DB2 for z/OS  Data Sharing for Everybody

Bart Steegmans

GSE meeting 3 June 2010
Agenda

• What is DB2 data sharing and why would you want it?
• DB2 Data Sharing Architecture
• Increased availability
• Systems management
• Workload management
• Summary

Many thanks to Michael Dewert, Jeff Josten and Mark Rader for allowing me to use parts of their material.
DB2 for z/OS Data Sharing – What and why?
The world has changed

• When DB2 Data sharing was introduced in 1995
  • Only for the “happy few”
  • At that time the major reason was capacity growth

• A lot has changed since
  • Increased availability needs with ALL customers
  • Fewer opportunities for planned outages
  • Unplanned outages can cause a ‘disaster’ for the business
  • Server consolidations

• If you are not using DB2 data sharing already, maybe this is a good time to look into what it can bring to your installation
DB2 Data Sharing Definitions

• DB2 data sharing – allows applications running on more than one DB2 subsystem to read and write to the same set of data **concurrently**

• DB2 data sharing – allows customers to provide highest level of **scalability, performance and continuous availability** to enterprise applications that use DB2 data
Alternative Parallel DBMS Architectures

Shared Nothing (SN)
- Database is partitioned
- No disks are shared amongst the nodes
- Distributed commit is necessary
- Data repartitioning necessary as nodes are added
- Susceptible to skewed access patterns

Shared Disks (SDi)
- No database partition necessary
  - But partitioning can give better performance
- Strong fail-over characteristics
- Dynamic load balancing
- Inter-node concurrency and coherency control mechanisms are needed
  - Messaging overhead limits scalability

Shared Data (SDa)
- Adaptation of SDi
- Coupling facility is used as hardware assist for efficient concurrency and coherency control
- Strong fail-over and load balancing as with SDi
- Flexible growth
- Messaging overhead minimized, excellent scalability
Why DB2 Data Sharing?

• General drivers
  • Capacity: outgrow single system size
    • Avoid splitting the databases
  • Continuous availability requirements
    • Protect against planned and unplanned outages
  • Easier growth accommodation and cost effectiveness
    • Need scalable, non-disruptive growth
  • Dynamic workload balancing
    • Effective utilization of available MIPS for mixed workloads
    • Handle unpredictable workload spikes
  • System consolidation for easier systems management

• Application Investment Protection
  • SQL interface is unchanged for data sharing
  • Applications do not need to become "cluster aware"
  • “Turbo Charger” for existing applications
Why DB2 Data Sharing -2?

• Leverage existing infrastructure
  • Enhanced Operational Capability
  • Enhanced Return On Investment
    Better leverage of your investment in the mainframe
  • Low Risk Implementation – as technology is proven
  • Low cost or high cost implementation choices available
DB2 for z/OS Data Sharing – Architecture
Parallel Sysplex (PSX)

- Scalable Capacity
- Flexible Configuration
- Workload Balancing
- 7x24 Availability
- Single System Image

PSX Components:

- Sysplex Timer Protocol (STP)
  - ‘Sysplex Timers’
- Coupling Facility (CF) LPARs
  - High-speed shared memory
  - CF Control Code (CFCC)
  - Structures (Lock, Cache, List)
- CF Links
- CF Resource Management (CFRM) Policy
- Cross-System Extended Services (XES), part of z/OS
DB2 Data Sharing Basics

• A DB2 Data Sharing Group consists of:
  • 2 or more DB2 members with a single Catalog/Directory
    • 1-way data sharing initially, for migration, etc.
  • Active and archive logs for each member
    • Use log record sequence number (LRSN) instead of RBA
  • DB2 and User data on shared disk

• For DB2 use, Coupling Facilities (CFs) contain:
  • 1 Lock structure per data sharing group
  • 1 Shared Communications Area (SCA) per group
  • Multiple Group Buffer Pools (GBP) per group
    • 1 GBP per Buffer Pool containing shared data
    • GBP0, GBP8K0, GBP16K0, GBP32K0 required
DB2 Data Sharing – Implementation Steps

Coupling Facilities

SCA LOCK1
Group BP

Sysplex Timer Protocol (STP)

1-way Data Sharing

2-way Data Sharing

DB2A
IRLM
Buffer Pools

DB2A Log

DB2B
IRLM
Buffer Pools

DB2B Log

Shared DASD

DB2A Log

DB2B Log

DB2 Cat/Dir

DB2 DBs
DB2 N-way Data Sharing

Coupling Facilities

SCA LOCK1 Group BP

Sysplex Timer Protocol (STP)

DB2A
IRLM Buffer Pools

DB2B
IRLM Buffer Pools

DB2n
IRLM Buffer Pools

Shared DASD

DB2A Log
DB2B Log

DB2n Log

DB2 Cat/Dir

DB2 DBs
Critical Performance Factors

• Two factors to preserving data integrity in a data sharing environment
  • Inter-system *concurrency* control - global locking
    • Multiple readers OR
    • One writer
  • Inter-system buffer *coherency* control – managing changed data
    • When one system changes data rows that also reside in other system(s)
• "Data sharing overhead" is attributable to the extra CPU cost needed to manage these two factors
  • Thousands to tens of thousands of messages per second – not a problem on a well designed system
Inter-system Concurrency Control

Global locking using
PSX Coupling Technology:

• Cost of obtaining lock does not increase when adding (3rd through nth) DBMS instances (scalability)
Inter-system Buffer Coherency Control

Managing changed data using PSX Coupling Technology:

* Cross-invalidate (XI) to other member without interrupt.
CASTOUT processing

- **CASTOUT** is the process that ‘hardens’ changed pages in the GBP to disk.

- **CASTOUT** will occur when:
  - CLASST exceeded
    - Analogous to VDWQT
  - GBPOOLT exceeded
    - Analogous to DWQT
  - GBP checkpoint
  - No more inter-DB2 interest in the page set
  - GBP being rebuilt, but alternate GBP is not big enough to contain cached pages
Design Goals - Data Sharing Performance

• Little or no performance impact if data not actually shared, i.e. if no inter-DB2 R/W interest
  • Dynamic recognition of sharing

• Minimal and acceptable CPU overhead if inter-DB2 R/W interest exists
  • Overhead will vary based on individual workload characteristics

• Near-linear scalability when adding 3rd through n\text{th} nodes
  • Incur majority of data sharing overhead going from 1-way to 2-way
  • Beyond 2-way, additional overhead is minimal
DB2 Data Sharing OLTP Scalability

- IMS/TM with DB2 V4 OLTP workload
- 96.75% of ideal scalability from 2 to 8 nodes demonstrated
Data Sharing Performance Summary

- CPU cost of data sharing varies based on:
  - CF access intensity for locking and caching. This varies based on:
    - Percentage of CPU time in DB2
    - Degree of read/write sharing
    - Number of locks obtained
    - Access rate to shared data
    - Insert/delete intensity
    - Release of DB2
  - Hardware configuration
  - Lock contention rates

- Data sharing cost varies from one workload to another
  - 'Typical' 2-way data sharing overhead about 10%
    - Important: this is relative to the workload, not an address space
  - Individual jobs/transactions may have higher overhead
  - < 0.5% added cost per member past 2-way
Data Sharing Performance in Production

- Host CPU effect with primary application involved in data sharing
  - 10% is a typical average
  - Scalability and performance for real life customer workloads

<table>
<thead>
<tr>
<th>Industry</th>
<th>Trx Mgr / DB Mgr</th>
<th>z/OS Images</th>
<th>CF access per Mi</th>
<th>% of used capacity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pharmacy</td>
<td>CICS/DB2</td>
<td>3</td>
<td>8</td>
<td>10%</td>
</tr>
<tr>
<td>Insurance</td>
<td>CICS/IMS+DB2</td>
<td>9</td>
<td>9</td>
<td>10%</td>
</tr>
<tr>
<td>Banking</td>
<td>IMS/IMS+DB2</td>
<td>4</td>
<td>8</td>
<td>11%</td>
</tr>
<tr>
<td>Transportation</td>
<td>CICS/DB2</td>
<td>3</td>
<td>6</td>
<td>8%</td>
</tr>
<tr>
<td>Banking</td>
<td>IMS/IMS+DB2</td>
<td>2</td>
<td>7</td>
<td>9%</td>
</tr>
<tr>
<td>Retail</td>
<td>CICS/DB2+IMS</td>
<td>3</td>
<td>4</td>
<td>5%</td>
</tr>
<tr>
<td>Shipping</td>
<td>CICS/DB2+IMS</td>
<td>2</td>
<td>8</td>
<td>9%</td>
</tr>
</tbody>
</table>

Note: “Mi” stands for ‘million instructions’
DB2 for z/OS Data Sharing – Increased availability
Design for High Availability

• Most single points of failure eliminated:
  • DB2 subsystem or z/OS system
  • CPC (or CEC)
  • I/O path

• Goal: Continuous availability across planned or unplanned outage of any single hardware or software element

• Strategy:
  • Remove all causes for planned outages
  • Build on legacy of robust, fault tolerant MVS components
  • On a failure:
    • Isolate failure to lowest granularity possible
    • Automate recovery and recover fast
DB2 Member Outage – Planned

• "Rolling" maintenance
  • One DB2 stopped at a time
  • DB2 data continuously available via the N-1 members
  • Other members temporarily pick up work of the member that is down
  • Batch work can be offloaded to another member with more available capacity to reduce the batch window

• Applies to hardware and operating system changes too
  • Rolling IPLs

• KEY TO SUCCESS: Applications must be able to run on more than one DB2 member!
DB2 Release Migration

- DB2 Data Sharing Group can be available – and applications executing - across release migrations
- N/N+1 release levels can coexist
- Coexistence of mixed releases in a data sharing group can add complexity
  - Consult DB2 manual, *Data Sharing: Planning and Administration* for details

- Process:
  - Apply SPE to each DB2 member
  - Restart each DB2 member
  - Put in new release
  - Restart each DB2 member

- Catalog migration done once per data sharing group
  - Catalog migration locks some of the catalog
  - Recommendation: Perform during a period of low activity
DB2 Member Outage – Unplanned

- The other "surviving" members remain up and running
- The architecture allows all members to access all portions of the database
- Work can be dynamically routed away from the failed DB2 member – assuming applications can run on >1 DB2
- The failed member holds "retained locks" to protect inconsistent data from being accessed by other members
- MVS Automatic Restart Manager (ARM) can automatically restart failed DB2 members
- Restart ‘Light’ minimizes impact of LPAR failures
Coupling Facility Outages

- **Planned outages**: use the z/OS operator command to "rebuild"
  - REBUILD moves the structures to another CF LPAR
  - No outage to data sharing group
  - REALLOCATE to ‘move’ back
- **Unplanned outages**: the system automatically recovers the lost structures
  - Lock & SCA are dynamically rebuilt into alternate CF
    - Spare CF capacity required to house the structures
    - ‘White space’ part of CF capacity planning
  - GBPs must be duplexed for high availability
  - DB2 for z/OS allows duplexing of Lock and SCA but duplexing not necessary for high availability
    - Unless in a 2-ICF configuration.
Duplexing – 2 Kinds

• “User-managed” duplexing
  • Applies to GBPs
  • “User” = DB2; DB2 is responsible for managing two structures in different CF LPARs

• “System-managed” duplexing
  • Applies to LOCK1 and SCA
  • XES (z/OS) is responsible for managing two structures in different CF LPARS
    • DB2 is not aware of second structure
    • Everything happens on the primary, then on the secondary
  • Recommended if CFs run on Internal Coupling Facilities (ICFs) and co-reside on CEC with DB2 members
    • Avoid ‘double failure scenario’
    • More expensive than DB2’s user managed duplexing (more CF requests)
  • For best availability and performance, ensure that Lock and SCA are failure-isolated from the DB2 members
GBP Duplexing

• Allocate secondary GBP on alternate CF
• Write changed pages to both primary and secondary
• If loss of connectivity or loss of structure -
  • Switch to secondary (seconds)
  • No rebuild required; changed pages already in GBP
  • Cross-invalidate buffers and gradually repopulate directory entries
• No application outage unless both primary and secondary GBPs are lost
• Always enable GBP duplexing
  • Minimal performance overhead
  • Can avoid long outage time in case of CF failure
V5 GBP Duplexing Recovery

CF(GBP) Failure case

IRWW workload
Measured at 08/22/98

Transaction per 10 seconds

TIME (10sec interval)
Parallel Sysplex Configurations

A

CF01

DB2A

CF02

DB2B

B

CF01

DB2A

DB2B

ICF02

C

DB2A

ICF01

DB2B

ICF02

D

DB2A

ICF01

DB2B

ICF02
Parallel Sysplex Configurations

• **A**: “Traditional” configuration
  - LOCK1 and SCA in one CF
  - Duplexed GBPs spread across both CFs
    - Primary and secondary GBPs balanced based on load

• **B**: One Integrated CF (ICF), one external CF
  - LOCK1 and SCA in external CF
  - Duplexed GBPs spread across both CFs
    - Primary GBP in ICF has advantages for ‘local’ DB2

• **C**: Two ICF configuration
  - Lock1 and SCA duplexed; allocated in both CFs
    - “System-managed” duplexing
    - Performance implication for LOCK1 requests
  - Duplexed GBPs spread across both CFs
    - Primary GBP in ICF has advantages for ‘local’ DB2
Sysplex in a Box

- Some SAP customers use this configuration to avoid planned outages
  - Still at risk of unplanned outages, especially hardware
Disaster Recovery (DR) Planning

- Many disaster recovery techniques are available and used
  - PTAM
  - XRC
  - GDPS
  - ...

- Don’t forget to delete the DB2 CF structures before starting up at the remote site!!!

- HyperSwap to protect against DASD failures
  - Used by GDPS for site failover
  - z/OS 1.10 “Basic Hyperswap” for local site availability (not DR)

- Recommendation:
  - Practice Disaster Recovery processes and procedures
Geographically Dispersed Parallel Sysplex

Single System z → Parallel Sysplex → GDPS

Site 1 → Site 1 → Site 1

System z → Site 2+
DB2 for z/OS Data Sharing – Systems management
Systems management

- Requires a parallel sysplex
- New hardware
  - CF, CF links, Sysplex Timer Protocol (STP), ..
- New software components
  - XCF, XES, CFRM policy, ..
- Existing software behaves different
  - Eg. DB2 needs different structures in the CF, ..
  - Not just DB2, but also MVS, IMS, CICS, ...
- Don’t rush into it – Build knowledge!
  - Carefully design, set up, test
  - DR will be have to be adjusted as well
Systems Management -2

• **Goal:** Single - system image

• **DB2 data sharing "ease of use“ features:**
  - Command prefix support
  - Group attachment name
  - DDL, Bind, utilities, authorization are all "group scope"
  - Single DDF location name for the DB2 data sharing group
    • Subset: LOCATION ALIAS
  - "Group scope" on display output
  - "Group scope" for online performance monitors
  - Log merging for replication (IFCID 306)
Dynamic Workload Balancing

- Workload Manager (WLM)
- CICSPlex System Manager (CPSM)
  - Route workload between CICS TORs and AORs
- IMS Transaction Manager (TM)
  - Shared message queues
  - BMPs
- WebSphere
  - Connection pooling
  - Connection concentration
- Distributed access (DDF)
  - Dynamic Virtual IP Addressing (DVIPA)
- Sysplex Distributor
WebSphere

To exploit DB2 Data Sharing workload balancing and transparent failover, both application server connection pool AND connection concentrator/ connection pool should be used
Sysplex Distributor and DVIPA

First connection through Sysplex Distributor determines to which DB2 member the requester will attach - Workload Balancing - Availability

“DB2 Universal Driver JDBC/SQLJ” IBM Data Server Driver or DB2 Connect

CONNECT TO GroupIP

“DB2 Universal Driver JDBC/SQLJ” Or DB2 Connect

CONNECT TO GroupIP

Sysplex Distributor

Pooled connections to DB2 Data Sharing Group
DB2 for z/OS Data Sharing – Applications changes
Application Changes?

- SQL interface does not change
- However, locking and commit frequency may impact data sharing performance
  - Commit frequently – long time recommendation (Also applies to non-data sharing)
  - Take advantage of lock avoidance
    - ISO(CS) and CURRENTDATA NO
- New messages and return codes
- Applications should be able to run on more than one DB2 member for high availability
- Summary: “Turbo Charger” for existing applications
Does data sharing allow parallelism?

V3
- I/O Parallelism
  - Single Execution Unit
  - Multiple I/O streams
  - TCB (originating task)

V4
- CP Parallelism
  - Multiple Execution Units
  - Each has a single I/O stream
  - SRBs (parallel tasks)

V5
- Sysplex Query Parallelism
  - Sysplex Query Parallelism
  - Parallel tasks spread across Parallel Sysplex
Best Practices Can Help You Improve

• **Availability**
  - Tune for fast DB2 restarts
  - Tune RETLWAIT system parameter to wait for incompatible retained locks instead of immediate rejection
  - Use Restart Light for LPAR failures
  - Combine with z/OS Automatic Restart Manager (ARM)
  - Use GBP Duplexing
  - Utilize rolling software upgrades
  - Use advanced disaster recovery features

• **Performance**
  - High performance insert
  - Minimize lock contention
    - ISO(CS) CURRENTDATA(NO) or ISO(UR) to encourage lock avoidance and minimize impact of retained locks
  - Commit frequently!
  - Avoid “XIs due to directory reclams”
Best Practices Can Help You Improve -2

- Avoid “affinity routing”
  - Sysplex-enable applications to run on multiple DB2 members
  - Symmetry of group brings operational simplicity, makes it easy to scale out horizontally, and easier to roll software maintenance
  - Remove application affinities to specific DB2 members so that connections can be dynamically rerouted to other members
  - Achieve much better high availability if affinity routing is avoided
  - Avoid limiting access to an object from a single member
    - Makes continuous availability unworkable
    - Retained ‘X’ page set P-lock prevents access from other members until failed member completes forward phase of restart
DB2 Data Sharing Should Be in Your Life

- Data sharing technology provides the base to allow DB2 to deliver continuous availability and nearly unlimited scalability into the future
- DB2 Data Sharing is a proven technology
  - Many customers in DB2 data sharing production
- Work is ongoing to deliver further data sharing enhancements in future releases
DB2 for z/OS Data Sharing – Summary
DB2 data sharing and Parallel Sysplex

- **Grow capacity**
  - More tran/sec
  - Reduce ET for single query by exploiting sysplex CPU

- **Increase availability**
  - Unplanned outages
  - Planned outages
    - Rolling maintenance

- **Workload distribution**
Enhanced Operational Capability and Return On Investment

• Have you already invested heavily in system infrastructure?

• Not getting full value from that investment?

• Leverage existing investment and achieve step change in operational capability

• Significant improvement to in-site operational recovery capability

• New model for in-site failover can be enabled
  • vs. full Disaster Recovery invocation

• Better utilisation of existing system resources through workload balancing

• Removing scaling factors that might inhibit M&A activity
Low Implementation Risk

- Well proven technology
- Widespread deployment around worldwide global customer community
- Global experience and skills
- Evolve to target environment with incremental change
- Target applications with demanding requirements for continuous availability
You are in good company
Thank you

Questions?

bart_steegmans@be.ibm.com
Disclaimer

© Copyright IBM Corporation 2009. All rights reserved.
U.S. Government Users Restricted Rights - Use, duplication or disclosure restricted by GSA ADP Schedule Contract with IBM Corp.

THE INFORMATION CONTAINED IN THIS PRESENTATION IS PROVIDED FOR INFORMATIONAL PURPOSES ONLY. WHILE EFFORTS WERE MADE TO VERIFY THE COMPLETENESS AND ACCURACY OF THE INFORMATION CONTAINED IN THIS PRESENTATION, IT IS PROVIDED “AS IS” WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED. IN ADDITION, THIS INFORMATION IS BASED ON IBM’S CURRENT PRODUCT PLANS 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 PRESENTATION OR ANY OTHER DOCUMENTATION. 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 ANY AGREEMENT OR LICENSE GOVERNING THE USE OF IBM PRODUCTS AND/OR SOFTWARE.

IBM, the IBM logo, ibm.com, and DB2 are trademarks or registered trademarks of International Business Machines Corporation in the United States, other countries, or both. If these and other IBM trademarked terms are marked on their first occurrence in this information with a trademark symbol (© or ™), these symbols indicate U.S. registered or common law trademarks owned by IBM at the time this information was published. Such trademarks may also be registered or common law trademarks in other countries. A current list of IBM trademarks is available on the Web at “Copyright and trademark information” at www.ibm.com/legal/copytrade.shtml