z/VM 6.3: Memory Management

Bill Bitner – z/VM Customer Focus and Care – bitnerb@us.ibm.com

June 2014

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	System zEC12
Dynamic Infrastructure*	z9*	System zBC12
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.

Agenda

- Objectives and strategies of the z/VM Large Memory enhancement
- Key features of the z/VM Large Memory enhancement
  - Algorithmic concepts: new, changed, or obsolete
  - Basic flows and data structures
  - Tuning options
- Planning for z/VM Large Memory
  - Paging DASD calculations
  - Reminders about best practices with respect to paging I/O
- Workloads
- CP Monitor and z/VM Performance Toolkit
- Summary

Objectives and Strategies

- Objectives:
  - Support 1024 GB aka 1 TB of central memory in a partition
  - Support large guests in such a context
  - Retain ability to overcommit memory

- Strategies:
  - Repair or replace memory management algorithms that do not scale well
  - Repair or replace memory management algorithms that are grossly unfair

- Specifically:
  - *Page reorder* is a real problem area. Get rid of it.
  - *Demand scan* has scaling problems and frame ordering problems. Repair them.
  - Introduce a new *global aging list* concept to add accuracy to frame reclaim decisions.
  - Improve *fairness* of frame steal to spread the discomfort equitably when memory is constrained.
  - Improve effectiveness of keeping virtual machine memory specified by **SET RESERVED** resident in memory
  - Extend **SET RESERVED** to DCSSes such as MONDCSS.

New Algorithms and Behaviors

New Approach: Highlights

- Objective: keep the *available lists* populated just right
- New visit heuristic tries to improve occupancy fairness in the face of storage constraint
- The in-use frames are tracked by a new hierarchical data structure:
  - Valid, often-touched frames are at the top
  - Demand scan pushes frames downward as they seem to increase in reclaim appeal
  - Best reclaim candidates are at the bottom
- DASD use for paging is changed to be more friendly to reclaim and to storage subsystems
  - Pages valid on DASD are not rewritten anymore
  - Pages get written back to their same slots
  - Channel program can do fully discontiguous reads or writes
  - z/VM can prewrite pages to DASD

New Approach: Management of The Available Lists

Old way

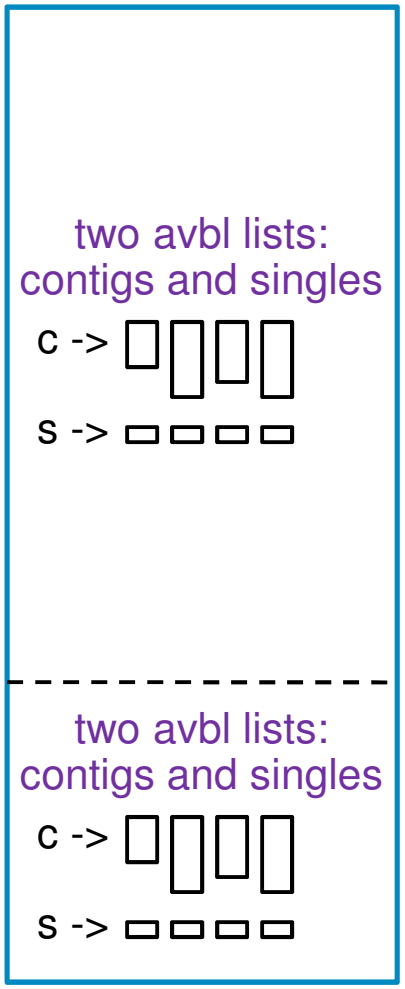
Each **list** had a low threshold and a high threshold

After every free storage request call, demand scan was kicked off if a **list** fell below its low threshold

The <2G lists were repopulated by demand scan

2 GB

<2G Use Policy:
Pre-6.2: *used <2G first*
In 6.2: *used <2G proportionally*
In 6.3: *uses <2G last*



New way

Each **kind of free storage request call** has a low and a high threshold:

- TYPE=ANY contigs
- TYPE=ANY singles
- TYPE=BELOW contigs
- TYPE=BELOW singles

Contig lists are protected from being completely raided by singles requests

After every request, the low threshold for **every type of request** is evaluated

If a TYPE=ANY low threshold is breached, demand scan is kicked off

If the <2G lists are empty, a frame table scan is kicked off

The Old Demand Scan Visit Policy

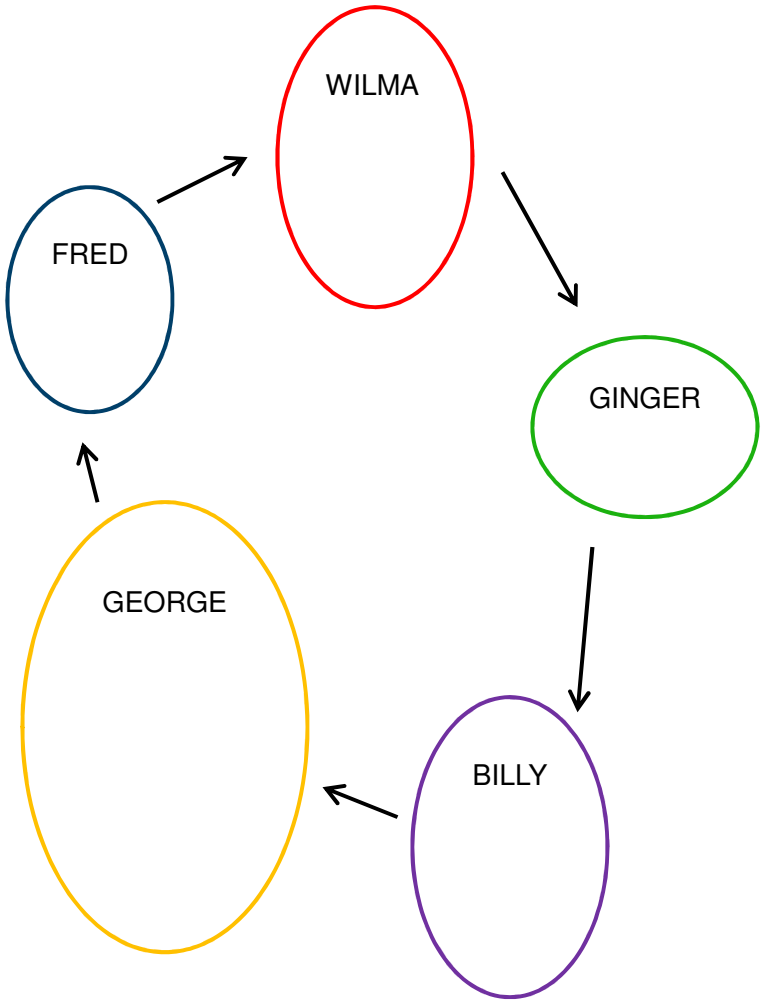
- It was a three-pass model:
  - Pass 1: tried to be friendly to dispatched users
    - Unreferenced shared-address-space pages
    - Long-term-dormant users
    - Eligible-list users
    - Dispatch-list users' unreferenced pages down to WSS
  - Pass 2: a little more aggressive… like pass 1 except:
    - Avoided shared address spaces
    - Would take from dispatch-list users down to their SET RESERVED
  - Pass 3: emergency scan
    - Anything we can find

The Old Demand Scan Problems

- We found a number of problems over time, to various degrees, such as:
  - Pass 1 tended to be too soft.
  - Scheduler lists tended not to portray “active” in a way usable by storage management.
  - Stole a lot from the first few users we visited.
  - **SET RESERVED** was not being observed.

- It used the System z page reference bit R to track page changes
  - Required lots of RRBE instructions to keep track of *recent* reference habits
  - RRBE can be an expensive instruction
  - (Large resident frame list) + (long RRBE instruction) = problems in Reorder

New Approach: The New Demand Scan Visit Policy



- Used to:
  - Visit according to scheduler lists
  - Take heavily at each visited user
  - Start over at list tops every pass
  - Take from private VDISKs nearly last
  - A “take” was truly a *reclaim* of a frame
- Now:
  - Cyclically visits the logged-on users
  - Keeps a visit cursor so it can resume
  - Takes a little and then moves to next
  - Takes from private VDISKs much earlier
  - A “take” is now just a push of in-use frames down toward eventual reclaim
- Effects
  - Better equalizing in the face of storage constraint
  - Better equalizing on the notion of “hot” vs. “cold” pages

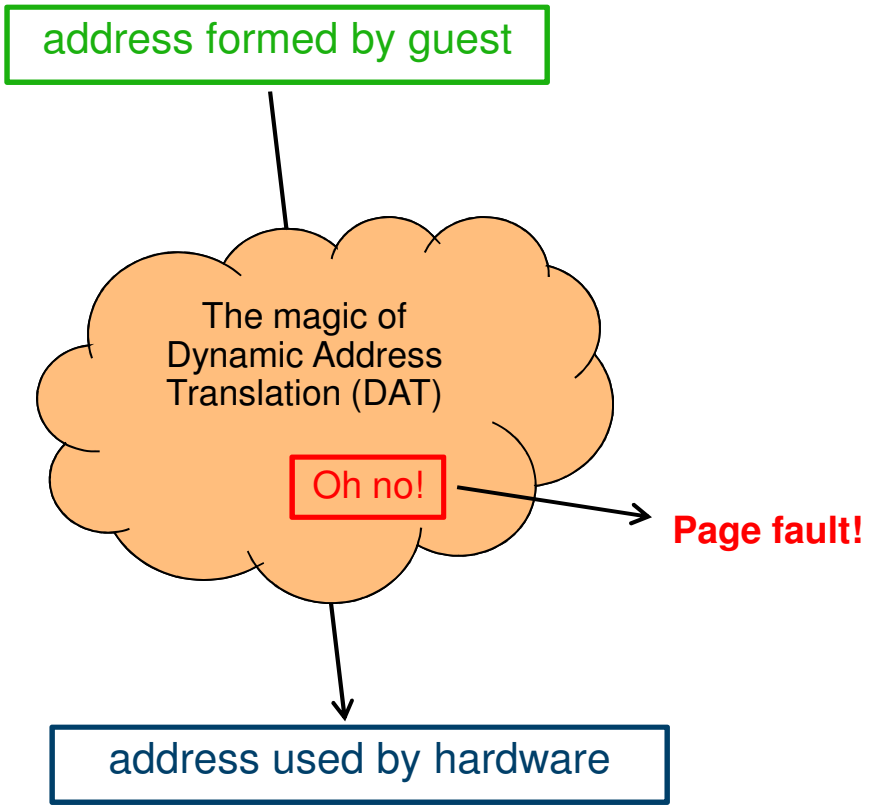
New Approach: Other New Things About Demand Scan

- Gives up control periodically
  - Lets other things happen
  - Avoids long-running “blackouts”

- Tries harder to be “fair” in the face of constraint.

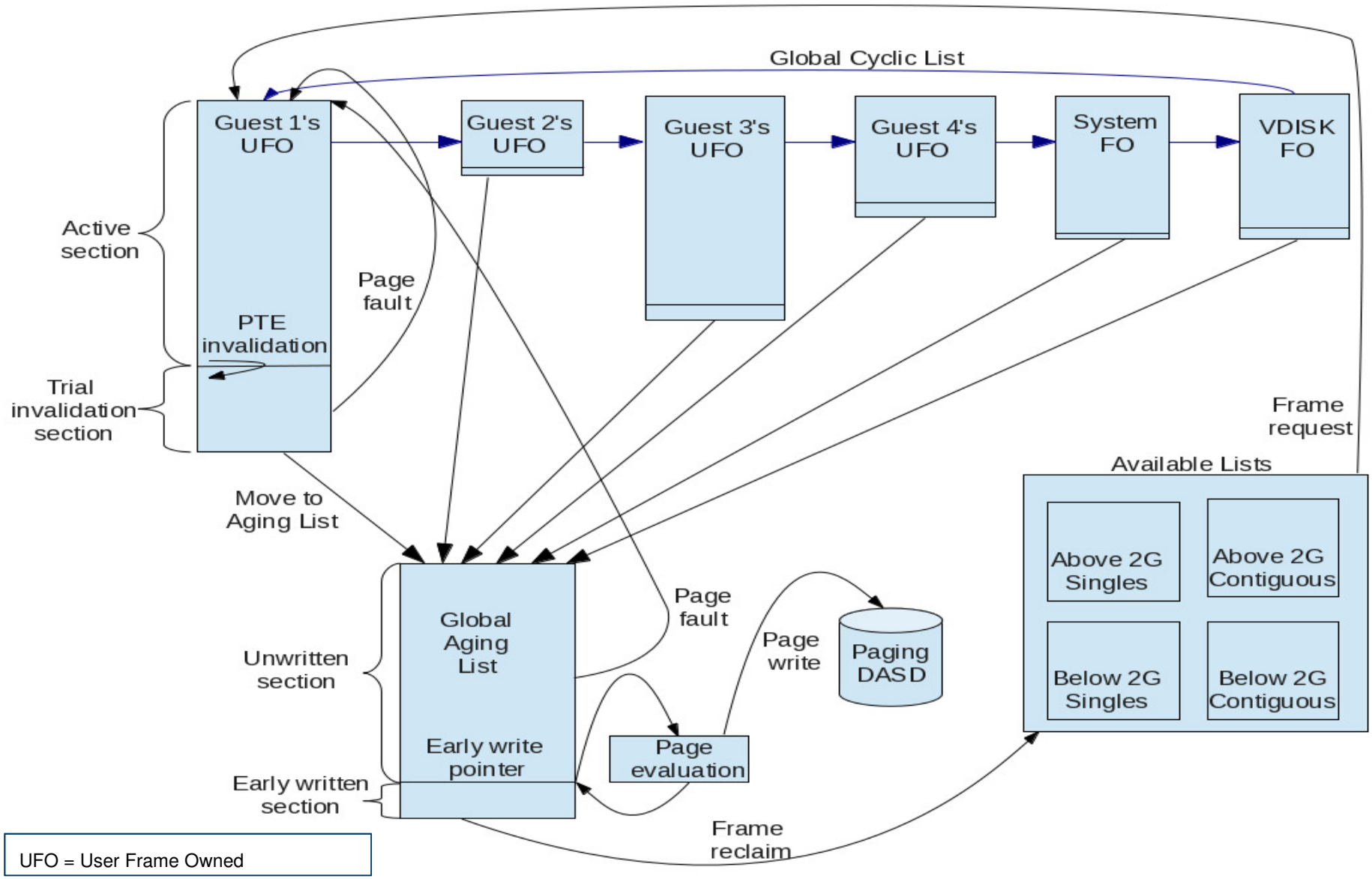
- Aspects of “fairness”:
  - Treat identical guests identically
  - Use a guest’s size and estimation of its page touch rate to decide how much to take
  - Take from large guests who touch their pages less often before taking from small guests who touch their pages a lot
  - Don’t take from a guest’s working set if another guest is not stripped to its working set
  - During startup (when page touch rate data is available) take an amount of pages proportionally to each guest’s size

New Approach: Trial Invalidation



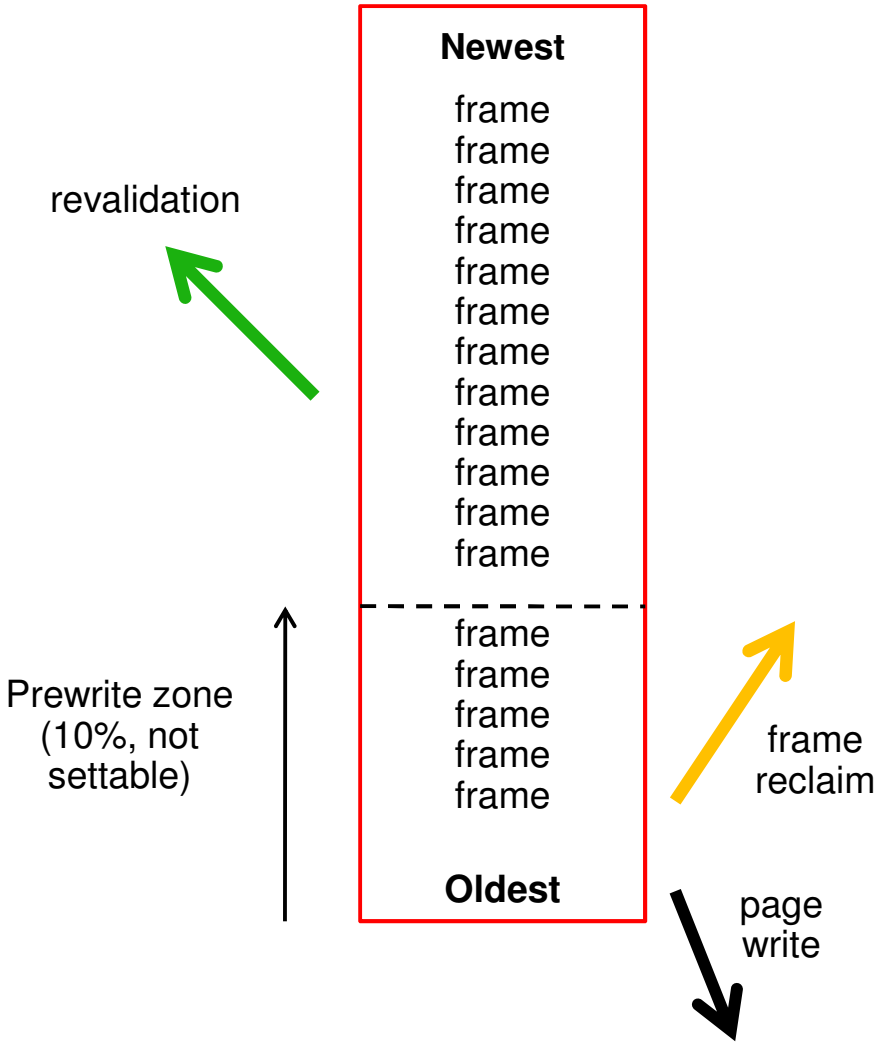
- Page table entry (PTE) contains an "invalid" bit
- What if we:
  - Keep the PTE intact but set the "invalid" bit
  - Leave the frame contents intact
  - Wait for the guest to touch the page
- A touch will cause a page fault, but…
- On a fault, there is nothing really to do except:
  - Clear the "invalid" bit
  - Move the frame to the front of the frame list to show that it was recently referenced
- We call this **trial invalidation**.

Memory Management Algorithm Visualization



UFO = User Frame Owned

New Approach: Global Aging List



- Size of global aging list can be specified…
  … but is best left to the system to manage

- All of the pages here are IBR

- Demand scan fills it from the top

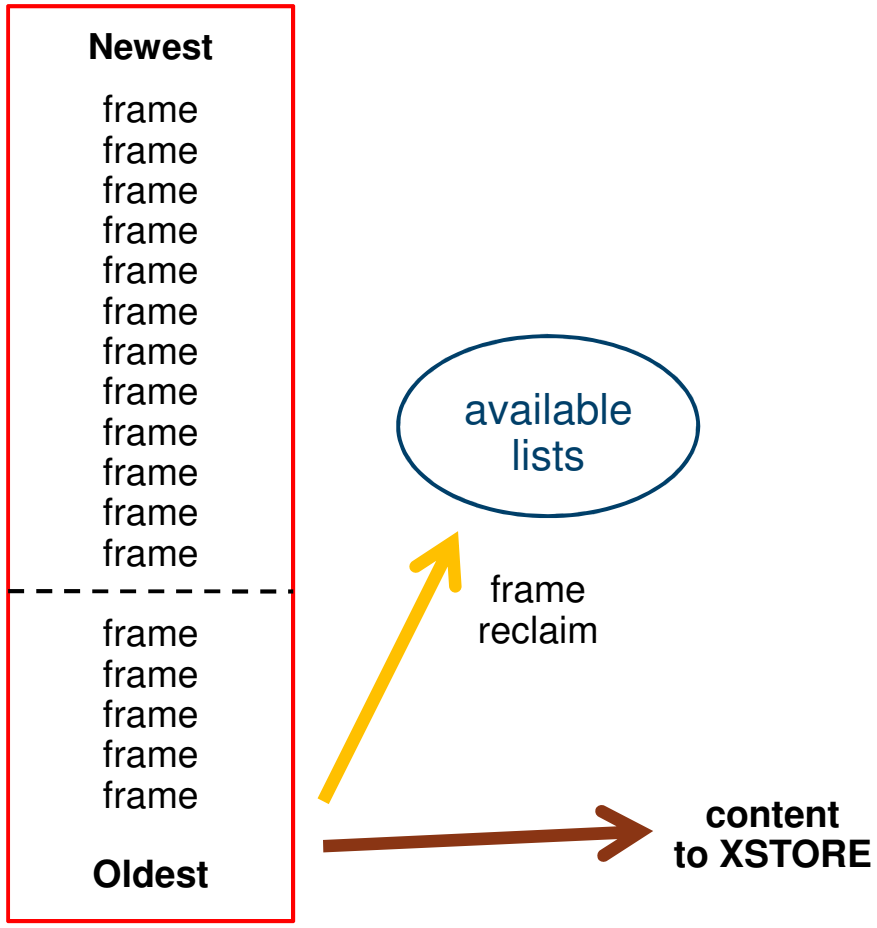
- Revalidated pages return to their owned-lists

- We prewrite changed pages up from the bottom of the list.

- The global aging list accomplishes the age-filtering process that XSTORE used to accomplish.

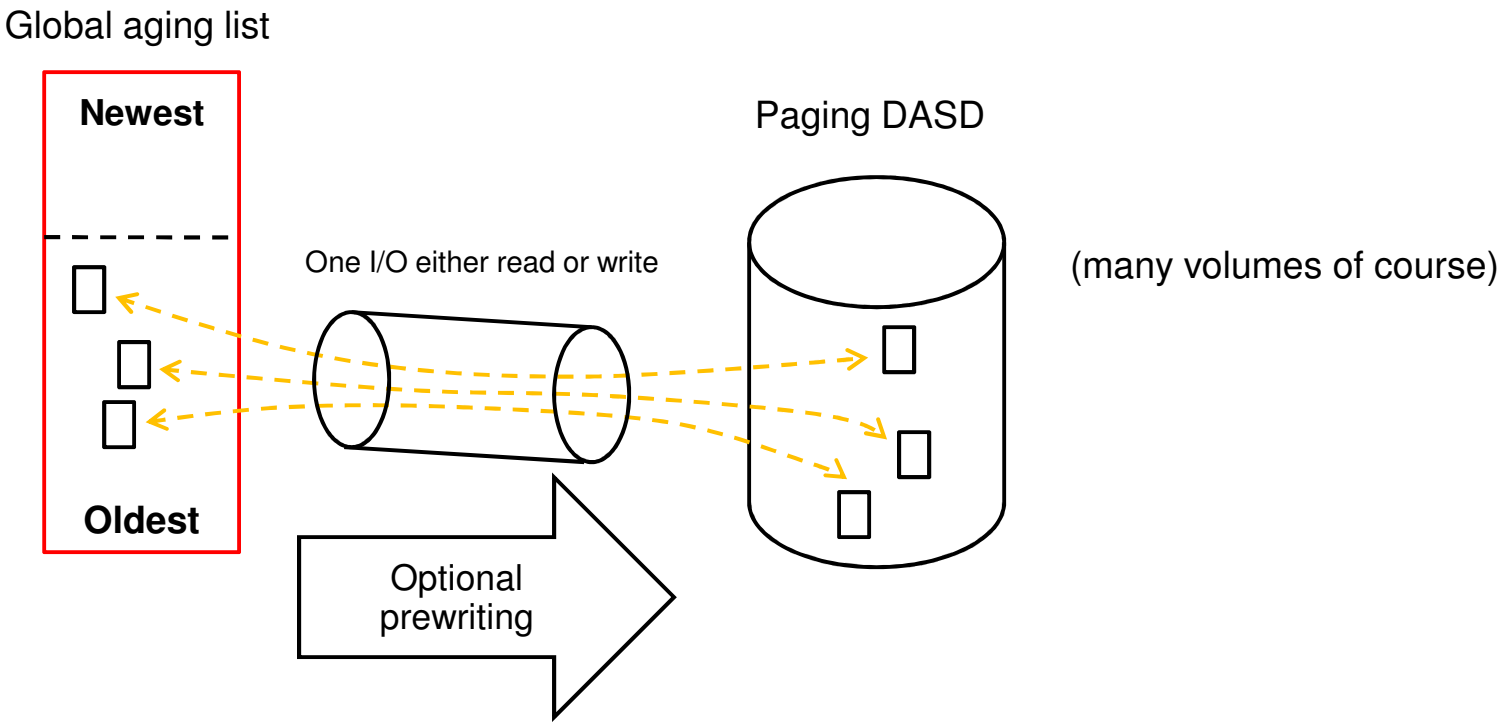
- We no longer suggest XSTORE for paging, but we will use it if it's there.

New Approach: What About XSTORE?



- We will use XSTORE if it is there.
- XSTORE is now the *second* line of defense.
- *When frame is reclaimed*, if XSTORE is present, we put a copy of the page there.
  - Even if the frame has already been prewritten
- On fault, if content is still in XSTORE, it comes back from there.
- If you decide to keep XSTORE, do NOT put MDC in XSTORE unless heavy CMS workload.

New Approach: How We Now Use Paging DASD



Highlights of new DASD techniques:

- A page almost always goes back to its same DASD slot.
  - *Exceptions: clogged or DRAINed volume*
- A page not changed since last read from DASD is almost never rewritten.
  - *Exceptions: DRAINed volume*
- The paging channel program can handle discontiguity on both ends, whether read or write.

New Approach: Large Real Implies Large Virtual, So…

- z/VM holds its DAT management structures in CP-owned pageable address spaces
- These *Page Table Resource Manager* address spaces are named PTRM0000, PTRM0001, …
- You will see them in the z/VM Performance Toolkit FCX134 DSPACESH report
- The number and size of these address spaces control how much logged-on guest real (aka virtual memory) the system can support
- In z/VM 6.2:
  - There were **16** of them: …, PTRM000F
  - We created them as we needed them
  - With 16 of these, we could address **8 TB** of virtual
- In z/VM 6.3:
  - There are now **128** of them: …, PTRM007F
  - We create them all at system initialization
  - With 128 of these, we can now address **64 TB** of virtual

New Behavior: CP SET RESERVED command

- We now do much better at honoring the setting
  - Revisit your uses to see whether you were trying to compensate

- Pages can be now be reserved for **NSS** and **DCSS** as well as virtual machines
  - Set ***after*** **CP SAVESYS** or **SAVESEG** of NSS or DCSS
  - Segment ***does not need to be loaded*** in order to **SET RESERVE** for it
  - A new instance of an NSS or DCSS ***does not*** inherit a pending-purge instance's RESERVED setting
  - Recommended for **MONDCSS**

- You can set a system-wide maximum (**SYSMAX**) on the number of reserved pages
- RESERVED settings ***do not*** survive IPL
  - Consider CP command in the CP directory (not for NSS or DCSS though)

Removed Behavior: Reorder

- z/VM no longer does Reorder processing
  - No longer a trade-off with larger virtual machines

- Commands remain for compatibility but have no impact
  - **CP SET REORDER** command gives RC=6005, “not supported”.
  - **CP QUERY REORDER** command says it’s OFF.

- You will no longer see reorder information in Monitor.

- Be aware of reorder settings when using LGR between z/VM 6.2 and z/VM 6.3

Changed Behavior: Eligible List

- One of the factors to the creation of an **eligible list** is the concept of “loading users”
  - Governed by **SET SRM LDUBUF**
  - A virtual machine is characterized as a “loading user” if its count of page faults in a dispatch slice exceeds a threshold
  - **SET SRM LDUBUF** attempts to keep the system from over-committing paging devices to the point of thrashing

- Changes in z/VM 6.3 paging algorithms can affect the number of virtual machines that are marked as “loading” users and therefore cause **eligible lists** to be formed where they had not formed prior to z/VM 6.3
  - Definition of page fault slightly different
  - Rate at which system can page fault has increased

- Recommend monitoring for eligible lists and adjusting the following as appropriate
  - **SET QUICKDSP**
  - **SET SRM LDUBUF**

- IBM is investigating improvements to avoid the unnecessary eligible list formation.

New or Changed Commands

Commands: Knobs You Can Twist

Concept	Knob	Comments
Size of the global aging list Whether early writes are allowed	Command: CP SET AGELIST … Config file: STORAGE AGELIST … Lookup: CP QUERY AGELIST	Sets the size of the global aging list, in terms of: - A fixed amount (e.g., GB) - A percent of DPA (preferred) The default is 2% of DPA. Seems OK. Sets whether early writes are allowed. (If storage-rich, say NO.)
Amount of storage reserved for a user or for a DCSS	Command: CP SET RESERVED … Config file: STORAGE RESERVED … Lookup: CP QUERY RESERVED …	You can set RESERVED for: - A user - An NSS or DCSS You can also set a SYSMAX on total RESERVED storage. Config file can set only SYSMAX.

Commands: Other Interesting "Queries"

Query or Lookup	Comments
CP INDICATE LOAD	The STEAL-nnn% field no longer appears in the output.
CP INDICATE NSS	Includes a new "instantiated" count. Number of pages that exist. Sum of locus counts might add to more than "instantiated".
CP INDICATE USER	Includes a new "instantiated" count. Sum of locus counts might add to more than "instantiated".
CP INDICATE SPACES	Includes a new "instantiated" count.

Required Planning

Planning for Large Memory

- Normal best practices for migrating from an earlier release certainly apply.
- Change your paging XSTORE into central
  - XSTORE gave us an aging function. It let us catch LRU mistakes.
  - The new IBR concept and global aging list provide the same function but do so more efficiently in central storage.
- Plan enough DASD paging space
  - The system now prewrites pages to DASD.
  - See space calculation on a later slide
- Plan a robust paging DASD configuration
  - Use plenty of paging volumes
  - Make the volumes all the same size
  - Put only paging space on the volumes you use for paging
  - Spread the paging volumes through your LCUs
  - Avoid LCUs that you know are hot on application I/O
  - Use plenty of chpids
  - Do not use ESCON chpids
  - Do not mix ECKD paging and SCSI paging
  - Leave reserved slots in the CP-owned list

Planning for Large Memory

- Look at your **CP SET RESERVED** settings to make sure they're right.
  - Revisit scenarios where you looked at this capability and it wasn't effective

- Add **CP SET RESERVED** settings for DCSSes or NSSes if you like
  - MONDCSS is a good one to consider

- If you increase central, make sure you also increase dump space
  - More guidance will be available on www.vm.ibm.com/techinfo/
  - Download updated *"Allocating Space for CP Hard Abend Dumps"*

Planning DASD Paging Space

- Calculate sum of:
  - Logged-on virtual machines' primary address spaces, plus…
  - Any data spaces they create, plus…
  - Any VDISKs they use, plus…
  - Total number of shared NSS or DCSS pages, … and then …
  - Multiply this sum by 1.01 to allow for PGMBKs and friends

- Add to that sum:
  - Total number of CP directory pages (reported by DIRECTXA), plus…
  - Min (10% of central, 4 GB) to allow for system-owned virtual pages

- Then multiply by some safety factor (1.25?) to allow for growth or uncertainty

- Remember that your system will take a PGT004 if you run out of paging space

- Consider using something that alerts on page space, such as Operations Manager for z/VM

Planning to Learn About Your System's Performance

- While you are still on the earlier release, collect measurement data:
  - Know what your key success metrics are and what their success thresholds are
  - Transaction rates – *only you* know where these are on your workloads
  - MONWRITE files – some tips:
    - When: Daily peaks? Month-end processing? Quarter-end processing?
    - Collection tips: http://www.vm.ibm.com/devpages/bkw/monwrite.html

- Then go ahead and try z/VM 6.3

- When you start running on z/VM 6.3, collect the very same measurement data

- Compare z/VM 6.3 back to z/VM 6.2 to see what the effect is on your workload

Planning to Keep Your System Maintained

- Additional service has shipped, current install media includes second RSU (6302)

- Keep listening:
  - www.vm.ibm.com
  - The IBMVM mailing list

- See also the PSP bucket for z/VM 6.3

Comments on Workloads

z/VM Large Memory: Amenable Workloads

- Best benefit: workloads highly affected by reorder or old demand scan
  - Large guests affected by reorder delays
  - Long demand scans looking for <2G frames

- Less benefit: workloads that were doing fine before
  - Storage-rich workloads
  - Running fine paging to only XSTORE
  - No problems with long demand scans
  - Small guests not affected by reorder

- Let's look at some examples

The "Sweet Spot" Workload

Our synthetic workload called *Sweet Spot* imitates behaviors we have seen in customer-supplied MONWRITE data.

	z/VM 6.2	z/VM 6.3	Delta	Pct. Delta
Cstore	256	384	128	
Xstore	128	0	-128	
External Throughput (ETR)	0.0746	0.0968	0.0222	29.8%
Internal Throughput (ITR)	77.77	105.60	27.83	35.8%
System Util/Proc	31.4	4.7	-26.7	-85.0%
T/V Ratio	1.51	1.08	-0.43	-28.5

By getting rid of both reorders and spin lock contention, we achieved huge drops in %CPU and T/V.

The "Sweet Spot" Workload

- Closer look at how the fairness and workloads may result in different results.
- Sweet Spot workload has four groups of virtual machines. Some benefit more than others.

	z/VM 6.2	z/VM 6.3	Delta	Pct. Delta
System External Throughput	0.0746	0.0968	0.0222	29.8%
User Group 1 ETR	0.0065	0.0128	0.0063	96.9%
User Group 2 ETR	0.0138	0.0236	0.0098	71.0%
User Group 3 ETR	0.0268	0.0264	-0.0004	-1.5%
User Group 4 ETR	0.0275	0.0341	0.0066	24.0%

Workload: The Apache Paging Workload

Our Linux-based workload called *Apache Paging* is built to page heavily to DASD almost no matter how much central or XSTORE we give it.

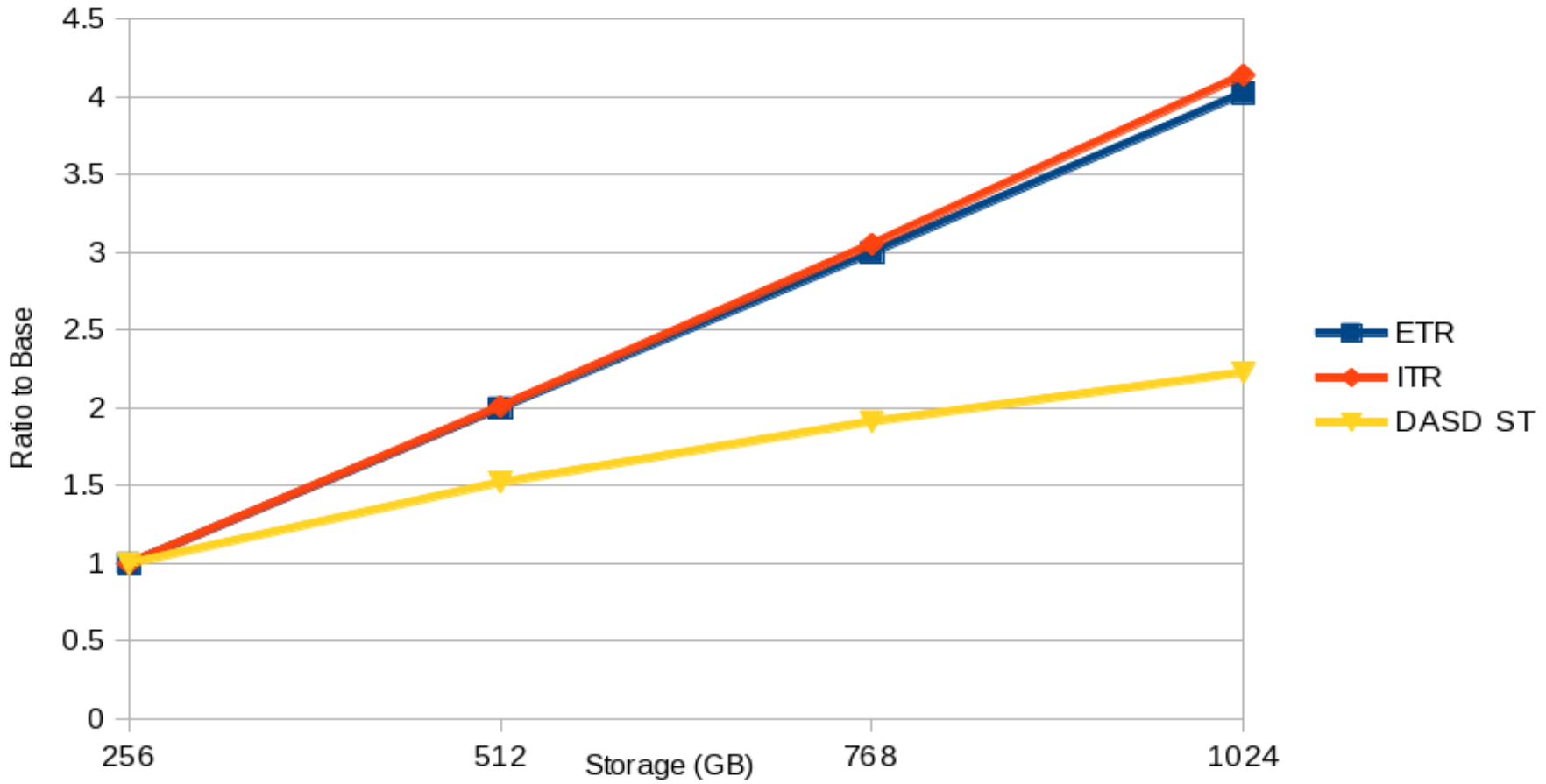
	z/VM 6.2	z/VM 6.3
Cstore (GB)	256	384
Xstore (GB)	128	0
External Throughput (ETR)	1.000	1.024
Internal Throughput (ITR)	1.000	1.017
Xstore paging / second	82489	0
DASD paging / second	33574	31376

This is an example of a workload where the limit comes from something large memory will not fix.

Large Memory Scaling Measurements

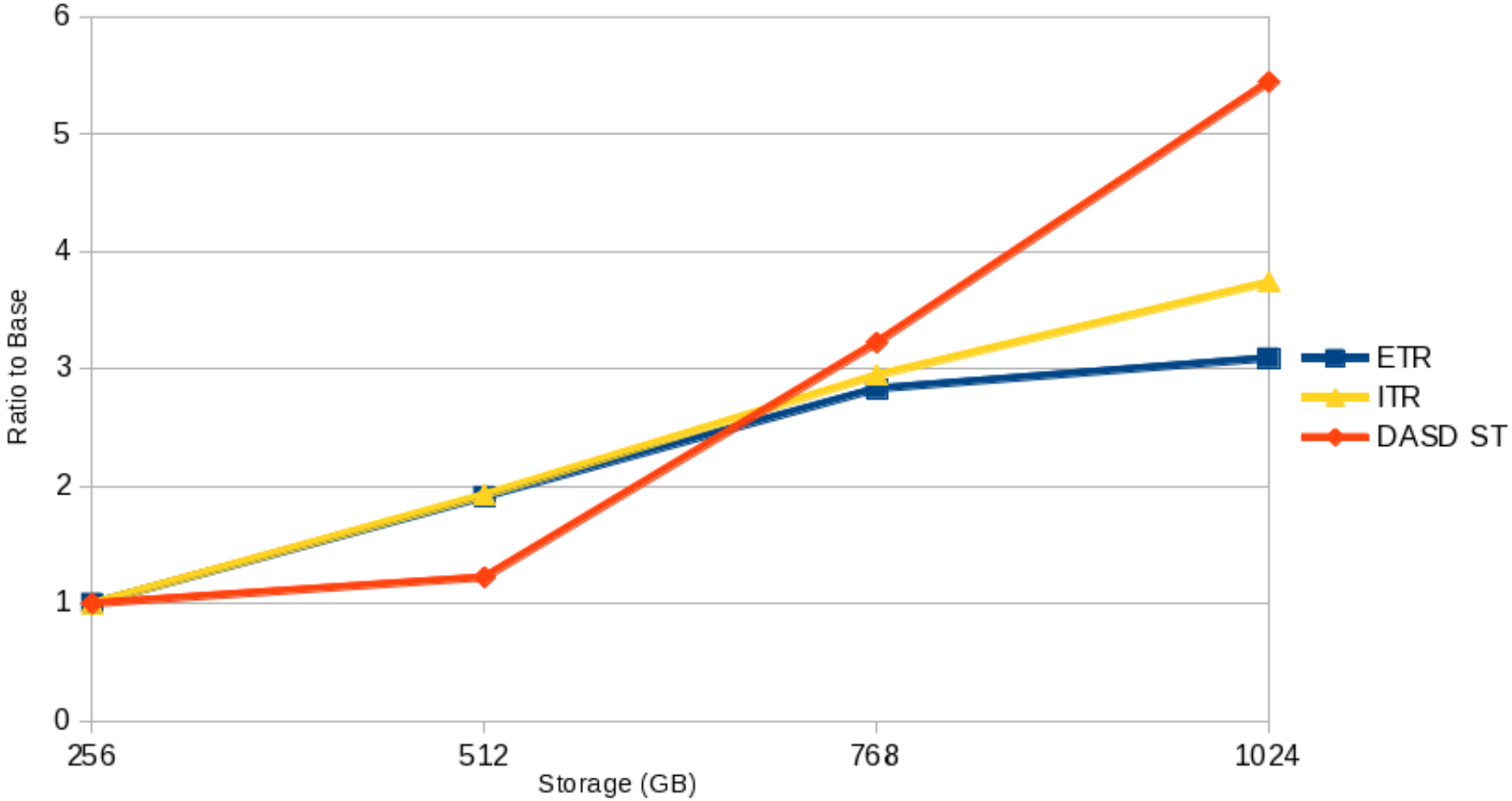
- VIRSTOR – Test case system started with CMS boot strap with controls over memory reference patterns and processor usage.
  - Create workload similar to resource usage from customer Monwrite data
- Linux Apache Static Web serving
- Measure and test levels of servers at peak usage for 256 GB in an overcommitted environment
- Scale up from there to 1 TB
  - All resources scaled up, though note that while additional DASD space was provided, it was on the same storage server.

VIRSTOR Workload in Overcommitted Environment



ETR = External Throughput; ITR = Internal Throughput; DASD ST = DASD Service Time

Apache Workload in Overcommitted Environment



ETR = External Throughput; ITR = Internal Throughput; DASD ST = DASD Service Time

CP Monitor and Performance Toolkit

Large Memory CP Monitor Changes

Domain	Record	Name	Type	Title	Fields, N / D / C
D0	R3	MRSYTRSG	sample	Real Storage Data (Global)	D C
D0	R4	MRSYTRSP	sample	Real Storage Data (Per Processor)	D
D0	R6	MRSYTASG	sample	Auxiliary Storage (Global)	N C
D0	R7	MRSYTSHS	sample	Shared Storage Data	D
D0	R23	MRSYTLCK	sample	Formal Spin Lock Data	N C
D1	R7	MRMTRMEM	config	Memory Configuration Data	N
D1	R15	MRMTRUSR	config	Logged on User	C
D2	R4	MRSCLADL	event	Add User to Dispatch List	D C
D2	R5	MRSCLDDL	event	Drop User from Dispatch List	D C
D2	R6	MRSCLAEL	event	Add User to Eligible List	C
D2	R8	MRSCLSTP	event	System Timer Pop	D
D3	R1	MRSTORSG	sample	Real Storage Management (Global)	N D C
D3	R2	MRSTORSP	sample	Real Storage Activity (Per Processor)	D
D3	R3	MRSTOSHR	sample	Shared Storage Management	N C
D3	R14	MRSTOASI	sample	Address Space Information Record	N C
D3	R15	MRSTOSHL	event	NSS/DCSS/SSP Loaded into Storage	N
D3	R16	MRSTOSHD	event	NSS/DCSS/SSP Removed From Storage	N C
D4	R2	MRUSELOF	event	User Logoff Data	N D C
D4	R3	MRUSEACT	sample	User Activity Data	N D C
D4	R9	MRUSEATE	event	User Activity Data at Transaction End	D C

z/VM Performance Toolkit: Highlights

- **Changed screens:**
  - FCX102 SYSTEM, Some Internal System Counters
  - FCX103 STORAGE, General Storage Utilization
  - FCX133 NSS, NSS and DCSS Utilization and Paging Activity
  - FCX146 AUXLOG, Auxiliary Storage Utilization, by Time
  - FCX147 VDISKS, Virtual Disks in Storage
  - FCX265 LOCKLOG, Spin Lock Log, by Time
- **Deleted screens:**
  - FCX254 AVAILLOG, Available List Management, by Time
  - FCX259 DEMNDLOG, Demand Scan Details, by Time
- **New screens:**
  - FCX290 UPGACT, User Page Activity — *page state transition rates*
  - FCX291 UPGACTLG, User Page Activity (benchmarks a user)
  - FCX292 UPGUTL, User Page Utilization Data — *page residency counts*
  - FCX293 UPGUTLLG, User Page Utilization Data (benchmarks a user)
  - FCX294 AVLB2GLG, Available List Data Below 2G, by Time — *available list counts*
  - FCX295 AVLA2GLG, Available List Data Above 2G, by Time
  - FCX296 STEALLOG, Steal Statistics, by Time — *steal algorithm activity*
  - FCX297 AGELLOG, Age List Log, by Time — *global aging list activity*

Key Considerations

- Do I have enough page space?
- Should Early Write be ON (default) or OFF?
- Do I have eligible lists forming?
- How much memory are virtual machines really using?
- How is SET RESERVE working?
- How effective is the local Invalid But Resident section?
- How effective is the global Age List?

z/VM Performance Toolkit: New Columns and Concepts

New Field	What this means
Inst	*Instantiations:* the rate at which valid memory is being created *Instantiated:* the amount of valid memory
Relse	*Releases:* the rate at which memory is being released
Inval	*Invalidations:* the rate at which demand scan is marking memory invalid as a way to determine whether it is being touched
Reval	*Revalidations*: the rate at which invalid pages are being made valid because somebody touched them
Ready	*Ready reclaims* or *ready steals:* the frame was found and selected for reclaim and had already been prewritten to auxiliary storage
Not Ready	*Notready reclaims* or *notready steals:* the frame was selected for reclaim but we had to wait for the auxiliary write (DASD) to finish before we could take it

z/VM Performance Toolkit: New Columns and Concepts

New Field	What this means
PNR	*Private, not referenced*: the page was read from aux as part of a block read, but it is still marked invalid because nobody has touched it yet
x<2G or *x*>2G	*Below 2 GB* or *Above 2 GB*: tells where the real backing frames are in real central
Sing	*Singles:* free frames surrounded by in-use frames (cannot coalesce)
Cont	*Contigs*: free frames in strings of two or more
Prot	*Protect threshold:* number of frames a singles-obtain must leave on a contigs-list

Page Utilization – FCX109 – DEV CPOWN

```
FCX109      Data for 2014/02/03  Interval 07:28:00 - 07:29:00    Monitor Scan

Page / SPOOL Allocation Summary
PAGE slots available      235865k          SPOOL slots available      4808160
PAGE slot utilization          17%         SPOOL slot utilization          59%
T-Disk space avail. (MB)  .......          DUMP slots available             0
T-Disk space utilization      ...%         DUMP slot utilization           ..%

____ .              .                          .     .     .     .     .      .
< Device Descr. ->                          <------------ Rate/s ------------>
            Volume Area   Area      Used    <--Page---> <--Spool-->          SSCH
Addr Devtyp Serial Type   Extent       %   P-Rds P-Wrt S-Rds S-Wrt Total   +RSCH
1020 3390-9 H2PG00 PAGE     5896620   17    23.4  13.2   ...   ...  36.6     5.7
1021 3390-9 H2PG01 PAGE     5896620   17    20.3  14.0   ...   ...  34.3     5.2
1022 3390-9 H2PG02 PAGE     5896620   17    20.5  13.1   ...   ...  33.6     5.2
1023 3390-9 H2PG03 PAGE     5896620   17    25.7  11.3   ...   ...  37.0     6.0
1024 3390-9 H2PG04 PAGE     5896620   17    26.2  11.7   ...   ...  37.9     6.5
1025 3390-9 H2PG05 PAGE     5896620   17    24.8  13.2   ...   ...  38.0     6.8
1027 3390-9 H2PG07 PAGE     5896620   17    22.7  12.0   ...   ...  34.7     5.8
1028 3390-9 H2PG08 PAGE     5896620   17    22.3  12.6   ...   ...  35.0     6.5
```

Page Utilization History – FCX146 - AUXLOG

```
FCX146      Data for 2014/02/03  Interval 07:28:00 - 07:33:00    Monitor Scan

           <Page Slots>  <Spool Slots>  <Dump Slots>  <----- Spool Files ----->
Interval     Total Used    Total  Used    Total Used  <-Created--> <--Purged-->
End Time     Slots    %    Slots     %    Slots    %   Total    /s  Total    /s
>>Mean>>   235865k   17  4808160    59        0   ..       0   .00      0   .00
07:29:00   235865k   17  4808160    59        0   ..       0   .00      0   .00
07:30:00   235865k   17  4808160    59        0   ..       0   .00      0   .00
07:31:00   235865k   17  4808160    59        0   ..       0   .00      0   .00
07:32:00   235865k   17  4808160    59        0   ..       0   .00      0   .00
07:33:00   235865k   17  4808160    59        0   ..       0   .00      0   .00
```

Early Writes? – FCX297 – AGELLOG (Age List Log)

```
FCX297      Data for 2013/10/15  Interval 09:28:00 - 09:29:00    Monitor Scan

                    <---------------------- Storage ----------------------->
                                        <-- Steal Ready ---> <--- Not Ready ---->
Interval  Size S E <-List Size--> <--RefOnly--> <--Changed--> <Evaluating->
End Time  %DPA Z W Target Current  NoWrt  Write  Write PndWrt   Refd Change
>>Mean>>   2.0 V Y  7787M   7787M   299M      0   480M  3884M    24K      0
09:29:00   2.0 V Y  7787M   7787M   300M      0   479M  3874M    48K      0
```

- Running with default 2% of DPA
- Early Writes is ON (“Y”)

Early Writes? – Write vs. Read – FCX143 - PAGELOG

```
FCX143      Data for 2013/10/15  Interval 09:28:00 - 09:29:00

<---------- Paging to DASD ------------>
                          <-Single Reads-->
Reads Write Total  Shrd Guest Systm Total
   /s    /s    /s    /s    /s    /s    /s
981.3 603.3  1585  46.9 302.2   1.1 303.3
```

- Compare Writes/Second to Reads/Second
  - Reads can be > Writes if pages aren't being changed
  - Writes can be > Reads if the pages aren't being re-referenced and sit idle on DASD
  - Writes can be >> Reads if written during early write, but revalidated before actually stolen

Early Writes Revalidated – FCX297 - AGELLOG

```
FCX297      Data for 2013/10/15  Interval 09:28:00 - 09:29:00    Monitor Scan
<------- Storage ---------------------->       <----- Revalidation ------>
<-- Steal Ready ---> <--- Not Ready ---->   %Of <------- Storage/s ------->
<--RefOnly--> <--Changed--> <Evaluating-> Pages <--RefOnly--> <--Changed-->
 NoWrt  Write  Write PndWrt   Refd Change  Eval  NoWrt  Write  NoWrt  Write
  299M      0   480M  3884M    24K      0    10 560742     .0  2303K  21026
```

- You see above that most of the revalidated pages are pages that were not written yet. Though the majority of those were ones that would have been written.

Eligible Lists Forming? – FCX145 - SCHEDLOG

```
FCX145      Data for 2013/10/15  Interval 09:28:00 - 10:05:00
<- In Eligible List -->
             <Loading->
 E1   E2   E3   E1   E2   E3
 .0   .0   .0   .0   .0   .0
 .0   .0   .0   .0   .0   .0
```

- Subtle changes in “Loading Users” in z/VM 6.3 can cause inadvertent eligible lists.
- Keep an eye on SCHEDLOG and the subset of users in eligible list that are “Loading Users”

Eligible Lists Forming? – FCX154 - SYSSET

```
FCX154      Data for 2013/10/15        System   Settings          Monitor Scan

Initial Scheduler Settings: 2013/10/15 at 09:27:50
DSPSLICE (minor) 5.000 msec.             IABIAS Intensity    90 Percent
Hotshot T-slice  1.999 msec.             IABIAS Duration      2 Minor T-slices
DSPBUF Q1        32767 Openings          STORBUF Q1 Q2 Q3   300 % Main storage
DSPBUF Q1 Q2     32767 Openings          STORBUF Q2 Q3      300 % Main storage
DSPBUF Q1 Q2 Q3  32767 Openings          STORBUF Q3         300 % Main storage
LDUBUF Q1 Q2 Q3    100 % Paging exp.     Max. working set  9999 % Main storage
LDUBUF Q2 Q3        95 % Paging exp.     Loading user         5 Pgrd / T-slice
LDUBUF Q3           85 % Paging exp.     Loading capacity    47 Paging expos.
```

- Review LDUBUF settings and Loading capacity
- From above example, 40 loading users in Q3 would cause an eligible list to form.
  - .85 x 47 = 39.95

Virtual Machine Memory Usage – FCX292 - UPGUTL

```
FCX292      Data for 2013/10/15  Interval 10:04:00 - 10:05:00    Monitor Scan
______        .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .
                <---------------------------------------- Storage ---------------------------------------->
                                  <------------------------- Resident ------------------------->
          Data                                                <------ Invalid But Resident ------>              Base
        Spaces                    <---- Total ----> <-Locked--> <-- UFO --> <-- PNR --> <-AgeList->              Space
Userid    Owned   WSS  Inst Resvd T_All  T<2G  T>2G  L<2G  L>2G  U<2G  U>2G  P<2G  P>2G  A<2G  A>2G XSTOR   AUX   Size
>>Mean>>     .9 1807M 2669M 86780 1529M 7588K 1522M  7567   504K  2378   550K 76557   11M  168K   33M    .0 2222M  3315M
DJSLA101      0 5120M 5113M     0 4404M   19M 4384M     0   208K     0   960K   16K   11M  280K   55M     0 3434M  5120M
```

```
 Data
         Spaces
Userid    Owned   WSS  Inst Resvd
>>Mean>>     .9 1807M 2669M 86780
DJSLA305      0 3100M 6728M     0
```

- "Inst" = pages guest has interacted with in some way which requires z/VM to back the page.
  - Up to the size of the virtual machine
  - Often less than sum of (Resident+XSTOR+AUX) because of pages kept on DASD and in real memory

Reserved? – FCX292 - UPGUTL

```
FCX292      Data for 2013/10/15  Interval 10:04:00 - 10:05:00    Monitor Scan
______        .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .
            <---------------------------------------- Storage ---------------------------------------->
                                <------------------------- Resident ------------------------->
          Data                                                <----- Invalid But Resident ------>              Base
        Spaces                  <---- Total ----> <-Locked--> <-- UFO --> <-- PNR --> <-AgeList->              Space
Userid    Owned   WSS  Inst Resvd T_All  T<2G  T>2G  L<2G  L>2G  U<2G  U>2G  P<2G  P>2G  A<2G  A>2G XSTOR   AUX   Size
WJBLA101      0 5120M 5113M   20M 4404M   19M 4384M     0  208K     0  960K   16K   11M  280K   55M     0 3434M  5120M
```

```
            Data
          Spaces
Userid     Owned    WSS  Inst Resvd
>>Mean>>      .9 1807M 2669M 86780
WJBLA101       0 5120M 5113M   20M
```

- “Resvd” = Amount of pages reserved. May be larger than number of resident pages if virtual machine has not instantiated that memory yet.
- Note that memory is now in bytes (suffixed) not pages.

Virtual Machine Activity – FCX292 - UPGUTL

```
FCX292      Data for 2013/10/15  Interval 10:04:00 - 10:05:00    Monitor Scan
______        .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .     .
            <------------------------------------------ Storage ------------------------------------------>
                                <-------------------------- Resident --------------------------->
          Data                                                <----- Invalid But Resident ------>              Base
        Spaces                  <---- Total ----> <-Locked--> <-- UFO --> <-- PNR --> <-AgeList->              Space
Userid    Owned   WSS  Inst Resvd T_All  T<2G  T>2G  L<2G  L>2G  U<2G  U>2G  P<2G  P>2G  A<2G  A>2G XSTOR   AUX   Size
WJBLA101      0 5120M 5113M   20M 4404M   19M 4384M     0  208K     0  960K   16K   11M  280K   55M     0 3434M  5120M
```

```
           <------------------------ Resident --------------------------->
                                              <----- Invalid But Resident ------>
           <---- Total ----> <-Locked--> <-- UFO --> <-- PNR --> <-AgeList->
Userid     T_All  T<2G  T>2G  L<2G  L>2G  U<2G  U>2G  P<2G  P>2G  A<2G  A>2G
WJBLA101   4404M   19M 4384M     0  208K     0  960K   16K   11M  280K   55M
```

- Get an understanding of where in the lists pages reside:
  - IBR = Invalid But Resident
    - UFO = User Framed Owned section
    - PNR = Private Not Referenced
    - AgeList = part of global age list, but still associated with virtual machine.

Reserved? – FCX290 - UPGACT

```
FCX290      Data for 2013/10/15  Interval 10:04:00 - 10:05:00    Monitor Scan
.            . _____     .     .     .     .     .     .     .     .     .     .
               <------------------------------ Storage ------------------------------>
                                                  <------- Movement/s -------->
           Stl <--- Transition/s ----> <-Steal/s->                        <Migrate/s>
Userid      Wt  Inst Relse Inval Reval Ready NoRdy  PGIN PGOUT Reads Write MWrit  Xrel
DJSLA329     1 64853 74069 38571 18978 15292     0     0     0  4506     0     0     0
```

- PGIN/PGOUT – zero due to not using expanded storage
- Reads would be what would be most important in relationship to Reserved.

- Also note rate of Invaliding and Revalidating
  - Reval / Inval = percentage of times trial invalidation leads to page moving back to top of user frame owned list.

- Note: FCX113 UPAGE still produced, but UPGACT is improved

z/VM Performance Toolkit: New Report FCX295 AVLA2GLG

```
FCX295  Run 2013/04/10 07:38:36          AVLA2GLG                                                    Page   25
                                         Available List Data Above 2G, by Time
From 2013/04/09 16:02:10                                                              SYSTEMID
To   2013/04/09 16:13:10                                                              CPU 2817-744   SN A6D85
For    660 Secs 00:11:00                 "This is a performance report for SYSTEM XYZ" z/VM   V.6.3.0 SLU 0000
```

	<---------- Storage ---------->				<--Times-->		<-Frame Thresh-->		
Interval	<Available>		<Requests/s>		<Returns/s>		<-Empty/s->	Sing	<-Contigs->	
End Time	Sing	Cont	Sing	Cont	Sing	Cont	Sing	Cont	Low	Low	Prot
>>Mean>>	23M	267M	47M	59M	47M	51M	.0	.0	1310	15	15
16:02:40	0	938M	32M	126M	502K	30310	.0	.0	1332	15	15
16:03:10	152K	4556K	50M	89M	49M	59M	.0	.0	1168	15	15
16:03:40	400K	4824K	68M	82M	71M	79M	.0	.0	1321	15	15
16:04:10	0	5896K	49M	72M	52M	70M	.0	.0	2409	15	15
16:04:40	0	2124K	40M	60M	41M	59M	.0	.0	1308	15	15
16:05:10	876K	3488K	54M	52M	55M	51M	.0	.0	1118	15	15
16:05:40	0	3624K	53M	58M	54M	57M	.0	.0	1409	15	15
16:06:10	2016K	4464K	49M	57M	51M	56M	.0	.0	1273	15	15

- Look for the new concepts: Singles Contigs Prot
- Amounts are in bytes, suffixed. Not page counts!
- FCX254 AVAILLOG is no longer produced.

Summary

z/VM Large Memory: Summary

- Objective was to get rid of algorithmic constraints that stopped growth
- Things we got rid of:
  - Reorder
  - Using the scheduler lists to visit users
  - Taking a large amount when we visit a user
  - Excessively favoring VDISKs as regards memory residency
  - Problems in evaluating depletion of available lists
  - Excessive or unnecessary rewriting of DASD
  - Dependency on long-running System z instructions
- Things we added:
  - Visiting all users round-robin
  - Taking only a little when we visit
  - Visiting VDISKs sooner
  - Detecting available list depletion a little more smartly
  - Scatter-to-scatter paging channel program
  - Using trial invalidation
- Effect: workloads constrained by z/VM 6.2 should go better on z/VM 6.3

References

- z/VM CP Planning and Administration
- z/VM CP Commands and Utilities
- z/VM Performance Report: www.vm.ibm.com/perf/