

# Db2 Z Hardware & z/OS Exploitation

Michael Lowe  
IBM UK Limited

November 2018  
Session IL



# Notices and disclaimers

## Legal Disclaimer

- © IBM Corporation 2017. All Rights Reserved.
- The information contained in this publication is provided for informational purposes only. While efforts were made to verify the completeness and accuracy of the information contained in this publication, 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 and strategy, which are subject to change by IBM without notice. IBM shall not be responsible for any damages arising out of the use of, or otherwise related to, this publication or any other materials. Nothing contained in this publication 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.
- 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. Product release dates and/or capabilities referenced in this presentation may change at any time at IBM's sole discretion based on market opportunities or other factors, and are not intended to be a commitment to future product or feature availability in any way. 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.
- If the text contains performance statistics or references to benchmarks, insert the following language; otherwise delete:  
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.
- If the text includes any customer examples, please confirm we have prior written approval from such customer and insert the following language; otherwise delete:  
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.
- Please review text for proper trademark attribution of IBM products. At first use, each product name must be the full name and include appropriate trademark symbols (e.g., IBM Lotus® Sametime® Unyte™). Subsequent references can drop "IBM" but should include the proper branding (e.g., Lotus Sametime Gateway, or WebSphere Application Server). Please refer to <http://www.ibm.com/legal/copytrade.shtml> for guidance on which trademarks require the ® or ™ symbol. Do not use abbreviations for IBM product names in your presentation. All product names must be used as adjectives rather than nouns. Please list all of the trademarks that you use in your presentation as follows; delete any not included in your presentation. IBM, the IBM logo, Lotus, Lotus Notes, Notes, Domino, Quickr, Sametime, WebSphere, UC2, PartnerWorld and Lotusphere are trademarks of International Business Machines Corporation in the United States, other countries, or both. Unyte is a trademark of WebDialogs, Inc., in the United States, other countries, or both.
- If you reference Adobe® in the text, please mark the first use and include the following; otherwise delete:  
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.
- If you reference Java™ in the text, please mark the first use and include the following; otherwise delete:  
Java and all Java-based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.
- If you reference Microsoft® and/or Windows® in the text, please mark the first use and include the following, as applicable; otherwise delete:  
Microsoft and Windows are trademarks of Microsoft Corporation in the United States, other countries, or both.
- If you reference Intel® and/or any of the following Intel products in the text, please mark the first use and include those that you use as follows; otherwise delete:  
Intel, Intel Centrino, 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.
- If you reference UNIX® in the text, please mark the first use and include the following; otherwise delete:  
UNIX is a registered trademark of The Open Group in the United States and other countries.
- If you reference Linux® in your presentation, please mark the first use and include the following; otherwise delete:  
Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both. Other company, product, or service names may be trademarks or service marks of others.
- If the text/graphics include screenshots, no actual IBM employee names may be used (even your own), if your screenshots include fictitious company names (e.g., Renovations, Zeta Bank, Acme) please update and insert the following; otherwise delete: All references to [insert fictitious company name] refer to a fictitious company and are used for illustration purposes only.

# Agenda

- zHyperWrite / zHyperLink
- > 4GB active logs
- zIIP Usage
- 16 TB Buffer Pools
- zEDC & LOB Compression
- Dataset open / close improvements
- Castout Accelerator

# zHyperLink – The Hardware



# zHyperLink

- What is the purpose:
  - Make synchronous I/O operations much faster: Reduce to ~20-40µs
- How to get it:
  - **z14 with zHyperLink Express adapter (FC #0431) installed**
    - zHyperLink Express uses optical cable
  - DS888x with I/O bay planar board and firmware level 8.3
  - z/OS 2.1: OA50653 + more (e.g. OA51450, ...)
  - Db2:
    - PI82575: Data base synch read I/Os which hit DS8K read cache
    - PI87072: Instrumentation support for data base read I/Os
    - PI89828: Additional support for Read I/O in Buffer Manager (V12)
    - Set new ZPARM ZHYPERLINK:
      - Turning on: ENABLE or DATABASE
      - Turning off: DISABLE

# zHyperLink – Additional considerations

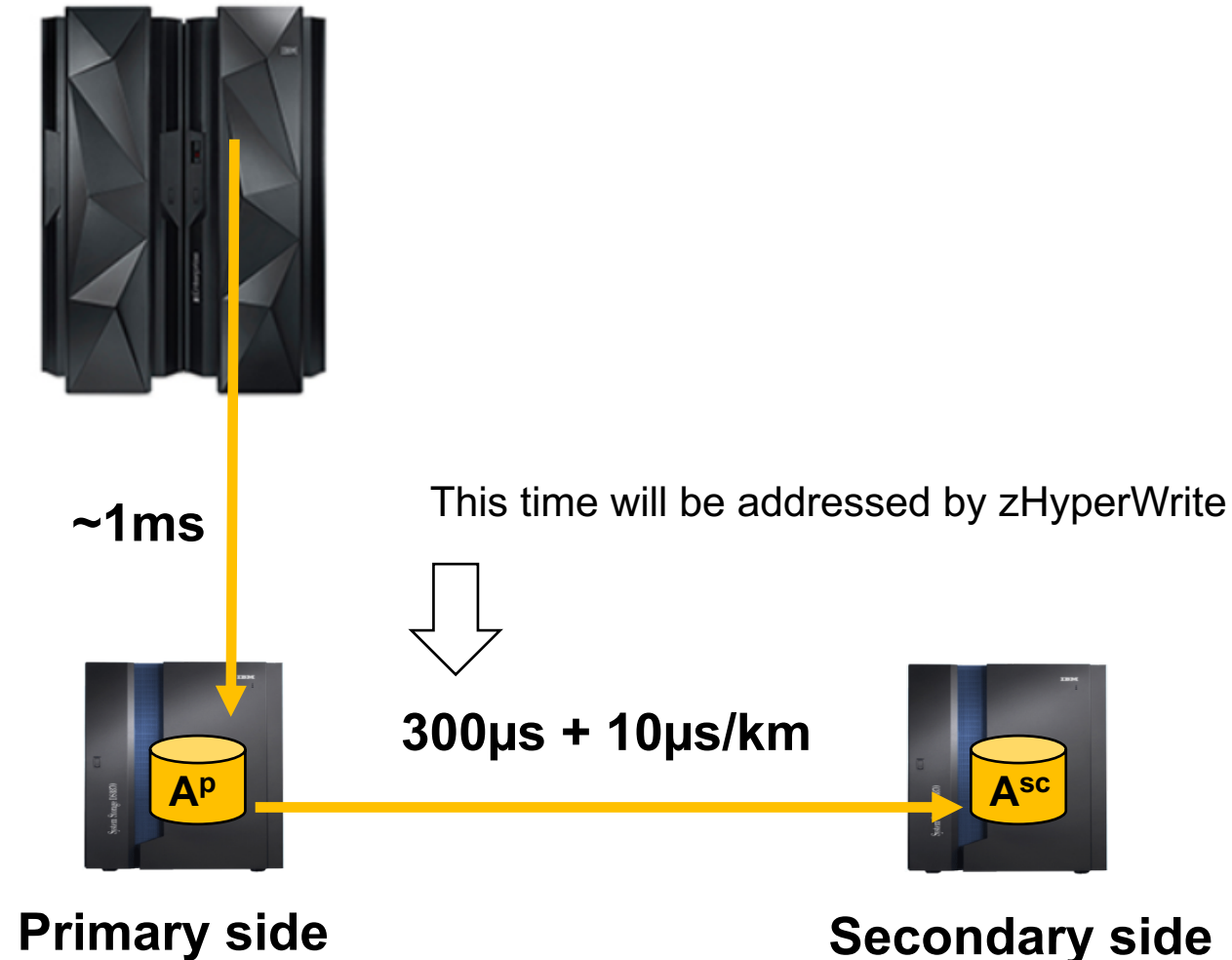
- Eligible if
  - Distance between z14 and DS888x <150m
    - If distance is longer, then I/O processing as usual
  - Data is in the disk controller cache
- How to plan for eligible workload:
  - zBNA tool (System z Batch Network Analyzer tool) at job level
  - Application level via Db2 accounting trace class 3
- PPRC
  - zHyperWrite enablement is required for PPRC configuration

# zHyperWrite

- The original purpose (2015):
  - Make synchronous log writes faster in a PPRC (aka Metro Mirror) environment
- How to get it:
  - IBM DS8870 with R7.4
    - Requires GDPS or TPC-R HyperSwap to initialize tables for PPRC relationships
    - Other vendors (e.g HDS or EMC): support since beginning of 2016
  - z/OS: OA45662 + more (e.g. OA45125, OA44973, ...)
    - Set new IECIOSxx parameter HYPERWRITE to YES
    - Check setting via new z/OS command DISPLAY IOS,HYPERWRITE
  - Db2: PI25747 + PI33569
    - Set new ZPARM REMOTE\_COPY\_SW\_ACCEL to ENABLE
    - Check setting via Db2 command -DISPLAY LOG

# Synchronous mirror, metro mirror, PPRC

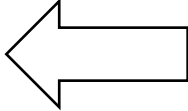
## Theoretical key numbers for *writes*





# What is the reason for this feature?

## Better response times for Db2 transactions

- Db2 **writes** against
    - Page sets (data base objects)
      - (Group) buffer pool thresholds
      - Mostly asynchronous (index page splits are synchronous)
    - Logs
      - Latest at commit
      - Synchronous
-  This operation will be addressed by zHyperWrite

# What is the reason for this feature? Does it help?

- Customer example

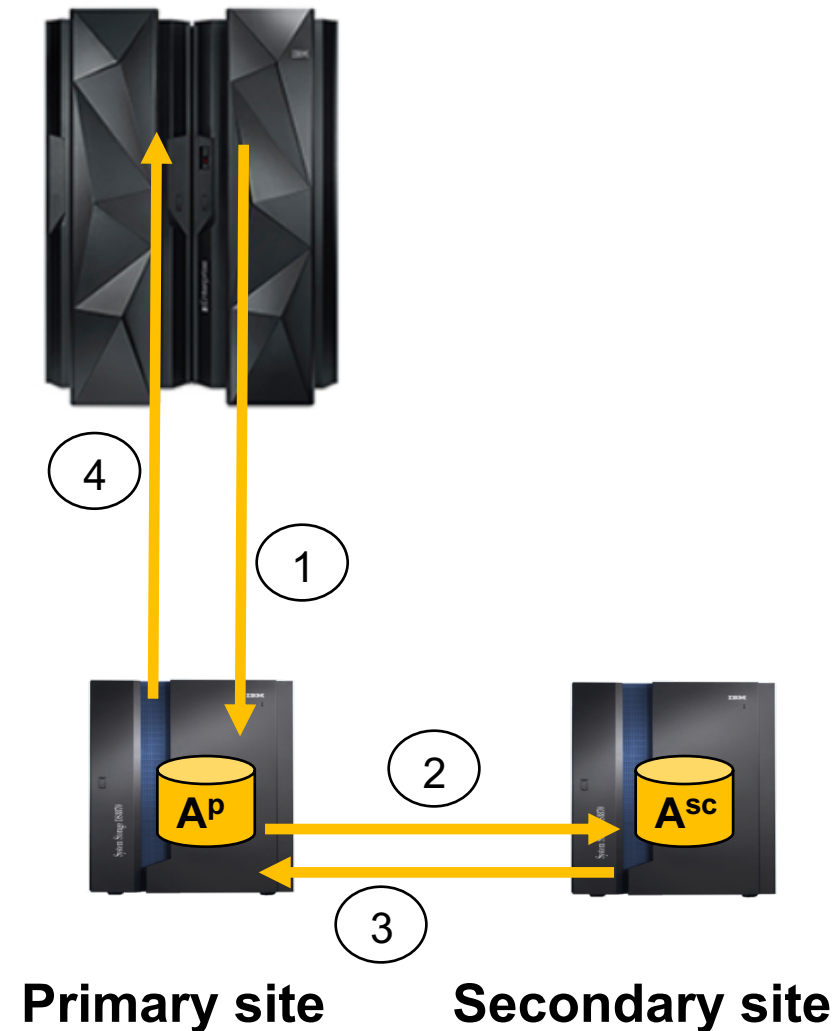
CLASS 3 WAIT	AVERAGE TIME	AVG EVNT
-----	-----	-----
...		
SYNC I/O	0:17:21.914547	1830753
DATABASE I/O	0:17:10.702058	1827090
<b>LOG/WRT I/O</b>	<b>0:00:11.212489</b>	<b>3663</b>
OTHER READ	0:00:48.918072	43348
OTHER WRITE	0:00:03.646287	721
...		

SQL DML	AVERAGE	TOTAL
-----	-----	-----
SELECT	351428.0	351428
INSERT	10284615	10284615
UPDATE	246009.0	246009
DELETE	208349.0	208349
...		

- Response time:
  - **LOG/WRT I/O:**     **avg ~3 milliseconds**
  - OTHER WRITE:     avg ~4.1 milliseconds

# How does it work? Without zHyperWrite

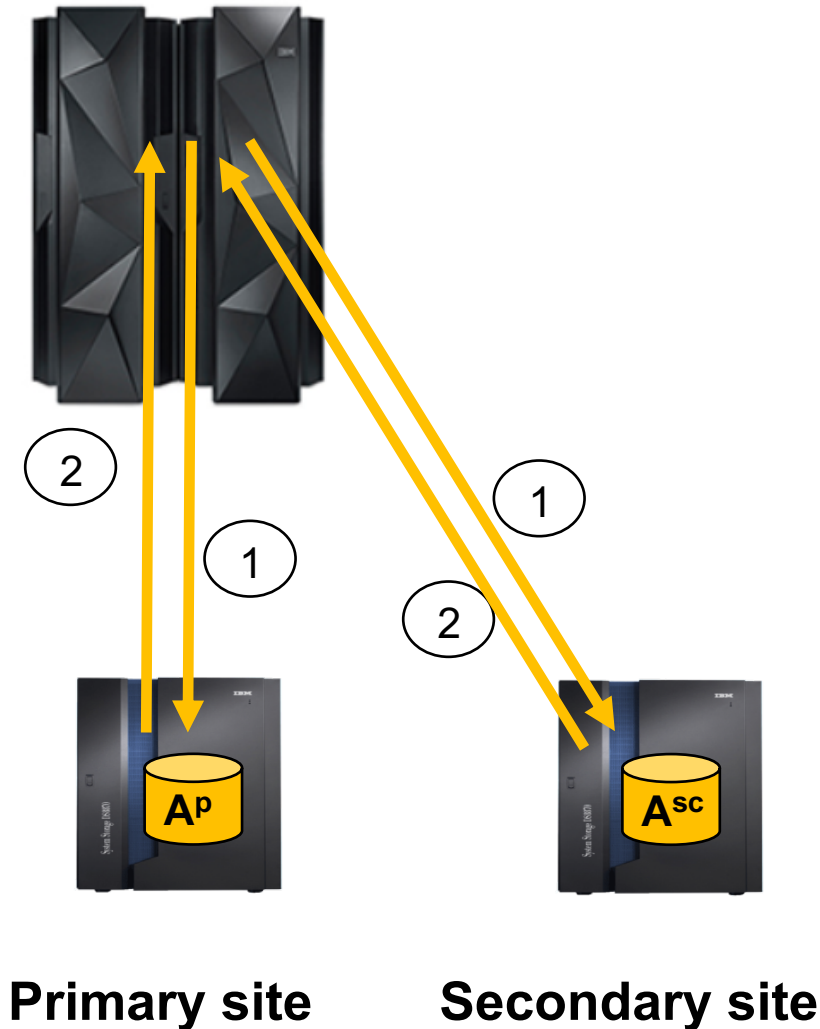
1. Db2 log write to Metro Mirror primary
2. Mirror write to secondary
3. Write acknowledgement to primary
4. Write acknowledgement to Db2



# How does it work? With zHyperWrite

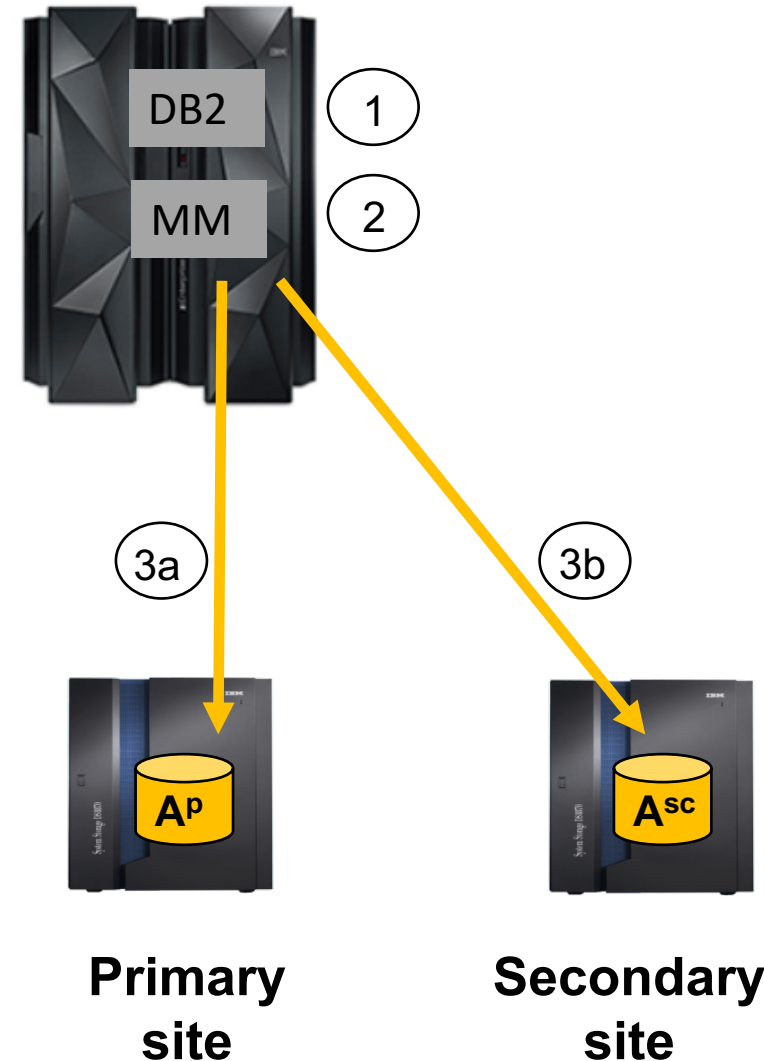
1. Db2 log write to Metro Mirror primary and secondary in parallel

2. Both sites independently write acknowledgement to Db2



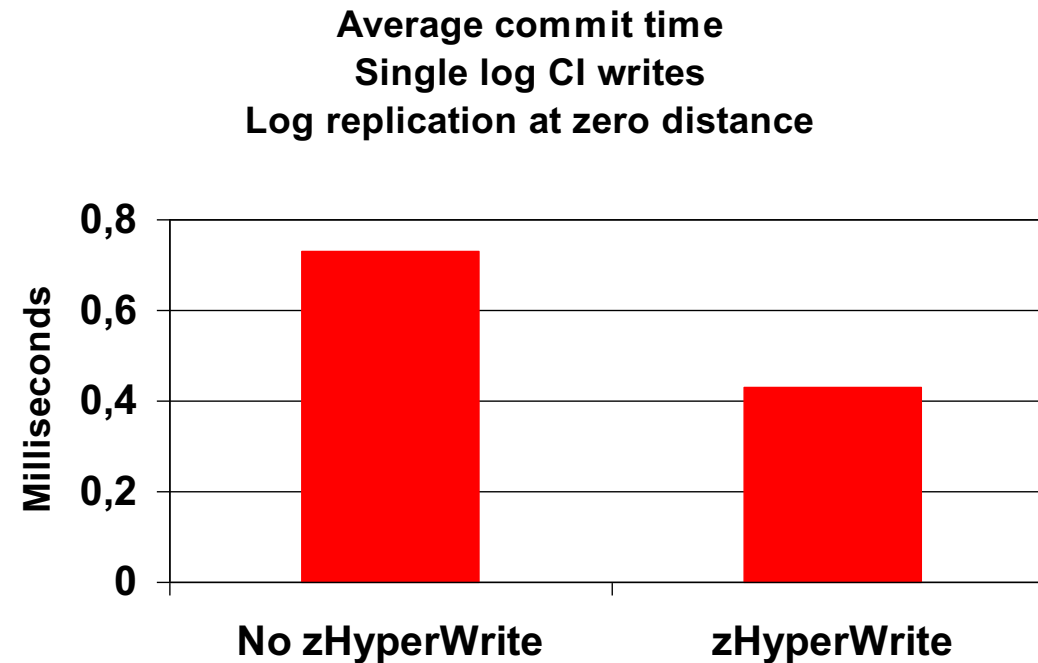
# How does it work? With zHyperWrite – details

1. Db2 indicates to Media Manager (MM) to use zHyperWrite for log writes
2. MM builds two unique I/O requests and CCW chains
3. Hardware specific information is sent to the control unit at the start of the CCW chains to the devices:
  - a) Primary device is told this is a zHyperwrite request and do not mirror the following writes
  - b) Secondary device is told this is a zHyperwrite request and do not block the subsequent writes



# zHyperWrite impact

- Writing a single log CI, zHyperWrite saves 300 microseconds per commit
- 40% reduction in commit response time



# Prerequisites – Db2 PI25747

- Applies to both Db2 10 and 11
- New ZPARM REMOTE\_COPY\_SW\_ACCEL
  - ENABLE / DISABLE (default)
    - ENABLE only means that Db2 indicates to Media Manager that zHyperWrite should be used, but the real usage is controlled by the disk storage boxes
- Enabling the function may improve active log write performance
  - If the active log data sets reside on PPRC volumes
  - Improved active log write performance may result in improved response time for Db2 operations that must ensure the log records have been externalized before continuing
  - When zHyperWrite is enabled, increased SRB time in the xxxxMSTR may be observed

# Prerequisites – Db2 PI33569

- New message which indicates that the ZPARM `REMOTE_COPY_SW_ACCEL` is enabled or disabled:

```
-DIS LOG
DSNJ370I  -DB2A DSNJC00A LOG DISPLAY
CURRENT COPY1 LOG = DSNC000.DB2A.LOGCOPY1.DS02 IS 30% FULL
CURRENT COPY2 LOG = INACTIVE IS 0% FULL
      H/W RBA = 000000000000958A8B4A
      H/O RBA = 000000000000952C2FFF
      FULL LOGS TO OFFLOAD = 0 OF 3
      OFFLOAD TASK IS (AVAILABLE)
      SOFTWARE ACCELERATION IS DISABLED
...
```



# >4GB log copies

- What is the purpose:
  - Keep a day or more of data in the active log data sets – performance benefit during recovery
- How to get it:
  - >4GB log copies can be added after activation of function level V12R1M500
- Additional remarks:
  - Maximum size of active log data set is 768GB (up from 4GB)
    - DSNJ159I is issued if size >768GB
  - Limit of 93 active logs per COPYn (n=1,2) remains in effect
  - Migration incompatibility:
    - V11 allows log copies >4GB (despite DSNJ158I): space >4GB is ignored
    - V12R1M100 does **NOT** allow log copies >4GB

# >4GB log copies

- Adding new active log data set >4GB can be done by
  - DSNJU003 standalone utility or
  - -SET LOG NEWLOG command
- Mix of <4GB and >4GB log data sets is allowed as you migrate from <4GB to >4GB
- SMS definitions
  - >4GB: data class with extended addressability (EA) set to YES
  - Archive log data sets if going to disk: data class with extended addressability enabled

# zIIP usage

- What is the purpose:
  - Reduce MLC (monthly license charge) by using zIIP instead of general purpose processors
- How to get it:
  - After installation/migration to Db2 12
  - Have **enough zIIP** resources available for usage by Db2
- The following additional workload is zIIP eligible:
  - RELOAD phase of
    - LOAD (up to 91%) and REORG (up to 59%)
    - Retrofitted to Db2 11 via APAR PI73882
  - Automatic retry of LPL/GRECP recovery
  - Daemon which monitors activity related to index fast traversal block
  - 100% zIIP eligibility for parallel query child tasks (up from 80%)

# zIIP usage – previous releases

- Db2 11
  - Cleanup of pseudo deleted index entries
  - Cleanup of XML multi-version documents
  - Log write and log read
  - Castout write processing
  - LOAD, REORG, REBUILD INDEX utility processing of inline statistics collection
  - RUNSTATS utility COLGROUP distribution statistics
  - LOAD REPLACE PART clearing of NPIs
- Db2 10
  - RUNSTATS
  - Prefetch and deferred write processing
  - Expanded use of query parallelism

# zIIP usage – previous releases

- Db2 9
  - DRDA calls to native stored procedures
  - XML parsing
- Db2 8
  - Utilities (index build)
  - DRDA requests
  - Parallelized queries
  - Return of result sets to DRDA callers of stored procedures

# 16TB buffer pools

- What is the purpose:
  - Leverage large real storage provided by current and future z hardware
- How to get it:
  - After installation/migration to Db2 12
  - z13 and z/OS 2.1 or higher allows 4TB per LPAR (10TB per CEC)
- Additional comments:
  - MEMLIMIT of DBM1 raised to 19TB (from 4TB)
  - Limits apply to keywords VPSIZE, SPSIZE, VPSIZEMIN, VPSIZEMAX of -ALTER BPOOL:
    - E.g. for a 4KB buffer pool the maximum can be “4 000 000 000”
  - Total sum of VPSIZE and SPSIZE (simulated pool size) can be up to 16TB (limit of 2x of real storage size still applies, example follows)

# 16TB buffer pools

## Example for 2x real storage limit

```
D M=STOR
IEE174I 22.29.57 DISPLAY M
REAL STORAGE STATUS
ONLINE-NOT RECONFIGURABLE
    OM-1024M
ONLINE-RECONFIGURABLE
    NONE
...
DSNB610I  -DB2A DSNB1BVP MAXIMUM ALLOCATABLE BUFFER
POOL
    STORAGE OF 2047MB HAS BEEN REACHED.
    UNABLE TO CREATE/EXPAND BUFFER POOL BP3 TO SPECIFIED SIZE
1000000.
    BUFFER POOL SIZE IS NOW 497787.
DSNB536I  -DB2A DSNB1BVP THE TOTAL VIRTUAL BUFFER
POOL STORAGE EXCEEDS THE REAL STORAGE CAPACITY
```

# 16TB buffer pools – lower and upper limits

- Apply to VPSIZE, SPSIZE, VPSIZEMIN, VPSIZEMAX:
  - 4KB        “2 000” -        “4 000 000 000”
  - 8KB        “1 000” -        “2 000 000 000”
  - 16KB      “ 500” -        “1 000 000 000”
  - 32KB      “ 250”        -        “ 500 000 000”

```

-ALT BPOOL(BP13) VPSIZE(3999999999)
DSNB508I  -DB2B THE TOTAL VPSIZE AND SPSIZE IS BEYOND THE MAXIMUM LIMIT
DSN9022I  -DB2B DSNB1CMD '-ALT BPOOL' NORMAL COMPLETION
  
```



# 16TB buffer pools – history

- Db2 for z/OS V1: 4 buffer pools explicitly named in ZPARM
- Db2 for z/OS V3: ALTER BPOOL 1.6GB (+8GB hiper pools)
- Db2 for z/OS V6: 1.6GB (+256GB in data space)
- Db2 for z/OS V8: 1TB
- Db2 for z/OS V12: 16TB
- LPAR limits:
  - z10, z196, zEC12:
    - 1TB per LPAR
    - Up to 3TB per CEC
  - z13:
    - 4TB per LPAR (since z/OS V2.1)
    - Up to 10TB per CEC

# zEDC and LOB compression

- What is the purpose:
  - Hardware-based compression and decompression of LOB data
- How to get it:
  - After activation of function level V12R1M500
  - z/OS 2.1 and higher
  - zEC12 and higher with zEDC installed – the following two items are separately priced:
    - Hardware (zEDC Express feature): zEDC Express feature (FC#0420)
    - Software (part of z/OS – enabled by IFAPRDxx specification): 5650ZOS - z/OS V2 zEDC

# zEDC and LOB compression

- Enabling LOB compression
  - CREATE / ALTER TABLESPACE COMPRESS YES...
  - ZPARM COMPRESS\_DIRLOB can be used for directory LOB tablespaces (not catalog)
    - e.g. SYSIBM.SYSDBD\_DATA contains a BLOB for SYSIBM.SYSDBR
- The LOB table space must be associated with a base table that is in a universal table space.
- The total length of the entire LOB must be larger than the defined data page size, otherwise the LOB is not compressed.

# zEDC and LOB compression: worth the investment?

- Pro:
  - Save disk storage for LOB page sets and associated logging
  - Buffer pool benefit for compressed LOBs
  - Save disk storage for other z/OS data sets that can be compressed via zEDC (e.g. BSAM files, ...)
    - Db2 image copies, archive logs
- Con:
  - Cost of the zEDC card (hardware)
  - Cost of the z/OS zEDC software (MLC)
  - Possible CPU overhead

# zEDC and LOB compression – z/OS commands

- Is hardware enabled ?

```

D PCIE,PFID=AA

IQP024I 07.29.15 DISPLAY PCIE 477
PCIE      0010 ACTIVE
PFID      DEVICE TYPE NAME          STATUS  ASID  JOBNAME  PCHID
000000AA Hardware Accelerator      ALLC    0011  FPGHWAM  FFFF
CLIENT ASIDS: NONE
Application Description: zEDC Express
Device State: Ready
Adapter Info - Relid: 00000B Arch Level: 03
                Build Date: 07/04/2011 Build Count: 02
Application Info - Relid: 010203 Arch Level: 02
    
```

- Is software enabled ?

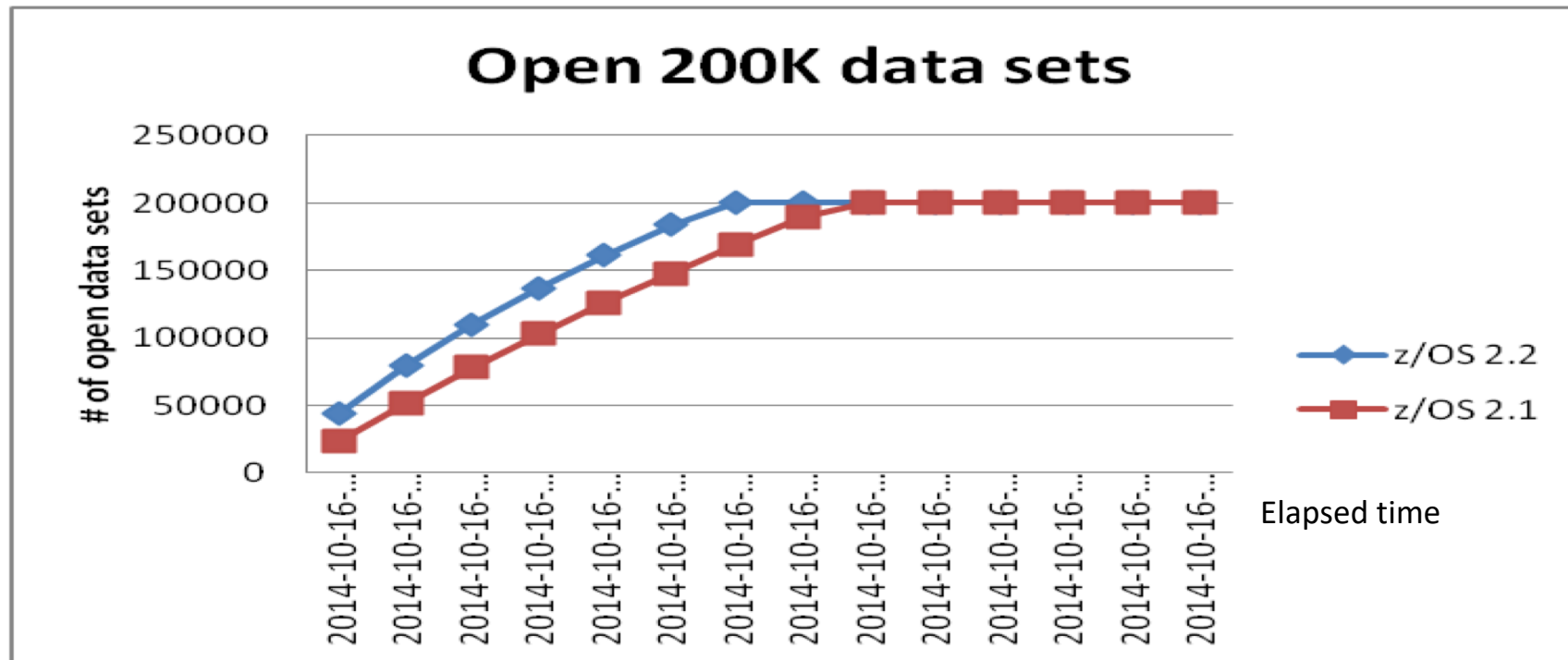
```

D PROD,REGISTERED
IFA111I 05.01.02 PROD DISPLAY 307
S OWNER          NAME          FEATURE          VERSION  ID
E IBM CORP      z/OS          zEDC          **.**.*  5650-ZOS
    
```

# Data set open/close performance improvement

- What is the purpose:
  - Large numbers of open Db2 data sets (DSMAX up to 200K) make single data set open and close performance important
- How to get it:
  - z/OS 2.2
  - Independent of Db2 maintenance or level
- Measurements of elapsed time to **open** 200K data sets:
  - z/OS 2.1: 8 minutes 23 seconds
  - z/OS 2.2: 6 minutes 53 seconds
- Additional comments:
  - The DSMAX limit remains 200K
  - Practical DSMAX limit can be lower due to DBM1 storage constraints below 2G bar

# Open of 200K data sets: z/OS 2.1 versus 2.2



- Elapsed time to open 200K data sets :
  - z/OS 2.1: 8 mins 23 secs
  - z/OS 2.2: 6 mins 53 secs
  - 18% decrease in elapsed time

# Cast Out Accelerator

- What is the purpose:
  - Improve performance for chained writes of non-contiguous pages, e.g. castout process in data sharing, deferred writes, ...
- How to get it:
  - Media Manager PTF's OA49684 and OA49685 (ensure OA51261 is on) for exploiting Cast Out Accelerator feature of the DS8880 R8.1 box
  - No Db2 PTF required
- Measurements (PPRC case) show following improvements:
  - I/O operations (IOOP) per second: increased by up to 49%
  - Response time (RT): reduced by up to 33%



# Cast Out Accelerator – measurements

	<b>IOOPs</b>	<b>IOOPs %improve</b>	<b>RT(ms)</b>	<b>CONN (ms)</b>	<b>DISC (ms)</b>	<b>RT improvement</b>
Base Code Simplex	2371		6.4	6.3	0	
Base Code with Metro Mirror	808		19.5	7.5	11.9	
Simplex with Cast Out Accelerator	4044	70%	3.6	3.6	0	43%
Metro Mirror with Cast out Accelerator	1209	49%	13	4.6	8.4	33%

See also White Paper at <http://w3-03.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/WP102605>

# We want your feedback!

- Please submit your feedback online at ....
  - <http://conferences.gse.org.uk/2018/feedback/IL>
- Paper feedback forms are also available from the Chair person
- This session is IL

