DB2 on S/390 Availability Features and Challenges

Jeff Josten
DB2 Development, IBM Silicon Valley Lab
e-mail: josten@us.ibm.com

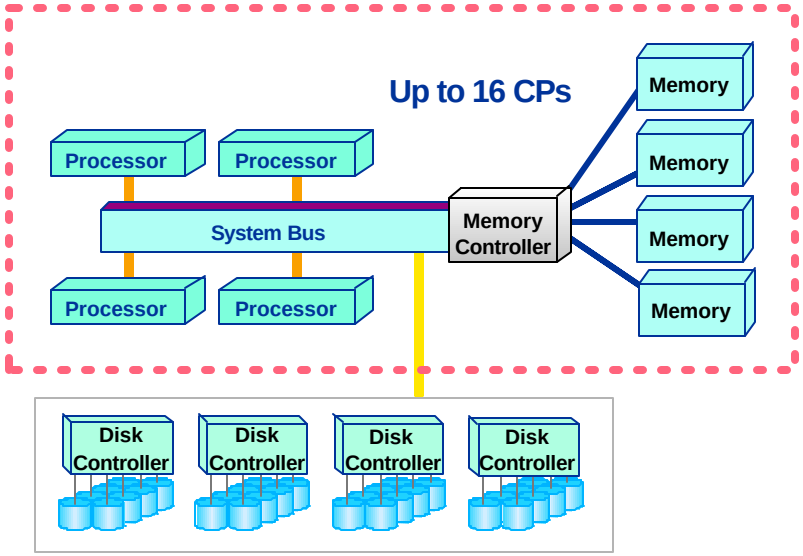
Oct. 15, 2001

Agenda

- System/390 and DB2/390 architectural overview
- Some unique things about System/390 design for availability
- Some key design elements for high availability
- How 390 handles various failure scenarios

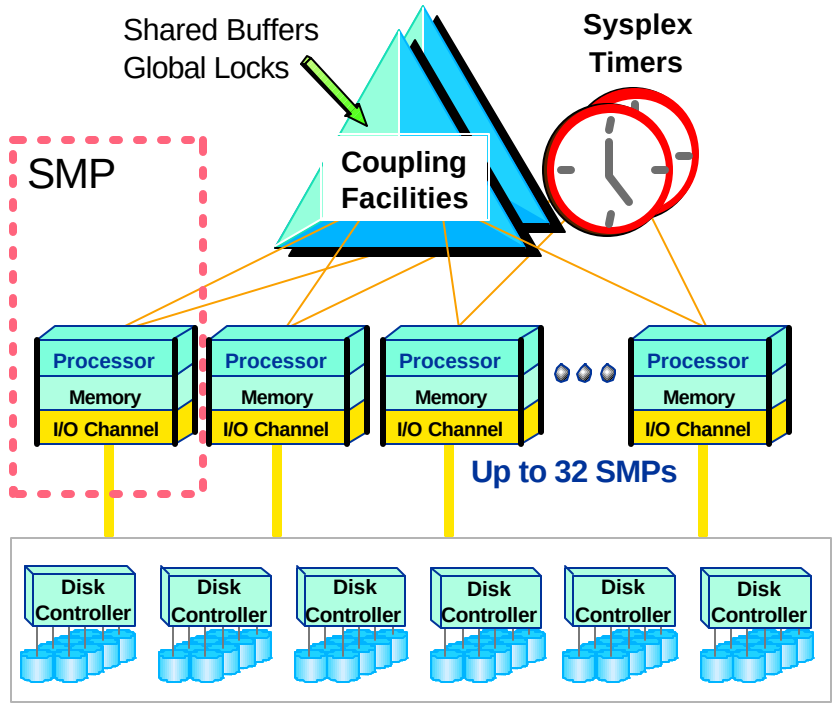
S/390 Architectural Overview

Symmetric Multiprocessor (SMP)



- ✓ *Ease of management*
- ✓ *"Shared memory"*
- ✓ *Single OS image or multiple OS partitions*

Parallel Sysplex Cluster (Data Sharing)



- ✓ *Proven, scaleable and ease of management*
- ✓ *"Shared data"*

DB2 Shared Data Clusters Advantages

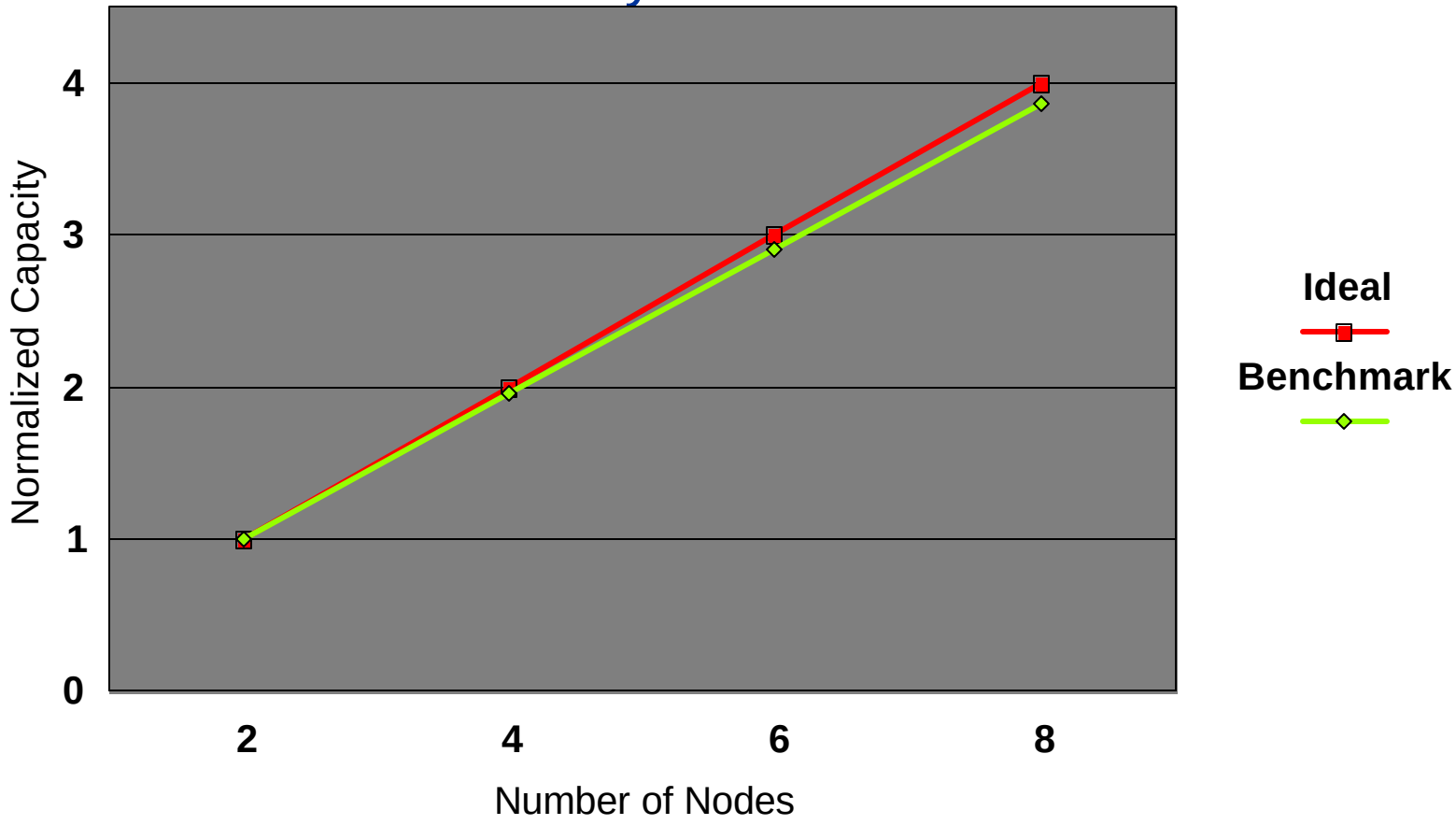
- 390 Parallel Sysplex offers major advantages for customers:
  - no need to balance data across data partitions
  - no need to rebalance partitions across servers
  - no outages to apply maintenance
  - system remains up and running on install of new DB2 release
  - dynamic workload balancing across the members of a parallel sysplex for both CPU and I/O processing
  - superior solutions for disaster recovery (hardware assisted remote site copy)
  - no outages required to modify stored procedures, application packages/plans, or most system parameters
  - highly scalable locking/data currency model
  - single system image for easy operations and application development
- Design point was OLTP and mixed workloads
  - "HPTS" stood for "Highly Parallel Transaction System"

Scalability: some facts and figures

- Up to 32 systems in a single cluster
- Up to 150K threads per system = 4.8M threads per cluster
- Max size of a single table = 16TB
- Max number of rows in a single table = 1 trillion (2**40)
- Max buffer pool size = 256GB
- Some sample DB2 customer stats:
  - 1.4B SQL stmts per day (retail, 4-way data sharing)
  - 24,000 inserts/sec. 150GB log/day (brokerage, 6-way)
    - 40,000 insert/sec. achieved on G4 ('97 vintage) processor
  - 167M inserts/day (UPS, single DB2)
  - 1,000 banking transactions per sec. (4-way data sharing)
  - 5800 stored procedures per sec. (brokerage)
  - Largest data sharing group = 14-way (brokerage)
  - Largest number of DB2 subsystems = 150 (bank)
  - 45 TB mirrored disk (bank)
  - >10 TB warehouses (banking, govt)

DB2 Data Sharing Scalability for OLTP

- 5-15% overhead typically observed for 1-way to 2-way
  - 8-10 CF acceess per million instr. is typical at customers
- <.5% additional overhead for each member past 2-way
- IMS/TM with DB2 V4 OLTP workload, 100% data sharing
- 96.75% of ideal scalability from 2 to 8 nodes demonstrated



What is a Coupling Facility?

- "Intelligent" shared electronic storage
- Three structure types supported
  - List: shared queues, messaging
  - Lock: inter-system concurrency control
  - Cache: inter-system buffer coherency control; host-side hardware bit vector for buffer invalidation
- DB2 uses all three structure types
- External or internal, single CP or multiple CPs
- Non volatility provided by battery backup
- Connectivity (can mix and match):
  - ISC (InterSystem Coupling) links: up to 200 MB/sec, 40 km
  - ICB (Integrated Cluster Bus): up to 1GB/sec., 10 m
  - IC (Internal Coupling channel): up to 1.25 GB/sec., "linkless"
- Typical observed command latencies: 10-50 usec.

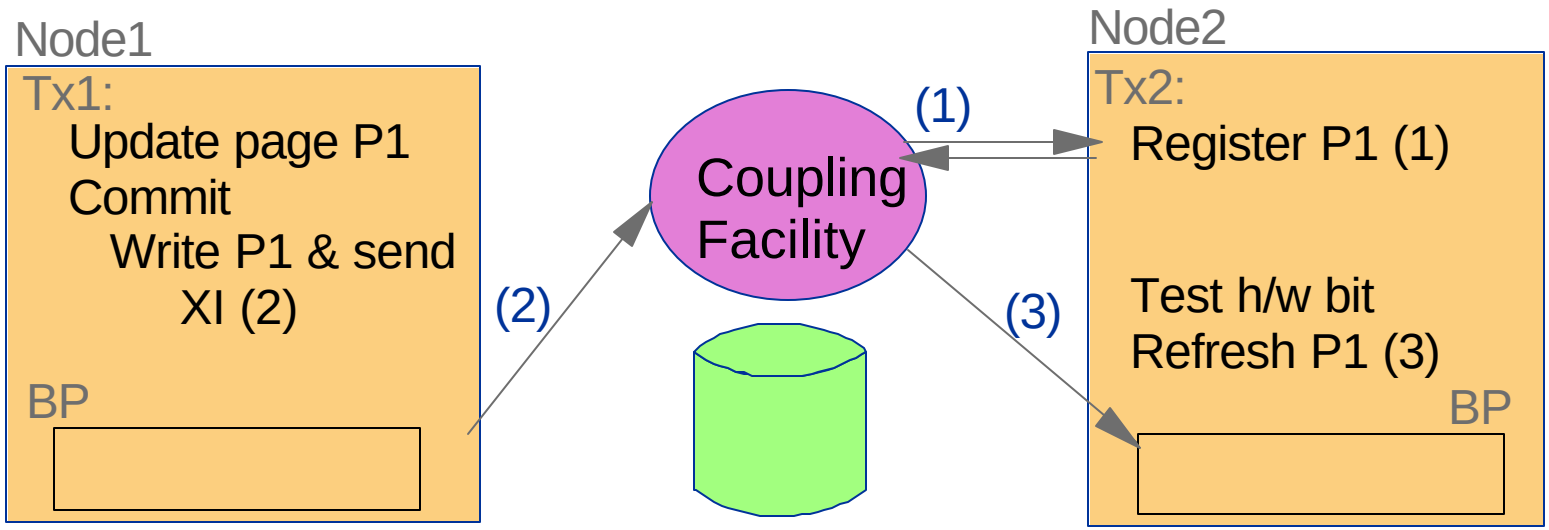
DB2's Use of the Coupling Facility

- CF Lock Structure
  - Fast inter-system transaction lock conflict detection
  - Fast inter-system page latch conflict detection for cases where sub-page concurrency is allowed (e.g. row level locking)
  - Inter-system read/write interest tracking for DB objects
  - Retained locks in cases of DB2 "instance" failure
- CF Cache Structure
  - Store-in cache by default; "no-data" option provided
  - Register buffers for cross-invalidate (XI)
    - "List" option provided for prefetch
  - Write changed buffers, send XI signals (bundled)
  - H/W instruction to test vector bits for buffer XI
  - Fast refresh of XI'ed buffers
  - Force-at-commit DB write protocol used for writes to CF
- CF List structure: shared status information

DB2's Use of the Coupling Facility...

- CF technology allows DB2 to do shared disk OLTP clusters avoiding synch. disk I/O and software messaging overheads
- Update example, transaction flow:
  1. Lock page X (0 s/w messages)
  2. Access page in BP (0 s/w messages, no add'l I/O)
     - If BP hit, test h/w bit and refresh page from CF if necessary
     - If BP miss, "read/register" page to CF
       - If CF hit, then pg returned from CF & disk I/O is avoided
       - If CF miss, then read pg from disk as before
  3. Update page in BP
  4. Commit (0 s/w messages, no I/O)
     - Write page to CF. XI signals are sent.
     - Unlock page

Buffer XI, Refresh Example



Key points:

- CF interactions are CPU synchronous
  - Avoid process switching overhead and CPU cache disruptions
- XI signals do not cause processor interrupts
- Force-at-commit fast page cleaning enables fast crash recovery

Implications of CF as a store-in cache

- Done for performance reasons. Avoid disk I/O under the Tx.
- Pages must be eventually be written to disk. Done asynch. via the "castout" process
  - CF provides feedback on every write request as to the number of changed pages that reside in the structure
  - When threshold is reached, DB2 triggers castout
  - Castout runs in the background
  - Reads changed pages from CF into "private" buffers, writes pages to disk
  - When write I/O completes, DB2 tells the CF.
  - These pages remain cached in the CF as "clean". Clean pages are subject to being stolen (via LRU)
  - Castout is non-blocking. i.e. new version of page can be written while castout is in progress
- If CF fails, pages must be recovered
  - Duplexing (mirroring) avoids recovery outages (later chart)

Other Important Optimizations

- Sysplex Timer provides efficient means of log record sequencing across systems
- Explicit Hierarchical Locking: transaction locks arranged in parent/child relationships
  - "Tablespace" is the parent; "page"/"row" are the children
  - Local lock managers dynamically track inter-system R/W interest on the parent, using the CF Lock Structure
  - If no inter-system R/W interest on parent, then child locks do not need to be sent to the CF
  - When inter-system R/W interest occurs, children must be "propagated" at that time.
  - R/R, R/W, W/W optimizations
- Buffer Mgr. dynamic tracking of inter-system interest
  - Local buffer managers dynamically track inter-system R/W interest at the "DB object" level using the CF Lock structure
  - If no inter-sys R/W interest, then CF not needed for cache coherency

The 24x7 Holy Grail

- Given: system h/w and s/w components will incur outages from time to time
  - Planned and unplanned outages
  - Processor, OS, DBMS, I/O devices, I/O paths, site disaster, ...
- Goal: end users / apps should perceive that the database is available for read/write 100% of the time
- Design Tenets:
  - Reliability: components should not fail often
  - Redundancy with failure isolation:
    - at least 2 of everything
    - isolate failure to lowest granularity - no sympathy sickness
  - Fast, automatic, non-disruptive recovery
  - Online maintenance and configuration changes
  - High concurrency locking/latching
- DB2/390 leverages Parallel Sysplex clusters for high availability

Some S/390 Availability Features

- Reliability: design point is 40+ years MTBF
- Redundant components (power supplies, AC power input, internal batteries, cooling units, etc.)
- Self-healing
  - Instruction retry: becoming more critical as densities increase
  - Dynamic processor/memory chip sparing
- Concurrent Upgrade
  - Processors
  - Memory, within card boundary
  - I/O, CF, and network adapters
- Concurrent Maintenance
  - uCode
  - Power/cooling
  - I/O, CF, and network adapters

IBM zSeries z900

z900 Today

- 64-bit z/Architecture
- 25 General Purpose Models
- Increased Uni-Processor Performance
- 16-Way (20 PUs)
- Up to 64 GB Memory
- Enhanced Parallel Sysplex Connectivity
- Intelligent Resource Director (IRD)
- Over 1000 units shipped

z900 Newly Available

- HiperSockets ("linkless TCP/IP") - esp. interesting with Linux/390
- IRD for Linux partitions
- FICON enhancements
- zSeries file system, new C++ compiler
- Capacity Upgrade on Demand for memory
- Capacity Backup nondisruptive return
- System managed CF structure mirroring

z900 Capacity Backup Upgrade

Who Needs It?

- Customers who have a requirement for Disaster Recovery

What Is It?

- Nondisruptive addition of one or more engines
- Contract between IBM and customer
- Needs "spare" PUs
- Must plan ahead for memory and connectivity requirements

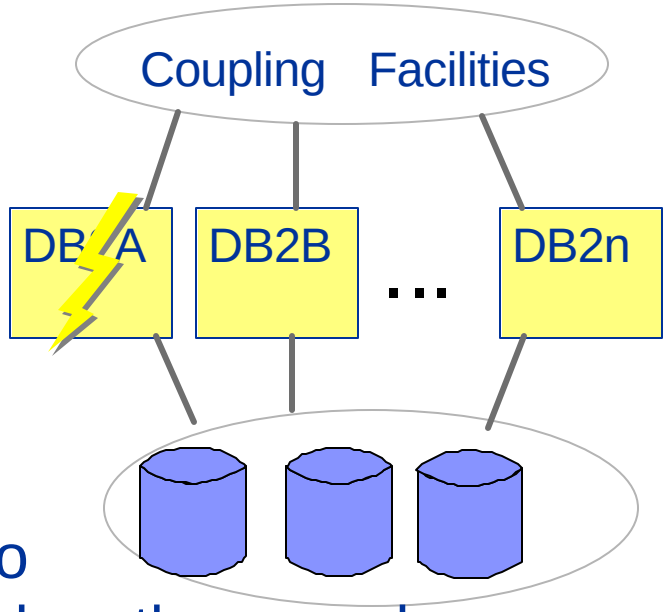
Improvements

- 1998 - Nondisruptive engine added
- 1999 - More logical CPs than physical CPs capability
- Keeps applications running while engines are added
- 2000 Fast activation
- 2001 Nondisruptive downgrade to original configuration

Process Failure

- OS/390 Recovery & Termination Manager
  - FRR and ESTAE routines intercept programming exceptions
  - "Percolate" or retry
- OS/390 SubSys Interface used for end-of-task (EOT), end-of-memory (EOM) monitoring
- Isolate failures (abends) to specific threads
- Collect diagnostic info for first failure data capture
  - Dumps, traces, error recording logs
- Maintain integrity of system control structures
- Cleanup shared resources such as latches
- Deferred EOT allows completion of commit/abort if the appl process goes away
- "Must complete" runs under internal, protected execution units
- Extensive debugging tools have been built around OS/390 Interactive Problem Control System (IPCS)
  - Allows for analysis of data on location at customer site

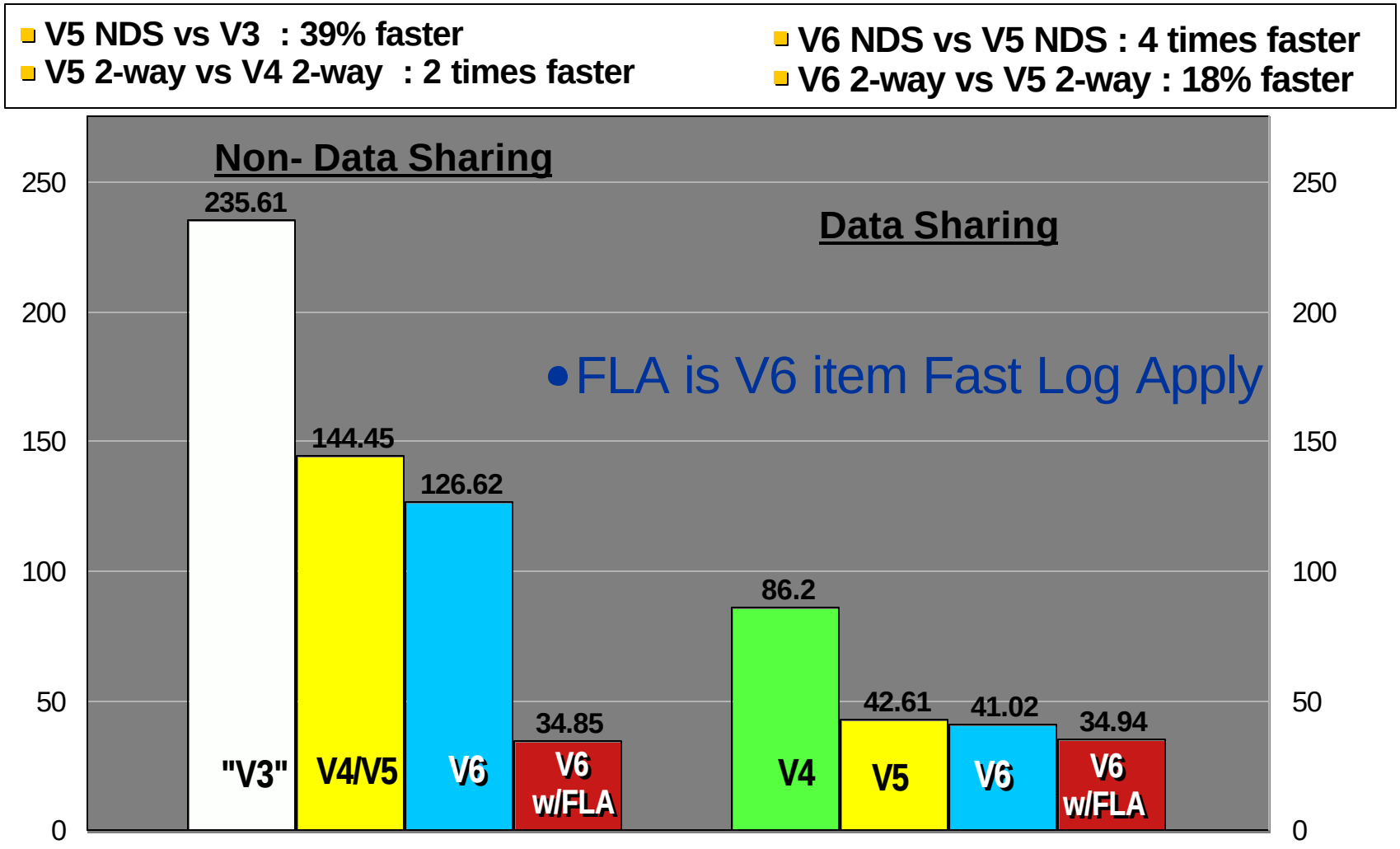
DB2 Member Failure Recovery



- The other "surviving" members remain up and running
- The architecture allows all members to access all portions of the data bases
- Work can by dynamically routed away from the down DB2 member
- The failed member holds "retained locks" to protect inconsistent data from being locked by other members
- OS/390 Automatic Restart Manager can automatically restart failed DB2 members
- Restart on same or different OS image
- Force-at-commit greatly reduces 'redo' log apply work
- Merged log never needed for restart
- "Restart light" option uses small storage footprint
- Option to defer backout

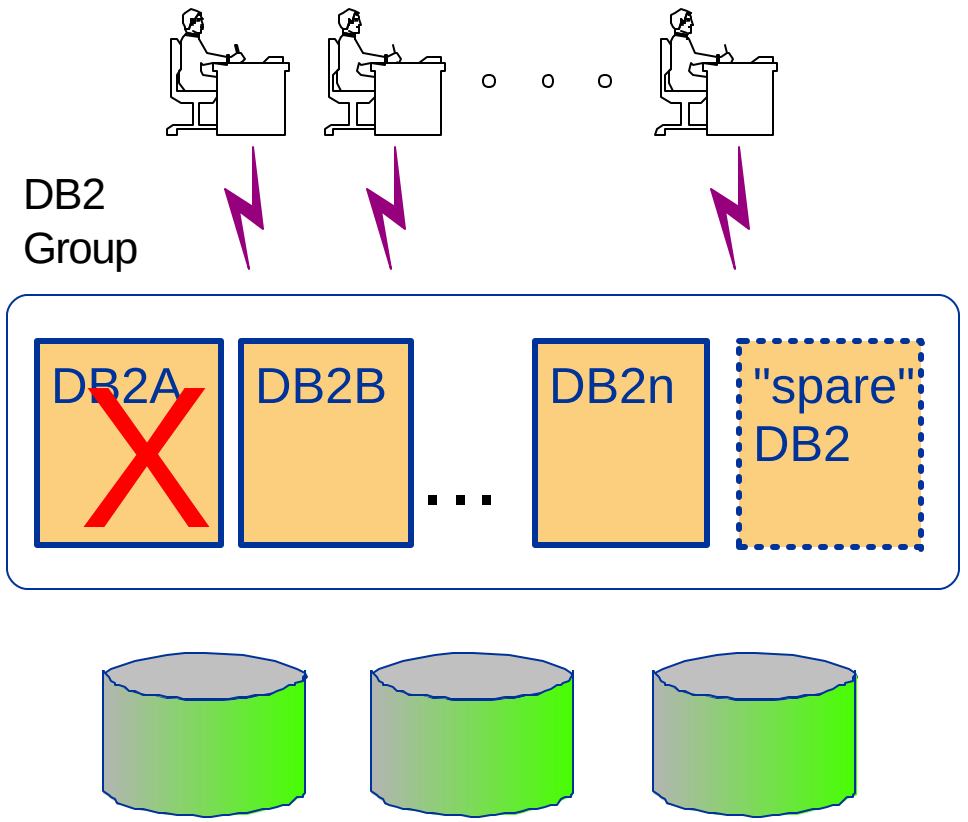
DB2 Restart Recovery Performance

- Laboratory performance measurements on G5 hardware



Online Software Maintenance

- "Rolling" maintenance with DB2 data sharing
- One DB2 member stopped at a time
- DB2 data is continuously available via the N-1 members
- "Spare" member can be optionally started to maintain full capacity
- Multiple release levels can coexist within a group
- Release fallback support



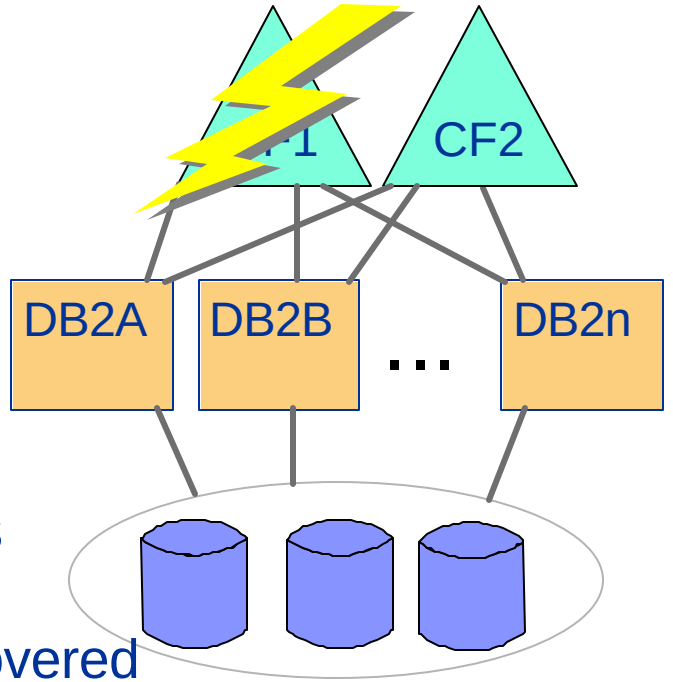
➡ Migrate software releases, apply software maintenance

➡ **Without** any user-perceived outage

➡ Catalog migration locks portions of the catalog

Coupling Facility Outages

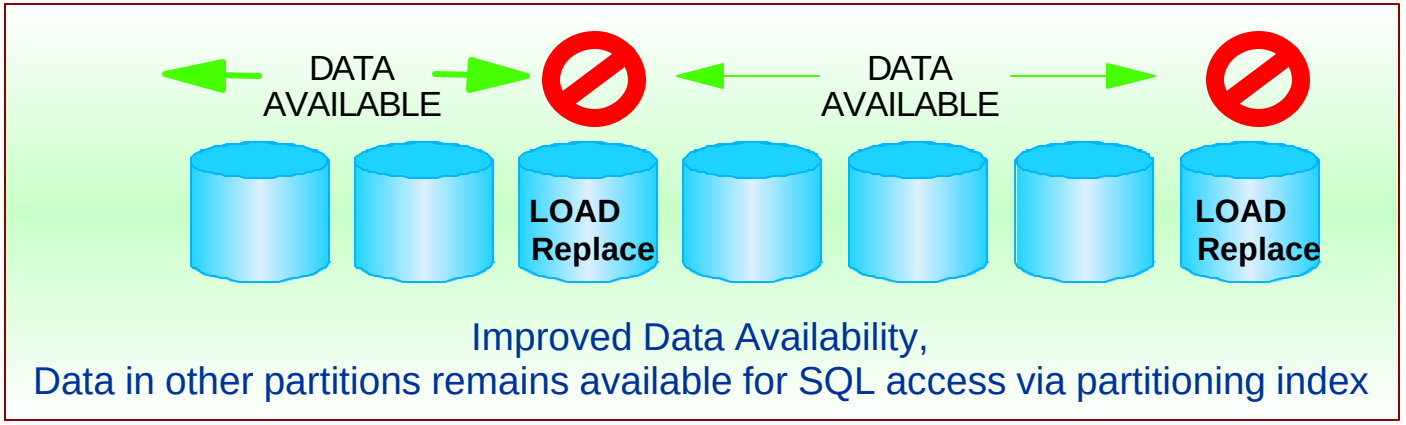
- Planned outages: use the MVS "rebuild" command to move the structures out of the CF that's targeted for maintenance
  - V4: Lock & SCA structures supported, but not group buffer pools (GBPs)
  - V5: support added for GBPs
- Unplanned outages: the system automatically recovers the lost structures
  - V4: Lock & SCA can be automatically recovered, but damaged GBPs must be recovered manually
  - V5: DB2 automatically recovers GBPs from logs
  - V6: GBP duplexing (retro-fitted to V5) via Apars



➡ CF failures are extremely rare, but.....

➡ As of V6, DB2 can recover from any CF failure within seconds

Partition Independence

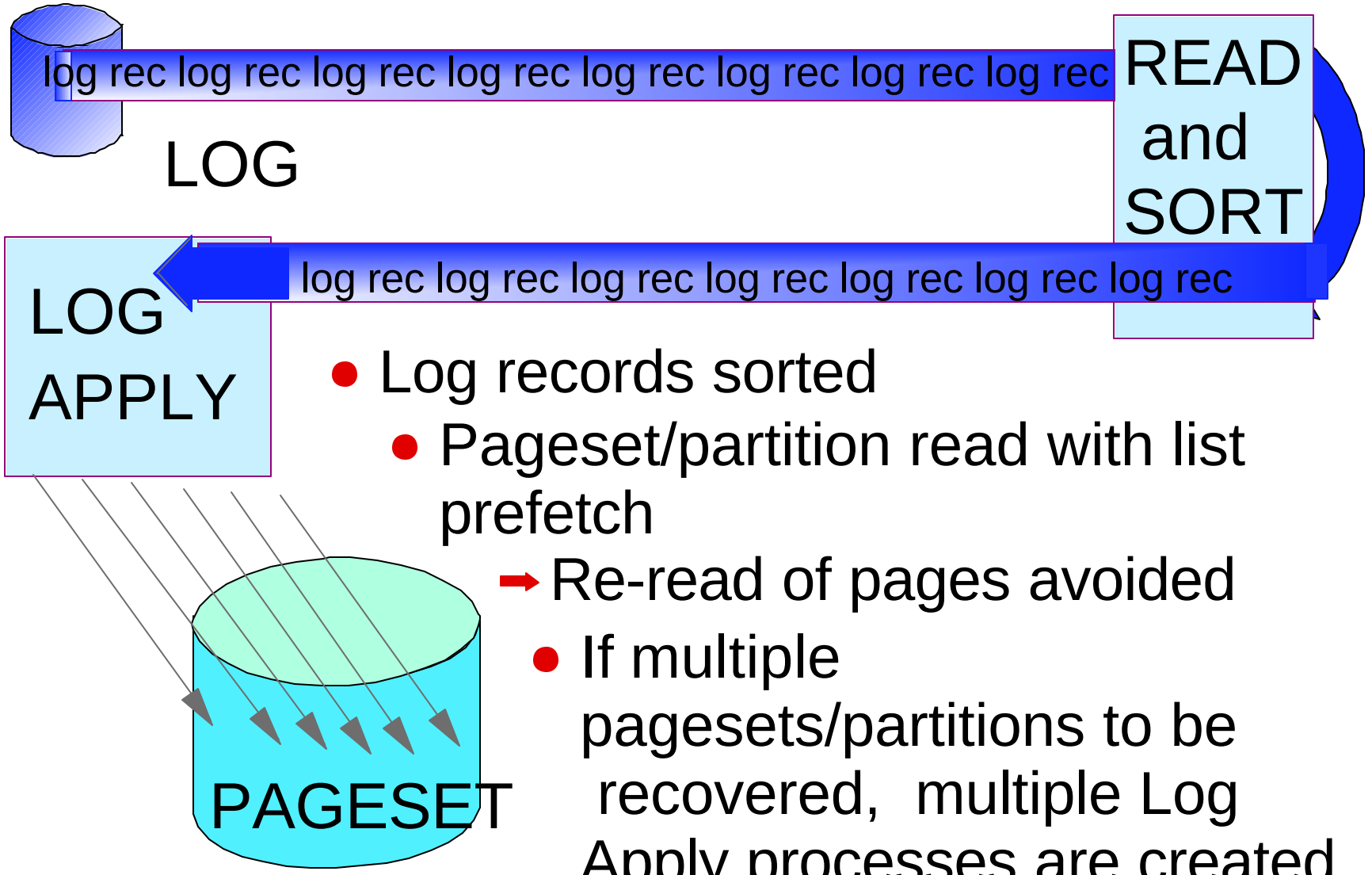


- Improves data availability for utilities and SQL processing
- Reduces elapsed time for utility processing
- Allows you to STOP or START DATABASE on partition level
- Different utilities can run concurrently on different partitions
- SQL runs on partitions not exclusively needed for utilities
- Version 5
  - Up to 254 partitions of 4GB each (1TB for a single table)
- Version 6
  - Up to 254 partitions of 64GB each (16TB for a single table)

Backup and Recovery

- Backup
  - Incremental or Full
  - Shrlevel(Ref) or Shrlevel(Change)
  - Exploits "fast copy" features of disk subsystems
- Recovery
  - To currency or to a point in time
  - SYSLGRNX directory table limits log scan needed
  - Parallel log apply process
  - LOGONLY recovery - use current database object on disk as the recovery base (instead of image copy)
- Wildcarding and dynamic allocation improves usability
- Partition level
- COPY/RECOVER parallelism for object lists
- Indices can be backed up & recovered independently of data
- System-level backup/recovery features also provided

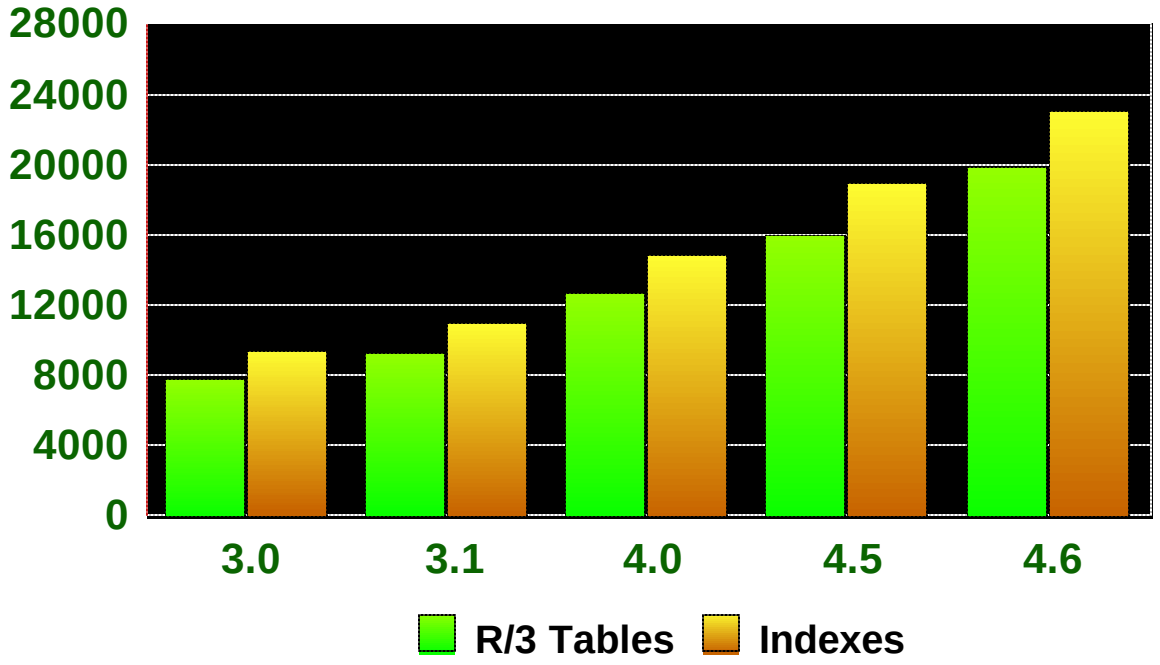
Fast Log Apply



- Log records sorted
  - Pageset/partition read with list prefetch
    - ➡ Re-read of pages avoided
  - If multiple pagesets/partitions to be recovered, multiple Log Apply processes are created

Online Schema Evolution

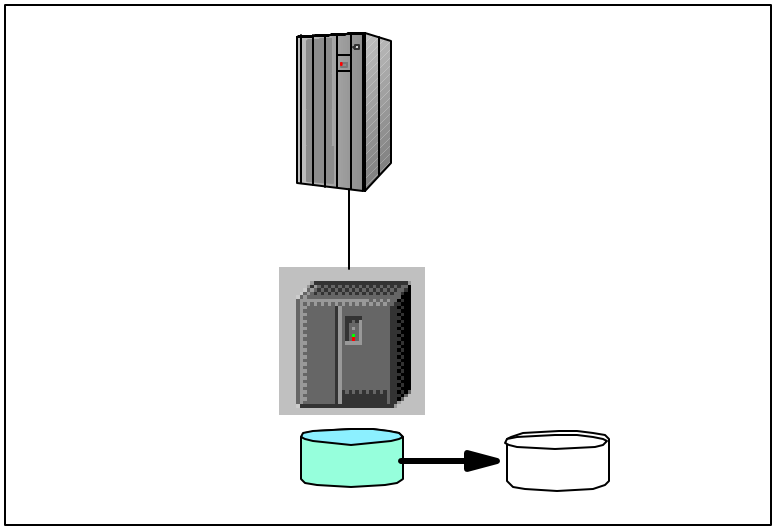
Example: *SAP R/3 Schema Object Number Growth*



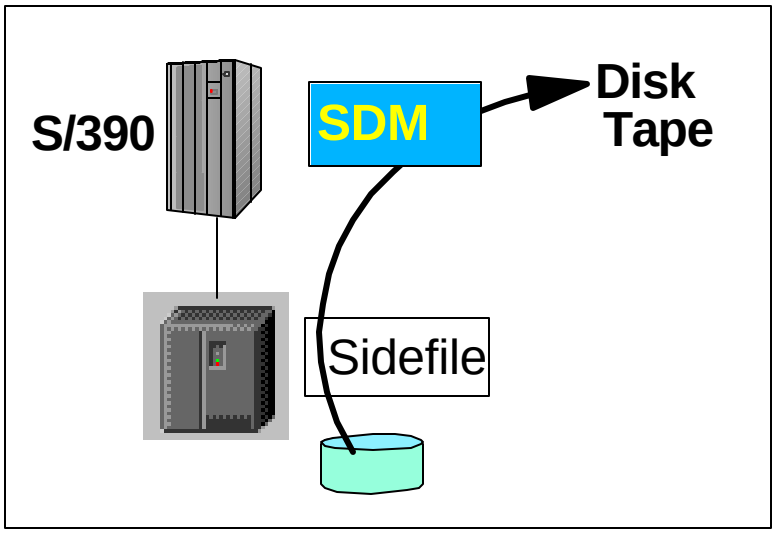
- Add / alter metadata without outages
- Do not invalidate related objects
- Multiple versions of records in same table
  - no changes to data when schema changes
- Change column data types, increase column lengths, ...

Shark Copy Services

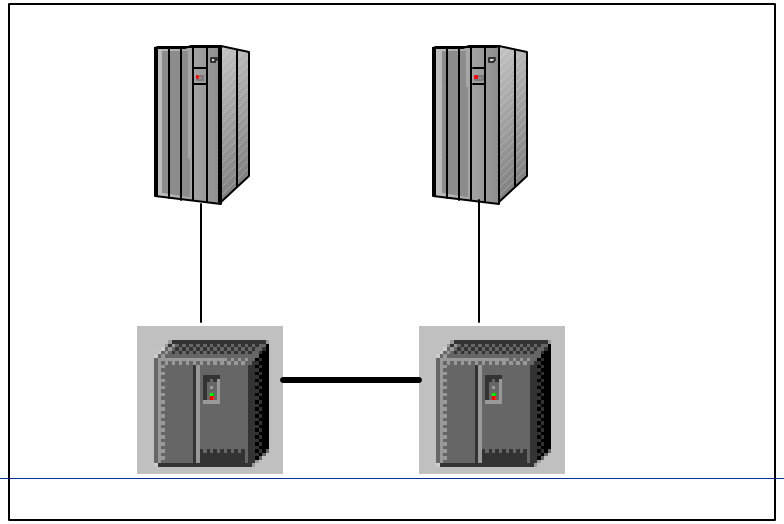
FlashCopy



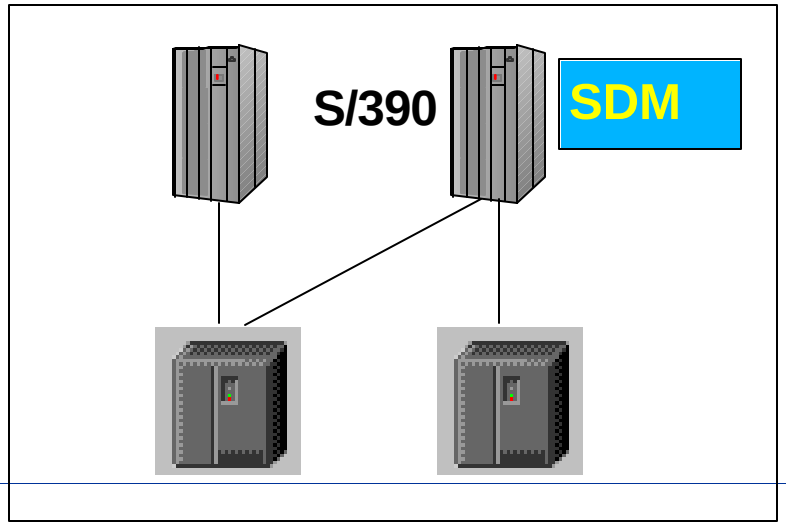
Concurrent Copy



PPRC



XRC



Shark Copy Services

FlashCopy

- S/390 and Open Systems
- Instant T0 copy of a volume for test, backup, temporary copy.....
- Source and Target must be in same logical Subsystem
- Target must have same capacity (or larger) as source
- Invoked by DFSMSdss full volume COPY, TSO command or ESS Copy Services

Concurrent Copy

- S/390 only
- Instant T0 copy of volume or data set for data backup
- Target can be on tape or DASD volume
- OS/390's System Data Mover is used to move the data
- Sidefiles in cache is used for the updates
- Invoked by DFSMSdss or applications internally call DFSMSdss as the copy program, such as DB2 COPY utility

PPRC

- S/390 and Open Systems
- Disaster recovery solution
- Synchronous mirror copies of volumes on remote ESS
- Direct connections between ESS systems using ESCON links
- Invoked by TSO command or ESS Copy Services

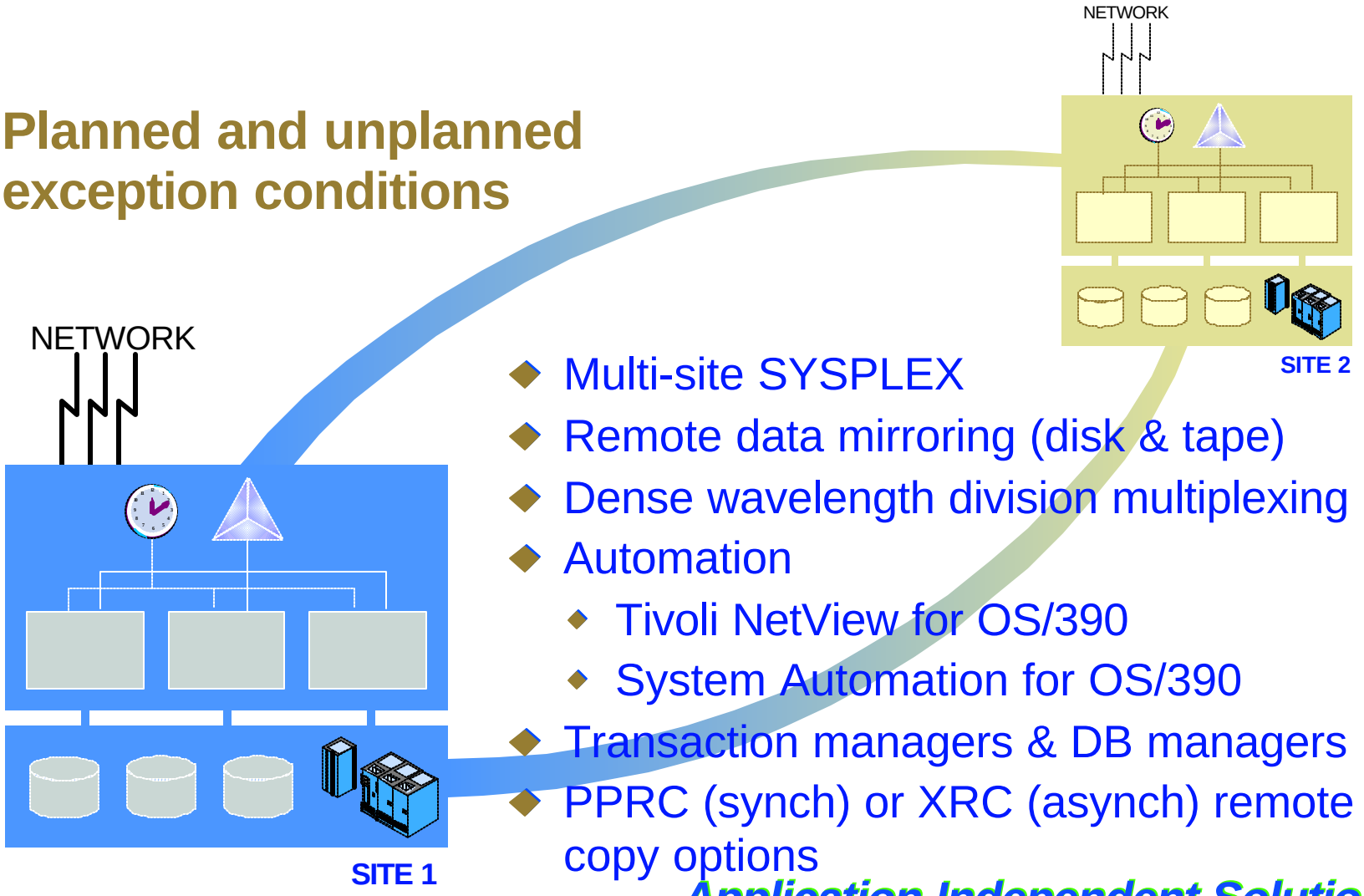
XRC

- S/390 only
- Disaster recovery solution
- Asynchronous remote mirror copies of volumes
- Supports large distances
- OS/390's System Data Mover is used to move the data
- Invoked by TSO command

GDPS Disaster Recovery Solution

GDPS (Geographically Dispersed Parallel Sysplex)

Planned and unplanned exception conditions



- Multi-site SYSPLEX
- Remote data mirroring (disk & tape)
- Dense wavelength division multiplexing
- Automation
  - Tivoli NetView for OS/390
  - System Automation for OS/390
- Transaction managers & DB managers
- PPRC (synch) or XRC (asynch) remote copy options

Application Independent Solution

GDPS Policy Options

FREEZE & GO

- Freeze secondary disk configuration
- Allow applications to continue
- ➡ Optimize for remote restartability
- ➡ Least impact on application availability
- ➡ May lose data in case of real disaster

FREEZE & STOP

- Freeze secondary disk configuration
- Stop all OS/390 images
- ➡ Optimize for remote restartability
- ➡ May impact application availability
- ➡ No data loss on primary site disaster

FREEZE & STOP Conditional

- Freeze secondary disk configuration
- Determine reason for Suspend
  - If secondary HW problem then FREEZE & GO
  - Other reason: FREEZE & STOP

Other Important Availability Features

- Online REORG utility
- Online LOAD RESUME utility
- Inline RUNSTATS and COPY utilities
- LOAD/REORG parallel index build
- Online system parameter change
- Stored procedures run in their own address spaces
  - Failure isolation
- Write I/O errors: isolate failure to page level
  - Logical Page List (LPL)

Some Challenges

- Dynamic schema change, ongoing work.
- Easier, less disruptive Point-in-time recovery
- Self-healing, self-managing systems

References

- DB2 for OS/390 www.ibm.com/software/db2os390
- Red Books www.ibm.com/redbooks
- Parallel Sysplex www.s390.ibm.com/pso/
- IBM zSeries ibm.com/servers/eserver/zseries/
- IBM Systems Journal "S/390 Parallel Sysplex Cluster", Vol. 36, Number 2, 1997