Advanced Technical Support (ATS), Americas

Why DB2 Data Sharing Should Be in Your Life

Mark Rader
IBM Dallas Systems Center
Advanced Technical Support (ATS Americas)

Disclaimer and Trademarks

Information contained in this material has not been submitted to any formal IBM review and is distributed on "as is" basis without any warranty either expressed or implied. Measurements data have been obtained in laboratory environment. Information in this presentation about IBM's future plans reflect current thinking and is subject to change at IBM's business discretion. You should not rely on such information to make business plans. The use of this information is a customer responsibility.

IBM MAY HAVE PATENTS OR PENDING PATENT APPLICATIONS COVERING SUBJECT MATTER IN THIS DOCUMENT. THE FURNISHING OF THIS DOCUMENT DOES NOT IMPLY GIVING LICENSE TO THESE PATENTS.

TRADEMARKS: THE FOLLOWING TERMS ARE TRADEMARKS OR ® REGISTERED TRADEMARKS OF THE IBM CORPORATION IN THE UNITED STATES AND/OR OTHER COUNTRIES: AIX, AS/400, DATABASE 2, DB2, e-business logo, Enterprise Storage Server, ESCON, FICON, OS/390, OS/400, ES/9000, MVS/ESA, Netfinity, RISC, RISC SYSTEM/6000, iSeries, pSeries, xSeries, SYSTEM/390, IBM, Lotus, NOTES, WebSphere, z/Architecture, z/OS, zSeries,

The FOLLOWING TERMS ARE TRADEMARKS OR REGISTERED TRADEMARKS OF THE MICROSOFT CORPORATION IN THE UNITED STATES AND/OR OTHER COUNTRIES: MICROSOFT, WINDOWS, WINDOWS NT, ODBC, WINDOWS 95

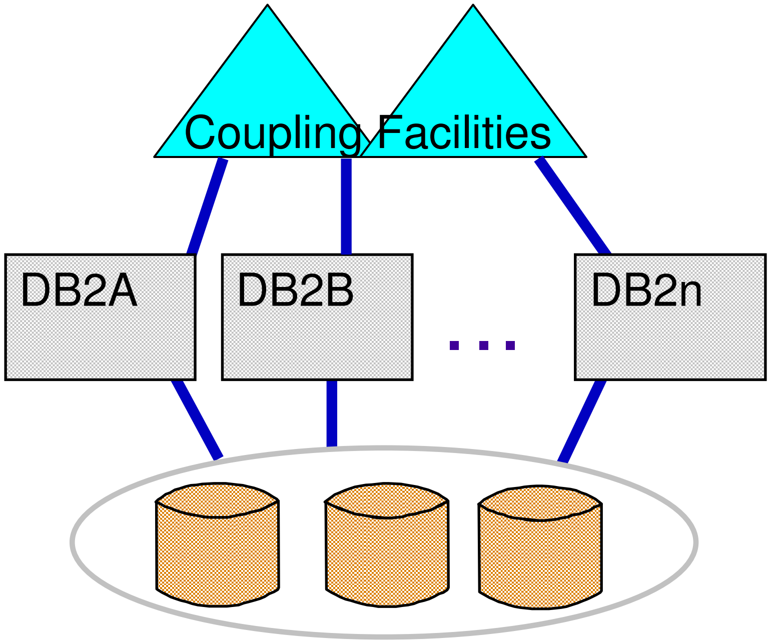
For additional information see ibm.com/legal/copytrade.phtml

Objectives

- Introduce DB2 data sharing concepts
- Describe how DB2 data sharing and Parallel Sysplex (PSX) provide the on-demand infrastructure for:
  - Increased availability
  - Non-disruptive scalability
  - Dynamic workload balancing
- Provide answers to common questions
- List additional sources of information

Agenda

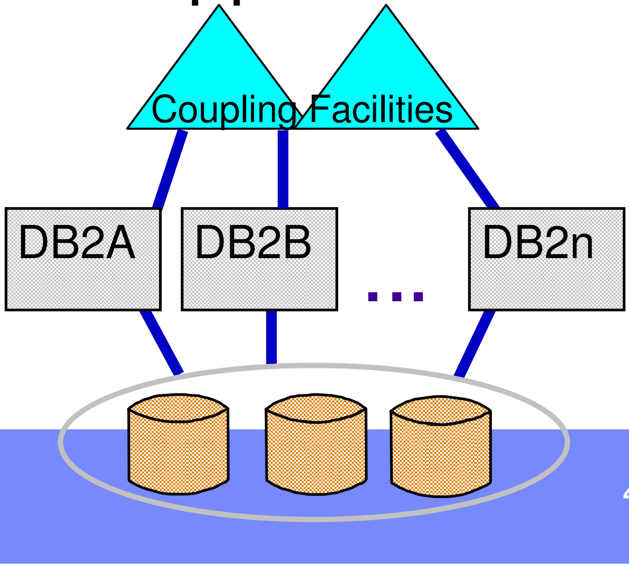
- DB2 data sharing – what is it? Why implement it?
- DB2 data sharing concepts
  - Coupling Facilities (CFs)
  - Performance and availability
  - Dynamic workload balancing
- Frequently asked questions



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.

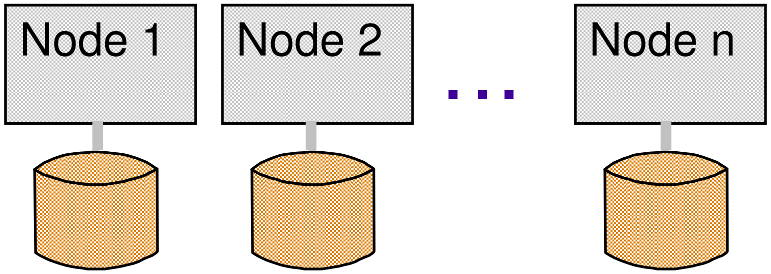
- DB2 data sharing – allows customers to provide highest level of scalability, performance and continuous availability to enterprise applications that use DB2 data.



Why Go to DB2 Data Sharing?

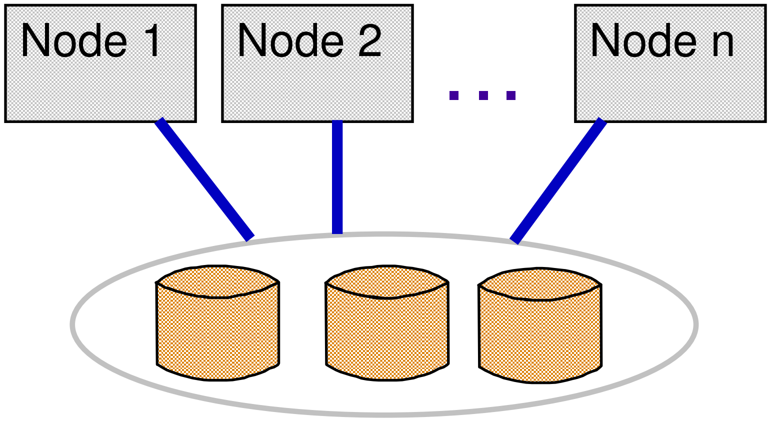
- Most common drivers:
  - Capacity: outgrow single system size
    - Avoid splitting the databases
  - Higher availability requirements
    - Protect against planned and unplanned outages
  - Easier growth accommodation
    - Need scalable, non-disruptive growth
  - Dynamic workload balancing
    - 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
  - Excellent scaling: applications do not need to become "cluster aware" as nodes are added
  - Most known DB2 members: 17-way

Alternative Parallel DBMS Architectures



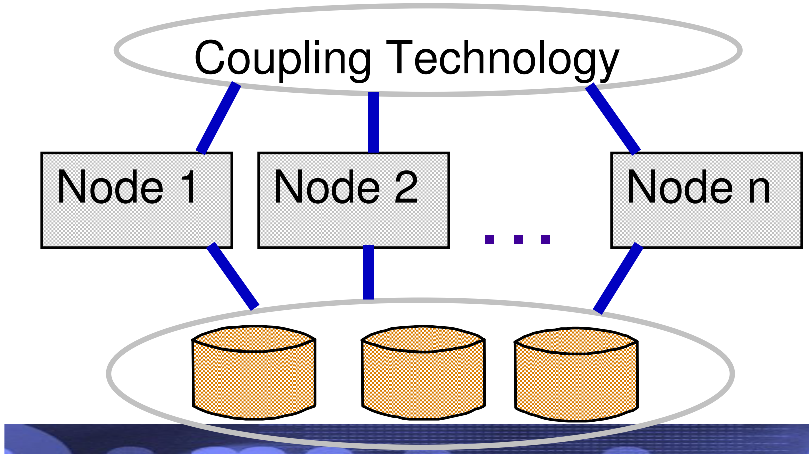
Shared Nothing (SN)

- Database is partitioned
- No disks are shared amongst the nodes
- Distributed commit is necessary
- Data repartioning 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

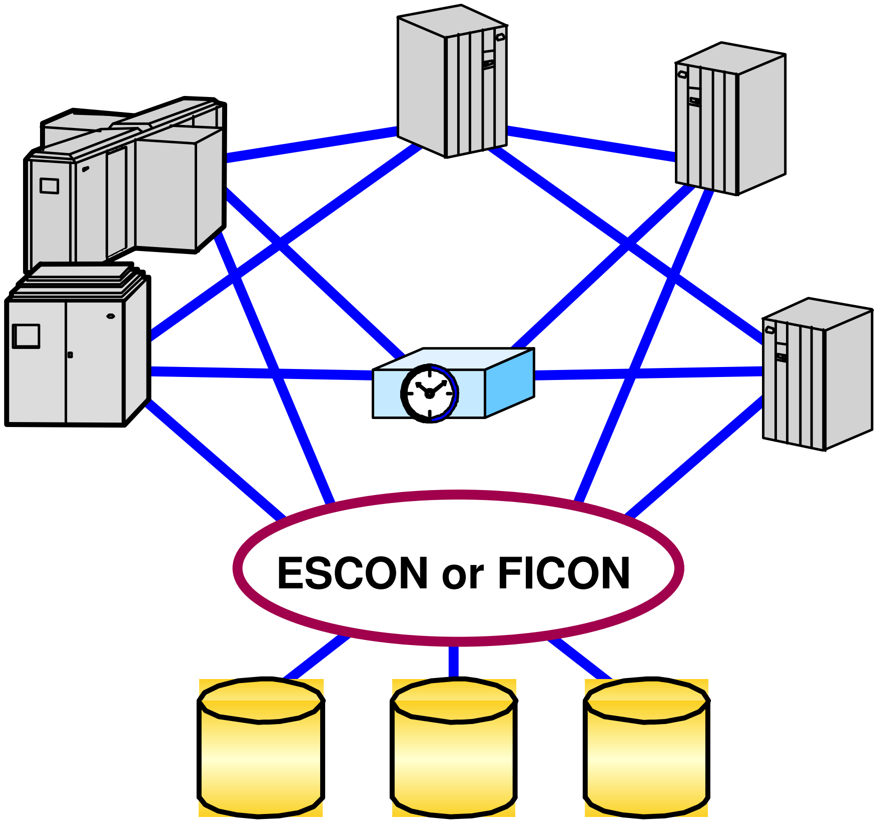


Shared Data (SDa)

- 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

Parallel Sysplex (PSX)

- Scalable Capacity
- Flexible Configuration
- Workload Balancing
- 7x24 Availability
- Single System Image
- PSX Components:
  - 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
  - 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, Coupling Facilities (CFs) contain:
  - 1 Lock structure per data sharing group
  - 1 Shared Communications Area (SCA) per group
  - Multiple Group Buffer Pools (GBP)
    - 1 GBP per Buffer Pool containing shared data
    - GBP0 required

Critical Performance Factors

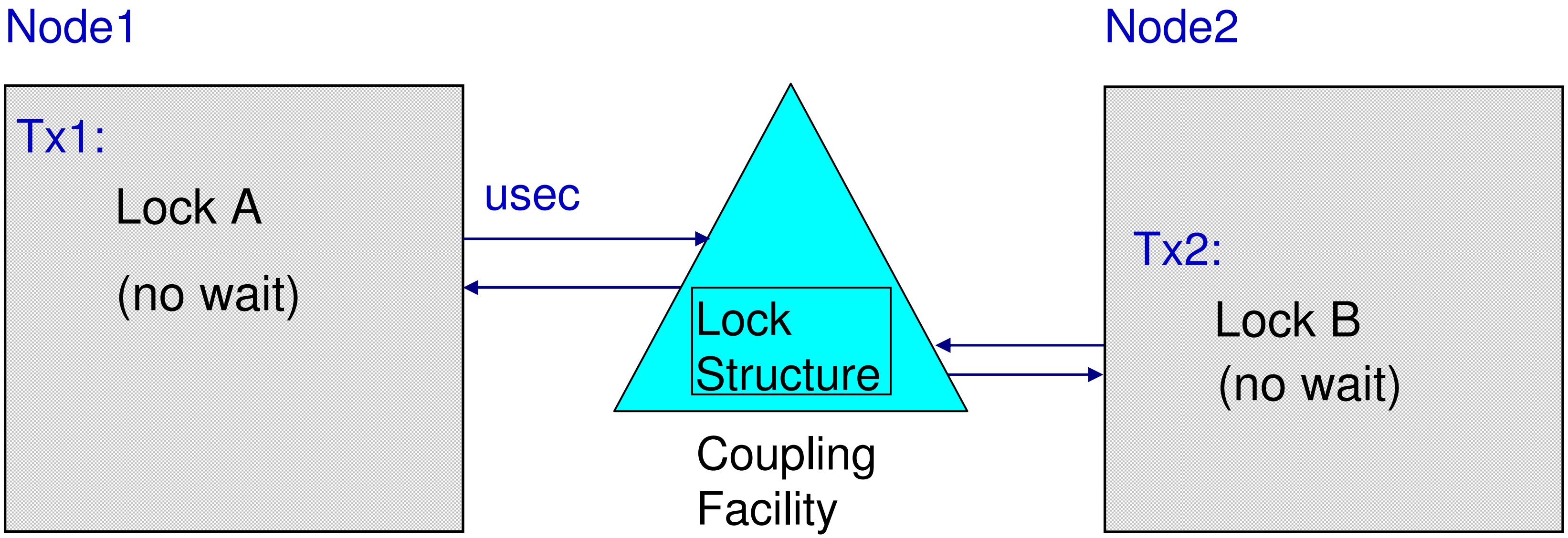
- Two factors to preserving data integrity in a data sharing environment
  - Inter-system ***concurrency*** control - global locking
    - Multiple readers OR
    - One writer
  - Inter-system buffer ***coherency*** control – managing changed data
    - When one system changes data rows that also reside 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

Data Sharing Performance Goals

- ► 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 nth nodes

Inter-system Concurrency Control

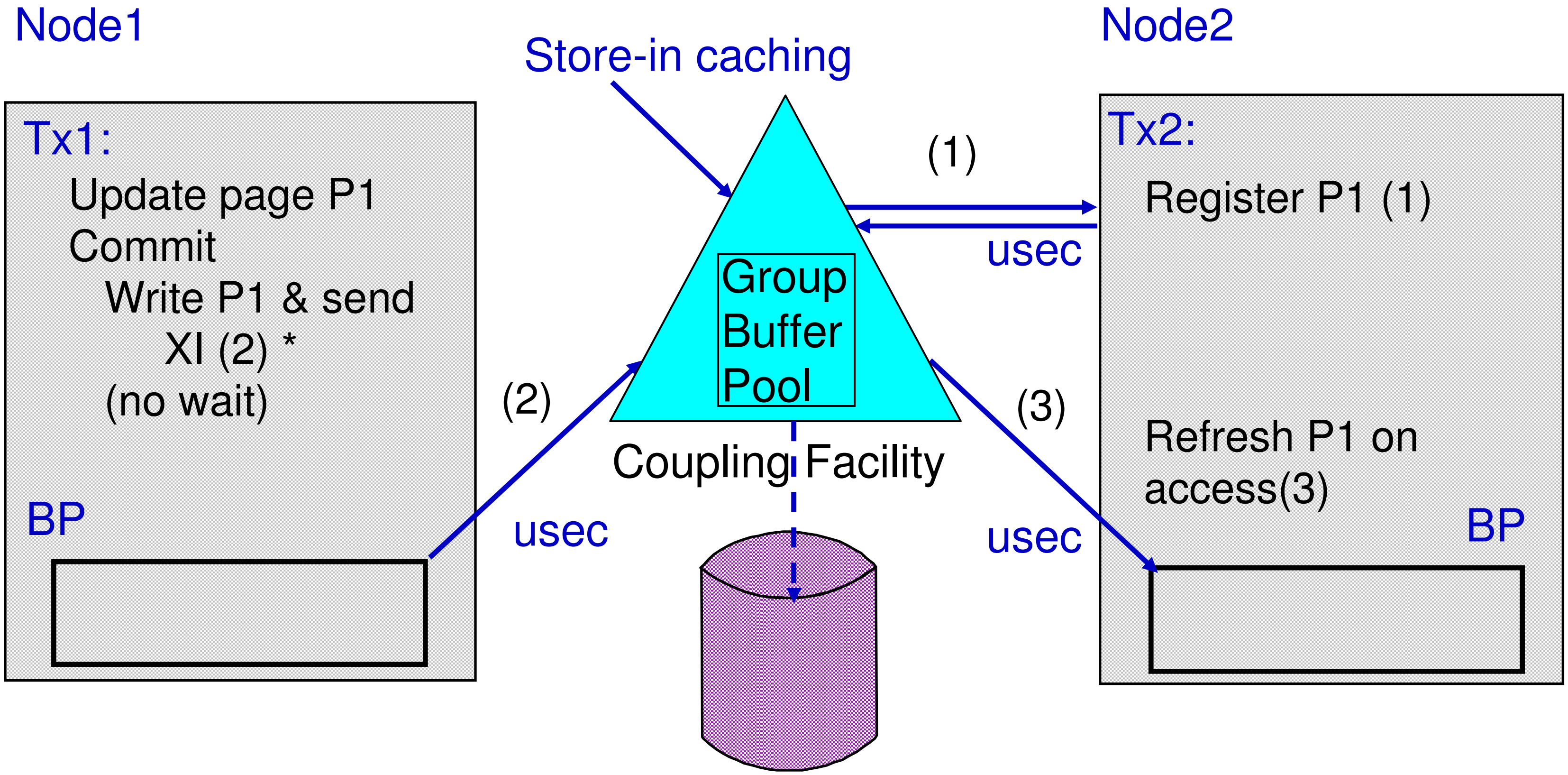
Global locking using
PSX Coupling Technology:



- ►Cost of obtaining lock does not increase when adding (3rd through nth) DBMS instances (scalability)

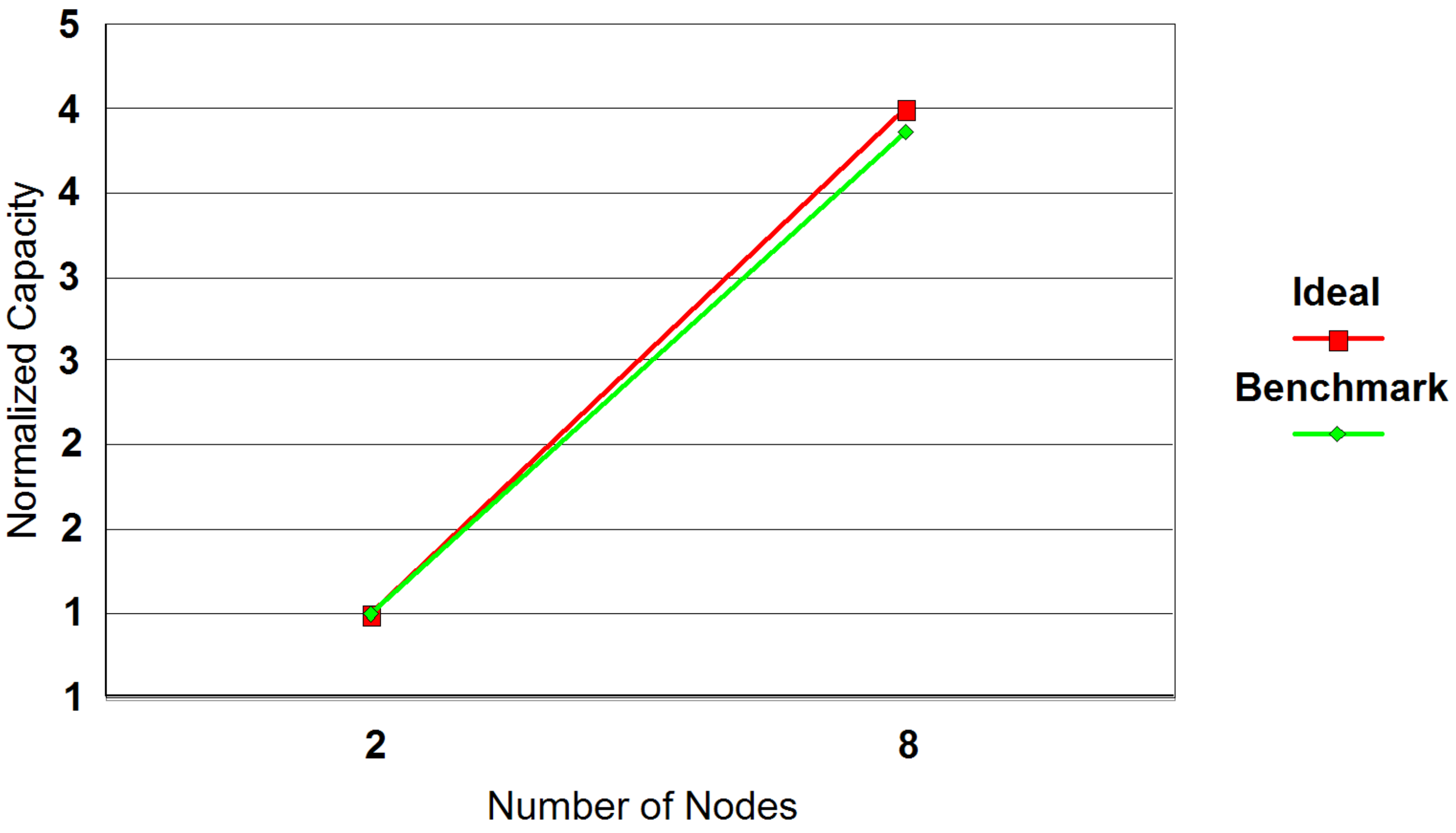
Inter-system Buffer Coherency Control

Managing changed data using
PSX Coupling Technology:



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

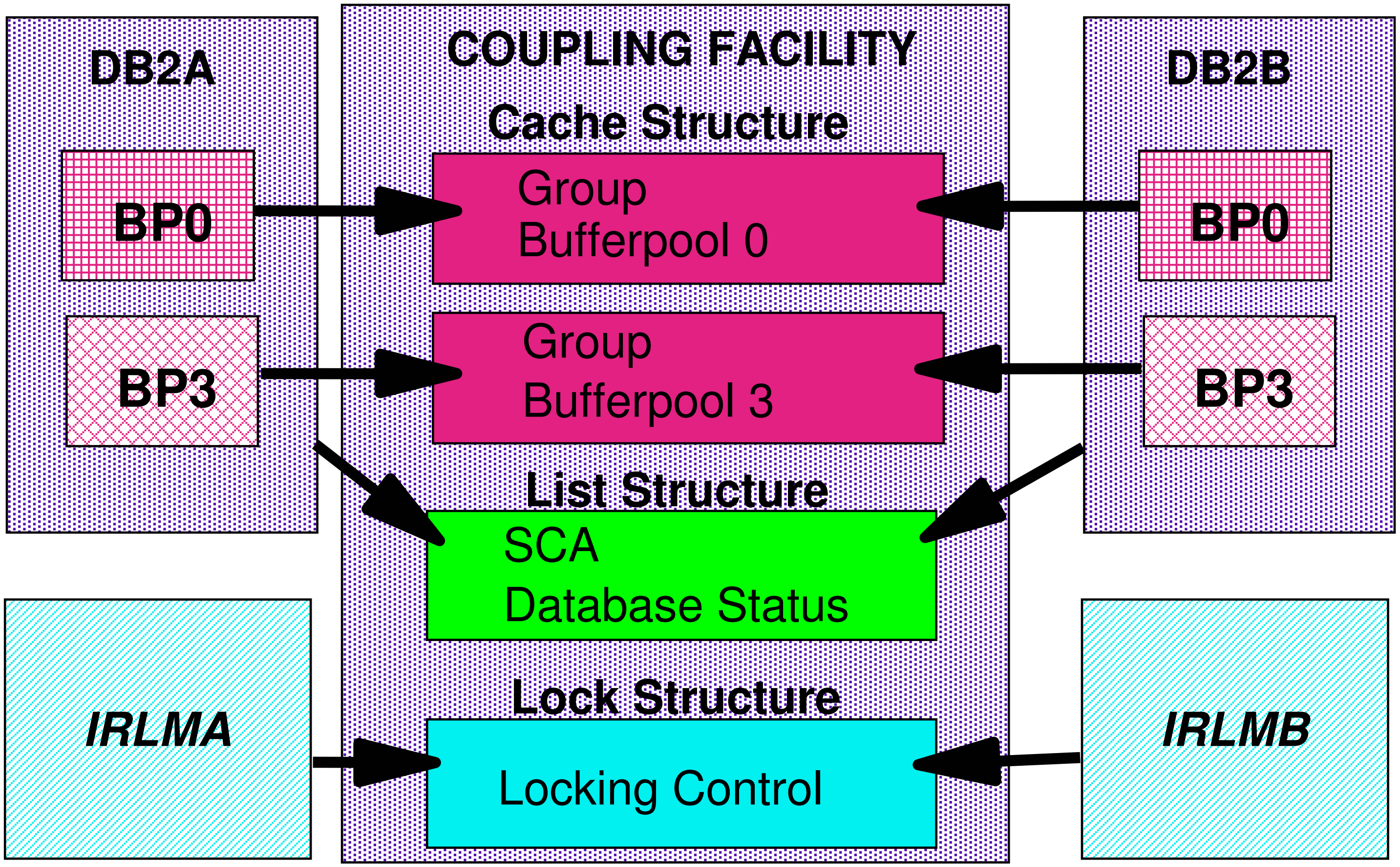
DB2 Data Sharing OLTP Scalability



►IMS/TM with DB2 V4 OLTP workload

►96.75% of ideal scalability from 2 to 8 nodes demonstrated

DB2 CF Structures



Shared Communications Area (SCA)

- ► Used by DB2 to maintain group-wide status information
  - Recovery Pending
  - Copy Pending
  - Write Error Ranges
  - Logical Page List
  - GRECP Status
  - BSDSs of all DB2s in data sharing group
  - System checkpoint intervals
  - Database Exception Table (DBET)
- ► SCA is generally not a performance concern

Lock Structure (LOCK1)

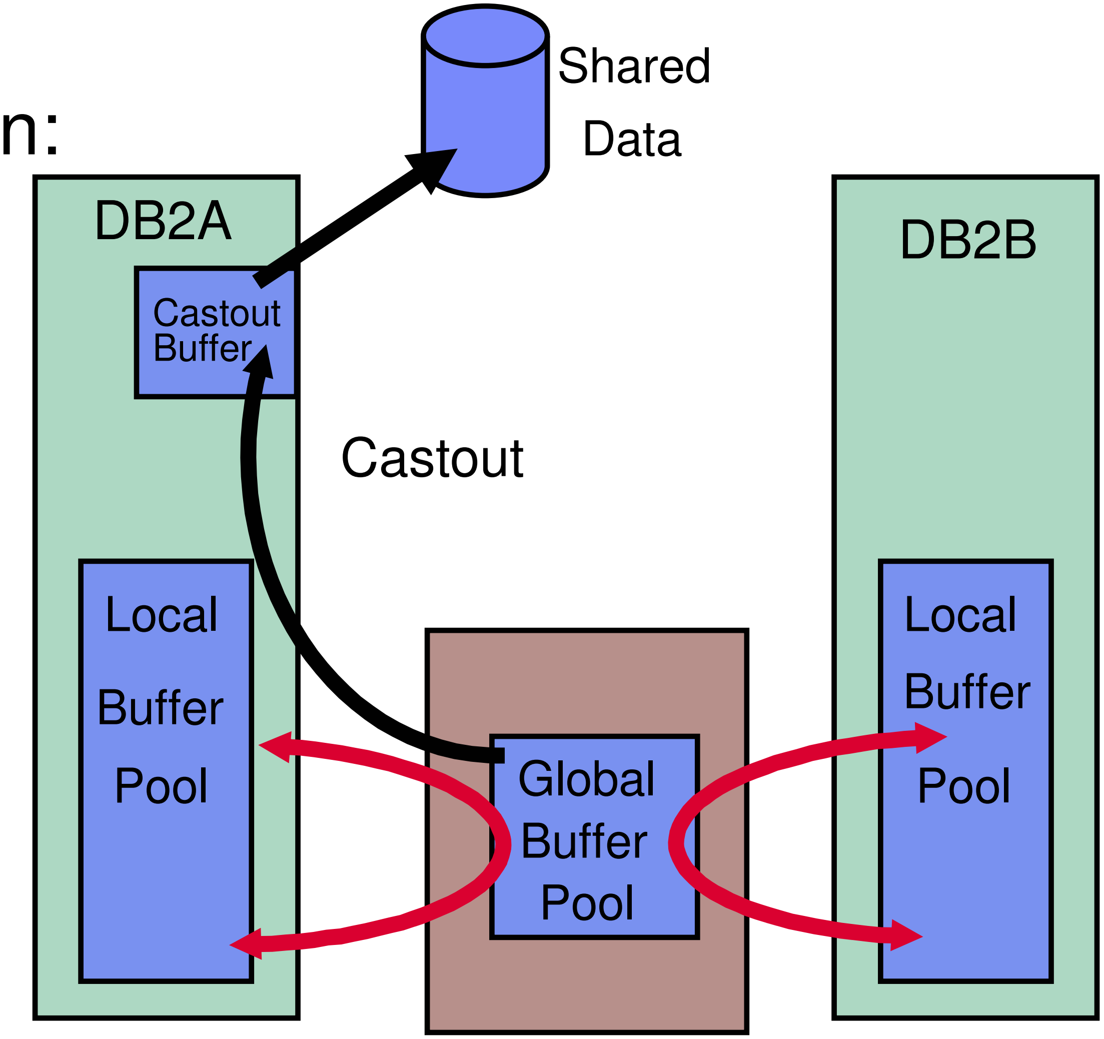
- Used by IRLM to manage global locking
- Holds L-locks and P-locks
  - L-locks to track concurrency
  - P-locks to track coherency
- Consists of a lock (or hash) table and a modify lock list
  - Lock table controls access to resources
  - Modify lock list contains retained locks in case of an IRLM or DB2 failure

Group Buffer Pools (GBPs)

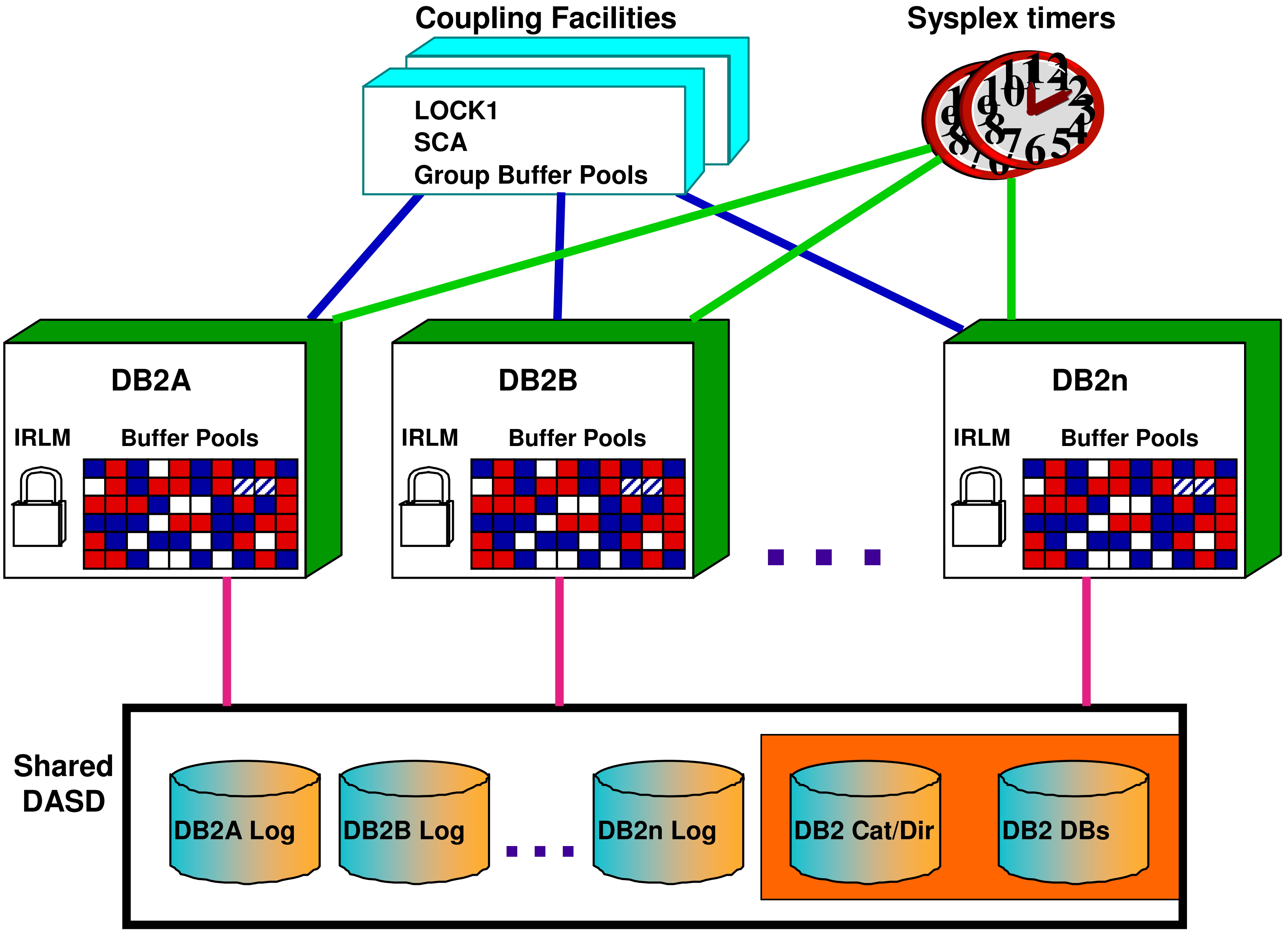
- DB2 uses GBPs to
  - Manage buffer coherency
  - Cache changed pages
  - Optionally cache read-only pages
- GBP consists of directory entries and data entries
  - Directory entries manage coherency by tracking interest in a data or index page by any DB2 member in the data sharing group
  - Data entries are the cached pages that a DB2 member changed

CASTOUT processing

- CASTOUT will occur when:
  - CLASST exceeded
  - GBPOOLT exceeded
  - 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



DB2 Parallel Architecture



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%
  - Individual jobs/transactions may have higher overhead
  - < 0.5% added cost per member past 2-way

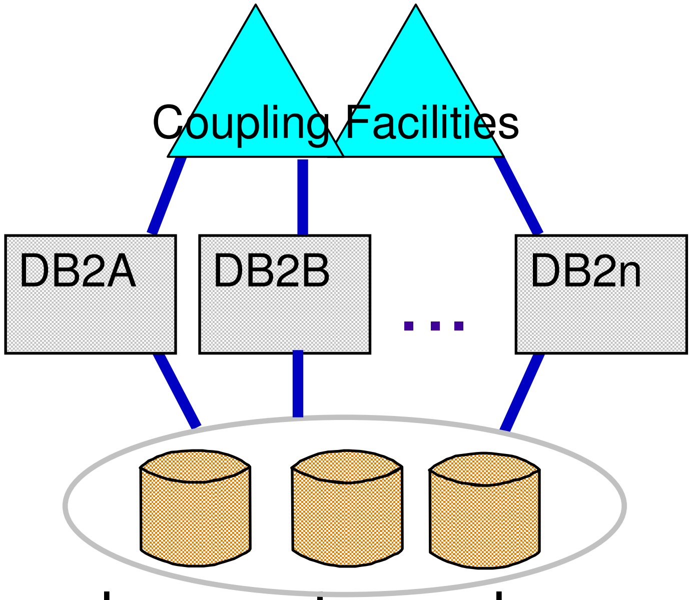
Data Sharing Performance in Production

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

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

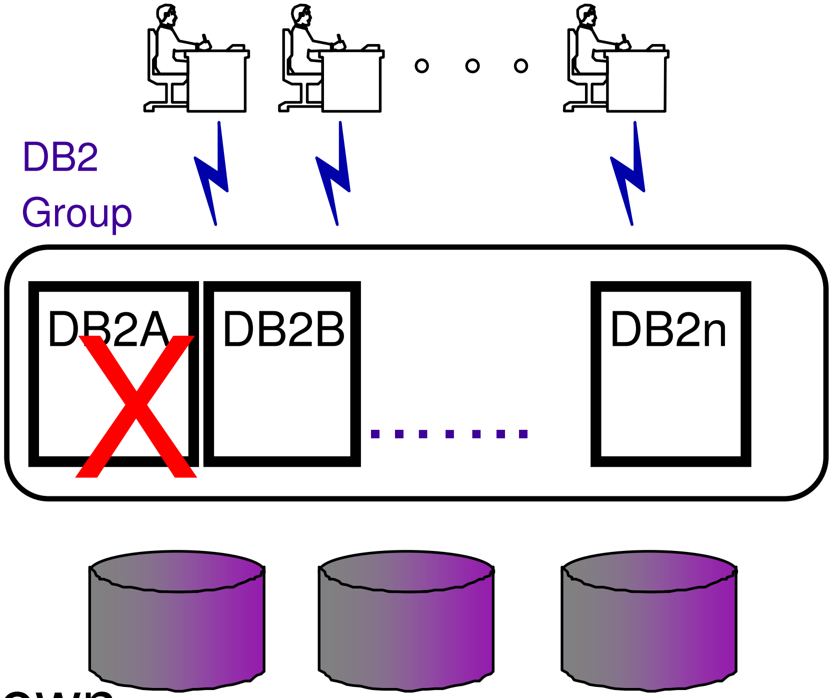
Note: “Mi” stands for ‘million instructions’

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

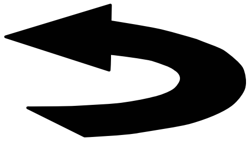
DB2 Member Outage – Planned



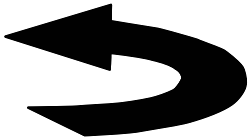
- "Rolling" maintenance
- One DB2 stopped at a time
- DB2 data continuously available via the N-1 members
- Other members temporarily pick up the work of the member that is down
- Batch work can be offloaded to another member with more available capacity to reduce the batch window
- Applies to hardware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

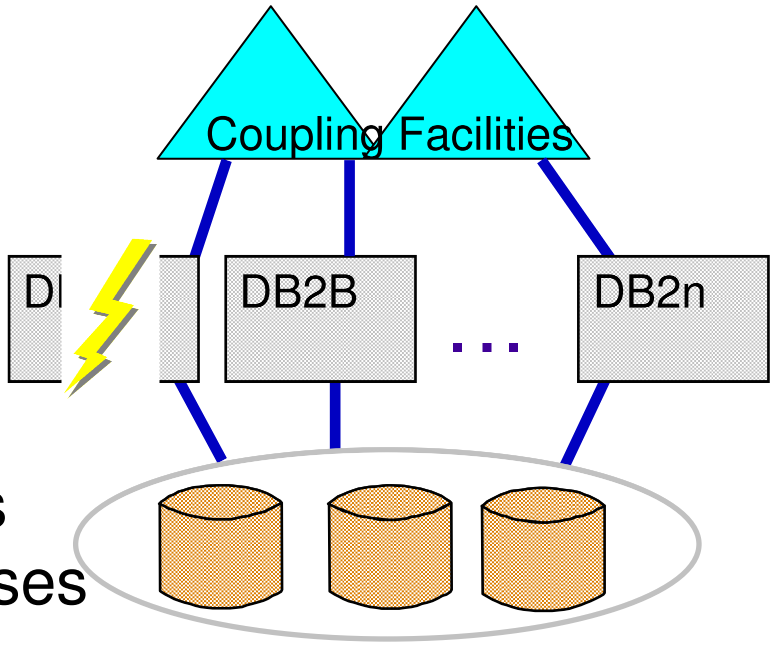


  - Put in new release
  - Restart each DB2 member



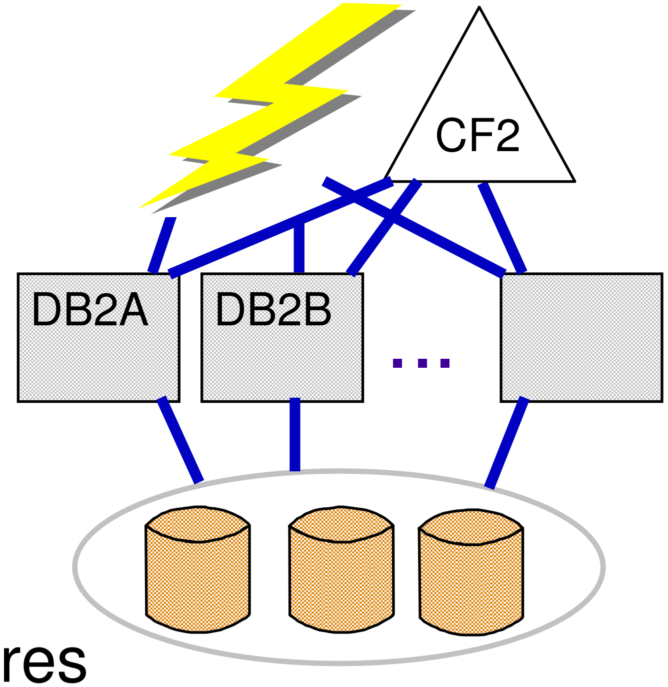
- ► Catalog migration done once per data sharing group

DB2 Member Outage – Unplanned



- The other "surviving" members remain up and running
- The architecture allows all members to access all portions of the data bases
- 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
- 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 V7 allows duplexing of Lock and SCA but duplexing not necessary for high availability

Duplexing – 2 Kinds

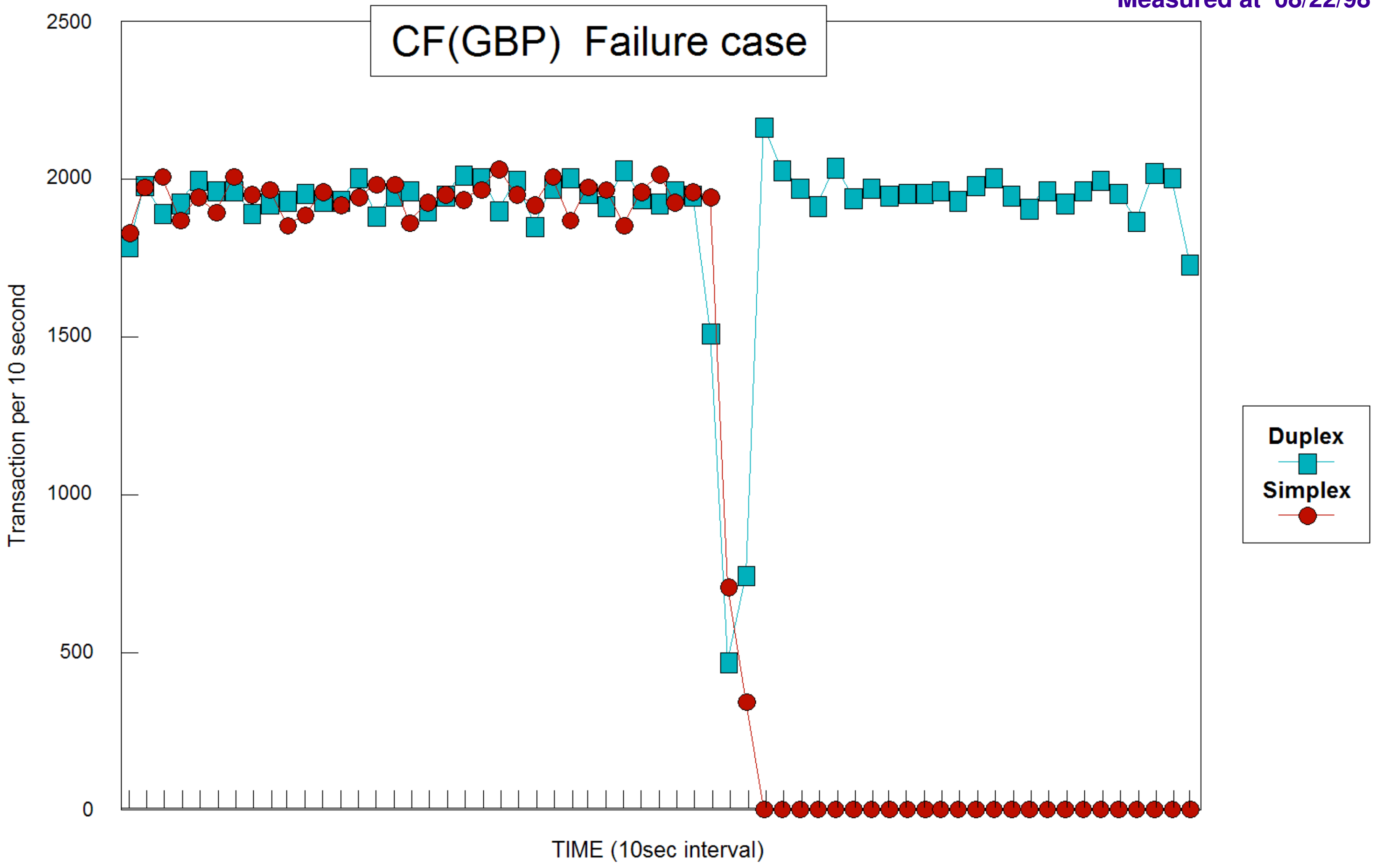
- “User-managed” duplexing
  - Applies to GBPs
  - “User” = DB2; DB2 is responsible for managing two structures in different CF LPARs
- “System-managed” duplexing
  - Applies to LOCK1 and SCA
  - XES (z/OS) is responsible for managing two structures in different CF LPARS
    - DB2 is not aware of second structure
  - Recommended if CFs run on Internal Coupling Facilities (ICFs) and co-reside on CEC with DB2 members
    - Avoid ‘double failure scenario’
    - But 4x cost for synchronous lock request

GBP Duplexing

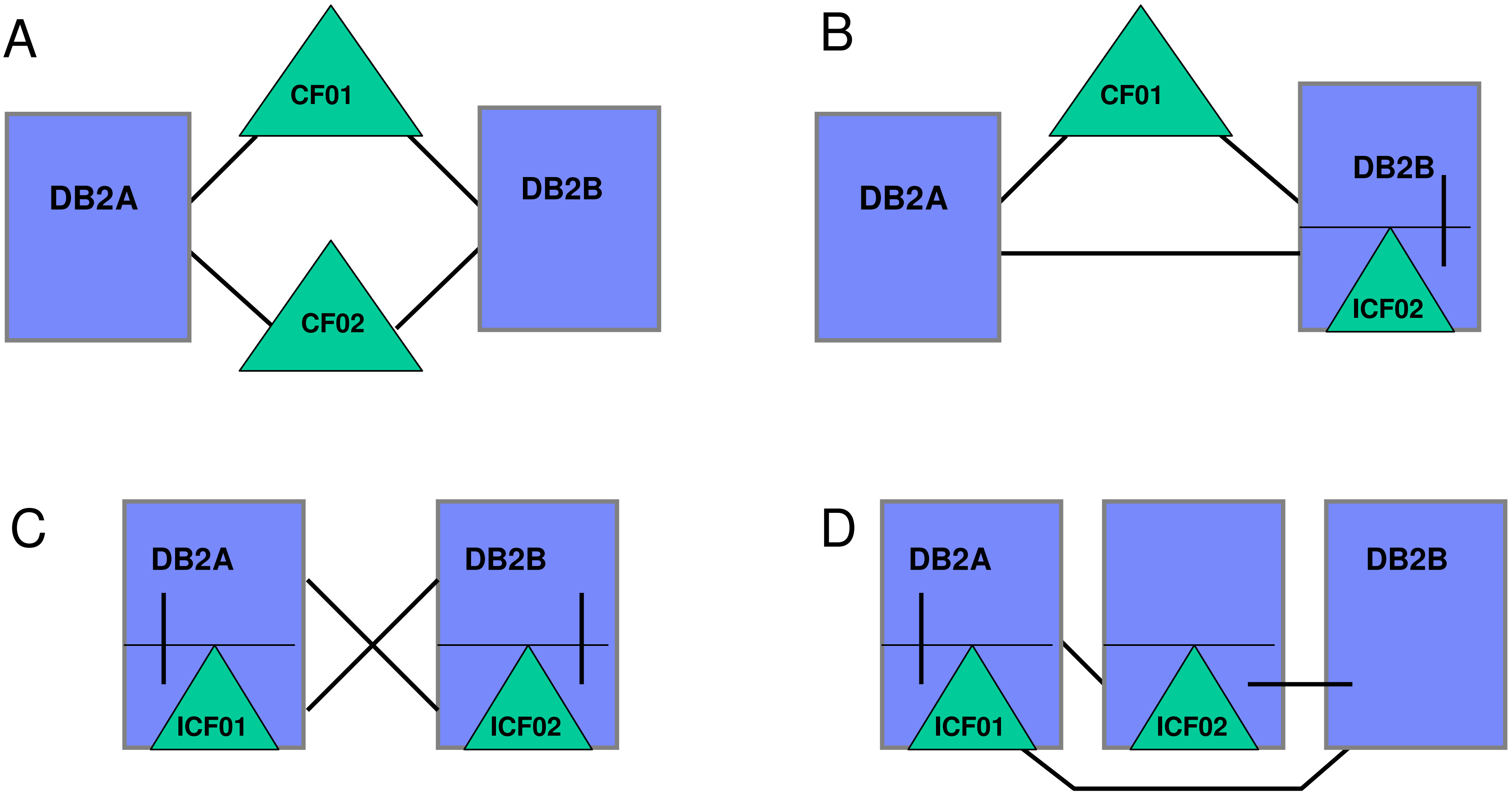
- Allocate secondary GBP on alternate CF
- Write changed pages to both primary and secondary
- If loss of connectivity or loss of structure -
  - Switch to secondary (seconds)
  - No rebuild required; changed pages already in GBP
  - Cross-invalidate buffers and gradually repopulate directory entries
- No application outage unless both primary and secondary GBPs are lost

V5 GBP Duplexing Recovery

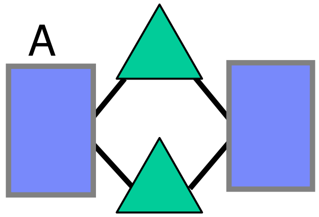
IRWW workload
Measured at 08/22/98



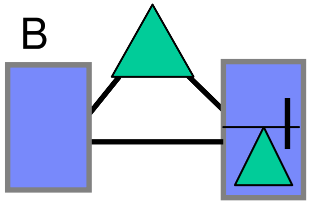
Parallel Sysplex Configurations



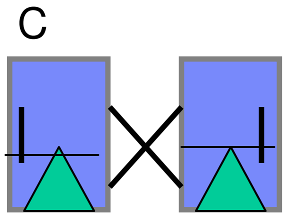
Parallel Sysplex Configurations



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



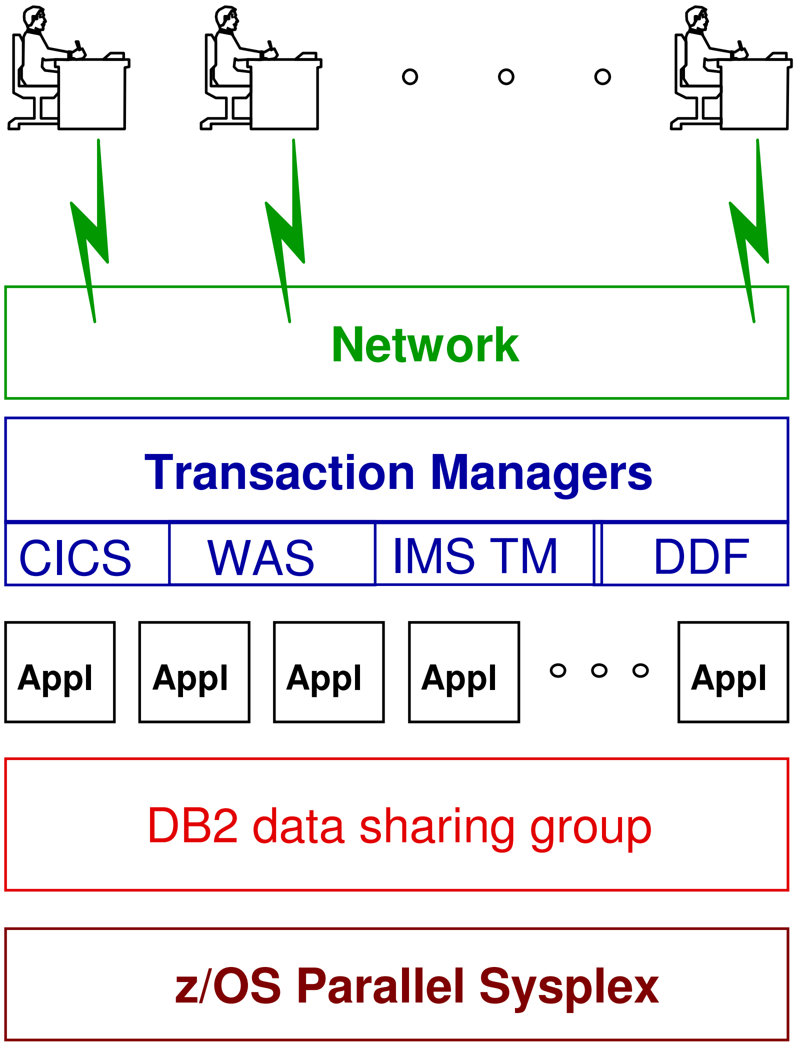
- ►B: One Integrated CF (ICF), one external CF
  - ▪ LOCK1 and SCA in external CF
  - ▪ Duplexed GBPs spread across both CFs
    - •Primary GBP in ICF has advantages for ‘local’ DB2



- ►C: Two ICF configuration
  - ▪ Lock1 and SCA duplexed; allocated in both CFs
    - •HW structure duplexing or system-managed duplexing
    - •Performance implication for LOCK1 requests
  - ▪ Duplexed GBPs spread across both CFs
    - •Primary GBP in ICF has advantages for ‘local’ DB2

Systems Management

- ► 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
    - V8 offers multiple location names
  - "Group scope" on display output
  - "Group scope" for online performance monitors
  - Log merging for replication (IFCID 306)



Network

Transaction Managers

CICS	WAS	IMS TM	DDF

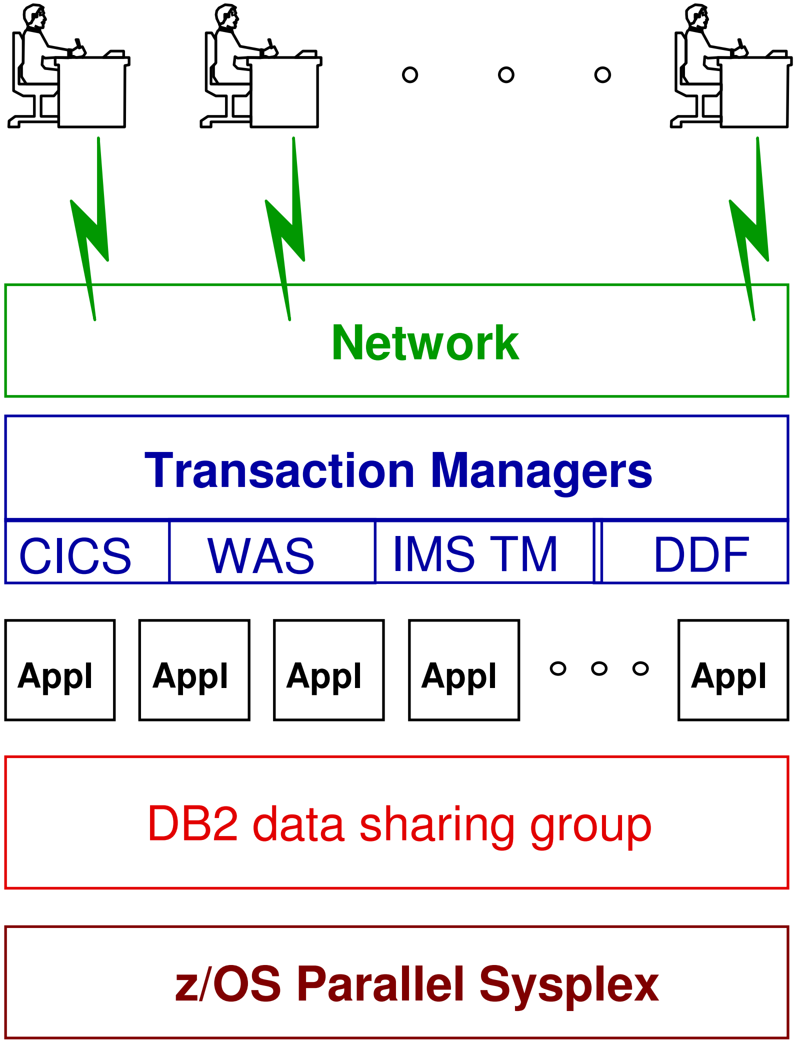
Appl	Appl	Appl	Appl	○ ○ ○	Appl

DB2 data sharing group

z/OS Parallel Sysplex

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 I/P Addressing (DVIPA)
- Sysplex Distributor



Network

Transaction Managers

CICS | WAS | IMS TM | DDF

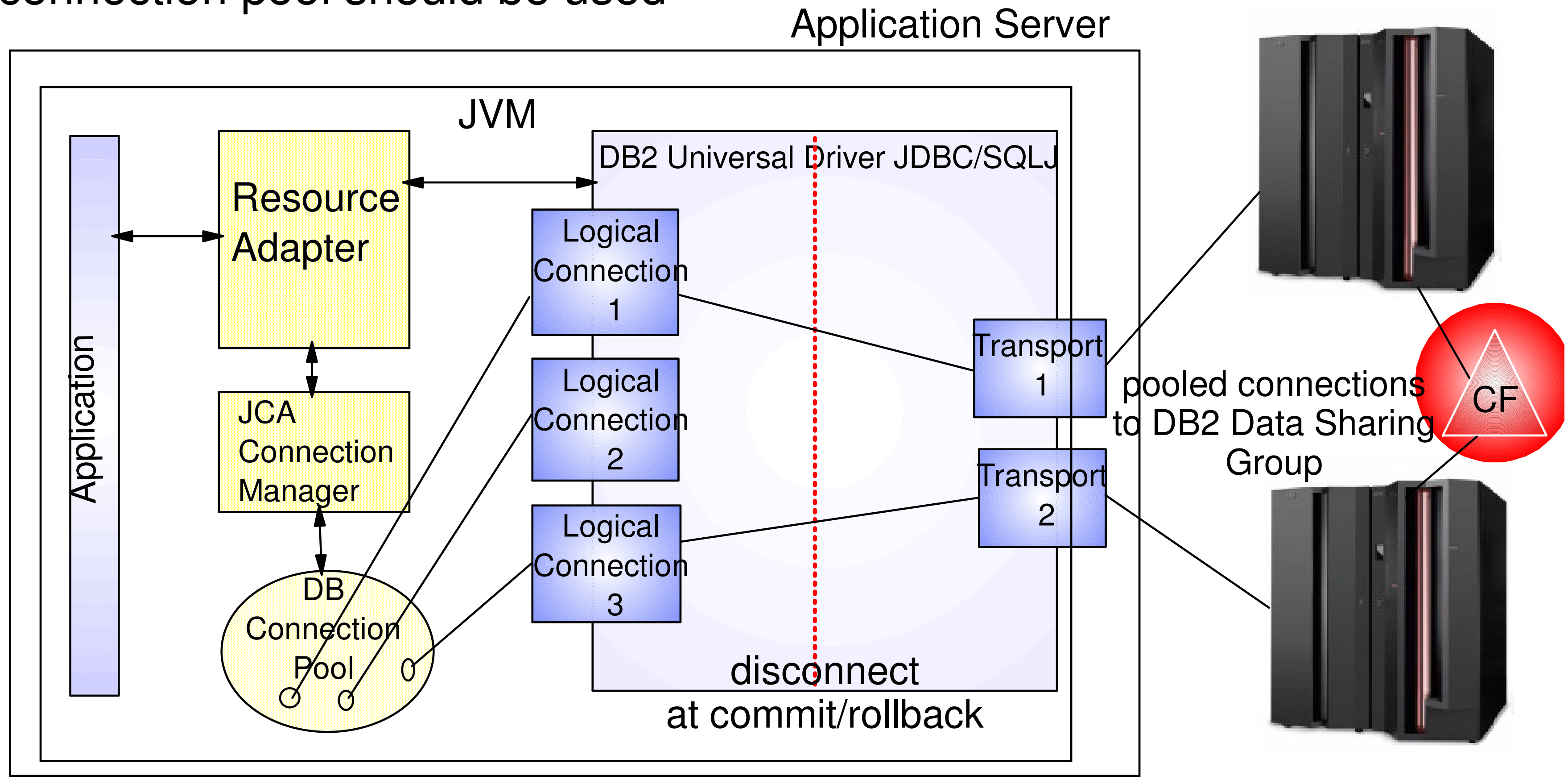
Appl Appl Appl Appl ○ ○ ○ Appl

DB2 data sharing group

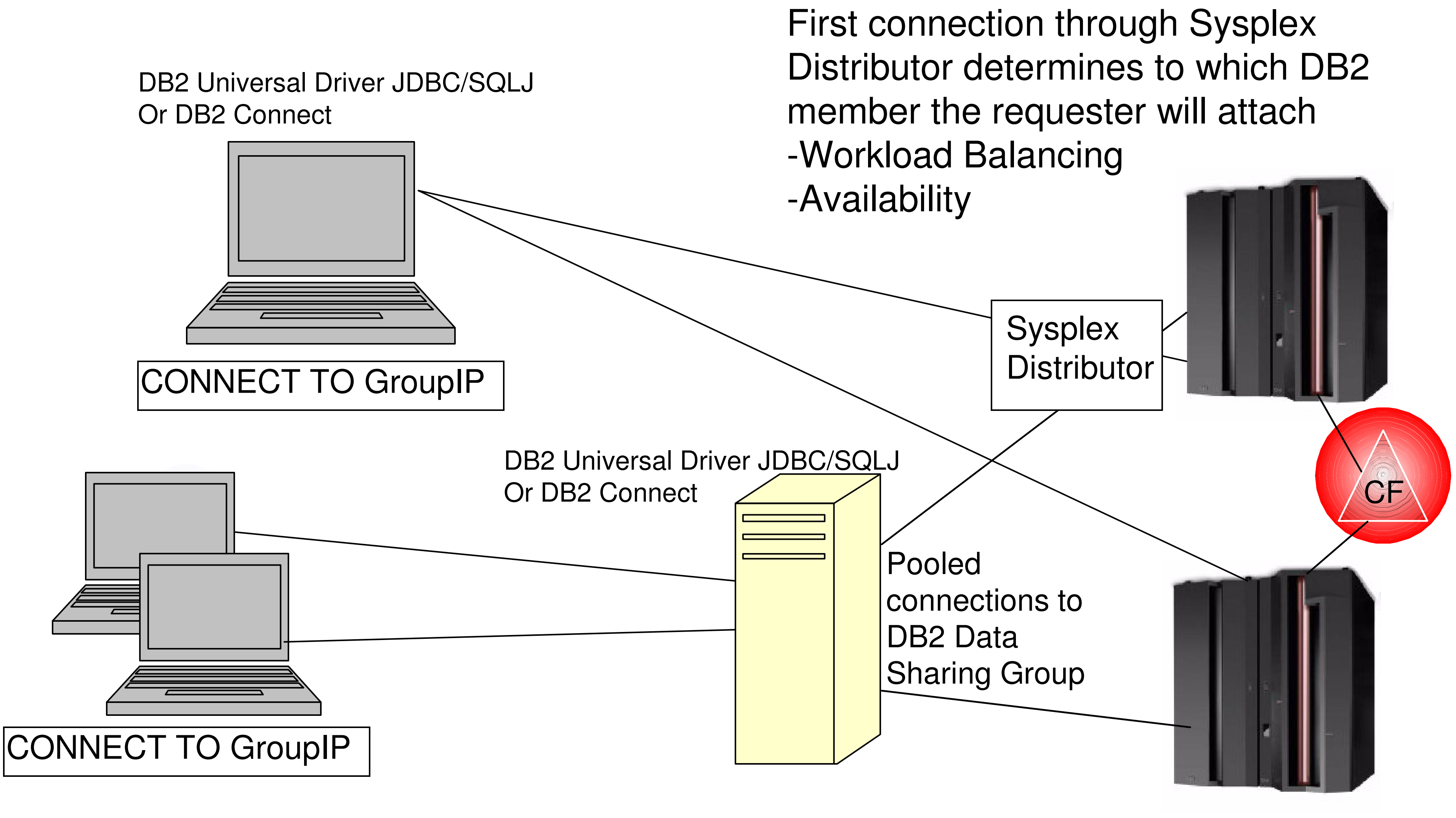
z/OS Parallel Sysplex

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
Or DB2 Connect

CONNECT TO GroupIP

Sysplex Distributor

DB2 Universal Driver JDBC/SQLJ
Or DB2 Connect

Pooled connections to DB2 Data Sharing Group

CONNECT TO GroupIP

Common Questions

- ► Overhead and capacity
- ► Application changes
- ► Parallelism
- ► Operational data vs. informational data
- ► Training

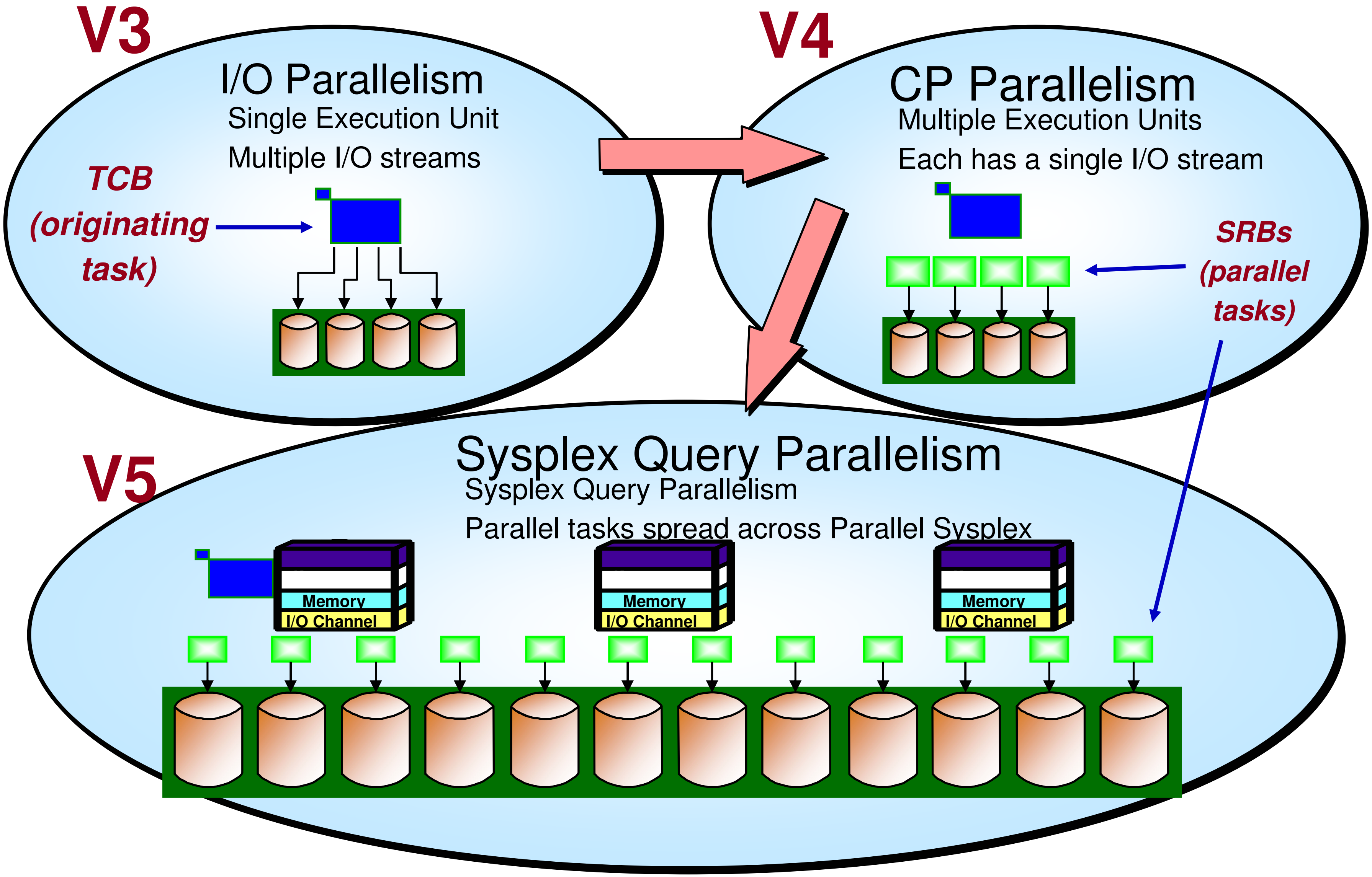
Overhead and capacity

- ► "I've heard as much as 50% overhead"
  - ▪ Not in well-defined PSX and DB2 data sharing environment
- ► "What can I expect?"
  - ▪ "It depends…"
- ► "How should I size my CF structures?"
  - ▪ CFSizer
    - http://www-03.ibm.com/servers/eserver/zseries/cfsizer/

Application changes

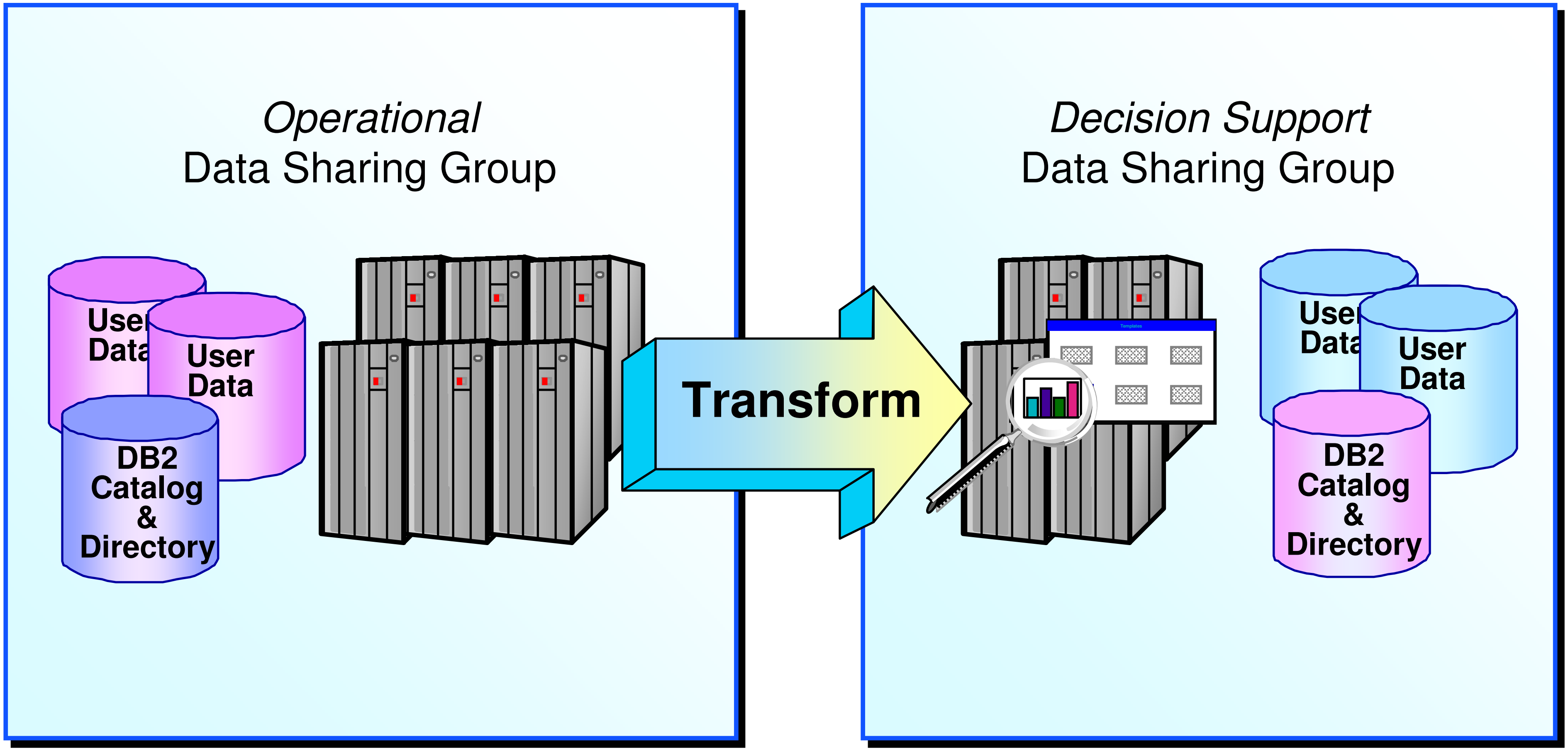
- ► “Don’t my applications have to change?”
  - SQL interface does not change
  - However, locking and commit frequency may impact data sharing performance
    - Commit frequently – long-time recommendation
    - Take advantage of lock avoidance
      - ISO(CS) or ISO(UR)
      - CURRENTDATA NO
  - New messages and return codes
  - Applications must be able to run on more than one DB2 member for high availability

Does data sharing allow parallelism?



Operational data vs. informational data

- ►A typical configuration for decision support:



Operational
Data Sharing Group

User Data

User Data

DB2 Catalog & Directory

Transform

Decision Support
Data Sharing Group

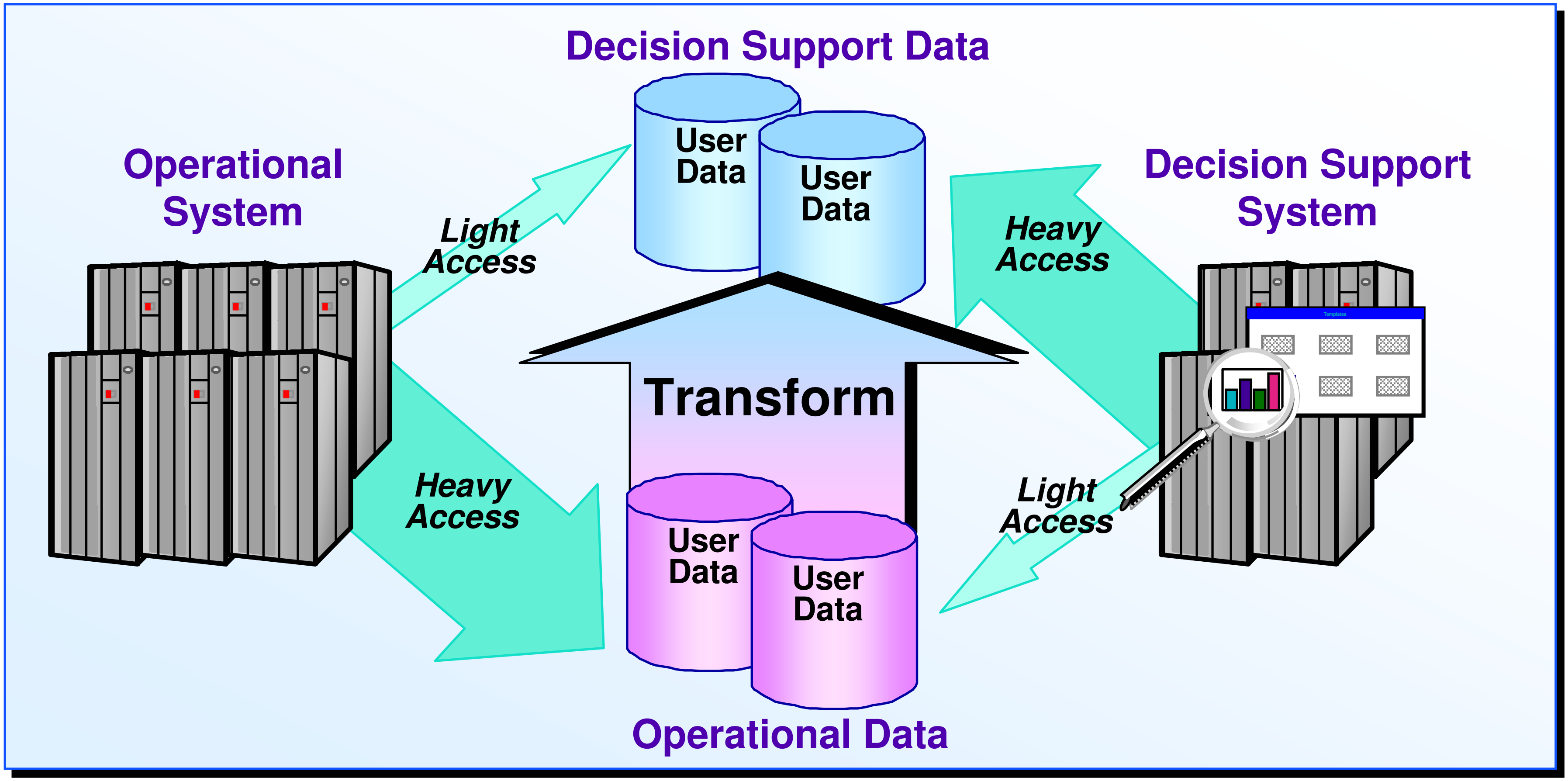
User Data

User Data

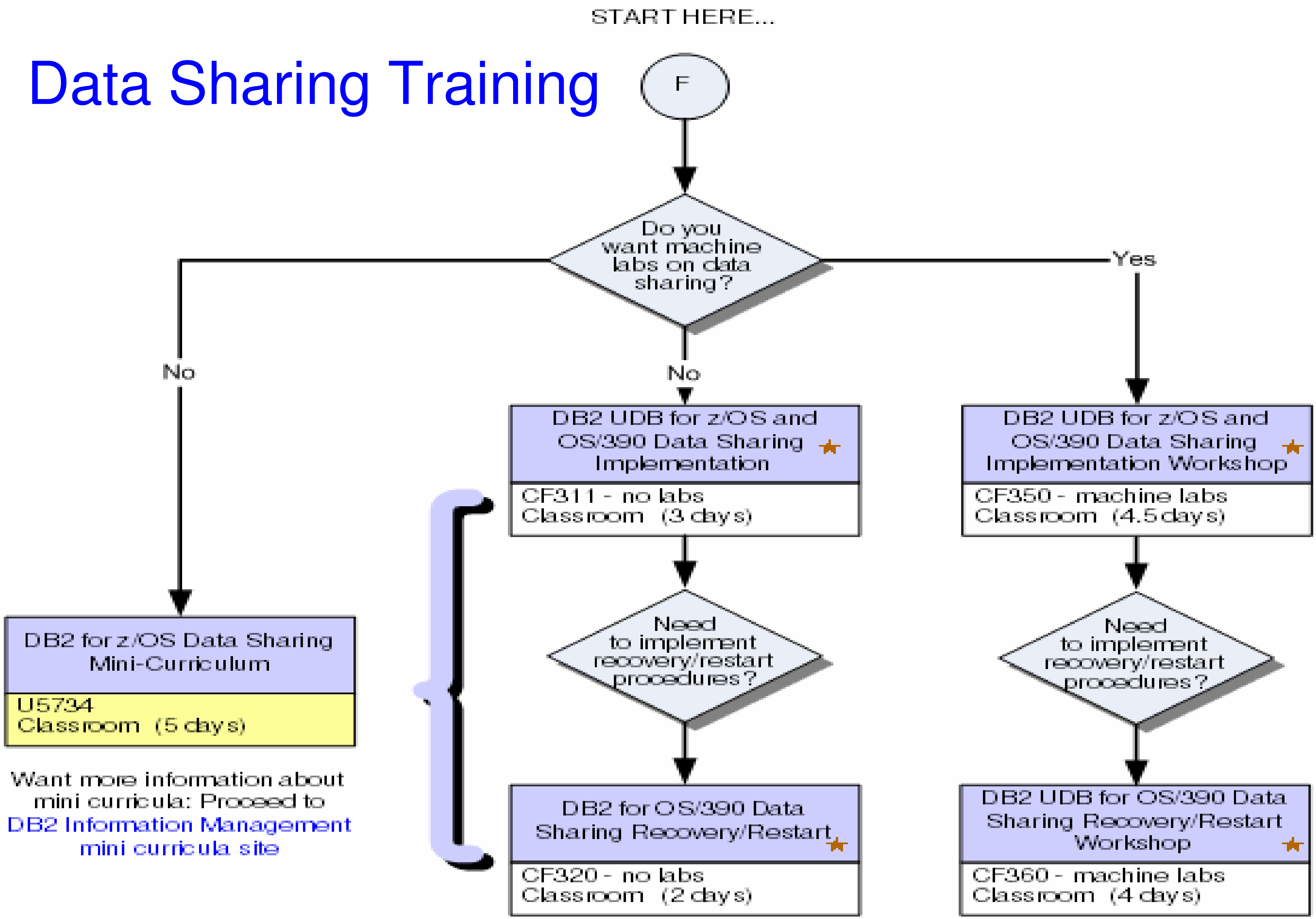
DB2 Catalog & Directory

Operational vs. informational…

- ►Another possible configuration can consist of one data sharing group with both operational and decision support work. This configuration has the advantage of having one set of data and the ability to query the operational data or even join it with the informational data.



Data Sharing Training



START HERE...

F

Do you want machine labs on data sharing?

No

No

Yes

DB2 for z/OS Data Sharing Mini-Curriculum

U5734
Classroom (5 days)

Want more information about mini curricula: Proceed to DB2 Information Management mini curricula site

DB2 UDB for z/OS and OS/390 Data Sharing Implementation ★

CF311 - no labs
Classroom (3 days)

Need to implement recovery/restart procedures?

DB2 for OS/390 Data Sharing Recovery/Restart ★

CF320 - no labs
Classroom (2 days)

DB2 UDB for z/OS and OS/390 Data Sharing Implementation Workshop ★

CF350 - machine labs
Classroom (4.5 days)

Need to implement recovery/restart procedures?

DB2 UDB for OS/390 Data Sharing Recovery/Restart Workshop ★

CF360 - machine labs
Classroom (4 days)

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

Bibliography: DB2 for z/OS & OS/390

- *Data Sharing: Planning and Administration*
  - DB2 UDB for OS/390 and z/OS V7: SC26-9935-05
  - DB2 UDB for z/OS V8: SC18-7417-03   (2/2006)
- Redbooks
  - *Distributed Functions of DB2 for z/OS and OS/390*: SC18-7417-02
  - *Achieving the Highest Levels of Parallel Sysplex Availability*: SG24-6061
  - *Parallel Sysplex Application Considerations*: SG24-6523
  - *TCP/IP in a Parallel Sysplex*: SG24-5235-02
  - Coming soon: new DB2 Data Sharing redbook

Questions
&
Answers