DB2 for z/OS
Real Storage Monitoring, Control and Planning

John Campbell
DB2 for z/OS Development

Session Code: 6006

Disclaimer

The information contained in this presentation is provided for informational purposes only. While efforts were made to verify the completeness and accuracy of the information contained in this presentation, it is provided "as is" without warranty of any kind, express or implied. In addition, this information is based on IBM's current product plans strategy, which are subject to change by IBM without notice. IBM shall not be responsible for damages arising out of the use of, or otherwise related to, this presentation or any other documentation. Nothing contain in this presentation is intended to, nor shall have the effect of, creating any warranties or representations from IBM (or its suppliers or licensors), or altering the terms and conditions of any agreement or license governing the use of IBM products and/or software.

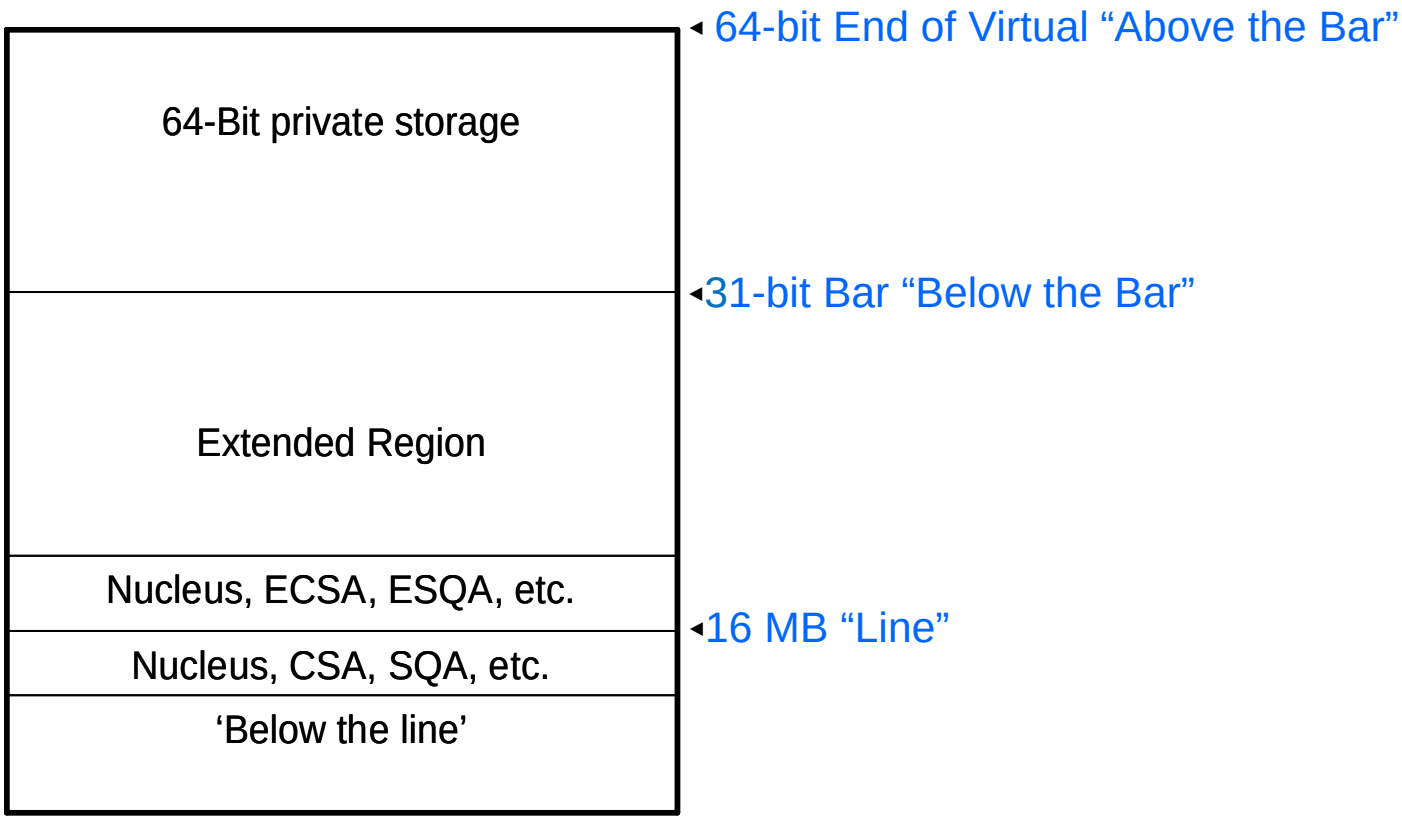
Information regarding potential future products is intended to outline our general product direction and it should not be relied on in making a purchasing decision. The information mentioned regarding potential future products is not a commitment, promise, or legal obligation to deliver any material, code or functionality. Information about potential future products may not be incorporated into any contract. The development, release, and timing of any future features or functionality described for our products remains at our sole discretion.

This information may contain examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious, and any similarity to the names and addresses used by an actual business enterprise is entirely coincidental.

Agenda

- 64-bit shared and 64-bit common
- Real storage controls prior to DB2 10
- Beginner's guide to DISCARD
- Large frame area
- Real storage control in DB2 10 and later
- Performance enhancements and SPIN avoidance
- Futures
- Monitoring REAL/AUX usage
- Summary

Storage Management – Terminology



Virtual Storage the big limiter

- Only 2G of 31-bit storage
- Common areas are taken away from this budget
- Effective area is about 800M to 1600M depending on the individual customer layout
  - Big IMS/TM users may have only as little as 800M region
- Answer is to scale horizontally not vertically
  - Add more members to a data sharing group
  - Add more stand-alone DB2 subsystems if data sharing not required

DB2 V8 the start of the 64-bit world

- Starting from Version 8, DB2 can allocate 64-bit memory objects
- Design of DB2 storage manager is the same as the Data Space manager
- Same types of DB2 storage as 31-bit allowed in 64-Bit (except stack)
  - Fixed
  - Variable
  - Getmained
  - Fixed and Variable are a subset of Getmained

Buffer pools and 64-bit private objects

- Objects are all 64 bit private
- No sharing across DB2 system address spaces
- Buffer pool objects have their own objects in the DBM1 system address space
  - Buffer manager does not use Storage manager to obtain storage so tracking buffer pools gets a little bit more difficult
- All objects that are tagged as DUMP YES will appear in a dump
  - If the buffer pools are in total more than 800M at DB2 start-up then ALL buffer pools will be tagged DUMP NO so no buffer pools will be dumped

DB2 9 for z/OS the start of 64-bit shared (private)

- New invention from z/OS
- 64-bit shared is available
  - Different than CSA/ECSA
  - Token required to connect to a memory object
  - Not prone to the overlays of other users of CSA/ECSA
  - DB2 gets a 128G object
- Used by DDF and some utilities
  - No high usage
- Difficult to track as the HVSHARE area only has a number for the LPAR, not the individual users e.g., DB2 subsystem
  - Usage is small enough to not be much of an issue

DB2 10 for z/OS the start of 64-bit shared use for almost all thread control blocks

- 64 bit shared is used in preference to private storage
  - Stack (save areas, working storage)
    - Allows one stack to be used by multiple system address spaces cutting down REAL footprint and complexity
  - Thread storage pools now shared, no need to use expensive cross memory moves if they were required before
    - Reduced complexity, addressable storage is always available
  - Use of Shared storage now dominates DB2
  - Downside
    - z/OS treats 64 bit Shared like common storage
    - Spin locks more often used for serialisation
      - See later charts

DB2 10 for z/OS the start of 64-bit common

- New invention from z/OS
- Allows DB2 to move some of the common areas
  - Blocks where it is awkward or impossible to know the connect token in advance
- Some usage in DB2 10 DB2 uses a 6G object for the common objects
  - Difficulty in starting DB2 if the default HVCOMMON area for LPAR is not enlarged to cope with multiple DB2 subsystems

Controlling REAL storage DB2 V8 and V9

- CONTSTOR
  - Contract thread storage regularly based on the number of times the thread commits
  - Very cheap, cost is amortised over at least 5 commits
- MINSTOR
  - Can be expensive
  - Orders the free chain in size order
  - A lot of CPU can be burned running through storage chains
- Heavy hitters for storage include Statement cache, EDM pool, Buffer pools
  - Buffer pools are usually densely packed virtual storage
  - Almost a V=R relationship
  - Can dominate the REAL storage landscape
  - Reducing buffer pool sizes can substantially reduce the REAL storage footprint
    - Down side is performance loss as read hit ratio degrades and page residency time decreases

REAL Storage control – DB2 10

- The frequency of contraction mode can also be controlled by system parameter REALSTORAGE_MANAGEMENT
- REALSTORAGE_MANAGEMENT options include:
  - OFF
    - Do not enter 'contraction mode' unless the REALSTORAGE_MAX boundary is approaching OR z/OS has notified DB2 that there is a critical aux shortage
  - ON
    - Always operate in 'contraction' mode
    - This may be desirable for LPAR with many DB2 subsystems or development/test systems
  - AUTO (the default)
    - When significant paging is detected, 'contraction' mode will be entered
    - But also under normal operating conditions with no paging DB2 will still perform DISCARD at thread Deallocation or after 120 commits
- Important note
  - Contraction mode is not exited immediately upon relief to avoid constant toggling in and out of this mode
  - New messages (DSNV516I, DSNV517I)

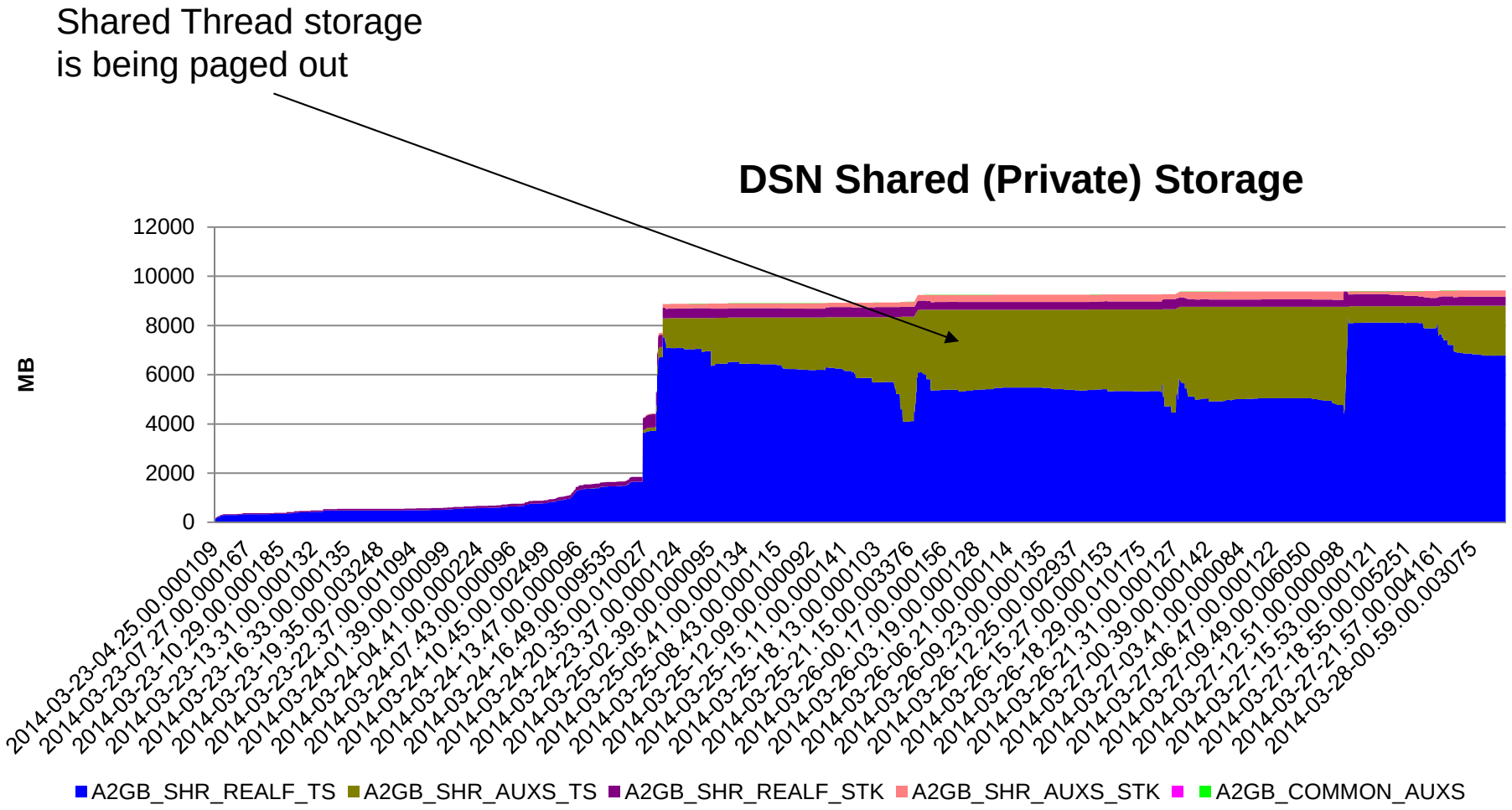
Beginners guide to DISCARD

- To KEEP REAL or not to KEEP REAL that is the question
- 64-bit pages are DISCARDED not freed
- First steps in DB2 V8 to control REAL storage
  - Add DISCARDDATA to RESET to tell z/OS to un-back the frames that are no longer needed
  - Did it work?
    - z/OS will not release frames until AVQLOW condition
    - DB2 is still charged for the storage unless the system begins to page
- When I say discard, I mean it …
  - New option in z/OS APAR OA15666
  - REQUEST=DISCARDDATA, KEEPREAL=NO (“Hard Discard”)
- Allows DB2 to discard real frames without hitting AVQLOW condition
- DB2 support in PK25427, DB2 uses KEEPREAL=NO and the statistics are now accurate

Large frame area

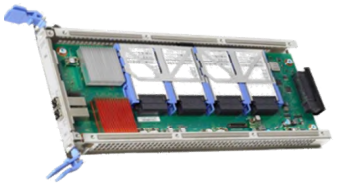
- Good for DB2 performance
- 1M size page large frames arrived in DB2 10
- Not good to use if the system is paging
  - z/OS converts unused 1M size large frames into 256 x 4K size small frames
  - DB2 is not allowed to use these frames since DB2 is non swappable and uses preferred storage
  - z/OS expects to recombine the small frames back into large frames later
    - Cannot be used for long term page fix
  - Could be that CICS ends up using the Large Frame Area

How to make large frames into small frames

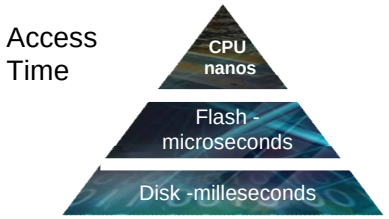


If you have no choice but to page – use Flash Express

- Flash Express is a PCIe IO adapter with NAND Flash SSDs
- Physically comprised of internal storage on Flash SSDs
- Used to deliver a new tier of memory- storage class memory
- Uses PCIe I/O drawer
- Sized to accommodate all LPAR paging
  - Each card pair provides 1.4 TB usable storage
  - Maximum 4 card pairs (4 x 1.4=5.6 TB)
- Immediately usable
  - Simplifies capacity planning
  - No intelligent data placement needed
  - Full virtualization across partitions
- Robust design
  - Delivered as a RAID10 mirrored pair
  - Designed for long life
  - Designed for concurrent firmware upgrade
- Secured
  - Flash Express adapter is protected with 128-bit AES encryption
  - Key Management provided based on a Smart Card



One Flash Express Card



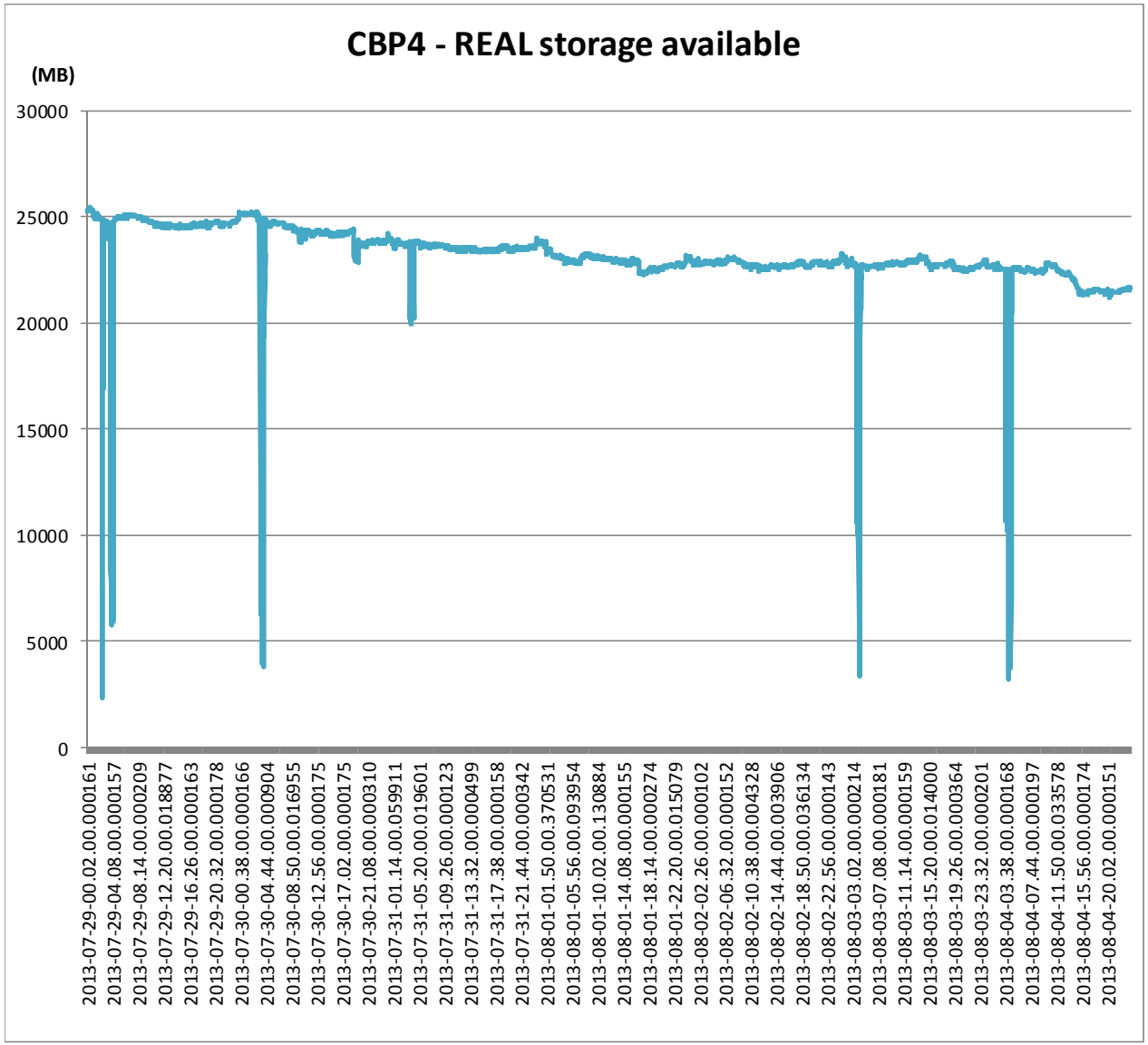
Real storage control

- Paging rate is a critical performance measure for any z/OS system
  - When shortage of REAL frames occurs, frames are moved to AUX (DASD)
  - Having DB2 paged out is a not good thing for performance
  - Paging should be minimised
- Page fixing buffer pools is a good idea for performance
  - It avoids page fix and page free for high activity buffer pools (heavy I/O)
  - Page fixing 1M size real storage page frames reduces TLB misses (saves CPU)
- But if insufficient REAL storage provisioned for the LPAR
  - LPAR begins to page and DB2 is a candidate for page stealing
    - Thread and EDM Pool storage is paged out
    - Performance problems as data is rapidly paged back in

Real storage control …

- "I have a large LPAR (128G) and my DB2 (6G) got paged out …"
- Why is that?
  - Shift in workload with REAL frames stolen by overnight batch processing
    - Poor response times in the first few minutes of the online day
    - A lot of rapid paging going on
    - Huge increase in number of threads causing application scaling issues (lock contention, global contention)
  - REAL frames stolen by DB2 utilities
    - REORG uses REAL storage for in memory sort e.g., 64G
    - DFSORT defaults
      - EXPMAX=MAX <<<<<< Make maximum use of storage
      - EXPOLD=MAX <<<<<< Allow paging of old frames
      - EXPRES=0 <<<<<< Reserved for new work
  - Dump capture
    - Excessive dump time caused by paging on the LPAR may cause sysplex-wide sympathy sickness slowdowns

Real storage usage and DFSORT settings …



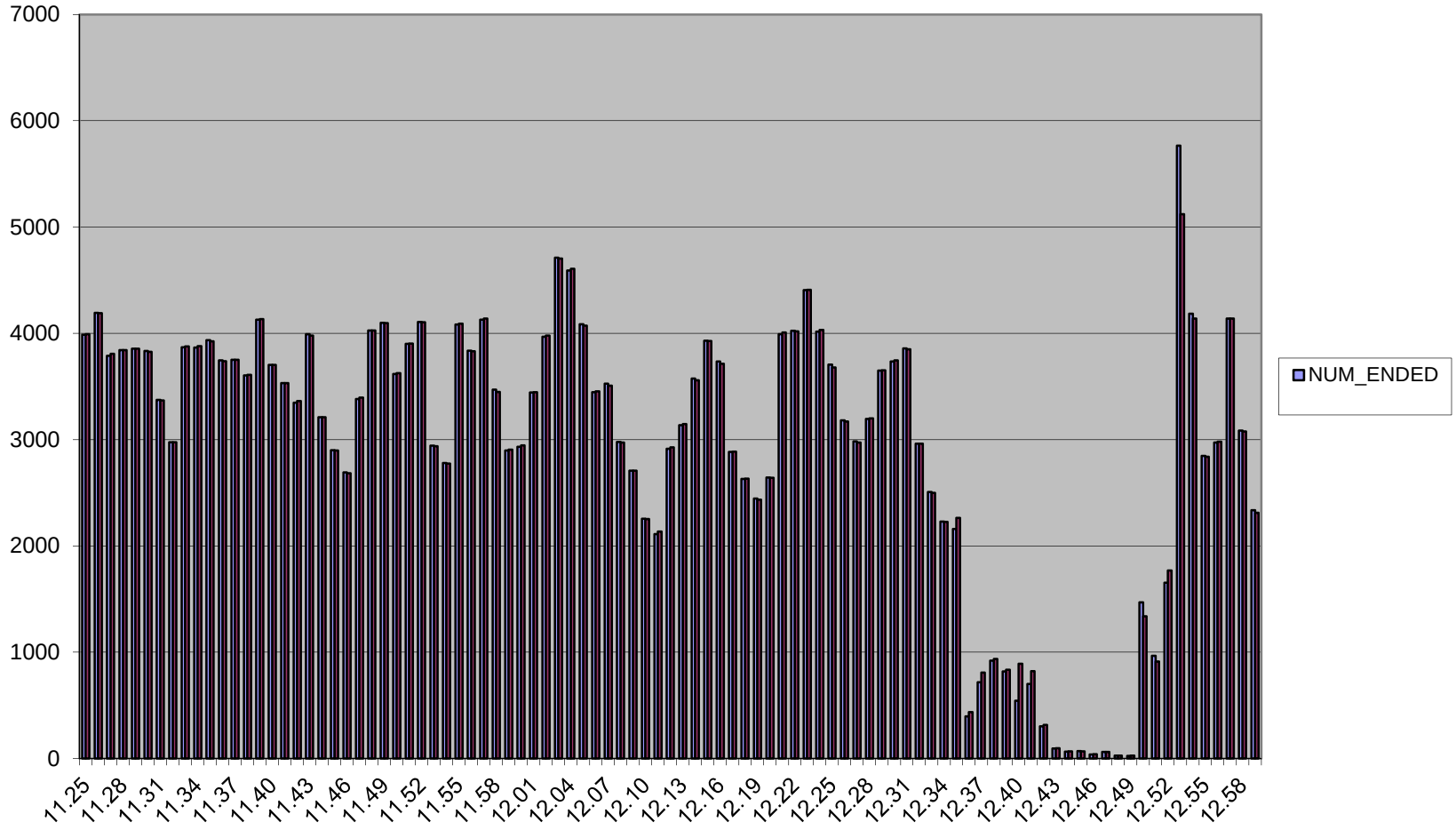
DB2 pages paged out and a dump happens?

- What is my exposure?
  - Increased MASTER CPU time, z/OS tries to steal frames to meet the excessive demand caused by the dump
  - Elongated dump times
- Auxiliary storage number > 0
  - In theory each page could have to be paged in
    - z/OS can have the page in both places, REAL and AUX
    - If total size across all bufferpools is more than 800MB then the bufferpools are not dumped
  - No prefetch on AUX storage, so all synchronous I/O
  - Worst case is the number of pages * page-in I/O time
    - For example 2GB of 4K pages * 3ms = 524288 * 0.003 = 26 mins)
  - Guinness world record for a dump 37 mins
  - Full Sysplex hang resulted

SYSPLEX sympathy sickness

- Slowdown and no apparent reason why?
- DB2 taking the dump may have TCBs non-dispatchable
- P-Lock negotiation affected
- Locks not released in a timely manner
- Excessive dump time caused by paging on the LPAR may cause massive sympathy sickness slowdowns
- How can a member slow down when there is plenty of CPU/Storage on the LPAR
  - May be the owner of a P-lock is being dumped or is paging a lot

The DUMP effect – no transactions being processed



Real storage control …

- Make sure LPAR has enough REAL storage
- REAL storage upgrade is the cheapest and easiest performance upgrade
  - REAL storage shortage not only can cause performance issues but if DUMPs are needed then it can cause a small issue to become a massive SYSPLEX failure
  - Cheapest because MLC and other charges do not factor in the amount of REAL storage
  - Vendors do not charge by the amount of REAL on the CEC/CPC processor
- Specify z/OS WLM STORAGE CRITICAL for DB2 system address spaces
  - To safeguard the rest of DB2
  - Tells WLM to not page these address spaces
  - Keeps the thread control blocks, EDM and other needed parts of DB2 in REAL
  - Prevents the performance problem as the Online day starts and DB2 has to be rapidly paged back in

Real storage control …

- Make sure MAXSPACE is set properly and defensively
  - Represents the total amount of storage for captured dumps for the entire LPAR
  - MAXSPACE value should not be set so high that paging can occur causing massive issues to the LPAR
  - If multiple DB2s on same LPAR can wildcard to the same dump, then MAXSPACE needs to be set appropriately
  - MAXSPACE=16G is a good start to cope with more than 90% of all cases
    - But there are MVS defects around which are inflating DUMP size
    - Fixing z/OS APARs available to handle and minimise DUMP size: OA39596, OA40856 and OA40015
  - MAXSPACE requirement should be
    - (DBM1 – Buffer pools) + Shared memory + DIST + MSTR + IRLM + COMMON + ECSA
  - Work is underway to get the exact formula based on all the new IFCID 225 fields
    - Once the formula is properly tested, will be posted on the various websites and Info APARs

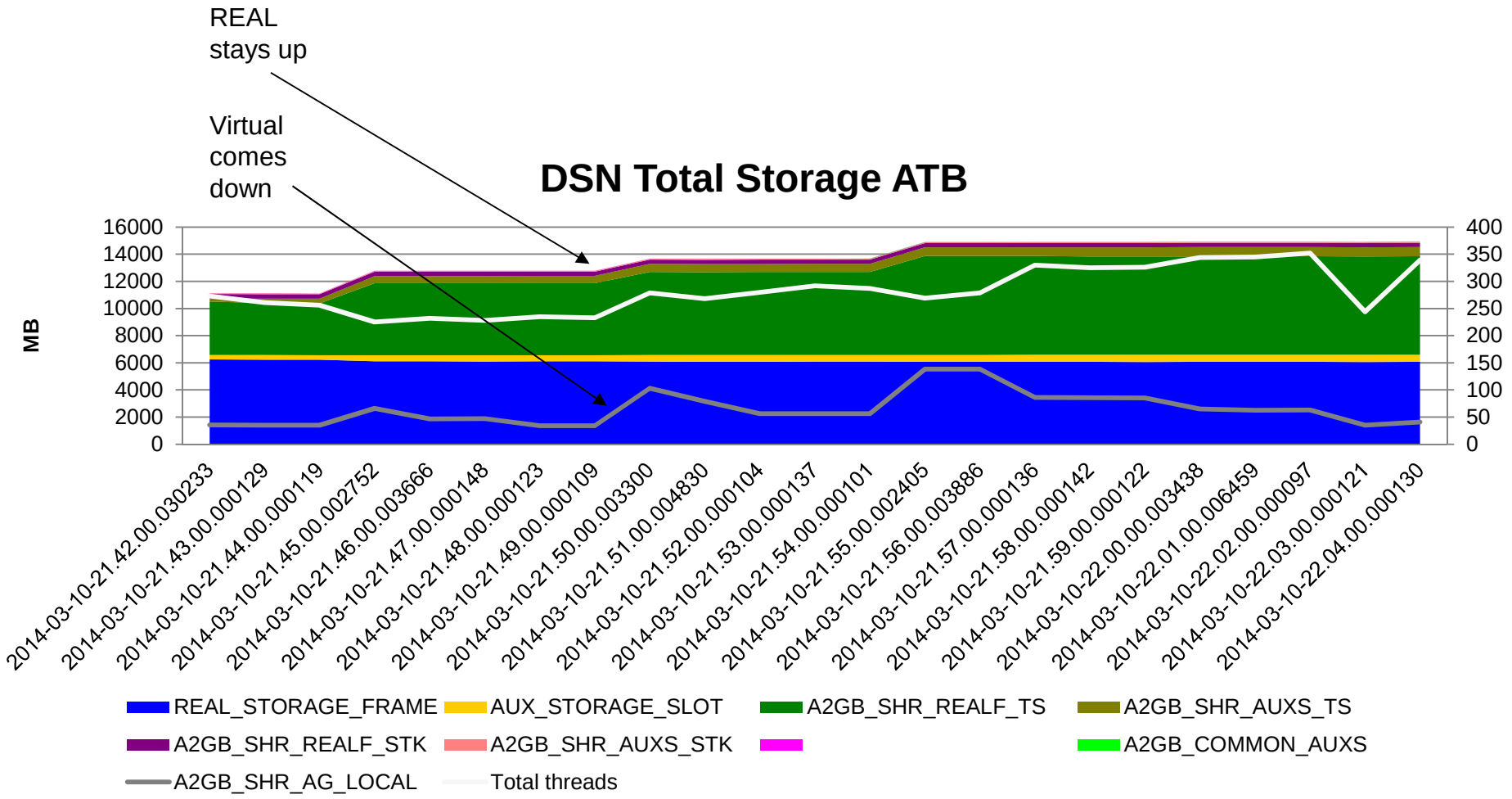
Real storage control …

- Make sure REALSTORAGE_MANAGEMENT=AUTO (default)
  - Particularly when significant paging is detected, “contraction mode” will be entered to help protect the system
    - “Unbacks” virtual pages so that a REAL frame or AUX slot is not consumed for this page
  - Use automation to trap the DSNV516I (start) and DSNV517I (end) messages
- As DB2 approaches the REALSTORAGE_MAX threshold
  - “Contraction mode” is also entered to help protect the system
- Control use of storage by DFSORT
  - Set EXPMAX down to limit maximum DFSORT usage
  - Set EXPOLD=0 to prevent DFSORT from taking "old" frames from other workloads
  - Set EXPRES=% or n {reserve enough for MAXSPACE}
- z/OS parameter AUXMGMT=ON
  - No new dumps are allowed when AUX storage utilization reaches 50%
  - Current dump data capture stops when AUX storage utilization reaches 68%
  - Once the limit is exceeded, new dumps will not be processed until the AUX storage utilization drops below 35%

Performance enhancements / SPIN avoidance

- SPIN locks used by z/OS RSM can cause performance issues when DISCARD processing is running on many CPUs at the same time
- Possible LPAR outage in severe cases
- DB2 APAR PM88804 fixes this outage issue
  - Reduce uncaptured time by freeing the virtual but not DISCARDing the REAL when DISCARD mode is off
    - REALSTORAGE_MANAGEMENT=OFF
  - No DISCARD processing when REALSTORAGE_MANGEMENT=AUTO and no paging
    - Immediate CPU performance reduction when transactions achieve poor thread reuse and there is a high rate of thread deallocation
    - Reduced CPU resource consumption, reduced exposure to SPIN locks
  - Customers using REALSTORAGE_MANAGEMENT=AUTO are essentially running OFF unless paging is occurring

After PM88804 Possible side effect of no DISCARD
Virtual flat, Increasing REAL



Virtual Contraction – DB2 APAR PM86952

- After APAR PM88804 some extreme customer cases saw some big REAL storage growth
- Very workload dependent
- New APAR PM86952 provides VIRTUAL contraction (NOT REAL)
- Contraction of Virtual has an implicit effect on REAL storage
  - Virtual Storage decreases
  - REAL Storage decreases

DB2 APAR PM99575

- Time to put back the DISCARD in normal processing to prevent runaway storage growth
- Keep the performance advantage when REALSTORAGE_MANAGEMENT=OFF
- Protect the system as much as DB2 can from SPIN lock contention and spin out
  - REQUEST=DISCARDDATA, KEEPREAL=YES (“Soft Discard”)
- Trade off accurate Statistics for system availability and reduced REAL storage use
  - Statistics are not accurate after PM99575 because frames will not be stolen back to reduce the count until paging occurs
  - Statistics will be a high water mark on most systems
  - Statistics will be fairly accurate on systems with small amounts of paging

Storage manager changes: ZPARM RSM REALSTORAGE_MANAGEMENT=xxx

After PM88804

- ENF 55 signal means DISCARD KEEPREAL=NO
- RSM=OFF means No DISCARD
- RSM=AUTO with no paging means no DISCARD at Thread Deallocation or 120 commits
- RSM=AUTO with paging or RSM=ON means DISCARD with KEEPREAL=NO at Deallocation or 30 commits. STACK also DISCARDED
- REALSTORAGE_MAX means DISCARD KEEPREAL=NO at 80%

After PM99575

- ENF 55 signal means DISCARD KEEPREAL=NO
- RSM=OFF means No DISCARD
- RSM=AUTO with no paging means DISCARD with KEEPREAL=YES at Thread Deallocation or 120 commits
- RSM=AUTO with paging or RSM=ON means DISCARD with KEEPREAL=YES at Deallocation or 30 commits. STACK also DISCARDED
- REALSTORAGE_MAX means DISCARD KEEPREAL=NO at 100%

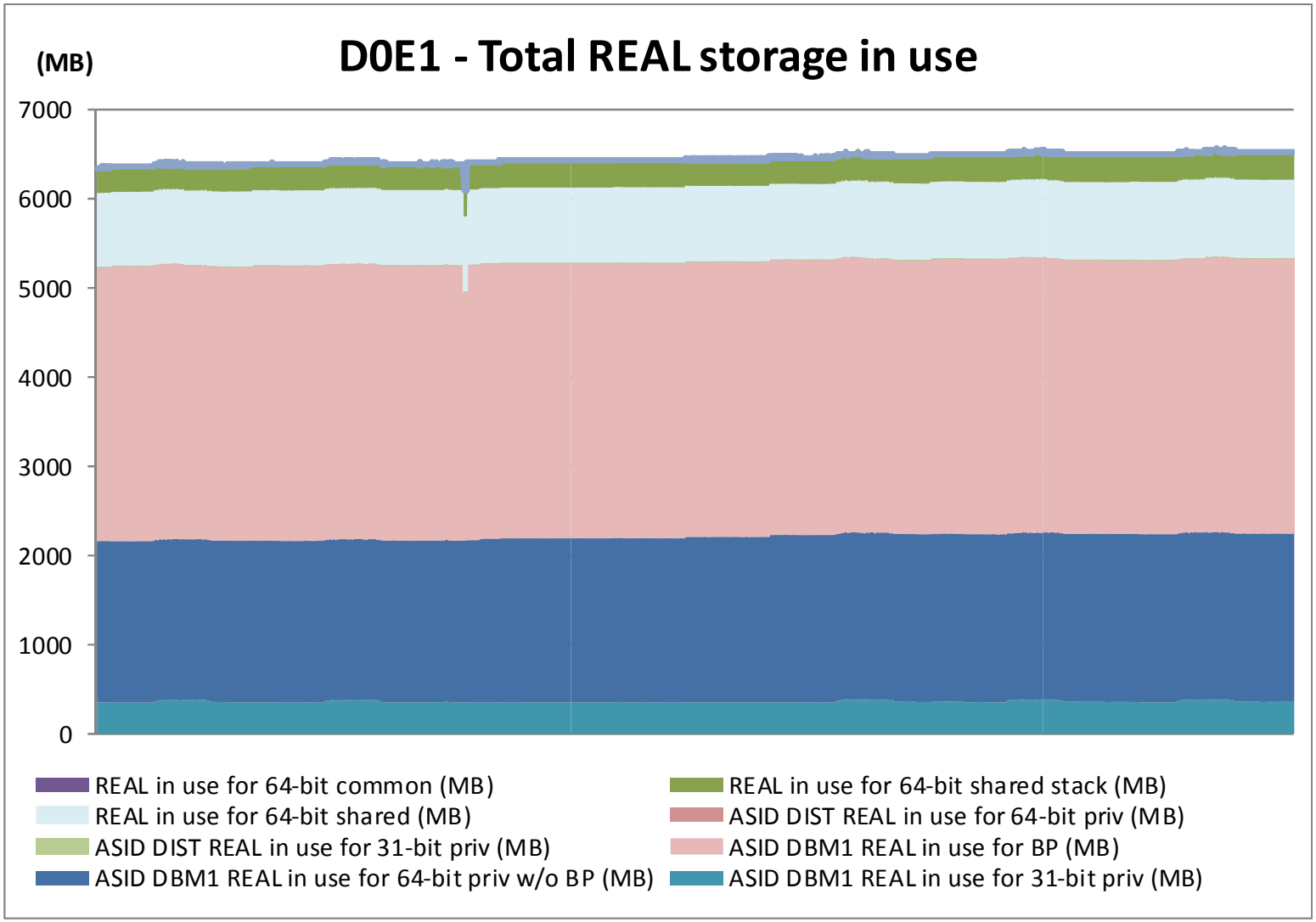
Just when we thought we were out of the woods - CRITICALPAGING

- If you have XCF CRITICALPAGING ENABLED, you also need to apply z/OS RSM APAR OA44913 or provision Flash Express on the LPAR
  - Without this APAR, the discarded shared frames with KEEPREAL(YES) are not stolen to replenish frame queues when paging occurs
  - Either APAR may be applied independently without the other
- If z/OS APAR OA44913 is not applied and Flash Express is also not provisioned then
  - No benefit from DB2 APAR PM99575 especially if running with RSM=AUTO, XCF CRITICALPAGING is enabled, and the system starts to page …
- Do I have it?
  - D XCF,COUPLE
- To Activate
  - Update COUPLExx with: FUNCTIONS ENABLE(CRITICALPAGING)

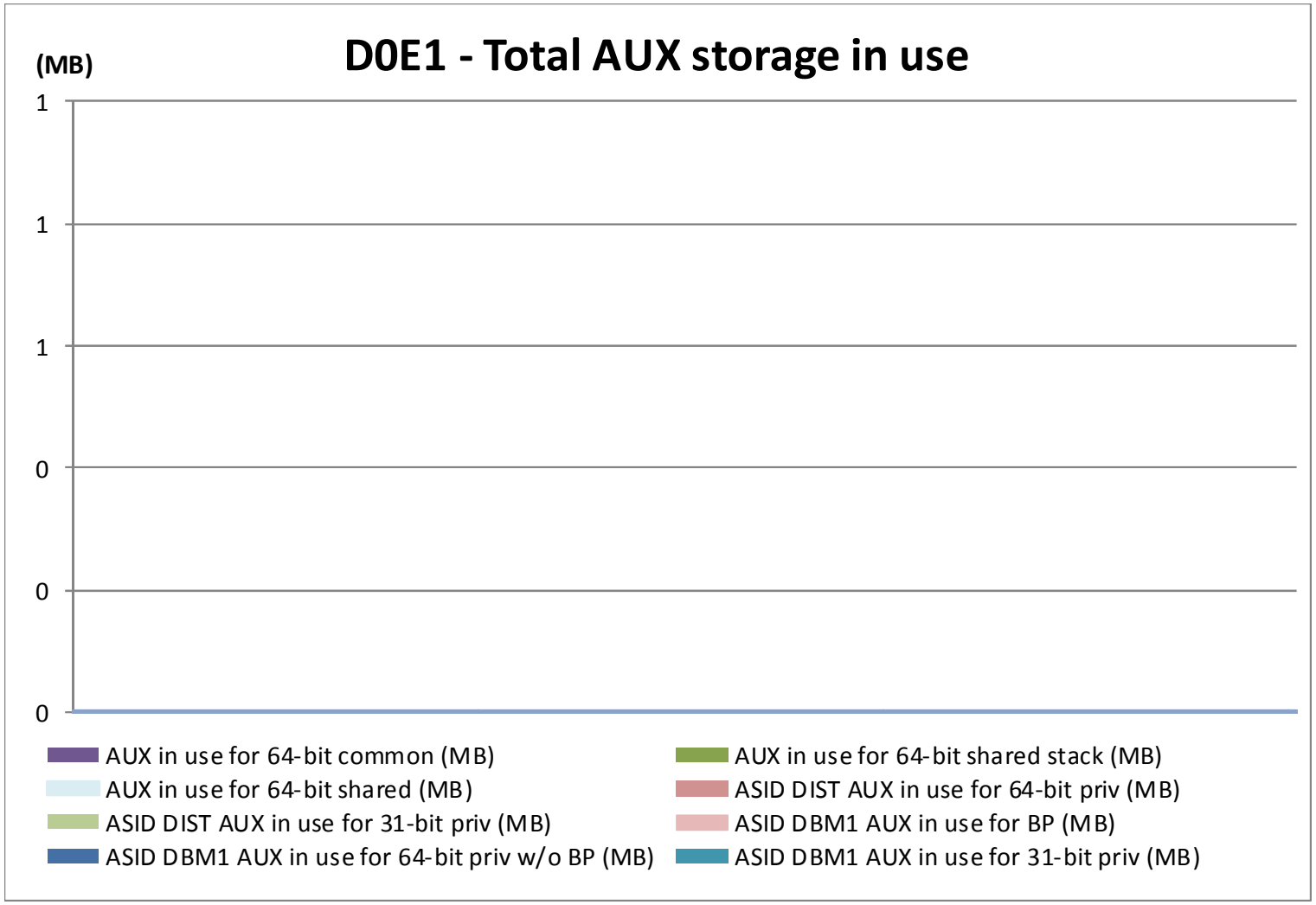
Futures

- Nothing more in plan for DB2
  - Storage manager now DISCARDING frames
- DUMP size
  - Inflated dump size and MAXSPACE requirement for full dump
  - DB2 and z/OS Development investigating – WATCH THIS SPACE
- REAL storage statistics
  - Since DISCARDED frames are not removed from the total, then the statistics are not as accurate as they could be
  - z/OS to look for a solution for this – WATCH THIS SPACE

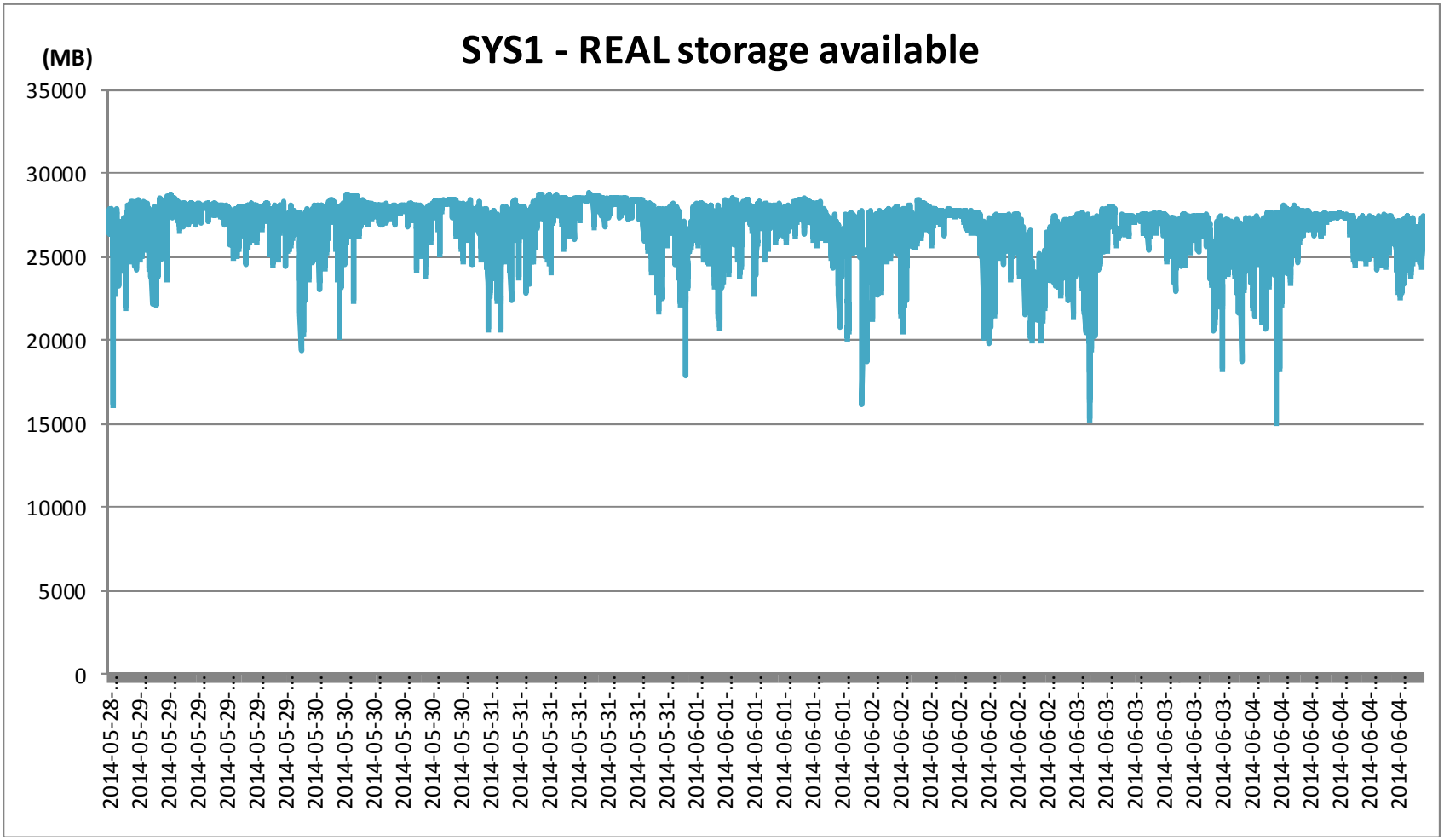
Monitoring REAL/AUX storage usage – Sample graph #1



Monitoring REAL/AUX storage usage – Sample graph #2



Monitoring REAL/AUX storage usage – Sample graph #3



Monitoring REAL/AUX storage usage – Based on OMPE PDB

- #1 - Stacked AREA graph - one for each DB2 member (one sheet per member)

```
(REAL_STORAGE_FRAME - A2GB_REAL_FRAME)*4/1024     AS  DBM1_REAL_PRIV_31BIT_MB
(A2GB_REAL_FRAME - A2GB_REAL_FRAME_TS)*4/1024     AS  DBM1_REAL_PRIV_64BIT_BP_MB
A2GB_REAL_FRAME_TS*4/1024                         AS  DBM1_REAL_PRIV_64BIT_XBP_MB
(DIST_REAL_FRAME - A2GB_DIST_REAL_FRM)*4/1024     AS  DIST_REAL_PRIV_31BIT_MB
A2GB_DIST_REAL_FRM*4/1024                         AS  DIST_REAL_PRIV_64BIT_MB
A2GB_COMMON_REALF*4/1024                          AS  REAL_COM_64BIT_MB
A2GB_SHR_REALF_TS*4/1024                          AS  REAL_SHR_64BIT_MB
A2GB_SHR_REALF_STK*4/1024                         AS  REAL_SHR_STK_64BIT_MB
```

- #2 - Stacked AREA graph - one for each DB2 member (one sheet per member)

```
(AUX_STORAGE_SLOT - A2GB_AUX_SLOT)*4/1024         AS  DBM1_AUX_PRIV_31BIT_MB
(A2GB_AUX_SLOT - A2GB_AUX_SLOT_TS)*4/1024         AS  DBM1_AUX_PRIV_64BIT_BP_MB
A2GB_AUX_SLOT_TS*4/1024                           AS  DBM1_AUX_PRIV_64BIT_XBP_MB
(DIST_AUX_SLOT - A2GB_DIST_AUX_SLOT)*4/1024       AS  DIST_AUX_PRIV_31BIT_MB
A2GB_DIST_AUX_SLOT*4/1024                         AS  DIST_AUX_PRIV_64BIT_MB
A2GB_COMMON_AUXS*4/1024                           AS  AUX_COM_64BIT_MB
A2GB_SHR_AUXS_TS*4/1024                           AS  AUX_SHR_64BIT_MB
A2GB_SHR_AUXS_STK*4/1024                          AS  AUX_SHR_STK_64BIT_MB
```

- #3 - Line graph - one for each LPAR

```
QW0225_REALAVAIL*4/1024                           AS  REAL_AVAIL_LPAR_MB
```

Monitoring REAL/AUX storage usage – Mapping for reference

IFCID FIELD	OMPE FIELD	OMPE PDB COLUMN NAME	MEMU2 Description
QW0225RL	QW0225RL	REAL_STORAGE_FRAME	DBM1 REAL in use for 31 and 64-bit priv (MB)
QW0225AX	QW0225AX	AUX_STORAGE_SLOT	DBM1 AUX in use for 31 and 64-bit priv (MB)
QW0225HVPagesInReal	SW225VPR	A2GB_REAL_FRAME	DBM1 REAL in use for 64-bit priv (MB)
QW0225HVAuxSlots	SW225VAS	A2GB_AUX_SLOT	DBM1 AUX in use for 64-bit priv (MB)
QW0225PriStg_Real	SW225PSR	A2GB_REAL_FRAME_TS	DBM1 REAL in use for 64-bit priv w/o BP (MB)
QW0225PriStg_Aux	SW225PSA	A2GB_AUX_SLOT_TS	DBM1 AUX in use for 64-bit priv w/o BP (MB)
QW0225RL	QW0225RL	DIST_REAL_FRAME	DIST REAL in use for 31 and 64-bit priv (MB)
QW0225AX	QW0225AX	DIST_AUX_SLOT	DIST AUX in use for 31 and 64-bit priv (MB)
QW0225HVPagesInReal	SW225VPR	A2GB_DIST_REAL_FRM	DIST REAL in use for 64-bit priv (MB)
QW0225HVAuxSlots	SW225VAS	A2GB_DIST_AUX_SLOT	DIST AUX in use for 64-bit priv (MB)
QW0225ShrStg_Real	SW225SSR	A2GB_SHR_REALF_TS	REAL in use for 64-bit shared (MB)
QW0225ShrStg_Aux	SW225SSA	A2GB_SHR_AUXS_TS	AUX in use for 64-bit shared (MB)
QW0225ShrStkStg_Real	SW225KSR	A2GB_SHR_REALF_STK	REAL in use for 64-bit shared stack (MB)
QW0225ShrStkStg_Aux	SW225KSA	A2GB_SHR_AUXS_STK	AUX in use for 64-bit shared stack (MB)
QW0225ComStg_Real	SW225CSR	A2GB_COMMON_REALF	REAL in use for 64-bit common (MB)
QW0225ComStg_Aux	SW225CSA	A2GB_COMMON_AUXS	AUX in use for 64-bit common (MB)
QW0225_REALAVAIL	S225RLAV	QW0225_REALAVAIL	REALAVAIL (MB) (S)

Note: All REAL/AUX storage fields in IFCID 225 and OMPE performance database are expressed in 4KB frames or slots – they should be converted to MB (conversion is already done in MEMU2)

Summary

- Apply the DB2 APARs
- Apply the z/OS APAR if XCF CRITICALPAGING is enabled
- Keep the LPAR well provisioned with REAL storage
- For optimal performance avoid paging to AUX or Flash Express
- For most customers set REALSTORAGE_MANAGEMENT=AUTO
- Only use REALSTORAGE_MANAGEMENT=OFF when
  - LPAR is generously over provisioned with REAL storage
  - Proactive monitoring is in place with alerting if the available REAL storage drops down encroaching on free cushion
- Do not over commit the LFAREA if the system may page and the large frames may be broken down and put back together again
- Watch out for MAXSPACE and large dump sizes that may cause the system to page
- Use ‘common currency’ for monitoring REAL and AUX usage, to determine what is the normal vs. abnormal system profile


Questions?

John Campbell
DB2 for z/OS Development
campbelj@uk.ibm.com

Session 6006
DB2 for z/OS Real Storage Monitoring, Control and Planning