IBM Software Group

DB2 Information Management Technical Conference :
Z30 Recent DB2 RAS Enhancements

Jeff Josten josten@us.ibm.com

RAS Enhancements

- Reliability
  - Components fail rarely
- Availability
  - Applications/users detect little or no downtime even when outages occur
  - Redundancy
  - Failure isolation
  - Fast recovery
  - Data sharing, Parallel Sysplex
- Serviceability
  - Sufficient diagnostic information exists to quickly determine why a failure occurred
    - User to take corrective action to restore service
    - IBM to provide a fix for the root cause
  - First Failure Data Capture (FFDC) philosophy
- RAS is a traditional strength of DB2 for z/OS
  - and continues to be a key area of future development
- It's also a key component of IBM's OnDemand strategy
  - Autonomic, self-configuring, self-healing systems

RAS Enhancements...

- DB2 has delivered several recent RAS enhancements
  - Service stream
  - Version 8
- Customers can leverage these enhancements to increase DB2 availability
- Many of these came as a result of analysis of customer outage situations
- Some of the enhancements extend DB2's "autonomic" capabilities
  - No action needed to exploit
- Key areas
  - DB2 virtual storage management
  - Data integrity checking
  - Problem diagnosis
  - Recovery enhancements
  - "Autonomic" support

DB2 Virtual Storage Management

- Background
  - DBM1 "out of storage" is one of the leading causes of customer reported outages
    - Abend 04E/00E20003 or 04E/00E20016
  - Several drivers:
    - Bigger machines, higher workload volumes
    - Increasing use of dynamic SQL
    - New Java and Websphere workloads
    - Over-allocation of buffer pools, threads
    - Massive number of open datasets (with compression)
  - Largest consumers of DBM1 virtual include:
    - Buffer pools
    - EDM pool
    - Local dynamic statement cache
    - Thread storage
    - Compression dictionaries
  - DB2 has several enhancements to allow you to control virtual storage usage

IFCIDs 225 and 217

- IFCID 225 - DBM1 storage summary
  - Snapshot at each DB2 statistics interval
  - Activated via Stats Class 6, or Zparm SMFSTAT
- IFCID 217 - DBM1 storage detail
  - Mostly used for serviceability
- Supported by DB2PM V7 or V8
- DB2 V7, and DB2 V6 Apar PQ47973
  - Ensure PQ76942 applied
  - Enhanced in V7 apars PQ73385 to reflect actual MVS memory map for DBM1

```
QW0225LO DS    F          MVS 24 BIT LOW PRIVATE
QW0225HI DS    F          MVS 24 BIT HIGH PRIVATE
QW0225EL DS    F          MVS 31 BIT EXTENDED LOW PRIVATE
QW0225EH DS    F          MVS 31 BIT EXTENDED HIGH PRIVATE
QW0225RG DS    F          MVS EXTENDED REGION SIZE (MAX)
QW0225EC DS    F          MVS EXTENDED CSA SIZE
```

IFCIDs 225 and 217...

- Use this information to monitor DBM1 virtual storage
  - General recommendation: keep at least 200MB of headroom
  - If running short, then consider the following actions:
    - Data space buffer pools (if running on 64-bit machine), or Hiperpools (if not)
    - Reduce CTHREAD or MAXDBAT
    - Scale back usage of RELEASE(DEALLOCATE)
    - Use IRLM PC=YES to reduce ECSA
    - Investigate DB2 Version 8 for 64-bit virtual storage support
    - Other.... depending on you individual usage profile
- Reports above-the-bar storage in V8
- Obtain via IFI READS for online monitoring (V8)

Automatic Storage Contraction

- Background:
  - Certain DB2 storage pools can become fragmented over time
  - Especially with long-running threads
  - DB2 triggers storage contraction when the amount of available storage hits the "storage cushion" threshold
    - Cushion = 2 * (40K + max. # of threads * 20K)
  - The "contraction" process gives storage back to the operating system
  - DB2 contraction was not aggressive enough to sufficiently relieve some storage shortages
- Apar PQ55346 makes fundamental improvments
  - Allows system agent into the storage cushion (avoids DB2 outages)
  - Triggers storage contraction sooner
    - Longer runway for contraction
    - 5% of DBM1 or old formula, whichever hits first
  - Frees up storage sooner instead of waiting until pool contraction
  - Large pools contracted first
  - V6, V7, V8. Retrofit to V5 with PQ70767

Virtual Storage - other enhancements

- RDS OP pool framentation
  - OP pool can become fragmented with long running threads
  - One recent customer had a 250MB OP pool with lots of RELEASE(DEALLOCATE) threads
  - Apar PQ78515 - compresses the OP pool on a 10 min timer instead of waiting until thread deallocation (V6, V7)
- CONTSTOR zparm
  - Tells DB2 to contract agent local pools. Triggered at commit when
    - Past # commit threshold (50) or
    - Past storage threshold (2MB)
    - Threshold lowered to 1MB in V8
    - Zparms SPRMSTH (stg threshold) and SPRMCTH (commit threshold)
  - Some performance overhead
  - Ensure that PQ71156 is installed before setting to YES
- IRLM storage cushion
  - Apar PQ52642 - maintains a storage cushion, more effective stg contraction
    - Only applies for PC=YES
  - PQ52651 - Raises max lock limits for NUMLKUS and NUMLKTS
  - PQ53071 - Allows BIND and PREPARE to tolerate IRLM out-of-storage condtion

Virtual Storage - some tuning tips

- If you are running short of DBM1 virtual, consider one or more of the following:
  - Monitor INFO APAR II10817 and apply recommended maintenance
  - Use data space buffer pools if running on a 64 bit machine
  - Use hiperpools if not on a 64 bit machine
  - Scale back usage of RELEASE(DEALLOCATE)
    - Version 8 removes need for REL(DEAL) to avoid XES lock contention
  - Set CTHREAD and MAXDBAT defensively
  - Set CONSTOR=YES
  - Reduce number of open compressed datasets
  - Run IRLM PC=YES, reduce ECSA
    - With Version 8 "PC=YES" is forced
  - Consider making more use of data sharing to spread the work
  - Consider implementing Version 8 64bit virtual

Version 8 64bit virtual

- Why 64-bit?
  - Needed for DB2 to continue to scale on a single OS image
  - More concurrent threads
  - Bigger buffer pools
  - Large memory exploitation
- DBM1 and IRLM are 64bit address spaces in Version 8
- The following data areas are moved above the 2GB "bar"
  - Buffer pools, BM control blocks
  - Castout buffers
  - RID pool
  - EDM DBD cache (OBDs)
  - Global dynamic stmt cache
  - Sort pool
  - Trace tables (Global, Lock, BB)
  - Accounting blocks
  - Compression dictionaries
  - IRLM locks
- Max buffer pool size is lifted to 1 TB
  - New PGFIX(YES) option can give significant performance benefits
- Data space pools, hiper pools are eliminated - easier management

Improved Data Integrity Checking

- LOBs
  - CHECK DATA to unmark invalid LOBs
    - Improved availability
    - PQ77366 (V7)
  - LOB space reuse tracking
    - Improved diagnostic logging for data corruption
    - PQ77368 (V7)
  - COPY CHECKPAGE enhanced for LOBs in PQ74793
- Do not set COPYP if CHECKPAGE detects an error
  - Currently marks data unavailable which can have availability consequences
  - Implemented in V8
- DSN1LOGP CHECK(DATA) option
  - Check for regressed data page problems by analyzing the log
  - Introduced in PQ62820 (V6, V7), enhanced since
  - V8 NFM: enhanced to also handle indexes
- CHECK INDEX option to check integrity of non-leaf pages (V8)
- Future: online, less disruptive data health checking

Problem Diagnosis

- Improved log formatting

From this:

```
0000  01080017 6CA10100 00000000 3F80D7C1 D7E24040 4040D7C1 E8C8C9E2 F0F10001
0020  00000000 1000A80D 001400FE 1394D5B7 FA240000 00000000 2A08BA5E 4F7E7791
0040  80000000 00000010 00010606 80000000 1394D5B7 FA240000 000000BA 5E4F7E77
0060  91BA5D88 6F4E77BA 5E0482B9 EC000000 CDEE0110 00010606 93400000 1394D5B7
0080  FA240000 000000BA 5E4F7E77 91BA5D65 210C93BA 5DA45ADF 23000000 C5A30210
00A0  00010606 93400000 1394D5B7 FA240000 000000BA 5E4F7E77 91BA5D65 21CF8DBA
00C0  5DAB40BC 6A000000 BB400310 00010606 93400000 1394D5B7 FA240000 000000BA
00E0  5E4F7E77 91BA5CB3 49630DBA 5D652288 34000000 B7D20410 00010606 93400000
```

To this:

```
PAPS    .PAYHIS01
BPID:   13
NO. OF PARTS:   254
PART      0) GBP-DEPENDENT: YES
             GBPCACHE: CHANGED
             TESTPAGE: NO
             P-LOCK HELD STATE: SIX
             P-LOCK CACHED STATE: SIX
             P-LOCK UPGRADED: YES
             INSTANCE: I
PART      1) GBP-DEPENDENT: YES
             GBPCACHE: CHANGED
             TESTPAGE: NO
             P-LOCK HELD STATE: SIX
             P-LOCK CACHED STATE: SIX
             P-LOCK UPGRADED: YES
             INSTANCE: I
```

Problem Diagnosis...

- Improved log formatting for DBET log records
  - Implemented in V8, retrofit to V7 with PQ81113
  - Speeds diagnosis, and customers can use to answer historical questions
- Improved internal lock tracing
  - V8
  - Wraparound traces increased from 250 to 500 entries
  - Trace entire IRLM input parm area
  - Improved performance, improved serviceability
- Improved buffer manager tracing
  - V8
  - BBTR wraparound trace extended, additional flags traced
  - Improved serviceability
- Diagnostic log records for online REORG problem analysis
  - V8 under DIAGNOSE utility
- Log diagnostic information for space reuse problems
  - PQ77330 (V6, V7)
- Improved DSN1LOGP performance
  - PQ72747 (V7)

Problem Diagnosis...

- Improved hung thread diagnosis
  - New SERVICE(WAIT) keyword to display where in DB2 a thread is suspended, and for how long
  - If suspended on a latch, then show the blocking holder & boost priority
  - PQ83649 (V7)

```
 =DIS THD(*) SERVICE(WAIT)
DSNV401I  = DISPLAY THREAD REPORT FOLLOWS -
DSNV402I  = ACTIVE THREADS -
NAME     ST A   REQ ID           AUTHID   PLAN     ASID TOKEN
BATCH    T  *     5 TABLESQ      SYSADM   DSNTEP81 0023    28
V491-LATCH 7F51FA38 ASID 0022 HELD BY TABLESP  ASID 0020 TABLESP
V490-SUSPENDED 03354-19:11:54.06 DSNVXLT0 +00000432 15.42
BATCH    T  *     5 TABLESP      SYSADM   DSNTEP81 0020    13
V490-SUSPENDED 03354-19:07:46.52 DSNJW001 +000005AA 17.51
DISPLAY ACTIVE REPORT COMPLETE
DSN9022I  = DSNVDT '-DIS THD' NORMAL COMPLETION
```

Problem Diagnosis...

- IRLM DXR167E handling
  - Problem: highlighted IRLM DXR167E message issued, but users don't know how to respond
    - Most 'NOTIFY-L' and 'NOTIFY-G' conditions are normal
    - e.g. Global Drain, -SET LOG SUSPEND
  - IRLM apar PQ73524 suppresses DXR167E for NOTIFY-L/G and drives DB2 status exit to alert DB2
  - DB2 V7 apar PQ81252 to respond to IRLM status exit notification when a "stuck notify" is detected
    - DB2 issues new message DSNT369I and takes an SVC dump if the notify message processing indeed appears hung
    - One response is to issue -DISPLAY THREAD SERVICE(WAIT) for DB2 to boost priority of any suspended latch holders
      - Effective for very high CPU utilization levels
      - Working on longer term solution to avoid need for manual intervention
- Other DXR167E conditions: check IRLM dispatching priority
  - Run IRLM in service class SYSSTC

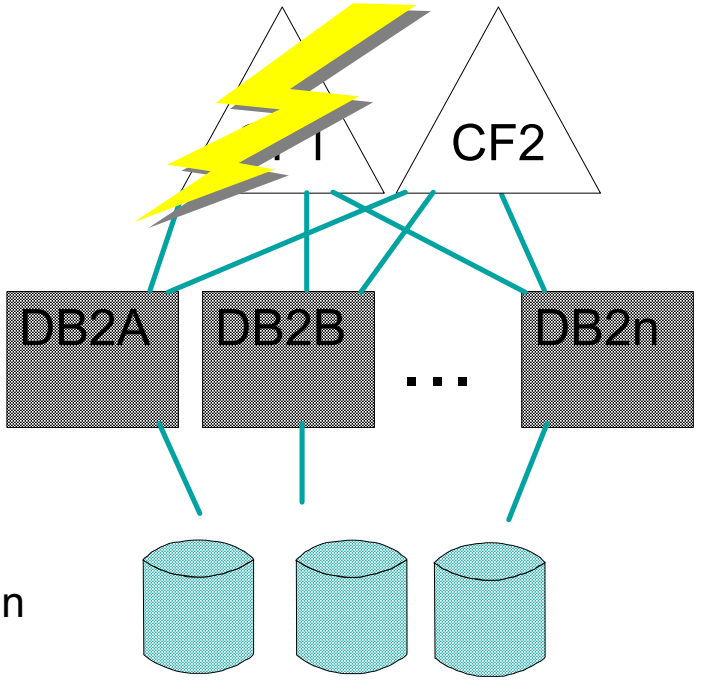
Restart

- Restart enhancement to avoid DB2 coldstart in error condition for postponed abort
  - PQ59410 (V6, V7)
  - If Zparm LBACKOUT changed to NO then Postponed Abort URs are changed to InAbort status and are backed out during restart
  - Also, Postponed Abort URs changed to InAbort for Restart Light
- DSNJU999 internal service aid to alter bufferpool size offline in BSDS
  - Useful when buffer pools mistakenly configured too large
  - PQ56034 (V6, V7)
- DB2 automatically resynch BSDSes if out-of-synch detected
  - Useful in DASD failure conditions, avoids need for manual intervention
  - Today DSNJ120I message: BSDS timestamp mismatch, DB2 restart fails
  - New message DSNJ131 issued to inform user that dual BSDS mode was restored and which BSDS copy was used as source
  - PQ73038 (V7)

Recovery

- Option to recover without SYSLGRNX
  - LOGRANGES NO option
  - Option of last resort if SYSLGRNX is damaged for some reason
  - Avoids need to stop SYSLGRNX which is highly disruptive to applications
  - PQ72669, PQ75397, PQ75398 (V7)
- Implicit utility restart
  - Avoids need to recode JCL to restart stopped utilities
  - PQ72337 (V7)
- V8 automatic LPL recovery
  - DB2 automatically initiates recovery when page added to LPL whenever possible
  - Avoid global drain of entire object when LPL pages are being recovered
    - Only pages in LPL are unavailable
    - Other pages remain available
  - DSNB250E message enhanced with more information about "why LPL?"
- V8 System Level PIT recovery
  - Two new utilities are introduced: BACKUP SYSTEM, RESTORE SYSTEM
  - BACKUP SYSTEM invokes HSM to take FlashCopy of database Copypool, and optionally Log Copypool
  - RESTORE SYSTEM restores a DB2 system to a prior point in time
    - LOGONLY option leaves the restore up to the user - don't need BACKUP SYSTEM

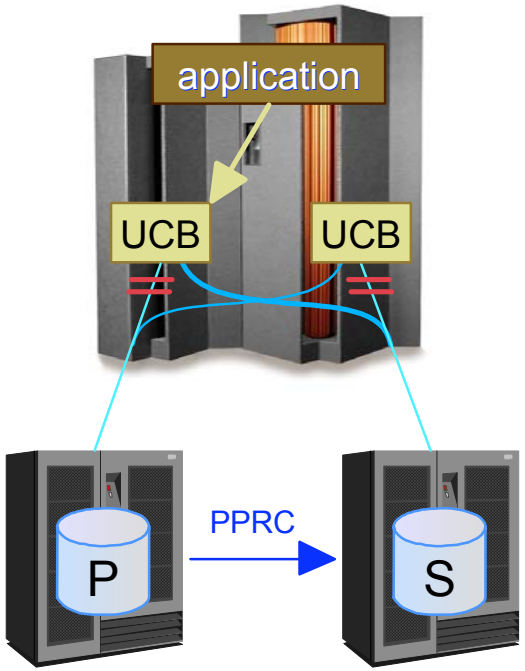
CF Duplexing



- GBP duplexing available since V5.
  - Recommendation: use it.
- Duplexing for Lock and SCA available in V7
  - Uses z/OS "system managed" CF duplexing which became GA in Feb '04
  - Useful to remove Internal Coupling Facility (ICF) as a single point of failure
  - Performance impact will depend mostly on intensity of CF lock requests for your application and distance between CFs
  - Pre-req's:
    - z/OS 1.2 or above plus maintenance
    - See CFDUPLEXING PSP Bucket
    - CFCC level 12 driver 3G (zSeries), or CFCC level 11 driver 26 (9672)
    - CF-to-CF links
  - http://w3-1.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/FLASH10272

GDPS/PPRC HyperSwap™...

Near continuous availability for DASD failures or site failures



- Site failover
  - DB2 data sharing group must still be recycled to recover CF structures
  - New technology under development to avoid having to recycle DB2 members
- DASD failover
  - DASD failure is transparent to DB2
  - DB2 detects no I/O errors
  - Dynamic switch to secondary and redriving of the I/O's is done by GDPS and z/OS
  - Consider for additional protection of DB2 catalog and other critical tables
- GDPS V2.8 HyperSwap prerequisites
  - New development APAR on z/OS 1.2 and up (not yet available as of this writing)
  - Parallel Sysplex with GRS Star
  - Disk subsystems that support PPRC level 3 (extended query)

DB2 "autonomic" support

- Goal: DB2 should automatically manage itself
- Some autonomic features already discussed. Here are a few more.
- Deferred End of Task monitoring, PQ82879 (V7)
  - dynamically create additional DEOT agents if needed (previous limit was 9)
  - Can avoid system hang when many threads are cancelled
- IRLM boosting logic
  - automatically boost priority of latch holder when waiter suspends
  - Avoids certain DXR167E conditions
  - Especially in systems with very high (>90%) CPU utilization levels
- V8 System Managed Extents
  - Avoid reaching "max number of extents reached" errors
  - DB2 automatically determines optimum primary/secondary extend sizes
  - To avoid running out of extents for large datasets
  - And avoid wasting space for smaller datasets
- V8 automatic lock priority management
  - Lower priority lock holder inherits the priority of the highest priority waiter
  - Priority is demoted once the lock is released
  - Helps with problem of low priority work getting in the way of higher priority work
  - Requires z/OS 1.4 or above

DB2 "autonomic" support...

- V8 automatic lock priority management
  - Lower priority lock holder inherits the priority of the highest priority waiter
  - Priority is demoted once the lock is released
  - Helps with problem of low priority work getting in the way of higher priority work
  - Requires z/OS 1.4 or above
- "Auto alter" of CF structures, OS/390 R10
  - XES monitors structure usage and dynamically adjusts size or directory/data ratio based on observations to avoid "structure full" and "XI due to directory reclaims"
  - ALLOWAUTOALTER(NO|YES) in CFRM policy, default=NO
  - Main value is for GBPs
  - Apply z/OS apar OW50397 (Oct '03) and DB2 PQ68114 (Aug '03) before enabling

Identifying Long Running Units of Recovery (URs)

- Problems caused by long running URs:
  - Long restart times (use of Postponed Abort can mitigate this)
  - Long lock hold time can impact concurrency
  - Reduces effectiveness of lock avoidance checking
  - Prevents online Reorg and other utilities from running
- Zparms control when DB2 issues long-running UR warnings:
  - UR CHECK FREQ , message DSNR035I (V5)
    - number of checkpoints that a UR has not committed
  - UR LOG WRITE CHECK, message DSNJ031I (V7)
    - number of log records that a UR has not committed
    - More granularity than UR CHECK FREQ
- IFCID 0313 written, if active, when long-runner detected
  - Use IFCID 0313 to keep a history of problematic URs
- Use UR CHECK FREQ for most accurate warning of restart impact
- Use UR LOG WRITE CHECK to accurately identify applications that write a lot of log records
  - DB2 Accounting reports can also help identify these guys
- V8: LONG-RUNNING READER THRESHOLD Zparm
  - Write IFCID 0313 when long-running reader detected

Other Version 8 RAS Enhancements include:

- Data Partitioned Secondary Indexes
- Online Schema Evolution
- 4096 Partitions
- Increase in active and archive logs
- Additional online Zparms
- Data sharing enhancements
- PARTKEYU no longer drains when partitioning key is updated
- Large VSAM Control Interval support

Summary

- DB2 has several recent RAS enhancements which users can exploit to improve the availability of their systems
- DB2 continues to add more "autonomic" features, with the goal being that DB2 should get to the point where it can manage itself
- Version 8 contains several key RAS enhancements
- We continue to assess whether it is appropriate to retrofit RAS enhancements to V7/V6 based on customer or IBM service benefit.