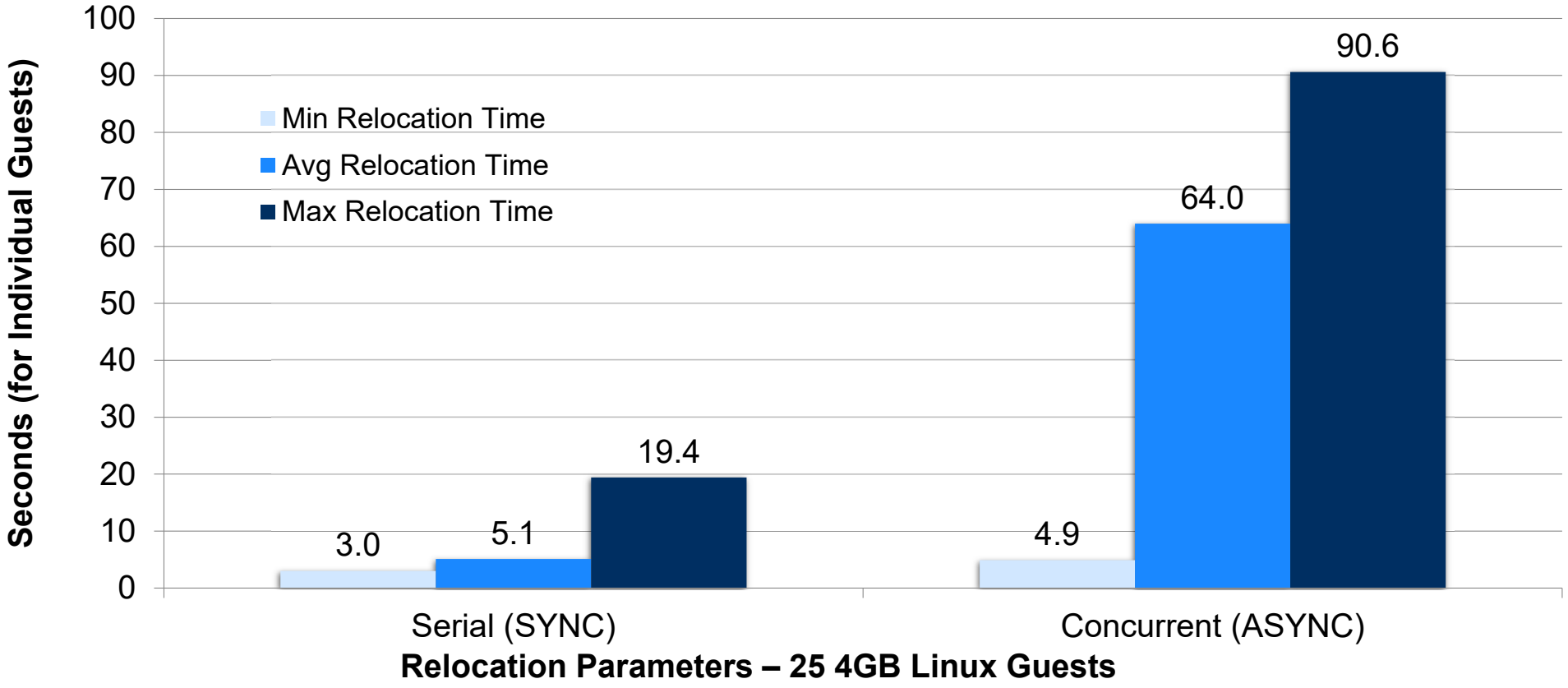
The best practice, though, is to run only one relocation at a time.

- QT and individual RT improves substantially when relocations are done serially
  - … and total RT elongates only slightly

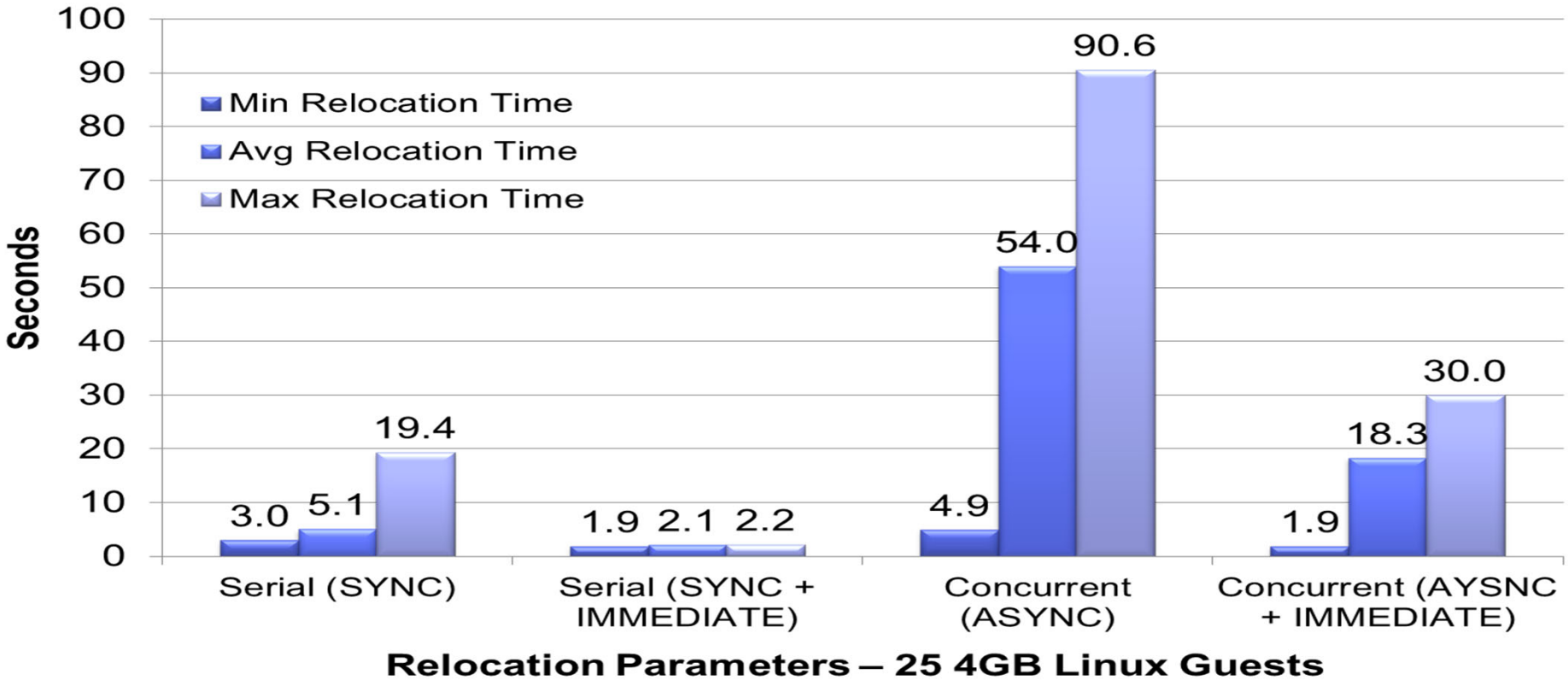
Effect of Serial vs. Concurrent on Quiesce Time



Effect of Serial vs. Concurrent on Relocation Time



Effect of IMMEDIATE option on Relocation Time



VMRELOCATE Options Summary

- Best total relocation time for all virtual machines
  - Concurrent (ASYNCH) + IMMEDIATE
- Best individual relocation time
  - Serial (SYNCH) + IMMEDIATE
- Best quiesce times
  - Serial (SYNCH)
- Worst quiesce times
  - Concurrent (ASYNC) + IMMEDIATE

Background on ISFC Capacity Test

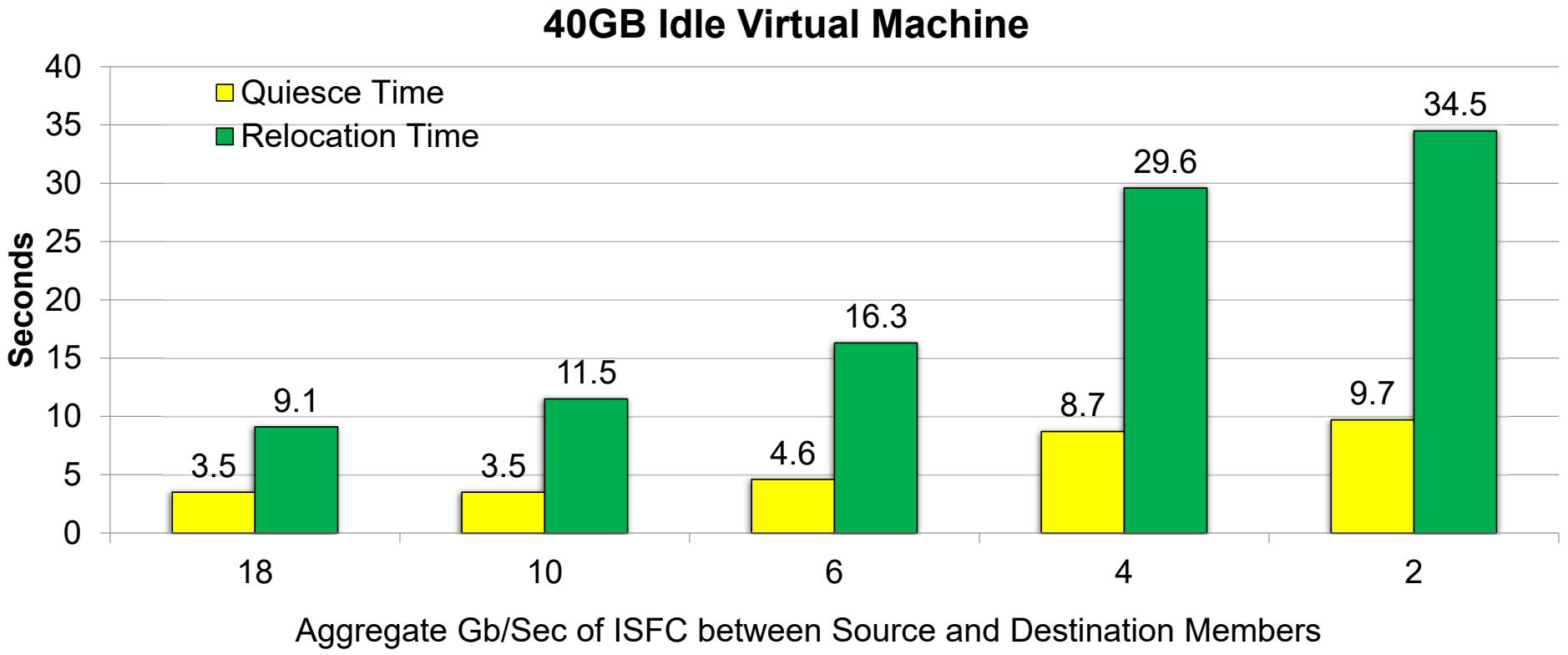
Table 3. Evaluated ISFC Logical Link Configurations.

ISFC Logical Link CHPIDs	ISFC Capacity Factor *	CTCs/FICON CHPID	Total CTCs
1-2Gb, 2-4Gb, 1-8Gb	18	4	16
1-2Gb, 2-4Gb	10	4	12
1-2Gb, 1-4Gb	6	4	8
1-4Gb	4	4	4
1-2Gb	2	4	4

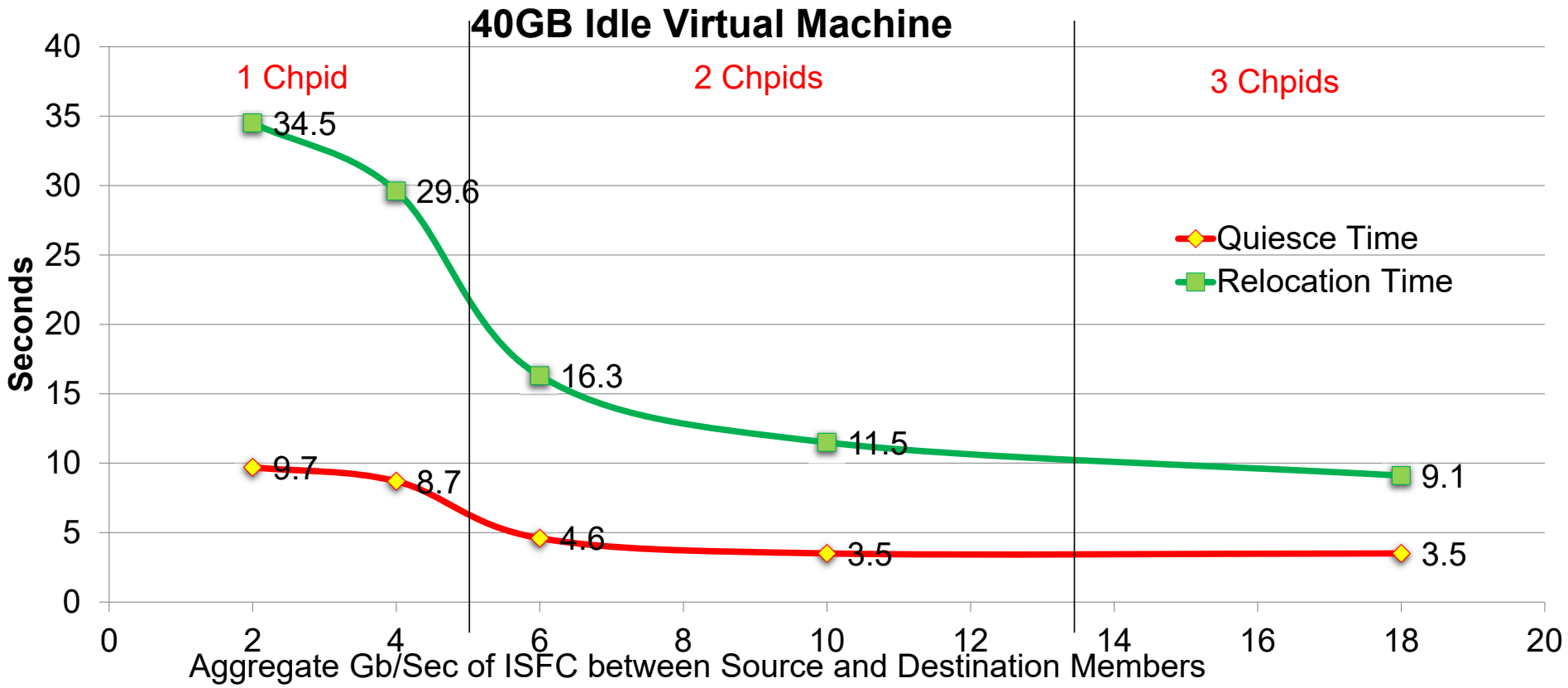
Note: * ISFC capacity factor is the sum of speeds of the FICON CTCs between the SSI member systems.

- "Logical Link" connects two members
  - Made up of up to 16 CTC devices
  - Spread across multiple FICON CHPIDs
- Recommend Chpids be of same type/speed
- Performance plateaus at 4 CTC devices per CHPID

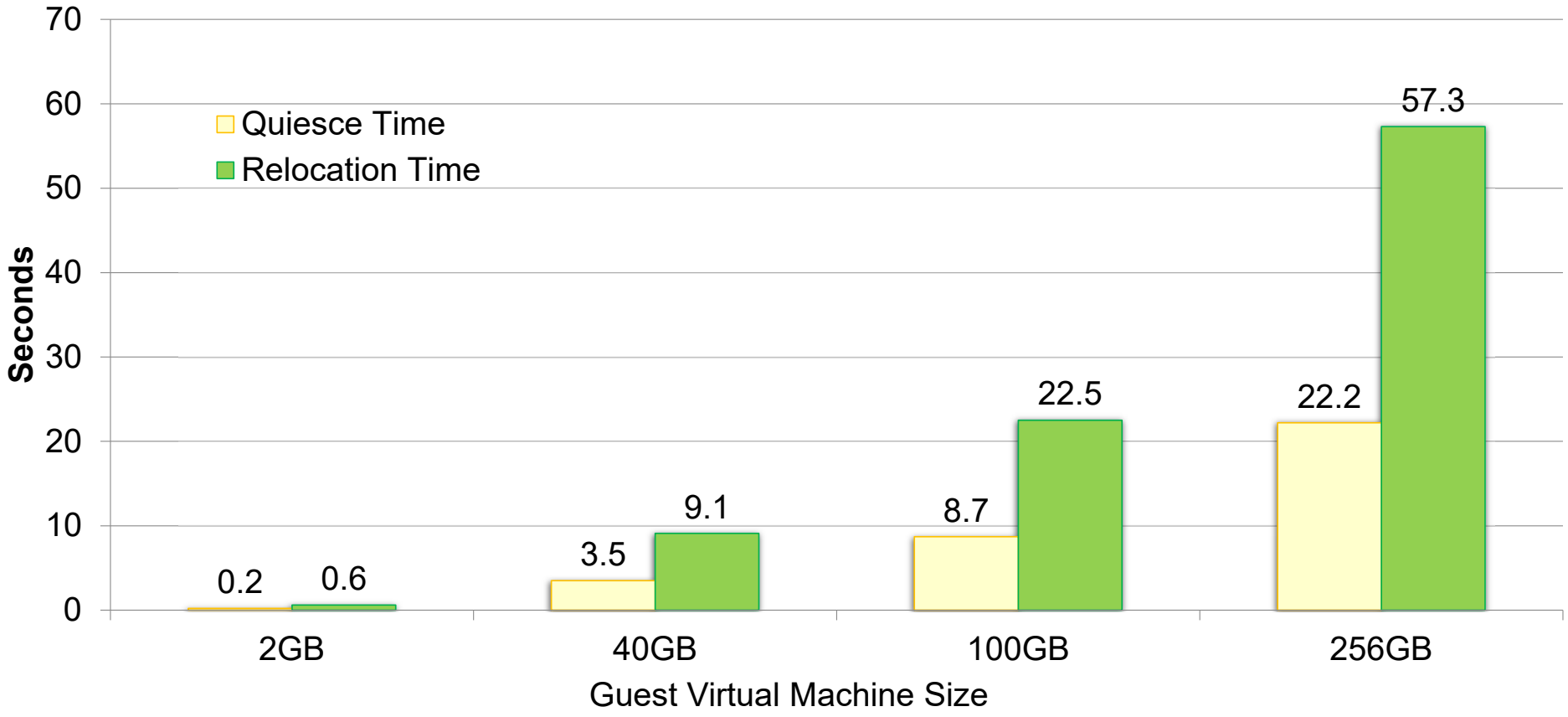
Effect of CTC Bandwidth on LGR



Effect of CTC Bandwidth on LGR

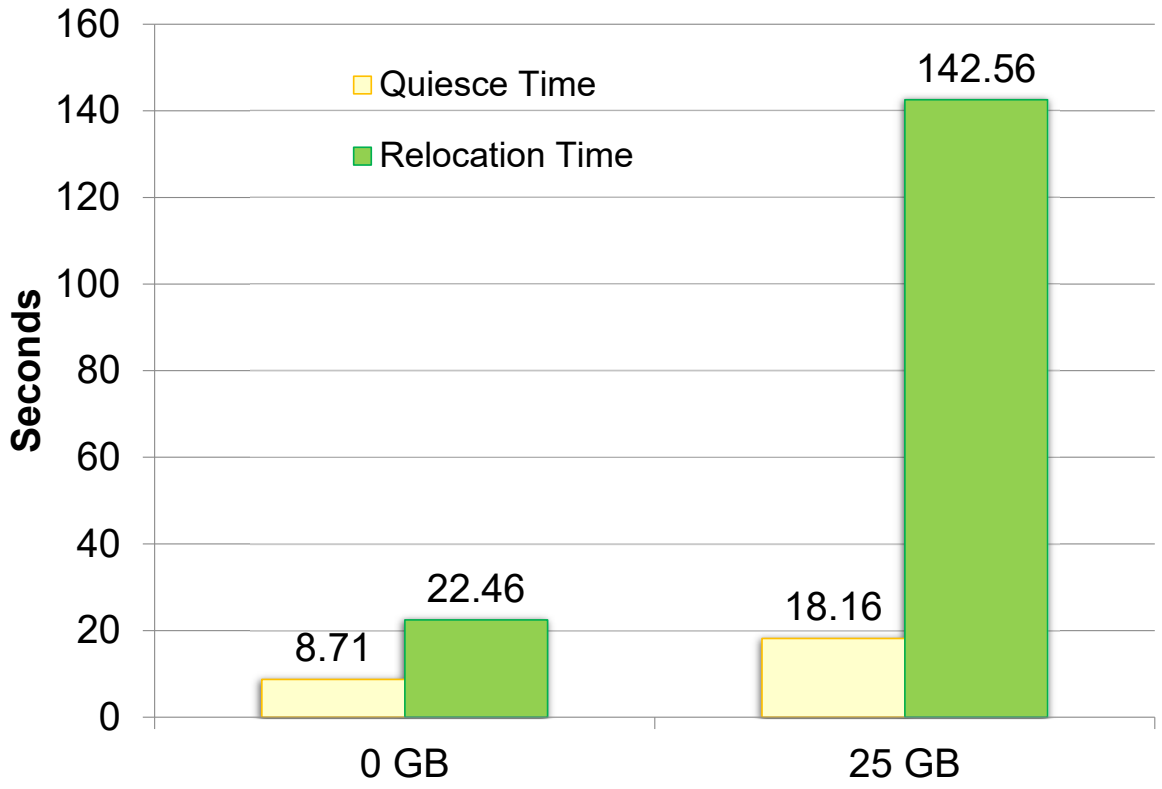


Effect of Virtual Machine Size on LGR



Impact of Virtual Machine Changing Memory on LGR

- Idle case (0GB changing) there is less memory to move and fewer Memory Move Passes
- Number of Passes
  - 0GB: 4
  - 25GB: 8
- Total Memory Moved
  - 0GB: 4.9GB
  - 25GB: 160GB



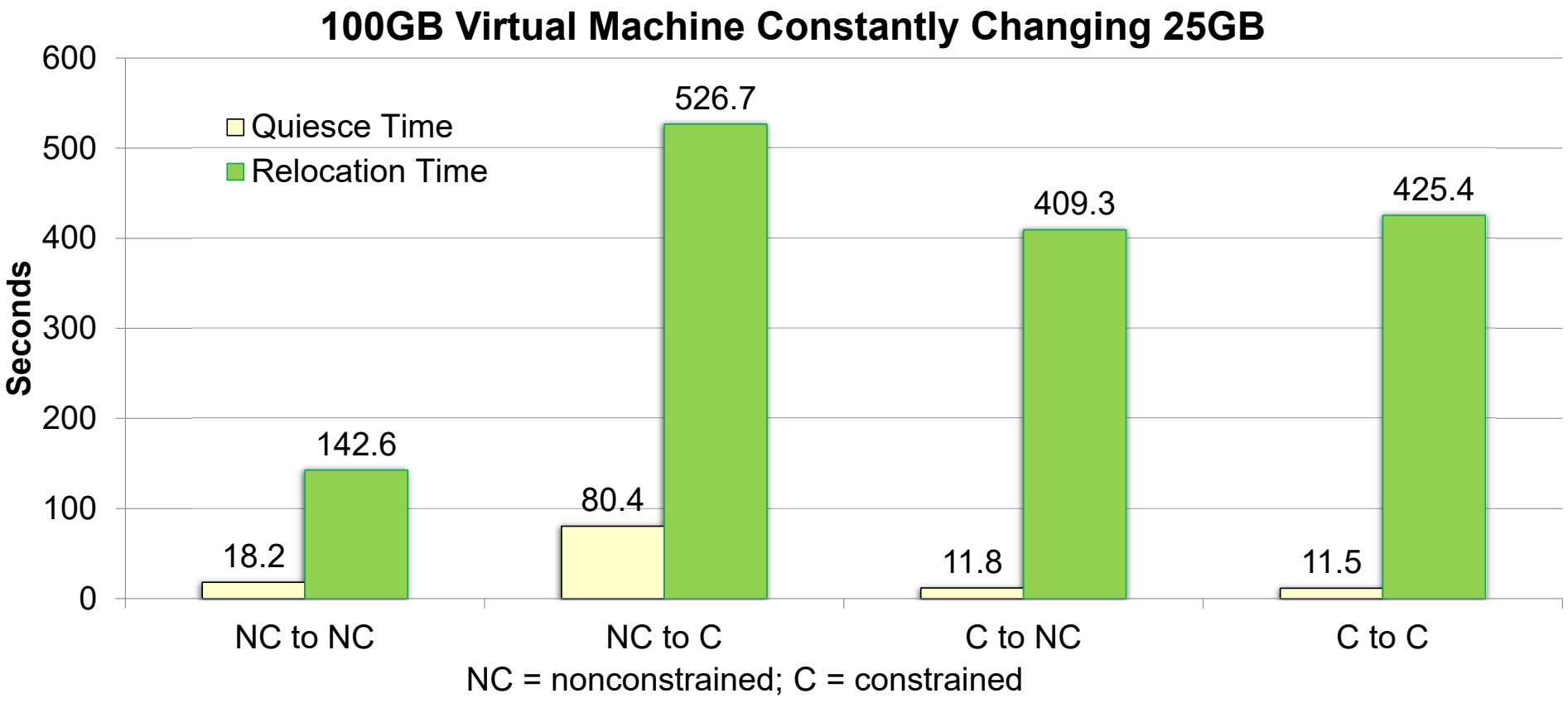
LGR: CPU and Memory Use Habits

- CPU: generally LGR gets what it needs
  - Taken “off the top” compared to your workload

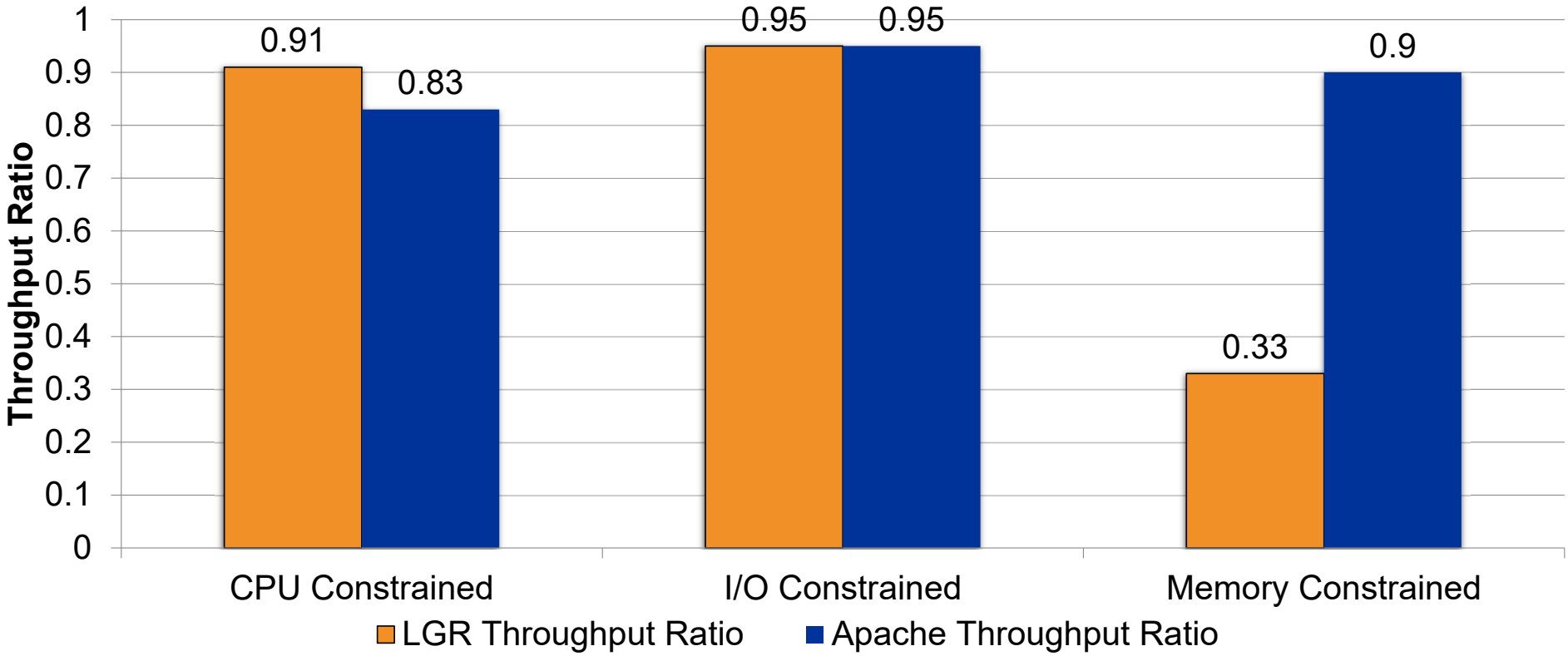
- Memory: CP tries really hard not to interfere
  - End-to-end throttling, ISFC buffer limits, …
  - Socket memory-move throttling – triggered by memory consumption
  - ISFC logical link throttling – triggered by ISFC running out of queued traffic buffers
  - Considers effect on paging, memory use for specific relocations, …

26

Effect of System Memory Constraint on LGR



Effect of LGR on Existing Workloads
LGR Bounce and Apache Web Serving Workloads



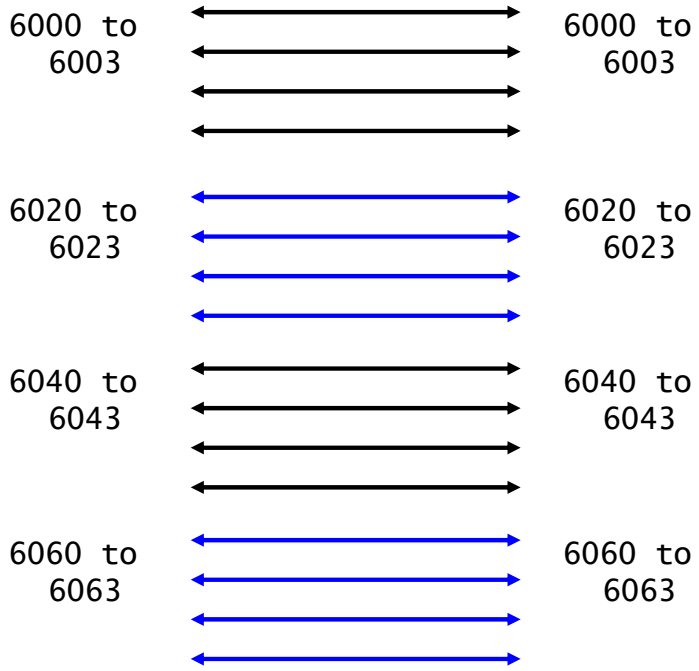
LGR: Keep These in Mind…

- Charge back: can your procedures handle guests that suddenly disappear and then reappear somewhere else?
- Second-level schedulers: do you have them? Can they handle guest motion?
- VMRM: if VMRM-A tweaks the guest and then the guest moves to system B, what happens? And then what happens when the guest comes back?
- System Management Tools: are the SSI aware?

Best practice is not to include relocating guests in VMRM-managed groups.

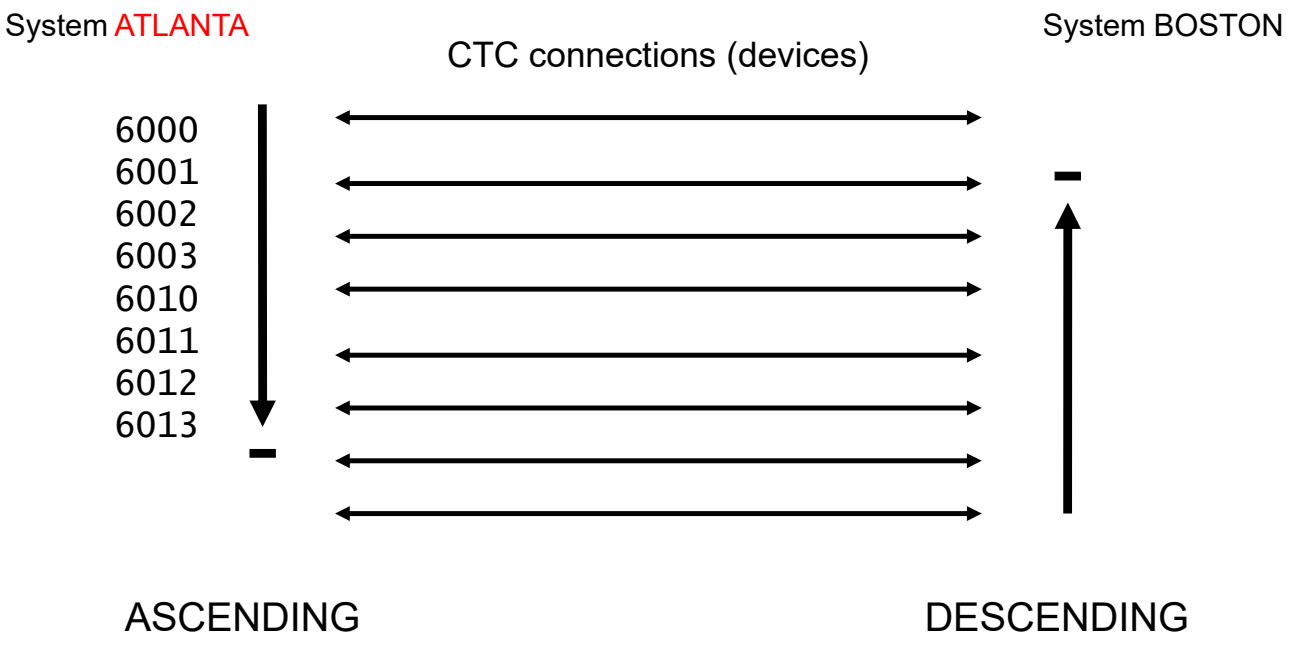
SSI: ISFC Logical Link Configuration Best Practices

- Maximum of 16 CTC devices per ISFC logical link
- Use multiple FICON chpids of all the same speed. Up to 4 chpids[1].
- Use four or fewer CTC devices per chpid
- Use same RDEV numbers on both ends
- Can share the chpids but requires capacity planning



[1] – You could use more chpids and fewer CTCs per chpid, but that having that many available chpids in a large SSI environment is not likely.

SSI: ISFC Logical Link Write Scheduling, under the covers



Moral: put the fast chpids in the middle of ATLANTA's RDEV range.
Selection of where to start in selecting write path is alphabetical.

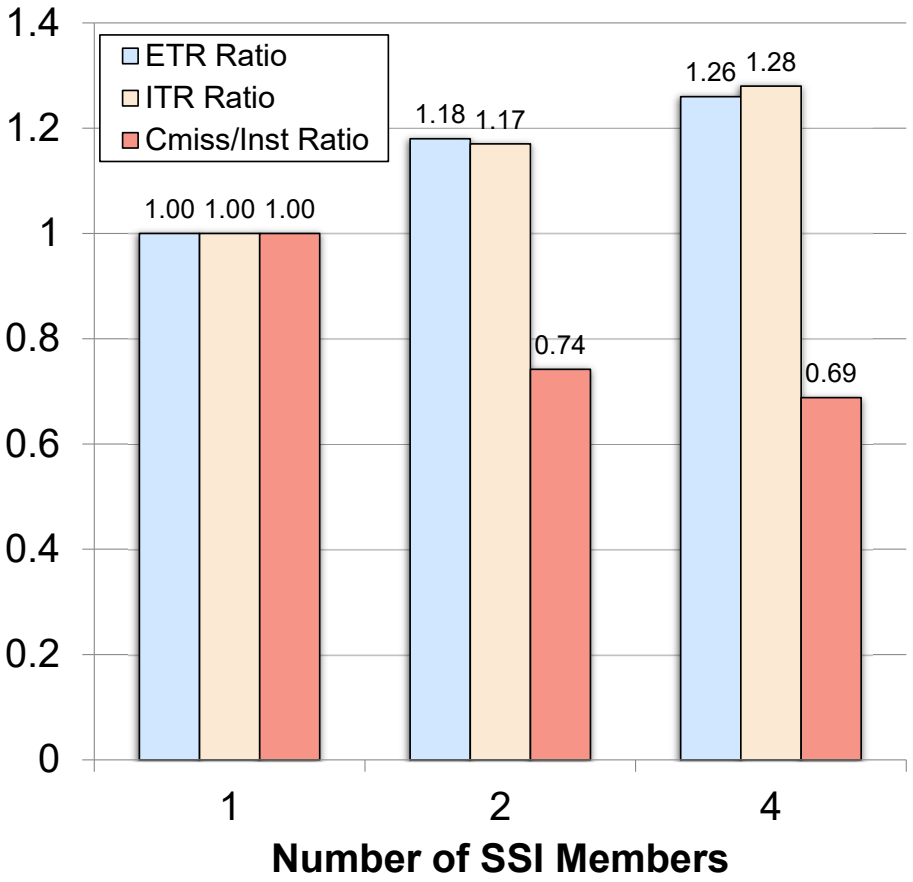
SSI Workload Distribution Measurements

Parameters	1 Member	2 Member	4 Member
Central Storage	43 GB	22 GB	11 GB
Expanded Storage[2]	8 GB	4 GB	2 GB
Processors	12	6	3

- Series of measurements to see how a workload spread across a number of members would run compared to one larger systems of just one member.
- Resources kept the same, as shown above.
- Apache workload where clients and servers were all virtual machines was used.
  - Varied number of client and servers and use of MDC to create different stress points.

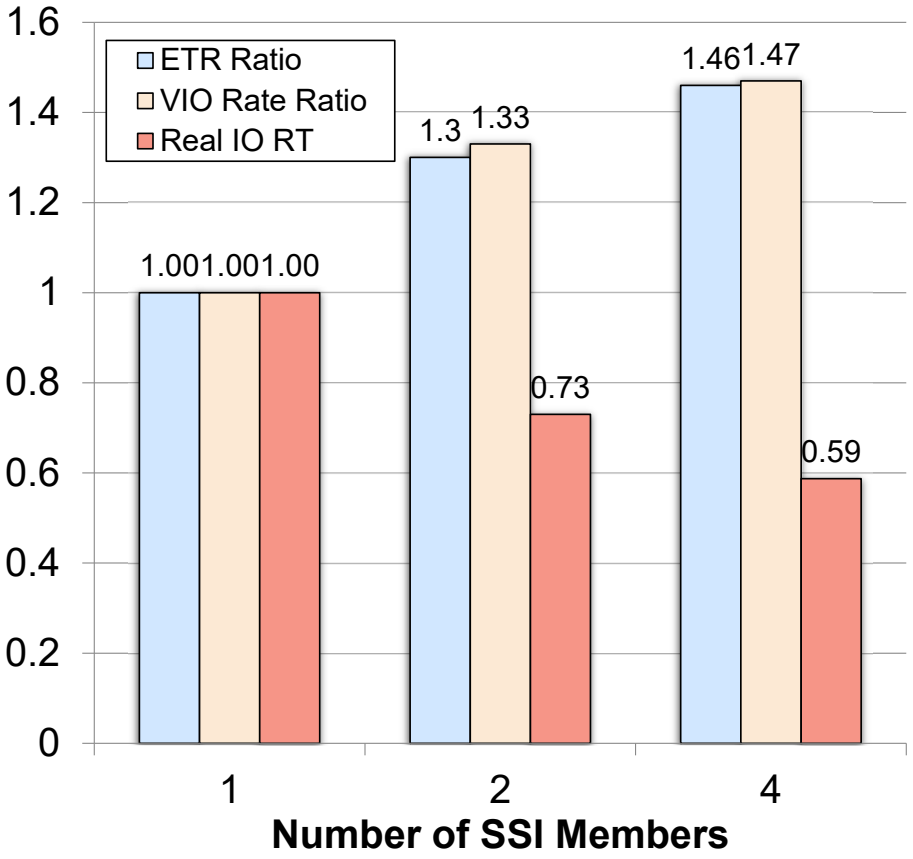
[2] – This was a z/VM 6.2 measurement when expanded storage was still supported.

SSI Distribution: CPU Constrained Measurement



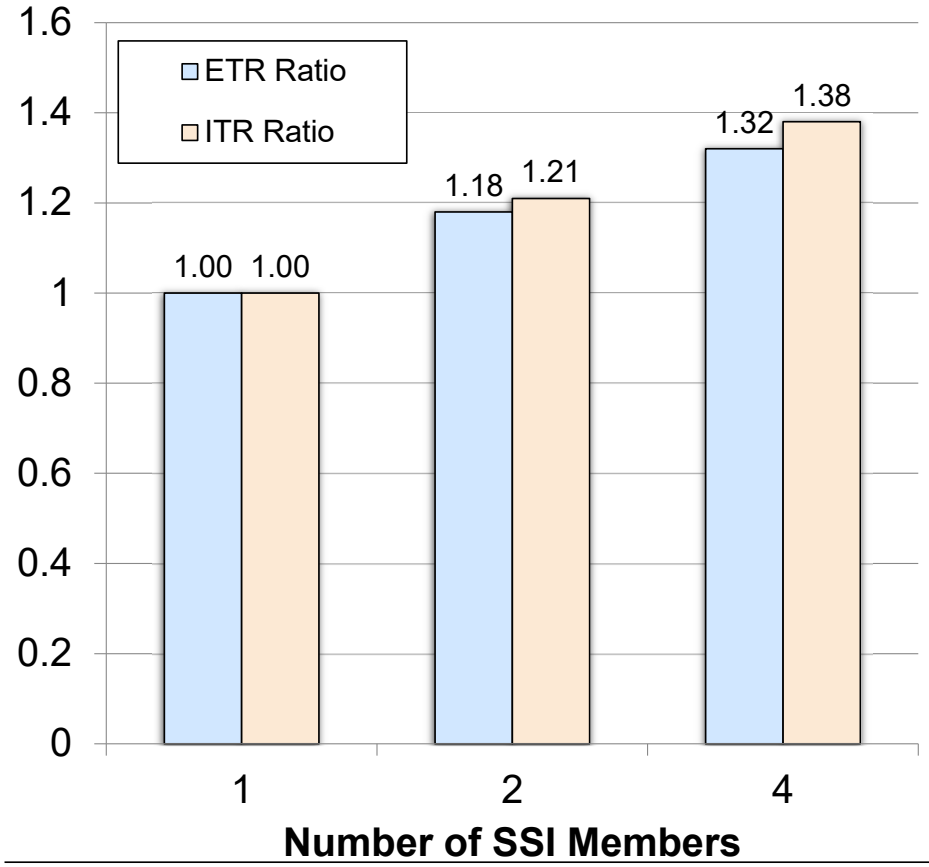
- Keep the physical resources the same, but distribute over 1, 2, or 4 members.
- Apache Web Serving with the configuration being CPU bound.
- Benefits from running smaller n-way partitions

SSI Distribution: Virtual I/O Constrained Measurement



- Keep the physical resources the same, but distribute over 1, 2, or 4 members.
- Apache Web Serving with the configuration being I/O bound due to virtual read I/O.
- PAV not used in base case, so SSI essentially gives PAV like benefits.
- Real I/O RT shown is for one of the shared Linux volumes containing files being served.

SSI Distribution: Memory Constrained Measurement



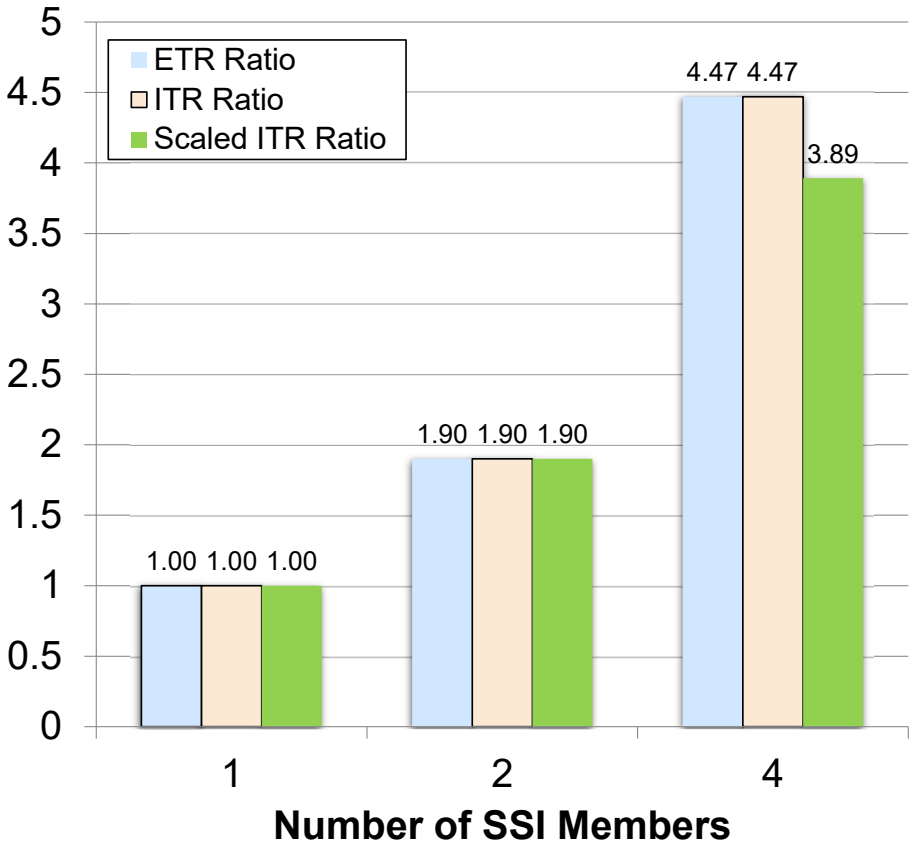
- Keep the physical resources the same, but distribute over 1, 2, or 4 members.
- Apache Web Serving with the configuration with there being memory constraint.
- Similar savings as in CPU bound measurement.
- Additional efficiencies in memory management.

SSI Workload Scaling Measurements

z/VM Total Resources	1 Member	2 Member	4 Member
Central Storage (Memory)	256 GB	512 GB	1 TB
IFLs	32	64	128

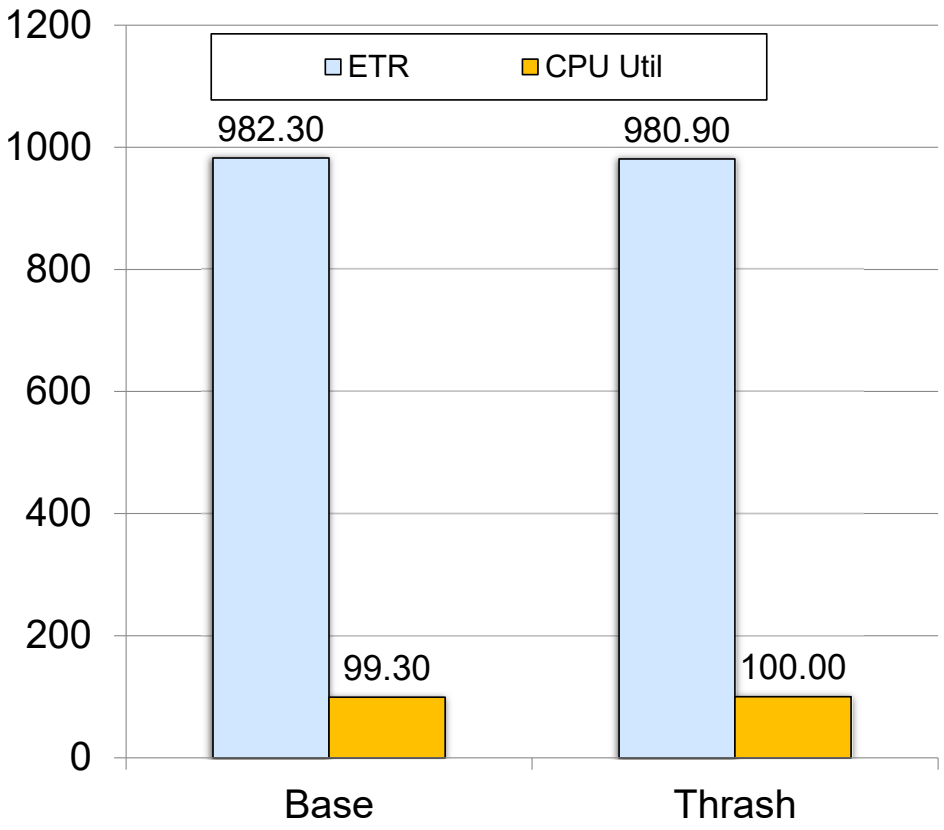
- Measurements were made to see how well z/VM scales within an SSI cluster.
  - z/VM 6.2 system
- Resources increased with each new member added to configuration.
- Apache workload where clients and servers were all virtual machines was used.
  - Apache clients and servers scaled accordingly.
- Needed to mix processor types to get 128 IFLs, so 1 & 2 Member runs are z10, 4 member adds in z196.
- Scaled down memory to make runs more feasible.

SSI Scaling Measurements



- The SSI Cluster overhead for a running environment is very low.
- Note: z196s were added to get the 3rd and 4th Member.
- “Scaled ITR Ratio is an estimate of the Ratio if the entire cluster were on z10 processors.

SSI Transition Measurement



- Measurement to determine if activity or Cluster management would influence performance.
- Four Member environment where 3 of the members are constantly transitioning through states (thrashing):
  - Joined
  - Leaving
  - Down
  - Joining
  - *repeat*

SSI: Performance Toolkit, Considerations

- Performance Toolkit continues to run separately on each member of the cluster
  - There continues to be a unique z/VM monitor data stream for each member.
  - There will be a PERFSVM virtual machine on each member

- Configuration and usage
  - Configure so that you will log onto or connect to a different PERFSVM on each system.
  - Configure Performance Toolkit to use the Remote Performance Monitoring Facility, which allows local and remote performance monitoring from a single screen.

- In general, Performance Toolkit does not produce “cluster view” reports
  - DASD device-busy view, for example

SSI: Performance Toolkit

- Reports for SSI
  - SSICONF: SSI configuration
  - SSISCHLG: SSI state change synchronization activity log
  - SSISMILG: SSI state/mode information log

- ISFC reports related to SSI
  - ISFECONF: ISFC end point configuration
  - ISFEACT: ISFC end point activity
  - ISFLCONF: ISFC logical link configuration
  - ISFLACT: ISFC logical link activity
  - ISFLALOG: ISFC logical link activity log

SSI: MONWRITE Considerations

- IBM often asks you to run MONWRITE
  - PMR diagnosis, for example
- You should be running MONWRITE anyway
- You should now be running MONWRITE on every member of the cluster
- Make sure it's easy to go find the MONWRITE data for all members for a specified time interval

SSI: Dump and PMR Considerations

- To solve your PMR,
- … IBM might need concurrently-taken dumps.

- Just be prepared:
  - Know how to take a SNAPDUMP. Practice.
  - Know the effect of SNAPDUMP on your workload.
  - Know how to take a restart dump.

SSI & LGR: Planning White Space

- Need white space for planned outages where you move work off of a given member.

- How will work move off the member?
  - Use existing HA solutions to redirect work to existing servers on other members or elsewhere in enterprise.
  - Use LGR to move to another member.
  - Log off and then logon to another member.
  - Shutdown non-critical virtual machine for duration of planned outage.

- To where do you move the virtual machines?
  - To a single member or multiple members?
  - To a member on same CEC or another CEC?
  - To a member held in reserve (such as a DR LPAR)?
  - It's not just one z/VM image anymore

Relocation Planning: Other Considerations

- "The bucket gets heavier as you add water."
  - Destination system may become more constrained as you continue to relocate virtual machines to it.
- "Get the big rocks in first."
  - In general, it is better to move the virtual machines generating the greatest memory load first.
    - Larger virtual machines
    - Virtual machines with higher page change rate

SSI & LGR: Planning White Space

- CPU
  - Shared logical processors?
  - Adjust LPAR weight settings?
  - Vary on additional engines?

- I/O
  - Ensure sufficient resources at all levels:
    - Channel, switch, control unit, device
  - Shared channels?

- Memory white space is not as easy to manage
  - Ensure sufficient paging space and concurrency or data rate capability
  - Increase real memory over commitment?
  - Temporarily decrease size of some virtual machines?
  - Use Dynamic Memory Upgrade?
    - No downgrade available (yet)

Summary

Summary

✓SSI is a low-overhead clustering technology

✓LGR performance is dependent on a number of factors:
- Activity of the virtual machine
- Activity of both the source and destination systems
- Configuration, especially of ISFC

✓ See the z/VM Performance Report for more details - http://www.vm.ibm.com/perf/reports/zvm/html/