How To Serve 1.3 Billion?
- Challenges in Chinese Banks

Session Number : 2907A and 2907B

IBM US, Silicon Valley Lab.
Akiko Hoshikawa
Haakon Roberts

IBM China
Miao Zheng, China Development Lab.
Wei Wan, GTS

IBM Software
Information On Demand 2011

Objectives and Agenda

- **Abstract**
  - With large number of population and account base in China, banks in China are facing the challenges that no other customer has before. This presentation describes operational and scalability performance challenges that Chinese banks are facing with DB2 for z/OS and how they address them.

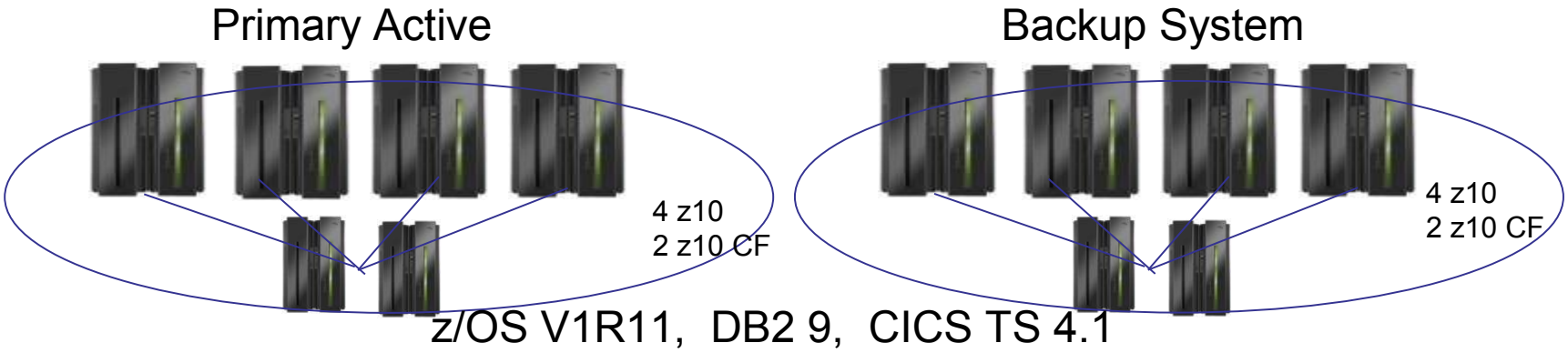
- **Agenda**
  - Environment & Business Requirements
  - Technical Challenges
    - Reducing application outage
    - Online REORG
    - Night Batch window vs. Online transaction response time
  - How DB2 9 or 10 can help the challenges
  - DB2 9 migration experience

Environment and Business Requirements

Today's Challenges in Chinese Banks

- **Huge data size based on the population**
  - Largest table space with 4096 partitions, 8TB data
  - Largest table space page set size as 64GB, 3billion account details
- **Fast growing business results in a rapidly expanding IT requirement**
  - The peak transaction rate grows from ~3000 TPS in 2007 to more than 5800 TPS in 2011, so ~20% increasing per year
- **Business innovation & expansion lead to frequent application change**
  - iPhone/Android mobile banking, iPad e-Banking
  - Thousands of packages rebound each quarter
- **High demand on performance**
  - Focus on elapsed time, any online transaction longer than 1 sec need to be reviewed

Bank A - System Configuration (Production)

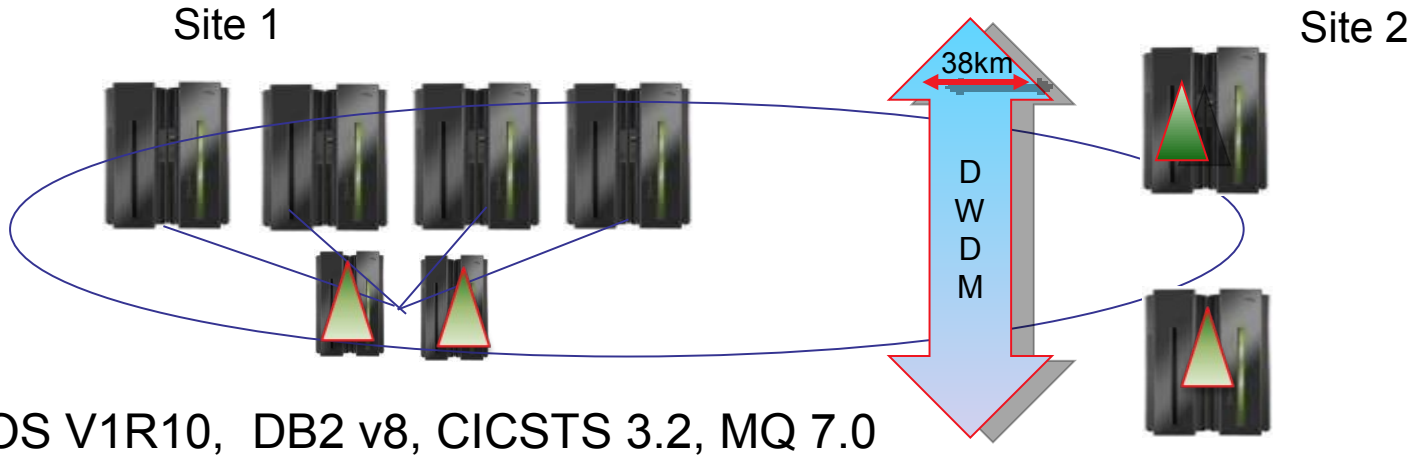


- Configurations
  - Production system include 10 DB2 data sharing groups, nearly 40 DB2 members, the biggest data sharing group is 12 members.
  - 10-way SYSPLEX, 4 way as primary (where all workloads running on), 4 way as standby (activated but nothing is running), and 2 way PPRC K system.

Bank A - DB2 Applications

- DB2 Applications
  - 50TB DB2 data.
  - Core banking for personal: approx. 8100 online packages, 11000 batch packages. 2100 tables.
  - Core banking for corporation: approx. 5900 online packages, 8600 batch packages. 1500 tables.
- Highest historical transaction rate is about 5800 TPS.
- 3 hours for daily core batch window.
- Daily image copy about 100 subtasks, REORG/LOAD with 500 - 1000 parallel tasks

Bank B - System Configuration (Production)



- 8-way parallel sysplex as product core banking system, 2 NN, 4EN and 2 K system. 4-way parallel sysplex as product batch split system.
- DB2 system configuration
  - Production : 2 8-way DS group +1 2-way DS group and 1 single subsystem
  - Test : 50 DB2 DS group
- 38KM GDPS/PPRC
- Online transaction 2500 TPS

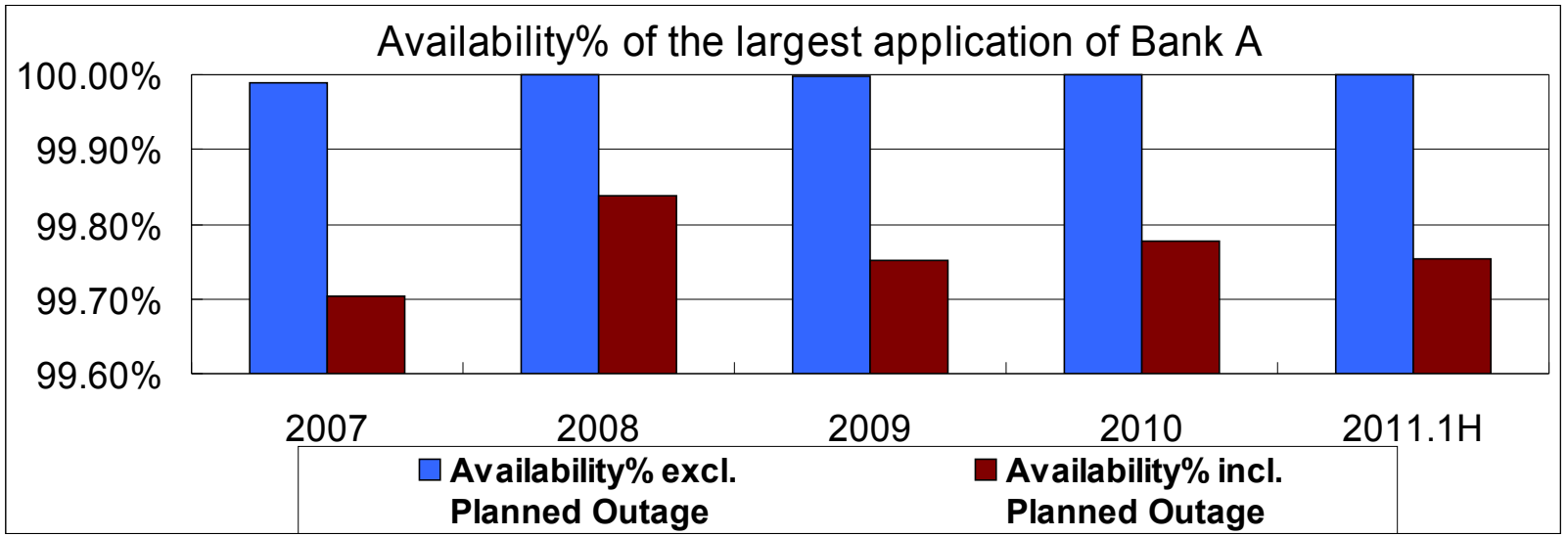
Bank B - DB2 Applications

- DB2 Applications
  - 30TB DB2 data.
  - 2255 online packages, 1632 night packages. 3960 tables with Night DB and DAY DB.
  - DBD SIZE is about 8.9MB for DAY DB, 2.8MB for Night DB
- Highest historical transaction rate is about 2500 TPS
- 26% transaction increasing per year
- 3 hours for daily core batch window. 4 hours for split-batch report batch window

Challenges
1. Application Outage

Application Availability Challenges

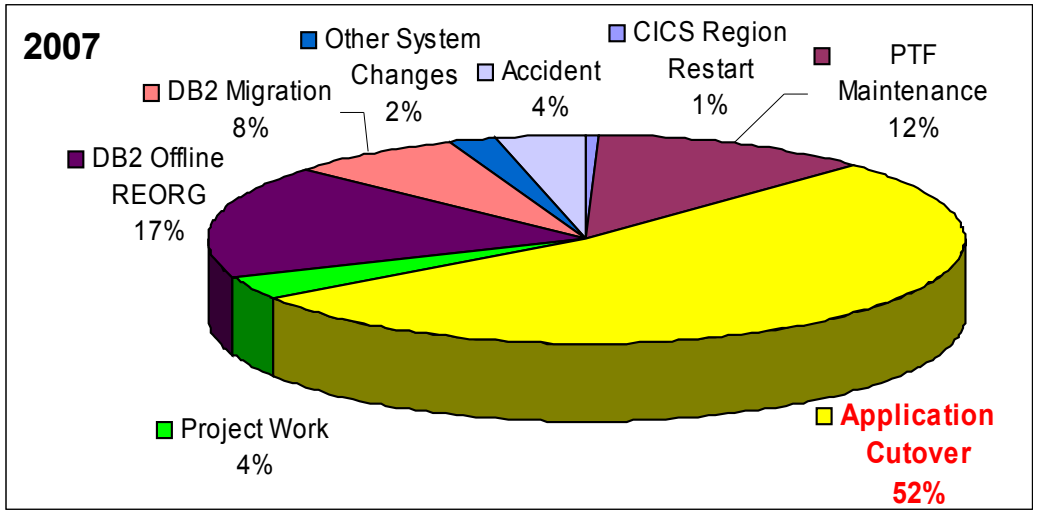
- Application availability
  - The availability excluding planned outage is very close to 100%, as a result of SYSPLEX, Data Sharing, PPRC/HyperSwap, SPOF elimination, etc.
  - The availability including planned outage is much lower, but it is the actual availability.

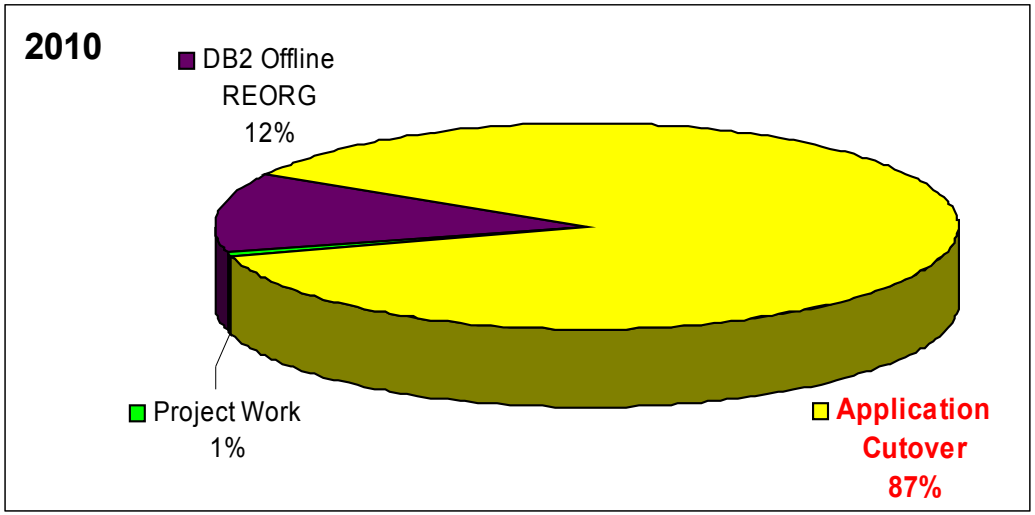


*Typically, availability of 99.75% means about 22 hours outage a year.

Focus On Application Cutover

- Bank A's application outage classification by causes





The largest application subsystem of bank A.

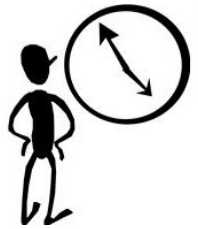
In 2007, the unplanned outage is 59 min, while the total planned outage is 1494 min.
The major causes are, Application Cutover – 827 min, DB2 Offline REORG – 271 min, PTF Maintenance – 183 min, etc.

In 2010, the unplanned outage is 0 min, while the total planned outage is 1174 min.
The major causes are, Application Cutover – 1023 min, DB2 Offline REORG – 146 min, etc.

Application Cutover Activities – Summary

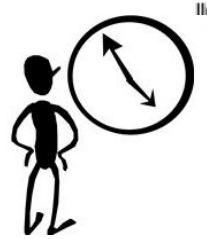
- DDL implementation
  - DROP, CREATE, ALTER, ROTATE, etc.
- Data migration
  - COPY, UNLOAD, Sort, LOAD, REORG, REBUILD, Application Batch, etc.
- Package/Plan update
  - BIND, REBIND, FREE, etc.
- Performance sustainment
  - RUNSTATS, REORG, Explain, Access Path comparison and tuning, etc.
- Outside support
  - Restart CICS regions, Restart DB2 (optional), Emergency reserved, etc.

Reduce The Window - DROP



- **Performance of DROP Table Space or Index**
  - It takes long time to delete VSAM data sets serially, to scan SYSCOPY for cleanup.
- **Actions**
  - MODIFY RECOVERY on the table spaces to be dropped, to delete related SYSCOPY before the cutover window.
  - STOP the table spaces and the corresponding index spaces
    - To flush out the page buffers in BP/GBP before DROP tables paces
  - Delete the VSAM data sets before DROP tables paces using IDCAMS DELETE jobs, which can be run in parallel

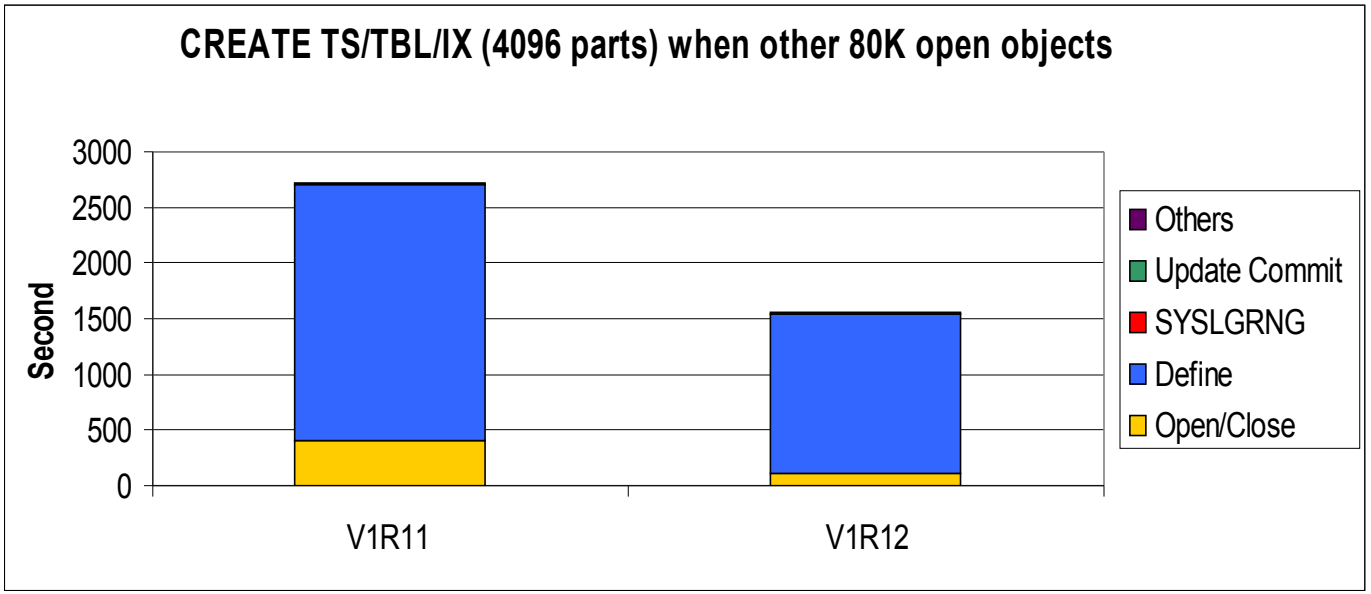
Reduce The Window - CREATE



- **CREATE Table Space or Index**
  - Takes long time to define and open the data sets
  - NPIs would increase the elapsed time of part level parallel LOAD data back to partitioned table space.
- **Actions**
  - CREATE new added objects beforehand, move it out of the cutover window.
  - Group DDL implementation jobs by database to run CREATE concurrently
  - Delay the creation of NPIs after table data get reloaded, then CREATE with DEFER YES followed by REBUILD INDEX.

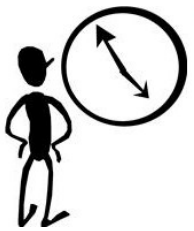
Create and Drop Performance with Large #of Data Set

- z/OS V1R12 shows significant improvement creating a table with 4096 partitions when other 80,000 DB2 objects are opened



- z/OS V1R12 has penalty in AMS SCRATCH (DB2 DROP) performance with large number of DB2 objects are opened
  - z/OS APAR OA36354 (July 2011) to reduce the cost of DROP
  - z/OS APAR OA37821 is opened for further enhancements on DROP performance

Reduce The Window - ROTATE

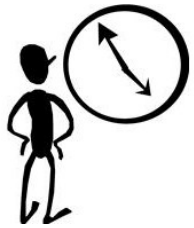


- **Long elapsed time in ROTATE**
  - Package invalidation, log write I/O, Notify messaging
- **Actions**
  - FREE packages depending on the table to be rotated before ROTATE, then BIND them when appropriate.
  - REORG INDEX on DSNDPX03 (index of SYSTABLEPART) to reserve sufficient free space
  - Keep only one DB2 member where ROTATE is running and shutdown other members to gain dramatic reduction in notify.
    - OR…
      - Remove GBP dependency of DSNDPX03 to reduce log write I/O,
      - Remove GBP dependency on the table space to be rotated
    - DB2 9 ACCESS DATABASE command with MODE(NGBPDEP) option to remove GBP-dependency easily.

Reduce The Window - ALTER TABLE ADD Partition

- **Long elapsed time on ALTER TABLE ADD Partition**
  - Notify message, SYSLGRNX sacn, EXT/DEL/DEF.
- **Actions**
  - Keep only one DB2 member where ADD PARTITION is running active and shutdown other members, this would dramatically reduce notify message time.
  - Better index for SYSLGRNX search : PM16655 (Aug 2010)
  - REORG SYSLGRNX before ADD PARTITION
  - Faster DFSMS CATALOG search : z/OS OA32612
  - Proactive action - Run MODIFY utility before dropping any tables to clean up the old SYSLGRNX entry.

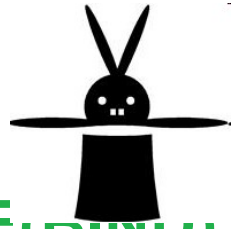
Reduce the Window - Data Migration and Package Update



- **Data migration window**
  - Increase the parallelism of COPY/UNLOAD by part level to reduce the total elapsed time.
  - Sort data during program unload (DSNTIAUL) or before load data back to tables, by clustering order.

- **Package update**
  - REORG PLAN_TABLE, SYSPACKAGE, SPT01, etc. beforehand to improve BIND/REBIND performance.
  - Bind online programs first, because batch programs can be bound after the cutover window.

Something Tricky



- **BIND VALIDATE(RUN) then REBIND VALIDATE(BIND)**

  Two major benefits from BIND packages with VALIDATE(RUN) option before formal BIND.

  - VALIDATE(RUN) does NOT check object existence and validity, which means packages can be preliminarily bound before DDL implementation is completed.
  - The elapsed time of REBIND VALIDATE(BIND) after BIND VALIDATE(RUN) is shorter than normal BIND, due to the package entry is created already.
  - Therefore, the first-step BIND can be run concurrently with DDL implementation with no contention, and the second-step BIND can save the total application cutover critical path time.

Access Path Control

- **Use DB2PLI8 to update stats information**
  - DB2PLI8 is a diagnostic tool provided by DB2 service team to gather DDL and catalog statistics
  - Collect stats info by DB2PLI8 from sandbox system and move the information into production system. This is to avoid RUNSTATS on production in the cutover window.
  - Access path can be analyzed and tuned in sandbox, before real cutover
- **DB2 9 New Function to model production environment**
  - PM26475 (DB2 9, PUT1103) provides the capability of modeling production environment
    - DSN6SPRM SIMULATED_CPU_SPEED
    - DSN6SPRM SIMULATED_CPU_COUNT
    - SYSIBM.DSN_PROFILE_ATTRIBUTES for Buffer pools, RID pool, Sort pool information

How Can DB2 9 Utilities Update Help

- **DB2 9 utilities update**
  - PM19584, LOAD PRESORTED option

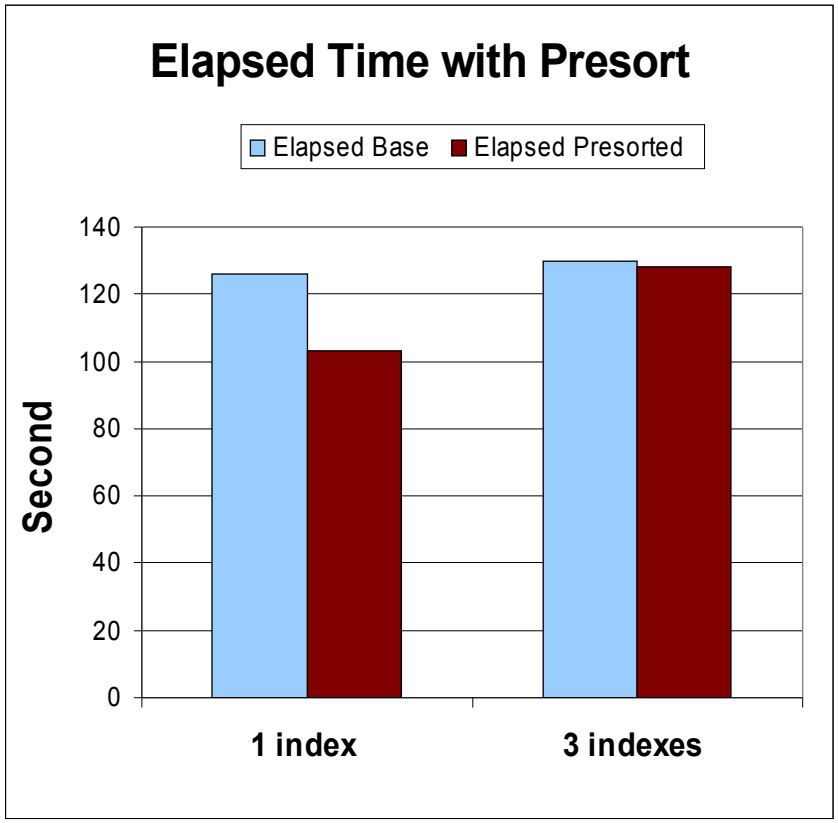
    When the input data is already in clustering order, LOAD PRESORTED option can help bypass the sort of clustering index, therefore reduce CPU time and elapsed time.
  - PM19584, LOAD and UNLOAD FORMAT INTERNAL option

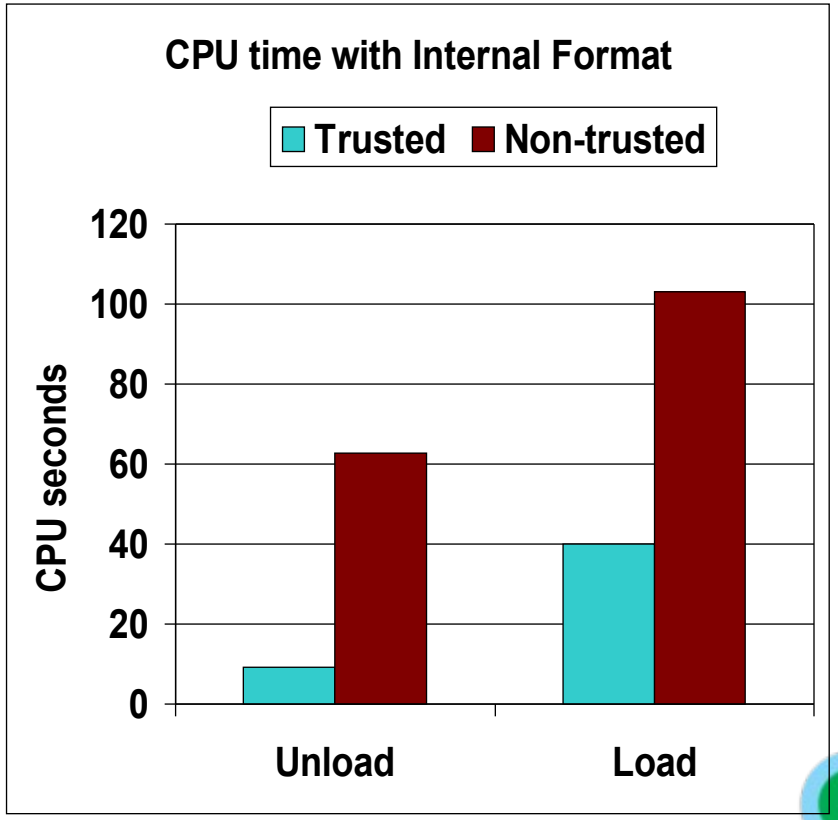
    The data will be unloaded and kept into DB2 internal format without any conversion, as well as load. This option can reduce CPU time and elapsed time because conversion and data type checking are bypassed.
  - PM27962, LOAD INDEXDEFER option

    For LOAD RESUME or part level LOAD REPLACE with NPIs, this INDEXDEFER option can help skip NPI processing. You can REBUILD NPIs when appropriate.

Load and Unload Improvement (IBM SVL Measurements)

- LOAD with PRESORT
  - Noticeable improvement with the object with one clustering
- Load and Unload with Internal (Trusted) format
  - More improvement with many small columns





How Can DB2 10 Help

- **Catalog restructure**
  - The DB2 10 NFM Catalog is restructured to dramatically reduce contention, include removing all links, using Row Level Locking
  - As a result, concurrent BIND/REBIND become possible.
- **Online schema enhancements**
  - DB2 10 supports more online schema changes, e.g. alter table space page size, data set size, segment size, table space type, index page size, MEMBER CLUSTER, etc.

Challenge
2. Online REORG

Online REORG Challenges

- **LOG APPLY is not catching up for hot table**
  - Hot table with heavy update
    - 1.8 Million changes per hour for 470 partitions
    - Through 12 members * 300 user applications
  - Online REORG log phase can't catch up during peak
    - LOG apply : 0.6 Million changes per hour
  - Solution : REORG at partition level to reduce log apply
- **UTSERIAL lock contention**
  - Running 300 parallel REORG jobs concurrent to reduce the total elapsed time, but need to rerun due to failure of UTSERIAL lock contention
  - Workaround: control the degree at 100 utilities running concurrent
  - Solution: DB2 10 eliminates UTSERIAL lock

Online REORG Challenges

- **DB2 9 (PK87762) uses parallel process for LISTDEF with multiple partitions**
  - OLTP impact because need to drain multiple partitions, the 1st partition drained take longer outage
  - Less chance to acquire drain for hot table
  - Storage constraints when process multiple huge partitions in parallel
- **Solution**
  - PM25525 adds a new PARALLEL keyword to the REORG
  - PM37293 adds a new REORG_LIST_PROCESSING zparm to control. Setting the zparm to SERIAL will serialize LISTDEF parts

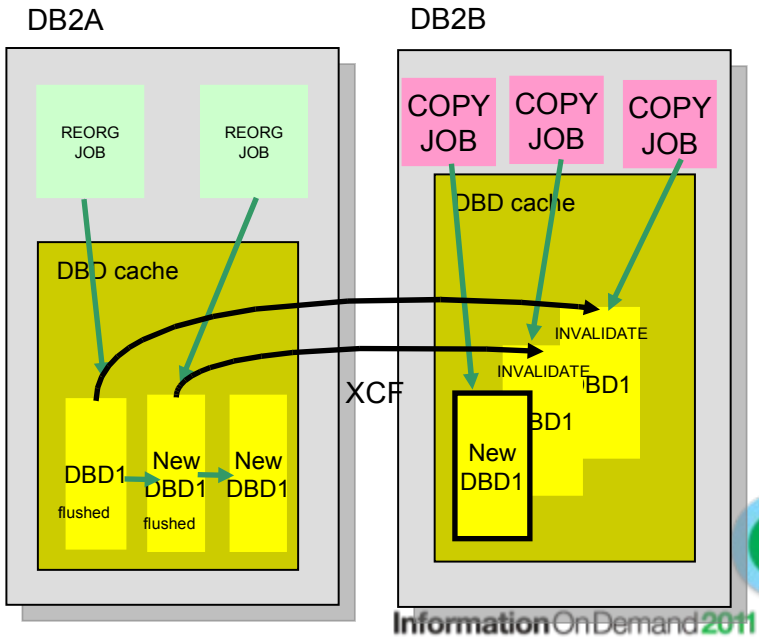
Online REORG Challenges

- **DFSORT usage on real storage**
  - Large Online REORG used more than 80GB real storage

    ICE199I 0 MEMORY OBJECT STORAGE USED = 0M BYTES
    ICE180I 0 HIPERSPACE STORAGE USED = 10482292K BYTES
    ICE188I 0 DATA SPACE STORAGE USED = 0K BYTES
  - DFSORT will use memory object OR Hiperspace but not both
  - To control the total Hiperspace+Memory Object+Dataspace storage in used by all sorts executing concurrently on a single system:
    - Set EXPMAX=5000 then the total used by all concurrent sorts will not exceed 5GB (However, that means if you have two utility jobs running concurrent the limit is 5GB total not 5GB for each one)
    - Or HIPRMAX=0, MOSIZE=MAX and MEMLIMIT of 5000M, this will limit DFSORT to only using memory object storage and each sort will evaluate the memory object storage already in use by the address space along with the MEMLIMIT to determine how much is available.

Online REORG Challenges

- Impact on EDM pool from Online REORG concurrent with COPY
  - REORG utility in data sharing modify DBD in EDM pool;DBD's Cross Invalidation occurs and user DBD is invalidated on other DB2 members. COPY utility job running on other member checks whether the DBD in the EDM pool is valid at each open dataset and holds the DBD in the EDM pool until it ends.
  - As a result, many DBD copies are temporarily kept in the EDM pool. User DBD size is large at 8.9MB and EDM pool full condition occurred
  - Options
    - Run REORGs on same member as COPY jobs
    - PM40388 : Avoid refreshing a new version when opening each data set, so COPY jobs do not keep DBD copies.



Online REORG Challenges

- **Sample REORG parameters used by Bank A**

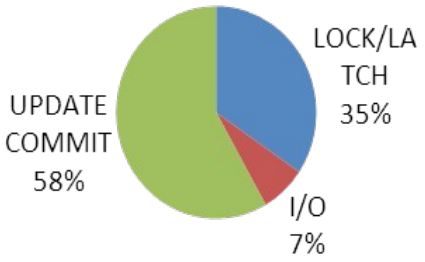
```
DRAIN_WAIT 20              RETRY 5    RETRY_DELAY 10
SHRLEVEL CHANGE            MAPPINGTABLE BANKA.MAPTAB
LONGLOG DRAIN MAXRO 25  DELAY 30  TIMEOUT TERM
FASTSWITCH YES             DRAIN ALL
SORTKEYS                   SORTDATA
SORTDEVT 3390              SORTNUM 5
COPYDDN SCOPY              UNLDDN SREC
DEADLINE CURRENT DATE + 23 HOURS + 30 MINUTES
```

Challenge 3 :
Night Batch Window vs. Night OLTP Response Time

Night Batch Performance – DB2 V8

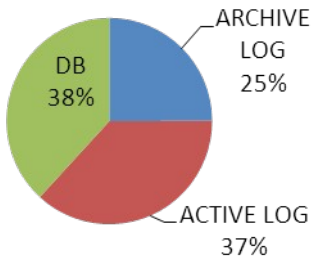
- **Heavy logging activity** due to concurrent batch update to process huge data within a narrow batch window
  - High DB2 latch wait (35% of suspension time)
    - LC19 124,000/s Log rate 32MB/s
  - Long update commit wait (58% of suspension time)
    - Log write /GBP write

2011-8-31 IN0800 suspension Second Distribution



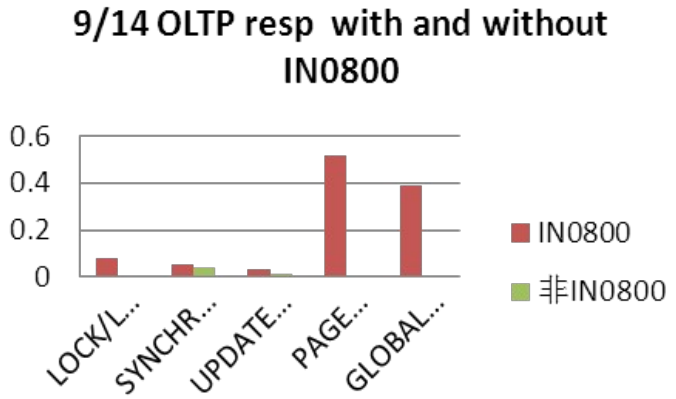
- **Actions**
  - 4 way extend to 8 way
  - Spread active log data set processing
  - Change page size from 4KB to16KB to reduce number of Get Pages, number of page locks/unlocks, reduce GBP reads and writes for batch update jobs.

2011/09/30 22:50 1.3GB/sec Write I/O Distribution



Night OLTP Performance – DB2 V8

- **Insert Issues – DB2 V8**
  - Synchronous log-Log force write when new pages are used in segmented table space when GBPD yes
  - With large partition (64GB) using MC, non zero FREESPACE, first insert after member planned outage, shows excessive getpages as it scans through all the space maps to reach the end
- **Page Latch**
  - S page latch for GBP write
- **Global Contention**
  - SPACEMAP/index split

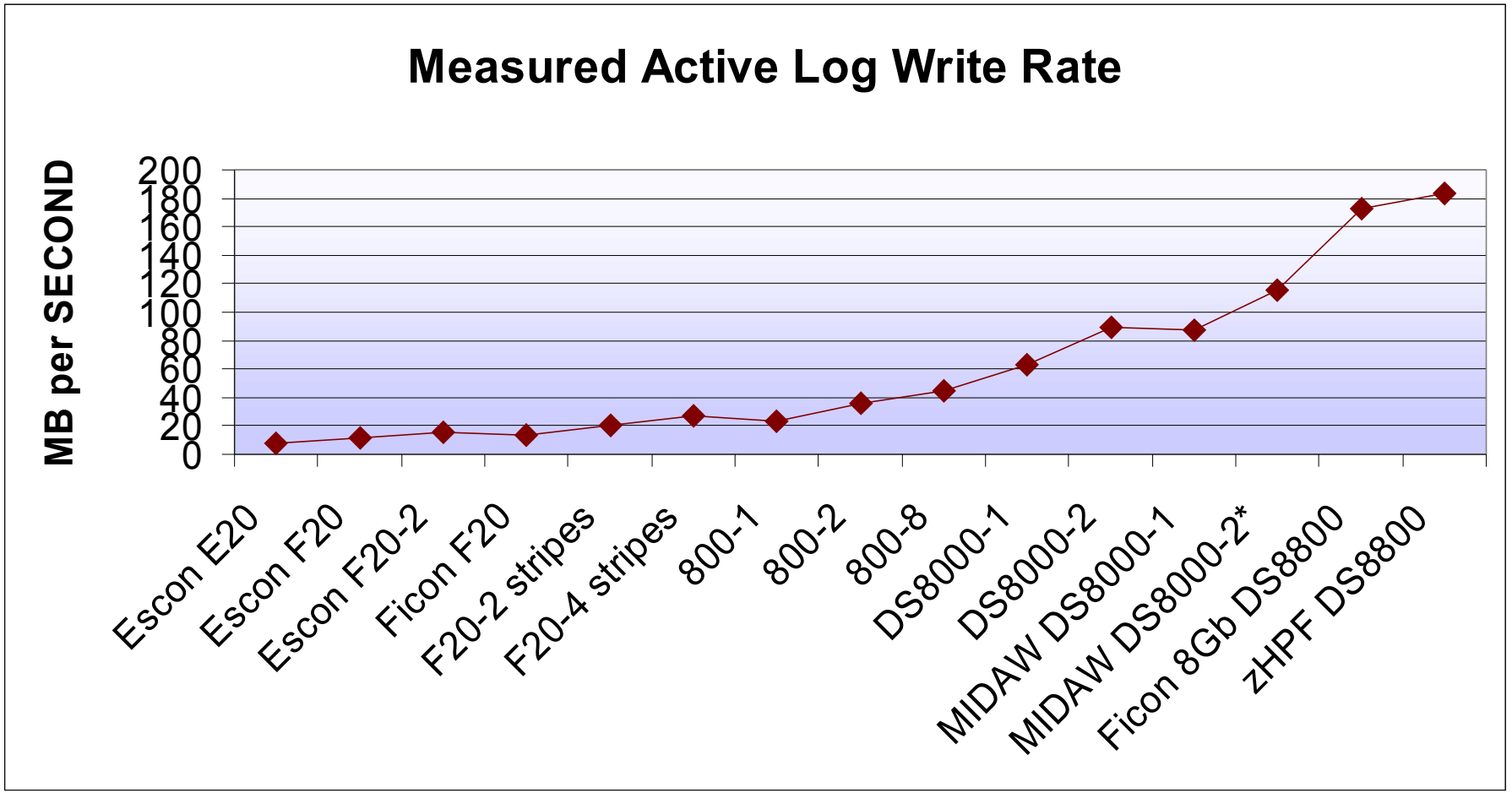


- **Options**
  - V9 removed the need of log force write when a new page is used in segmented table space under data sharing
  - Use MC00 to avoid insert spike
  - Increase hot night table page size, reduce S page latch for GBP write.
  - Partition and max 1row per page

DB2 9 and 10 Performance Improvement – DB2 Logging

- **Avoid log force write**
  - DB2 9 PM17817 during new data page format for UTS and Segmented table space
  - DB2 10 during CACHE processing in IDENTITY column
- **Log latch reduction in both data sharing and non data sharing**
  - DB2 9 Log latch reduction in data sharing
  - DB2 9 does not hold log latch during LRSN spin
  - DB2 10 reduction in both non data sharing and data sharing
- **DB2 10 Parallel log I/Os for dual logging**
  - Shorter LOG write I/O wait as I/O requests are done parallel for LOGCOPY1 and LOGCOPY2
- **DB2 10 Long Term Page Fix for output log buffers**
  - Reduce MSTR SRB time during log I/Os

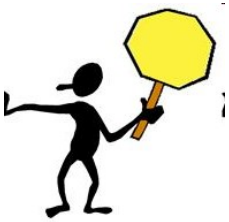
The Latest Disk Technology To Help Log I/Os



DB2 9 Migration Experience

Note:
This section applies to DB2 V8 -> DB2 10 skip migration, too.

DB2 9 Migration Experience 1



- **CPU time increase for some of packages**
  - Without REBIND, subsystem CPU increase from V8 to V9 is +10%
  - SELECT Procedure invalidation by release boundary
  - After REBIND top 20% frequently used packages, CPU increase from V8-> V9 dropped to +1%

	Without REBIND	After REBIND
BYPASS COL (QISTCOLS)	267, 965	12,085

- **ECSA**
  - Observed small increase per agent usage (5K)
  - But consider we have 200 agents per DB2 subsystem and 10 DB2 subsystems on 1 LPAR, then it is 10M increase which is visible now
  - > DB2 10 ECSA reduction

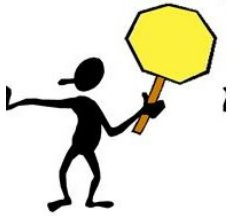
DB2 9 Migration Experience 2

- **Implicit DB**
  - DB2 9 will create DSN00001 to DSN60000 if you do not specify the IN clause on your CREATE TABLE statement
    - More DBA work to maintain so many databases in the test/dev. systems
  - Workaround: ALTER SEQUENCE SYSIBM.**DSNSEQ_IMPLICITDB MAXVALUE 1**
    - SYSIBM.DSNSEQ_IMPLICITDB can be altered by IDs with the installation SYSADM authority using the ALTER SEQUENCE statement. If the new maximum is lower than the number of existing implicitly created databases, DB2 will try to create or reuse database DSN00001 the next time an implicitly created database is needed and continue up to the new maximum.

DB2 9 Migration Experience

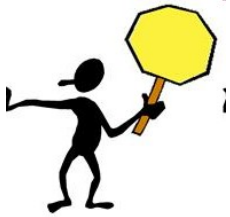
- **Application packages acquire s-lock on plan table**
  - LOAD REPLACE on plan table timeout because application has s-lock on plan table
  - Packages have both dynamic SQL & static SQL. If we use HINT on static SQL, dynamic SQL will check whether it should apply HINT by CURRENT OPTIMIZATION HINT register value, which will inherit from HINT name used to BIND package
  - Although the check result is none, it need access plan table anyway and acquire s-lock on it
  - Workaround is to set CURRENT OPTIMIZATION HINT to dummy at the beginning of the package

DB2 9 Migration Experience 3



- Test Center : All of work files are DB2 managed w/ secondary = 0.
  - Hit -904 for DGTT process
    - SQLCODE = -904, ERROR: UNSUCCESSFUL EXECUTION CAUSED BY AN UNAVAILABLE RESOURCE. REASON 00E7009A, TYPE OF RESOURCE 200
  - Created the work files with SECQTY > 0
- Production Center : All of work files are user managed w/ secondary = 0 except one workfile with SECQTY > 0
  - Hit -904 for DGTT process
    - SQLCODE = -904, ERROR: UNSUCCESSFUL EXECUTION CAUSED BY AN UNAVAILABLE RESOURCE. REASON 00D70025, TYPE OF RESOURCE 00000220
  - Created the DB2 managed workfile with SECQTY > 0

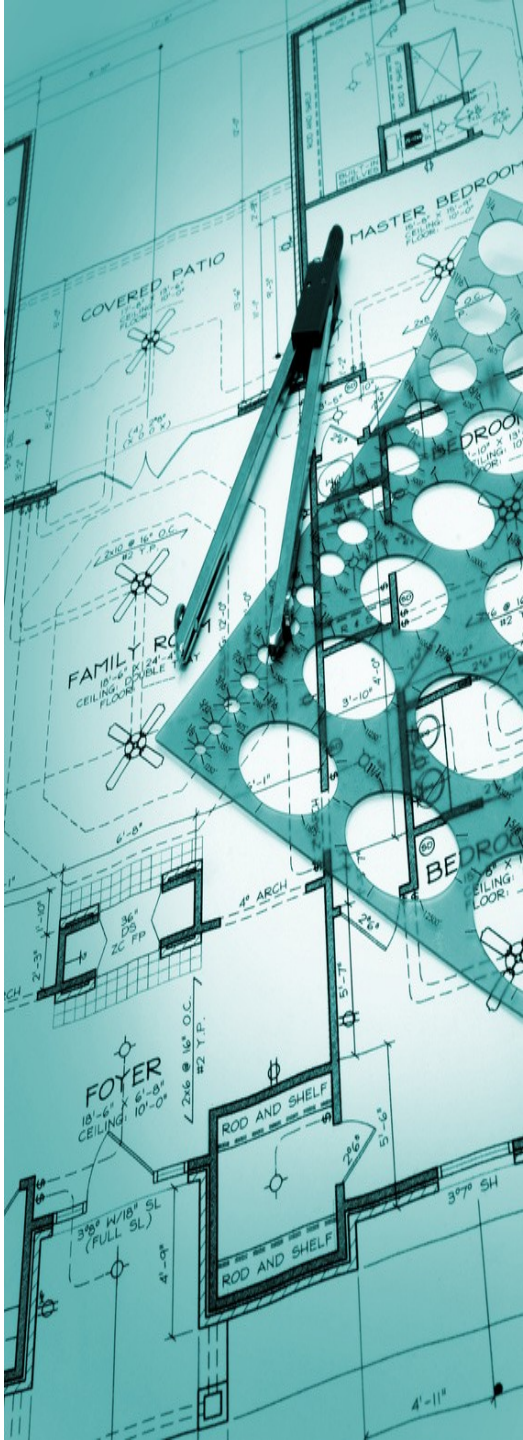
DB2 9 Workfile Separation



- DSN6SPRM WFDBSEP YES (default NO)
  - Physical separation of work file usage by DGTT and sort operations
- SECQTY = 0 work files for sort operation
  - Can be either user managed or DB2 managed
  - Requires more 32K work files starting DB2 9
- SECQTY > 0 work files for DGTT
  - Needs at least one DB2 managed 4K table space and DB2 managed 32K table spaces
  - Each space can grow to 64 GB
    - For high concurrent small DGTT operation, space growth is possible
    - Using smaller SEGSIZE can help to control storage growth
      - DB2 V8 TEMP SEGSIZE 4 -> DB2 9 SEGSIZE 16
      - SEGSIZE controls prefetch on DGTT

Summary

- **Frequent application upgrade**
  - Needs to be creative to fit in outage window
  - Help is expected from DB2 and z/OS enhancements
- **Online REORG**
  - Follow REORG recommendation and latest REORG enhancements
- **Insert Performance**
  - Great relief in DB2 10 and new disk technology
- **Experience from DB2 9 migration**
  - Most items apply to DB2 8->10 skip migration



Acknowledgement

Special Thanks to those who had contributed to this presentation

- James Teng (IBM SVL)
- Jiong FAN (IBM China)

Thank You… Leveraging the Best of z!

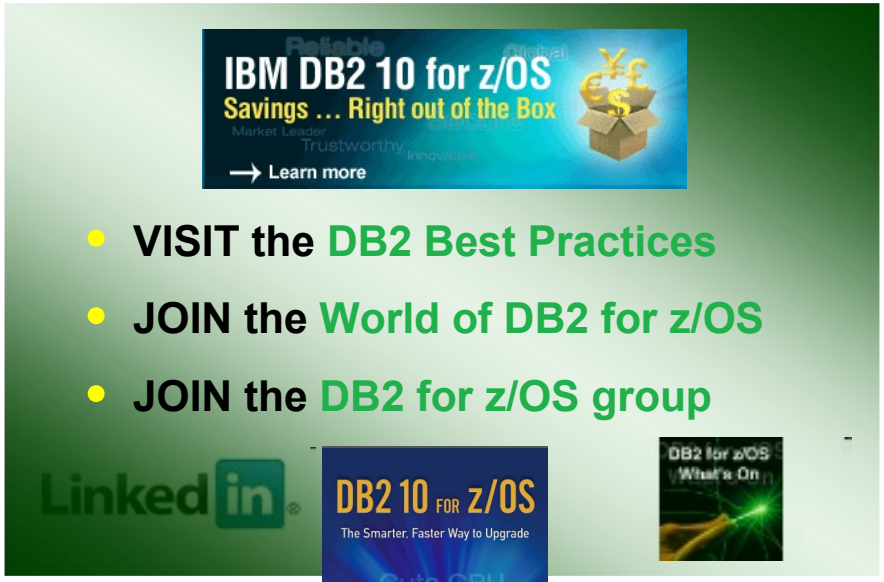
"The benchmarks and analysis of functionality and performance have exceeded our expectations. So far our upgrades have gone smoothly and we are looking forward to completing our successful rollout of DB2 10"

–Verizon

"We had migrated five sub-systems to DB2 10 and have had no reported application issues running on this release to date."

–

LabCorp



"The 'overall performance' in DB2 10 is better compared to DB2 9."

–HUK Coberg

[The Temporal Data] feature will drastically save developer time, test time … and improve business efficiency and effectiveness …

–Bankdata

Our regression tests showed performance improvements just by running the workload on a DB2 10 CM member…

–Dillards

An Exclusive Invitation for System z Attendees

ROCK THE MAINFRAME

at the

HOUSE OF BLUES

Music Hall

Wednesday, October 26th 7:00 pm - 10:00 pm

Enjoy a night of southern hospitality with cocktails and cajun hors d'oeuvres.

Keep the party rockin' by taking a turn on the Rock Band video game.

Join your colleagues, conference speakers and key members from your IBM System z team.

The House of Blues Music Hall is next door to the restaurant on the casino level across from the Mandalay Bay Hotel.

Wear your IOD badge and Z pin to get in

Thank You!
Your Feedback is Important to Us

- Access your personal session survey list and complete via SmartSite
  - Your smart phone or web browser at: iodsmartsite.com
  - Any SmartSite kiosk onsite
  - Each completed session survey increases your chance to win an Apple iPod Touch with daily drawing sponsored by Alliance Tech

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.

Acknowledgements and 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.