DB2 10 and 11 for z/OS Pitfalls Lessons Learned from DB2 360 (Health Check) Studies

Please Note

- IBM's statements regarding its plans, directions, and intent are subject to change or withdrawal without notice at IBM's sole discretion.
- 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.
- Performance is based on measurements and projections using standard IBM benchmarks in a controlled environment. The actual throughput or performance that any user will experience will vary depending upon many factors, including 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 results similar to those stated here.

Agenda

- Real storage control
- Online release migration in a 24*7*365 environment
- Steps to investigate CPU performance
- zIIP capacity
- Preventative Service
- Mass data recovery

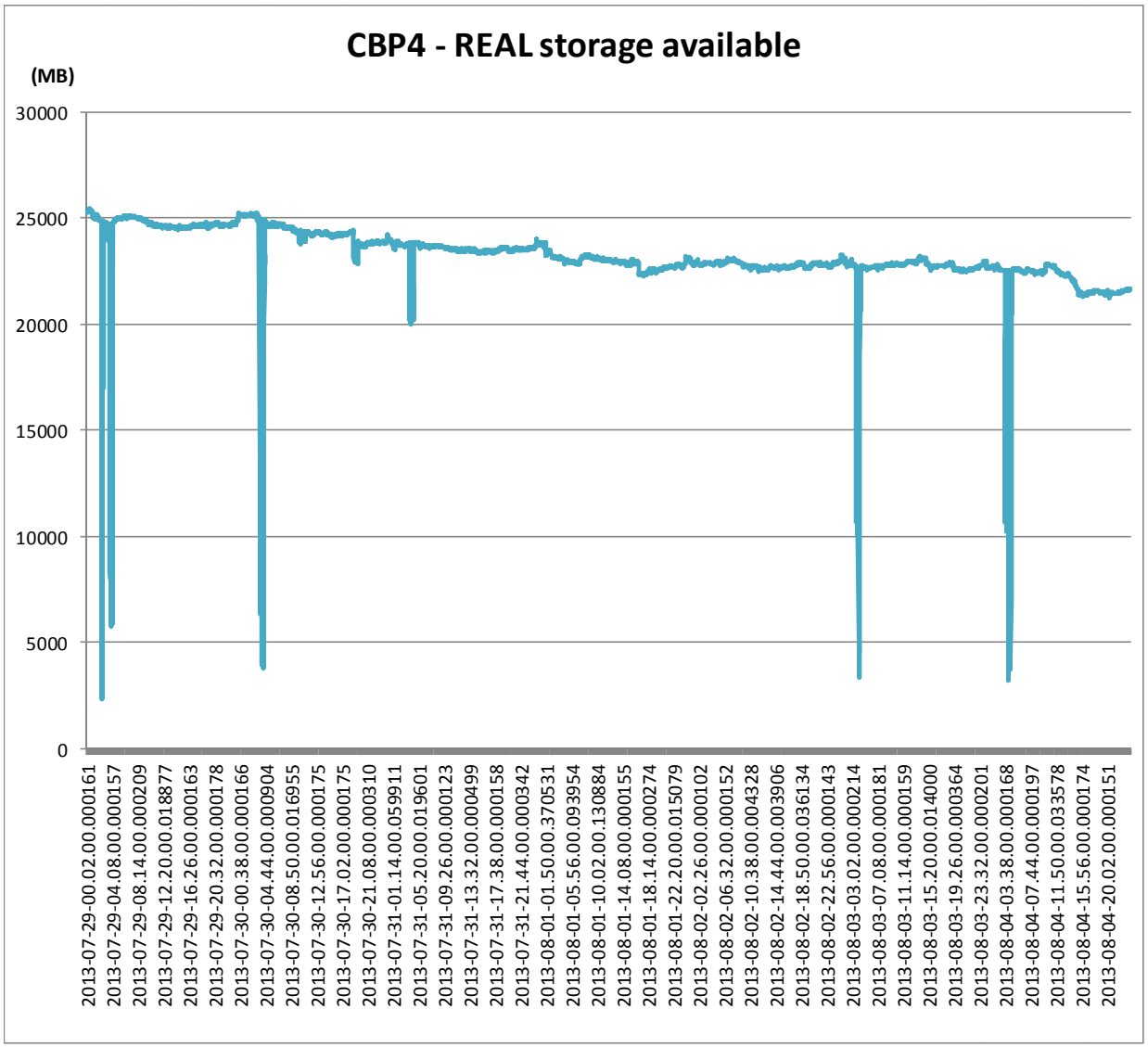
Real storage control

- 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

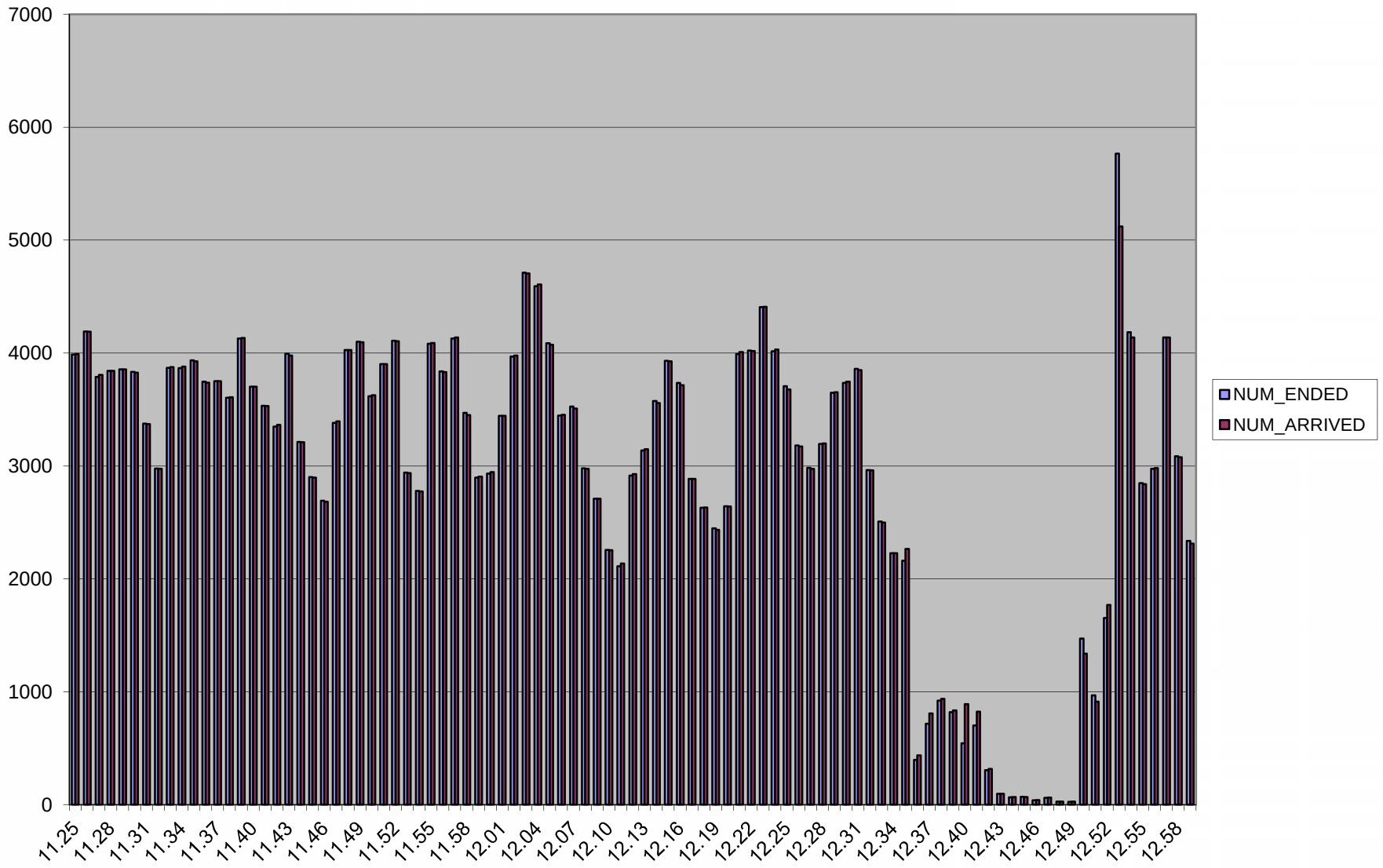
Real storage control …

- "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 non-DFSORT work
  - Dump capture
    - TCBs non-dispatchable
    - P-Lock negotiation delayed affecting other members
    - Locks not released in a timely manner
    - Excessive dump time caused by paging on the LPAR may cause massive sysplex-wide sympathy sickness slowdowns

Real storage usage/ DFSORT



The DUMP effect



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 DSN517I (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 usage
  - Set EXPOLD=0 to prevent 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%

Monitoring REAL/AUX

- **Done using OMPE Performance Database fields
- #1 - Stacked AREA graph - one for each DB2 member (1 sheet/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 (1 sheet/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

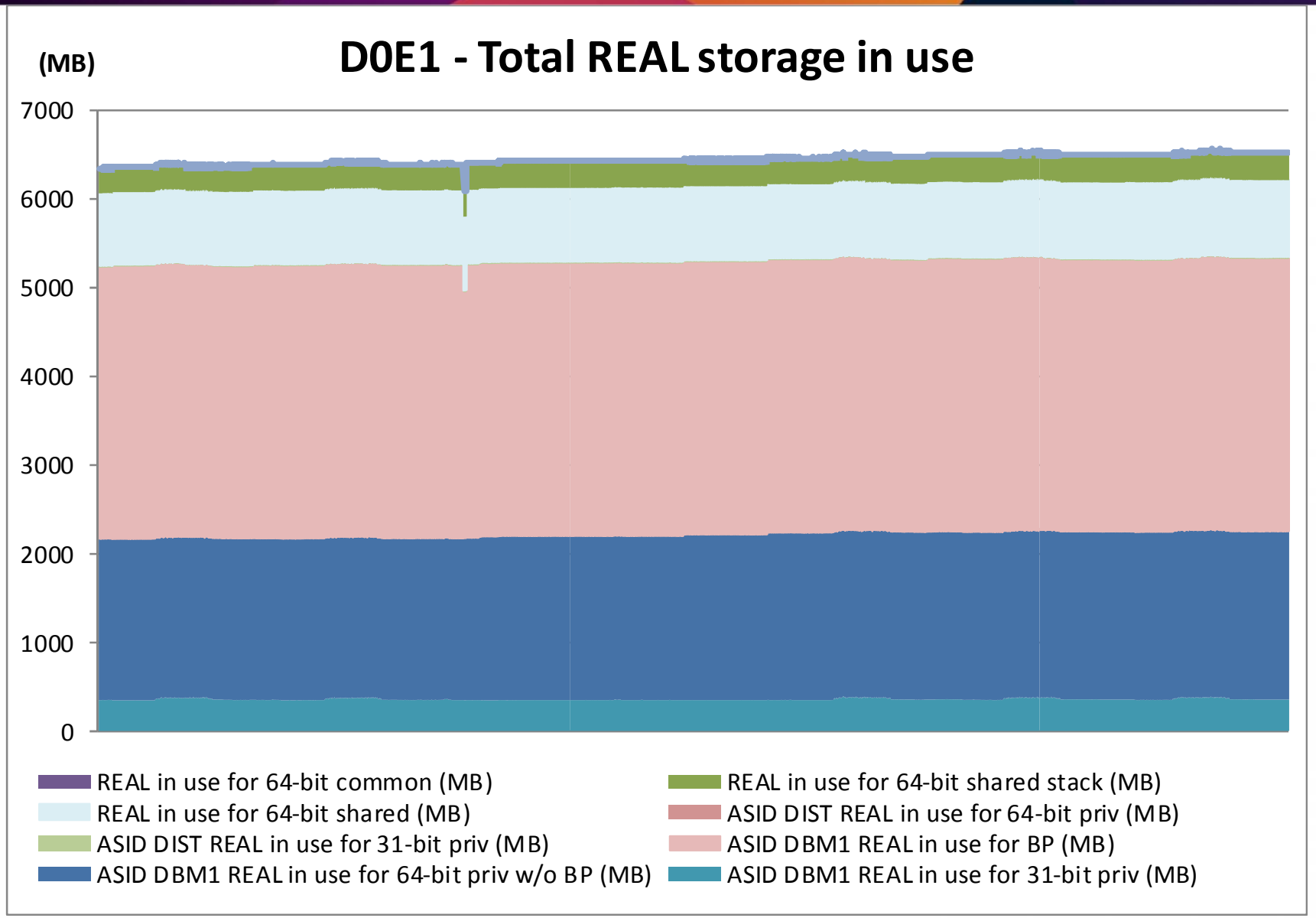
```
QW0225_REALAVAIL*4/1024                          AS REAL_AVAIL_LPAR_MB
```

Monitoring REAL/AUX

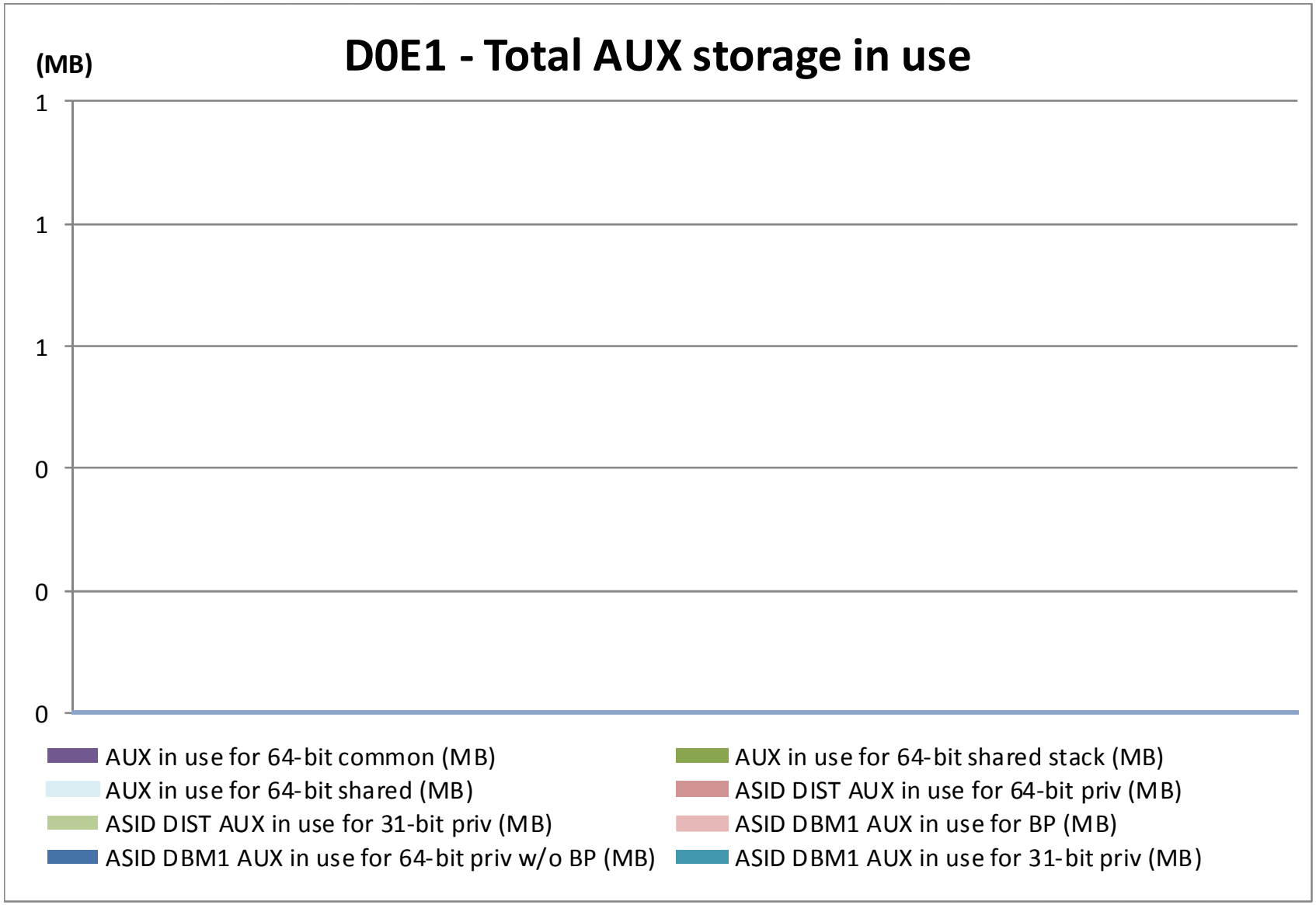
IFCID FIELD	OMPE FIELD	OMPE PDB COLUMN NAME	MEMU2 Description
QW0225RL	QW0225RL	REAL_STORAGE_FRAME	DBM1 REAL in use for 31 and 64-bit priv (MB)
QW0225AX	QW0225AX	AUX_STORAGE_SLOT	DBM1 AUX in use for 31 and 64-bit priv (MB)
QW0225HVPagesInReal	SW225VPR	A2GB_REAL_FRAME	DBM1 REAL in use for 64-bit priv (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)

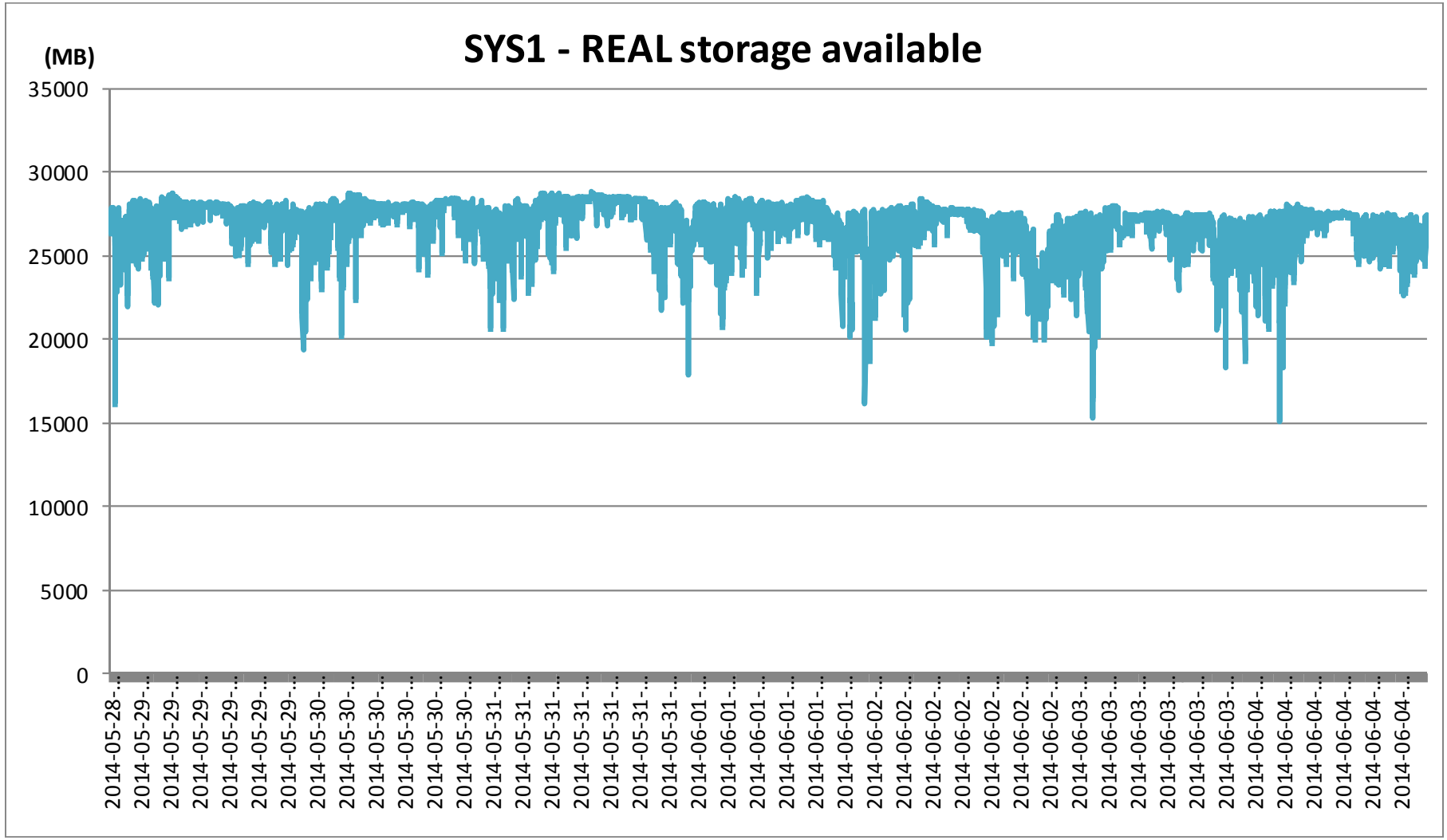
Monitoring REAL/AUX



Monitoring REAL/AUX



Monitoring REAL/AUX



Online release migration

- Technically possible to run DSNTIJTC and DSNTIJEN alongside well-behaved online workloads
  - Jobs use SQL DDL with frequent commit and REORG SHRLEVEL(CHANGE|REFERENCE)
  - Designed to fail gracefully leaving DB2 catalog fully operational
  - After problem determination is complete, the respective job can be corrected and resubmitted
  - The respective job will restart from where it left off

Online release migration…

- But some ‘rules of the game’ and you must be prepared to play
  - DSNTIJTC and DSNTIJEN jobs should be scheduled during a relative quiet period
  - If non data sharing
    - Must stop all application workload when DSNTIJTC job is running
  - If data sharing
    - Must route work away from the DB2 member where DSNTIJTC job is running
    - Must temporarily change workload balancing and sysplex routing scheme
  - Should synthetically stop all of the following workload types from running
    - SQL DDL, Grants & Revokes, BIND/REBIND, utilities, monitors
  - Set Utility Timeout Factor (UTIMOUT) down to 1
    - Limit impact on online transactions waiting behind DSNTIJTC/EN
  - All essential business critical workloads that are running be well behaved and should commit frequently
  - Must be prepared to watch and manually intervene if needed
  - Strong recommendation to perform Pre-Migration Catalog Migration Testing
  - Must be prepared for DSNTIJTC and/or DSNTIJEN jobs to possibly fail or for some business transactions to fail

Steps to investigate CPU performance

- Comparing CPU performance across release boundary e.g., V11 vs V10 vs V9
  - When looking at relative performance across DB2 releases or maintenance boundaries, need to factor in the CPU consumed by the DB2 system address spaces and normalise by transactions
  - Very difficult to do in real customer production environment
    - Uncertainty caused by promoted application changes
    - Fluctuation in the daily application profile especially batch flow
  - Must try to normalise things out to ensure workloads are broadly comparable
    - Broadly similar in terms of SQL and getpage profile
    - Usually have to exclude the batch flow
    - Factor out extreme variation
    - Need to look at multiple data points

Steps to investigate CPU performance …

- Check that you have the same pattern across releases from a DB2 perspective based on combined view of DB2 Statistics and Accounting Traces
- Validate that there have been no access path regression after migration or from application changes going on at the same time as the migration
- Use as a starting point look at
  - Statistics Trace
    - MSTR TCB, SRB & IIP SRB; DBM1 TCB, SRB & IIP SRB; IRLM TCB & SRB CPU times
    - Split of CP vs. zIIP for DBM1 and MSTR is likely to be very different across V9, V10, V11
  - Accounting
    - For each CONNTYPE
      - Class 2 CPU times on CP and zIIP, numbers of occurrences and commits/rollbacks
  - Workload indicators:
    - DML (split by type: select, insert, update, fetch, etc...),
    - Commits, rollbacks, getpages, buffer update
    - Read and write activity (#IOs, #pages)

Steps to investigate CPU performance …

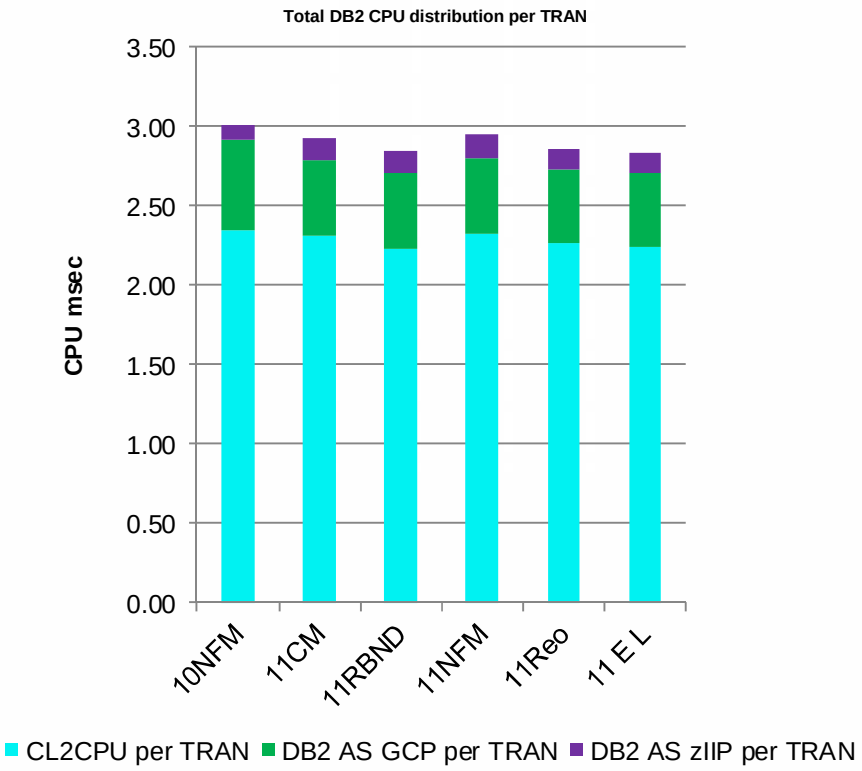
- It is a real challenge to get an 'apple-to-apple' comparison in a real production environment
- Best chance is to find a period of time with limited batch activity, and to look at the same period over several days in V9 and several days running on V10 or V11
- Make sure that the CPU numbers are normalized across those intervals i.e., use CPU milliseconds per commit
- Easy to combine statistics and accounting by stacking the various components of CPU resource consumption:
  - MSTR TCB / (commits + rollbacks)
  - MSTR SRB / (commits + rollbacks)
  - MSTR IIP SRB / (commits + rollbacks)
  - DBM1 TCB / (commits + rollbacks)
  - DBM1 SRB / (commits + rollbacks)
  - DBM1 IIP SRB / (commits + rollbacks)
  - IRLM TCB / (commits + rollbacks)
  - IRLM SRB / (commits + rollbacks)
  - Average Class 2 CP CPU * occurrences / (commits + rollbacks)
  - Average Class 2 SE CPU * occurrences / (commits + rollbacks)

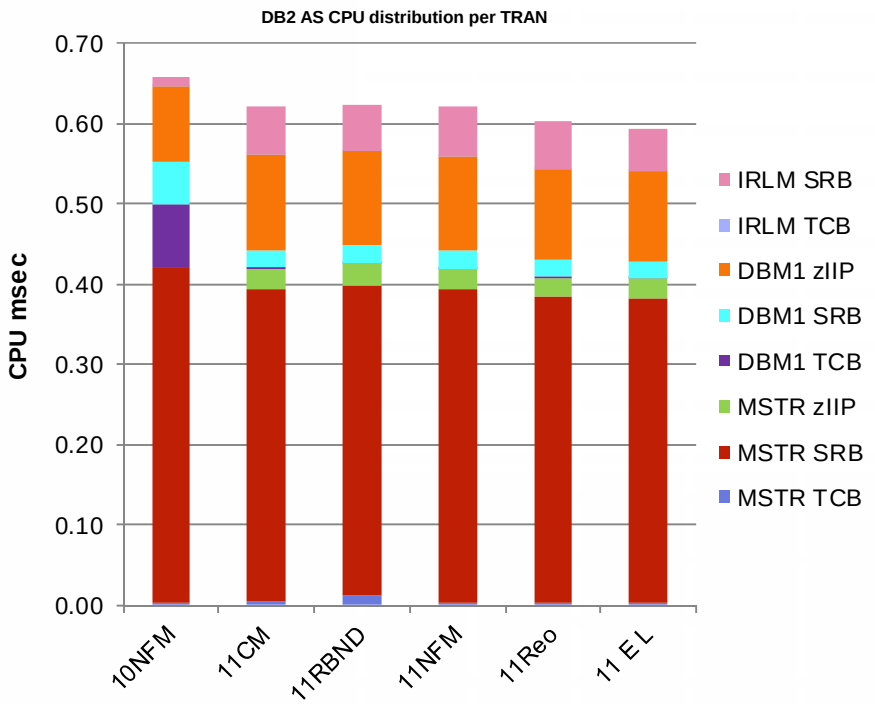
Steps to investigate CPU performance …

- Need to check the workload indicators for the chosen periods
- Similarities between data points for a given version, but big variations across V9, V10, V11
  - Sign that something has changed from an application or access path perspective
  - More granular analysis of accounting data will be required to pin point the specific plan/package

Steps to investigate CPU performance …

- Example comparing V10 vs. V11





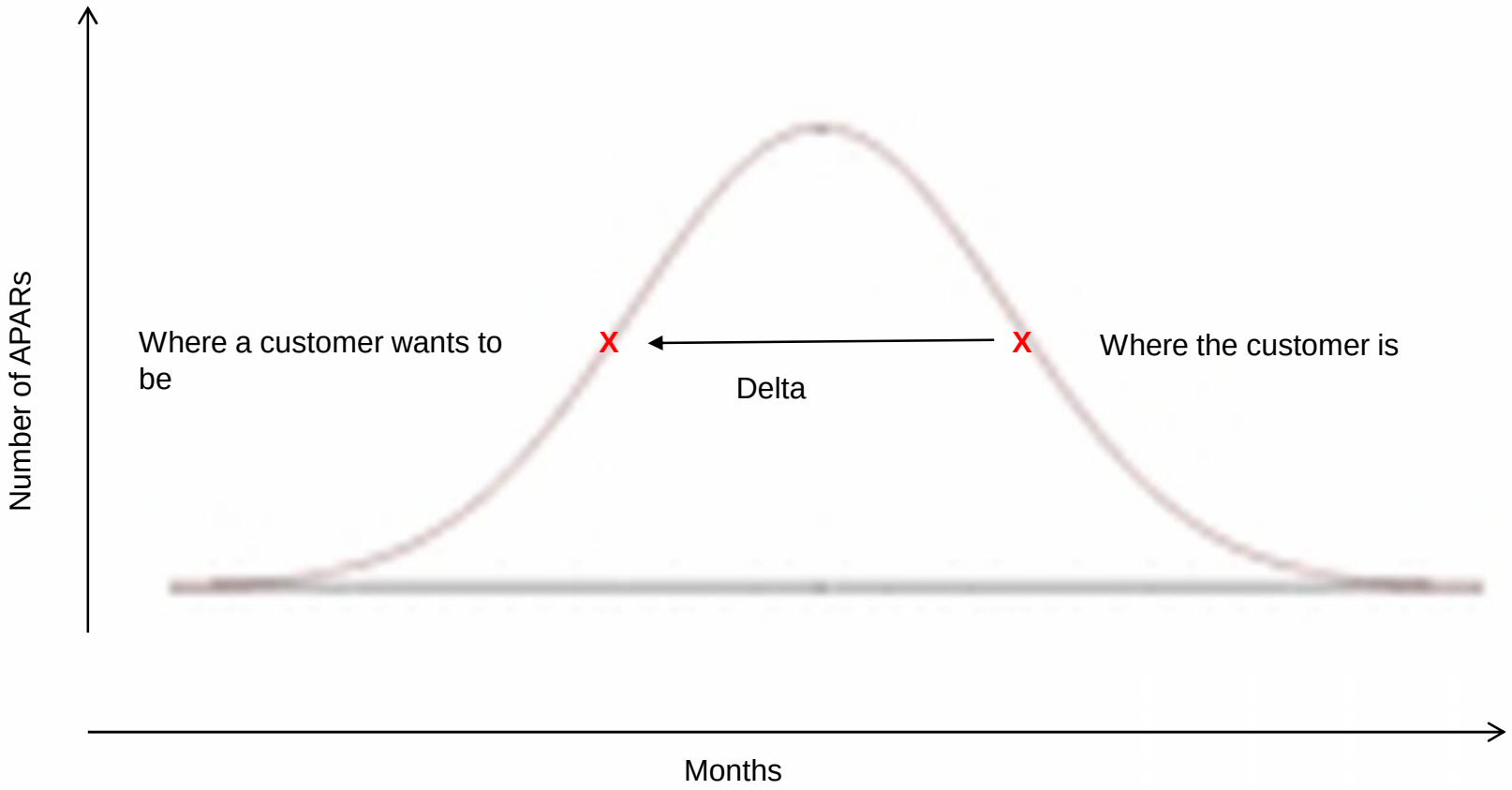
zIIP Capacity

- In DB2 10, prefetch and deferred writes engines are now eligible for zIIP offload – in DB2 11, all SRB-mode system agents (except p-lock negotiation) are eligible for zIIP offload
  - These DB2 tasks must be dispatched very quickly
  - Any delays could result in
    - Significant elapsed time degradation for some batch and DB2 utility jobs
    - Very high count for 'prefetch engine not available' in the DB2 Statistics Trace
- Many installations running with default settings IIPHONORPRIORITY=YES, HIPERDISPATCH=YES and ZIIPAWMT=3200 in IEAOPTxx parmlib member
  - zIIP processors can get help from standard processors (GCP)
  - zIIP needs to be running for 3.2 msec before it checks to see if it should ask for help from GCPs ('alternate wait management')
    - Many requests can be flowing through the zIIP during this time period. But if the zIIP has been running for the length of time specified by ZIIPAWMT, the queue of work is still growing, and all the zIIPs are busy, then the zIIP signals a GCP for help to process the work.

zIIP capacity ...

- With the above default settings and if the zIIP capacity is under-configured
  - DB2 prefetch engines can end up queuing for up to 3.2 msec for a zIIP before they are dispatched on a GCP
  - Of course, this could be much worse if the zIIP processors were not allowed to ask GCPs for help (IIPHONORPRIORITY=NO)
- Tuning knobs available which will impact zIIP offload for all workload
  - Disable ‘alternate wait management’ by setting HIPERDISPATCH=NO and reducing ZIIPAWMT value to get the zIIP to ask for help from GCP sooner
  - However using GCPs to do zIIP-eligible work will negatively impact TCO
- Watch for occurrences of Utilities or parallelism which saturate the zIIP
- Correct technical solution is to add more zIIP capacity
  - zIIPs are assist processors and not intended to be run as hard as GCPs
  - zIIPs usage should in the 30-50% CPU busy range on average (peaks higher)

Preventative maintenance – ‘Bell Curve’



Preventative maintenance – some observations

- Preventative maintenance strategy for any given customer needs to adapt to changing circumstances
  - Needs to be dynamic and flexible
  - Consider being more aggressive in maintenance strategy
    - If the number of problems experienced are fixed by PTFs that are not applied
    - If an early adopter of a new version or some new function
  - Consider being less aggressive and more conservative if continually hitting PEs
- Some customers are traditionally conservative and on the trailing edge on current version and preventative maintenance
  - Version upgrade triggered based on end of currency date of prior release
  - Cycle of preventative maintenance upgrades is also based on being on the trailing edge of current version
    - Reliance on PEs being found by other customers
    - No pro-active checking for HIPERs and PEs
  - Often resistant to and inflexible on
    - Application of critical HIPERs during and after roll-out period
    - Investigation of PEs (always back off which takes the base back in time)

Preventative maintenance – some observations

- Some customers are traditionally conservative and trailing edge on releases and preventative maintenance ...
  - Current process model may have served them well in terms of maintaining system availability and stability
  - Some of these same customers made the decision to migrate early to newer DB2 release to accrue benefits of reduced performance cost
  - The decision in some cases was not a balanced decision
    - Predicated on
      - Accruing CPU cost savings from DB2 10 earlier
      - Cost savings from performing only one version rather than two
    - No additional risk mitigation actions taken as a result of moving up the adoption curve of new version
      - Continued with the old behaviour model on applying preventative service
      - No adequate plans for more frequent application of preventative service
      - No enhanced pre-production QA testing
  - Without regular preventative service upgrades the base could be months old
    - Current run rate is ~30 HIPERs per month
  - HOLD actions
    - Infrequent preventative maintenance piles up the research and actions to be carried out

Preventative maintenance – recommendations

- As a priority implement a continuous scheduled program for applying regular DB2 preventative service using CST/RSU method
  - Paramount importance to maintain system availability and stability
  - Plan on 3-4 major preventative maintenance upgrades a year over next 12+ months to catch up and continue until world wide customer production adoption of new DB2 release becomes trailing edge
  - Pull and review Enhanced HOLDDATA on at least a weekly basis
    - Pro-active checking of all HIPERs and PEs looking for critical problems e.g.,
    - Data loss, NCORROUT, overlays, crashes, bad restart/recovery, etc
  - Introduce management process and procedure for expediting the apply of the most critical HIPERs after 1-2 weeks in Test
    - During the rollout of new preventative service package on way to Production
    - Production thereafter
  - Enhance QA testing to provide better coverage and ‘keep fires away’ from production
  - Application of preventative maintenance can become less aggressive as new DB2 release becomes trailing edge
    - Perform at least 2 major preventative maintenance upgrades per year

Mass Data Recovery

- DB2 log-based recovery of multiple objects may be required when…
  - Catastrophic DASD subsystem failure and no second copy
  - Plan B for disaster recovery
    - Mirror is damaged/inconsistent
    - Bad Disaster Restart e.g., using stale CF structures in data sharing
  - Data corruption at the local site caused by…
    - ‘Bad’ application program
    - Operational error
    - DB2, IRLM, z/OS, third-party product code failure
    - CF microcode failure, DASD subsystem microcode failure
- Scope of the recovery may be more or less extensive
  - One application and all associated objects
  - Part of the system (including a random list of objects across multiple applications)
  - Or, in the worst case, the ‘whole world’

Mass data recovery …

- DB2 log-based recovery of multiple objects is a very rare event
  - But statistically, it is more frequent than a true DR event (flood, fire, etc.)
- Taking regular backups is necessary but far from sufficient for anything beyond minor application recovery
- If not prepared, practiced through testing and optimised, will lead to extended application service downtimes – possibly many hours to several days

Mass Data Recovery …

- Common issues
  - Lack of planning for, intelligent design, practice/testing, optimisation & maintenance
  - No prioritised list of application objects and inter-dependencies
    - Limited use of DB2 referential integrity
    - Data dependencies and integrity management are buried in the applications
    - Any attempt heavily dependant on application knowledge and support
  - Procedures for taking backups and executing recovery compromised by lack of investment in technical configuration
  - Unintelligent generation of recovery jobs to flood the system
  - Use of tape including VTS/tapeless
    - Cannot share tape volumes across multiple jobs
      - Archive logs
      - Image copy backups stacked on the same volser
    - Relatively small number of read devices (VTS)
    - Concurrent recall can be a serious bottleneck (VTS)

Mass Data Recovery …

- Major recommendations
  - Keep at least 48 hours of recovery log data on DASD
    - Keep at least 6 hours in active log configuration during peak periods
    - Write both copies of each archive log pair to DASD
      - Keep archive log COPY1 on DASD for 48-72 hours before migrating it to tape/VTS
      - Archive log COPY2 can be migrated to tape/VTS at any time
  - Intelligently design, generate and schedule the recovery jobs
    - Design for sensible level of parallelism avoiding contention on tape and group multiple objects per job
    - Spread around members of data sharing group to exploit fast log apply bandwidth
      - Submit only 51 RECOVER jobs per member
      - Limit the number of objects per RECOVER to 20-30
    - Optimise job scheduling to avoid ‘dead times’
      - Objects with longest end-to-end recovery time should be recovered first
    - Use index RECOVER instead of REBUILD for very large NPIs
    - Consider use of FCIC for very large table parts and very large NPIs

Mass Data Recovery …

- Major recommendations ….
  - Consider shortening the full image copy cycle time for critical application data which is heavily updated to reduce log apply processing
  - Essential to perform regular testing to practice, optimise, ensure procedures are in good working order, and demonstrate actual service level
  - Application and data life cycle considerations
    - Prioritise most critical applications making sure to carefully understand application and data interdependencies
    - Separate active/operational data from inactive/historical data
    - Perform regular aggressive archiving to historical
    - Allow application toleration of unavailable historical data
    - Look at creating 'fire walls' between applications

Summary

- Real storage control
- Online release migration in a 24*7*365 environment
- Steps to investigate CPU performance
- zIIP capacity
- Preventative Service
- Mass data recovery

THANK YOU

Speaker Name
Speaker Title
Speaker Email

Acknowledgements/Disclaimers

Availability. References in this presentation to IBM products, programs, or services do not imply that they will be available in all countries in which IBM operates.

The workshops, sessions and materials have been prepared by IBM or the session speakers and reflect their own views. They are provided for informational purposes only, and are neither intended to, nor shall have the effect of being, legal or other guidance or advice to any participant. 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. IBM shall not be responsible for any damages arising out of the use of, or otherwise related to, this presentation or any other materials. Nothing contained 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 the applicable license agreement governing the use of IBM software.

All customer examples described are presented as illustrations of how those customers have used IBM products and the results they may have achieved. Actual environmental costs and performance characteristics may vary by customer. Nothing contained in these materials is intended to, nor shall have the effect of, stating or implying that any activities undertaken by you will result in any specific sales, revenue growth or other results.