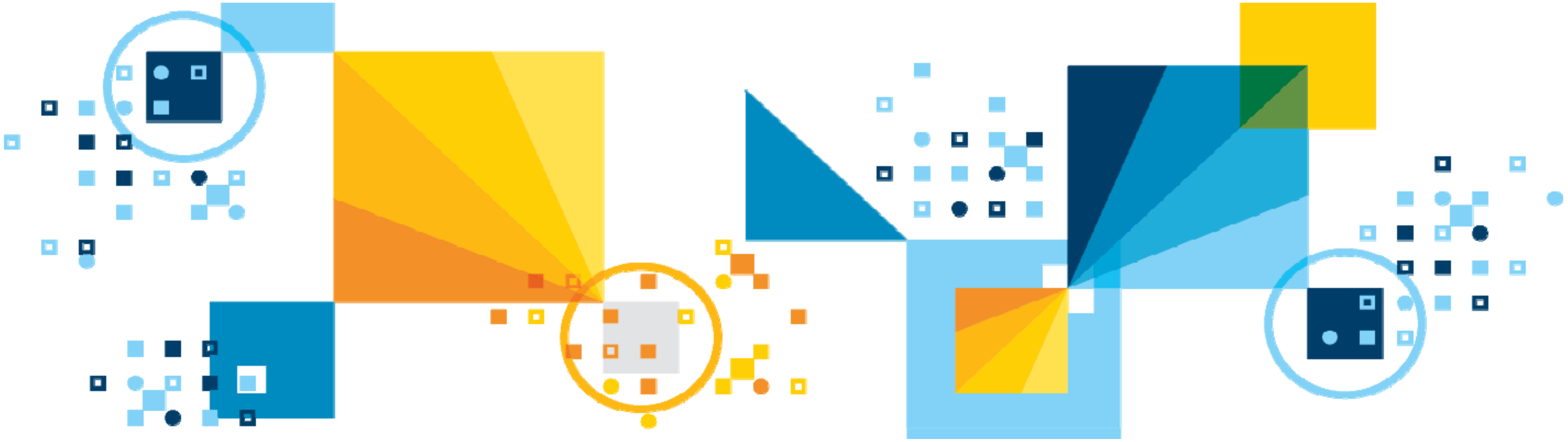
Information Management Skills Development

DB2 pureScale

Module ID	10113
Length	1 hour



For questions about this presentation contact askdata@ca.ibm.com

January 22, 2015

Disclaimer

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 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

Other company, product, or service names may be trademarks or service marks of others.

Module Information

- You should have completed or acquired the necessary knowledge for the following modules in order to complete this module:
  - DB2 Fundamentals
  - High Availability and Disaster Recovery

- After completing this module, you should be able to:
  - Describe the pureScale feature
  - Explain the concepts of continuous availability and application transparency

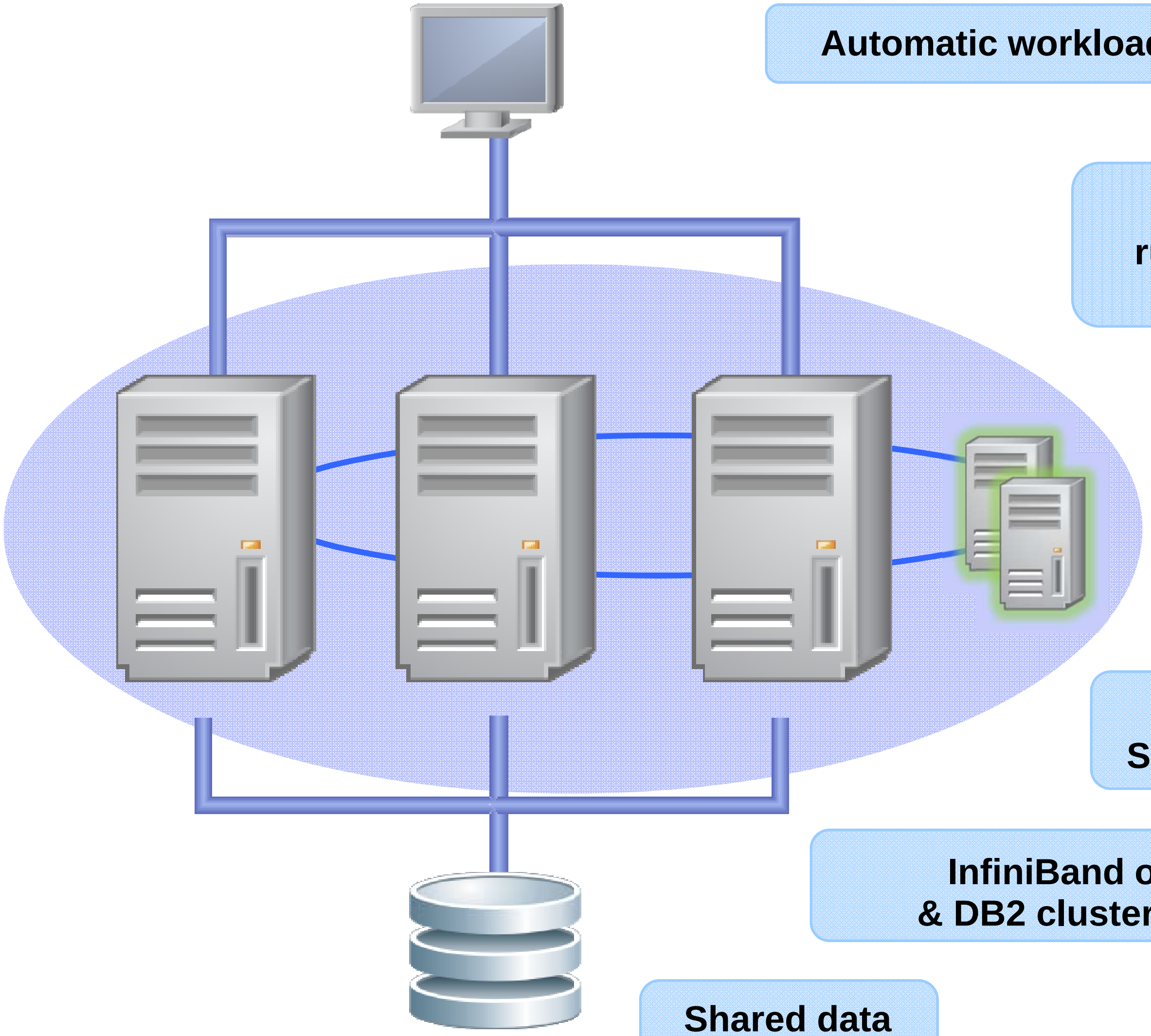
Module Content

- **Idea behind DB2 pureScale**
- **pureScale on System X and System P**
- **A deeper look into pureScale**
- **pureScale scalability**
- **Deploying DB2 pureScale**
- **Geographically dispersed pureScale cluster**
- **Enhancements**
- **What's new for pureScale in DB2 10.5**

DB2 pureScale – Designed for OLTP

- **Extreme Capacity**
  - Buy only what you need, add capacity as your needs grow
  - Handle key capacity spikes with **pay by the day pricing**
- **Application Transparency**
  - Easy to implement, easy to grow
- **Continuous Availability**
  - 24x7 availability so key database systems never go down, even if multiple servers fail

DB2 pureScale Feature



Automatic workload balancing

Cluster of DB2 nodes running on Power or x86 based servers

Leverages the global lock and memory manager technology from the z/OS platform

Integrated Tivoli System Automation

InfiniBand or RoCE & DB2 cluster services

Shared data filesystems

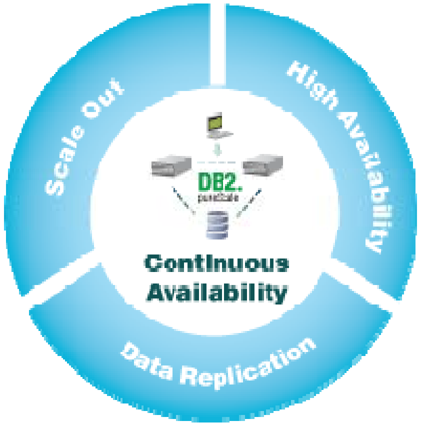
Extreme Capacity and Application Transparency

- Take advantage of extra capacity instantly
  - You can easily add/remove members and cluster powerHA servers to meet peak demands
  - No need to modify your application code
  - No need to tune your database infrastructure
  - No need to know hardware specific commands to add additional storage capacity

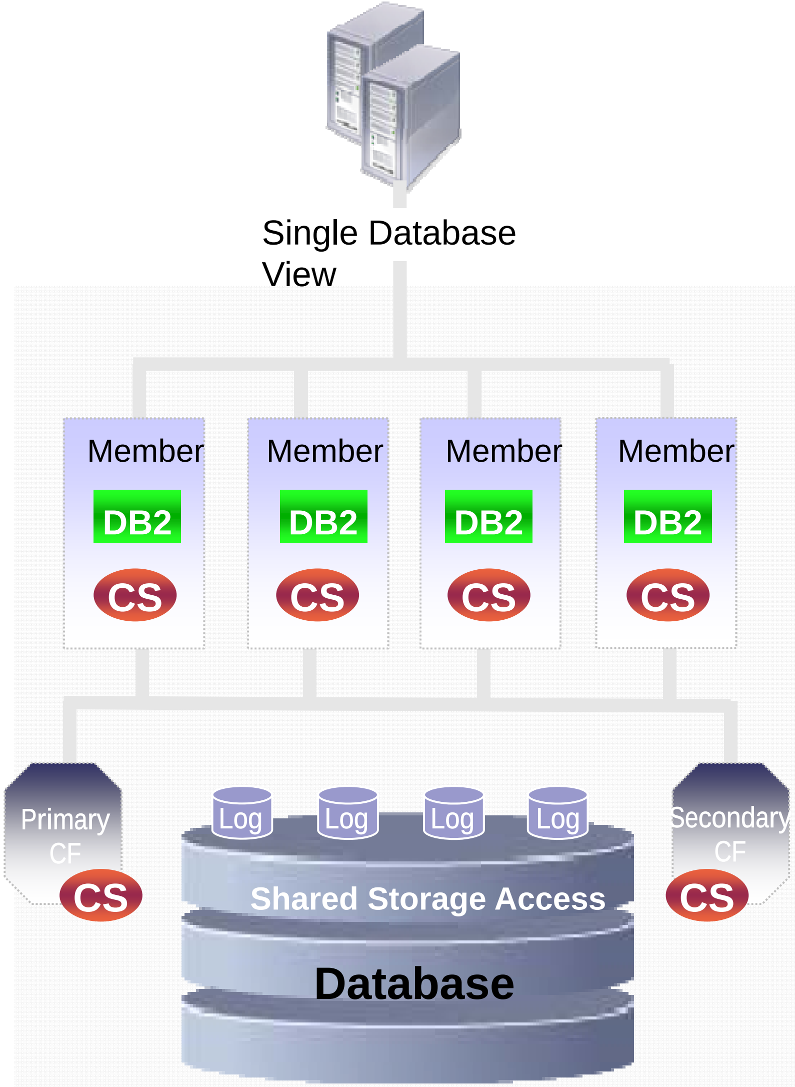
Your DBAs can add capacity without re-tuning or re-testing
Your developers don't even need to know more nodes are being added

High Availability vs. Continuous Availability

- **High Availability:**
  - Data is available **MOST** of the time
  - Planned and unplanned downtimes can affect availability
- **Continuous Availability:**
  - Data is available **ALL** of the time
  - Unaffected by system maintenance and unplanned events, such as host outages
  - Elimination of any single point of failure
  - **Zero downtime**
  - DB2 pureScale combines high availability with true transparent application scaling

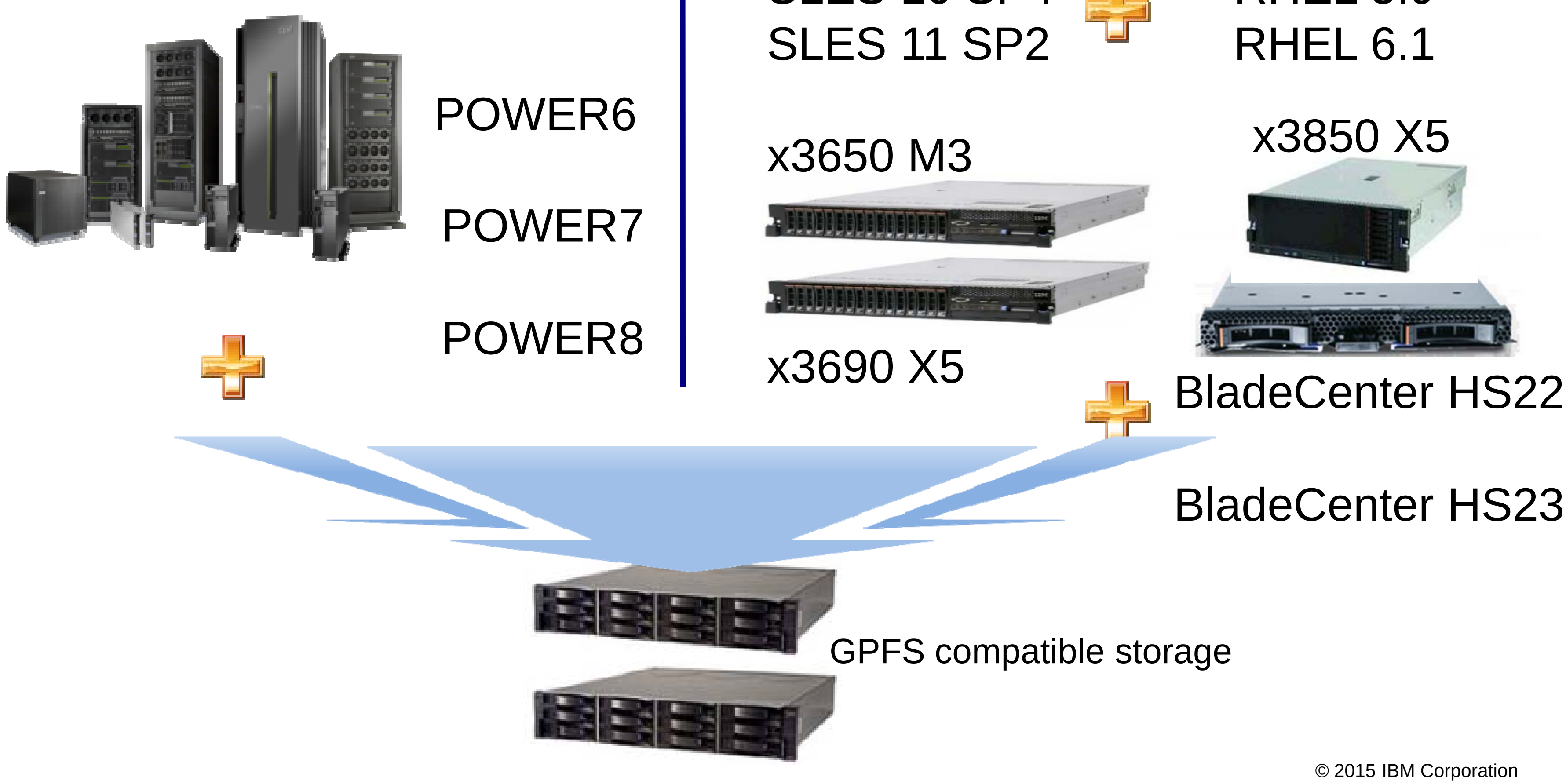


DB2 pureScale Architecture Overview



- **Clients connect anywhere, see single database**
  - Clients connect into any member
  - Automatic load balancing and client reroute may change underlying physical member to which client is connected
- **DB2 engine runs on several host computers**
  - Co-operate with each other to provide coherent access to the database from any member
- **Integrated cluster services**
  - Failure detection, recovery automation, cluster file system
  - In partnership with STG and Tivoli
- **Low latency, high speed interconnect**
  - Special optimizations provide significant advantages on RDMA-capable interconnects (InfiniBand and RoCE)
- **Cluster caching facility (CF)**
  - Efficient global locking and buffer management
  - Synchronous duplexing to secondary ensures availability
- **Data sharing architecture**
  - Shared access to database
  - Members write to their own logs on shared disk
  - Logs accessible from another host (used during recovery)

Supported Configurations – DB2 10.5



AIX

POWER6

POWER7

POWER8

SLES 10 SP4
SLES 11 SP2

OR

RHEL 5.9
RHEL 6.1

x3650 M3

x3690 X5

x3850 X5

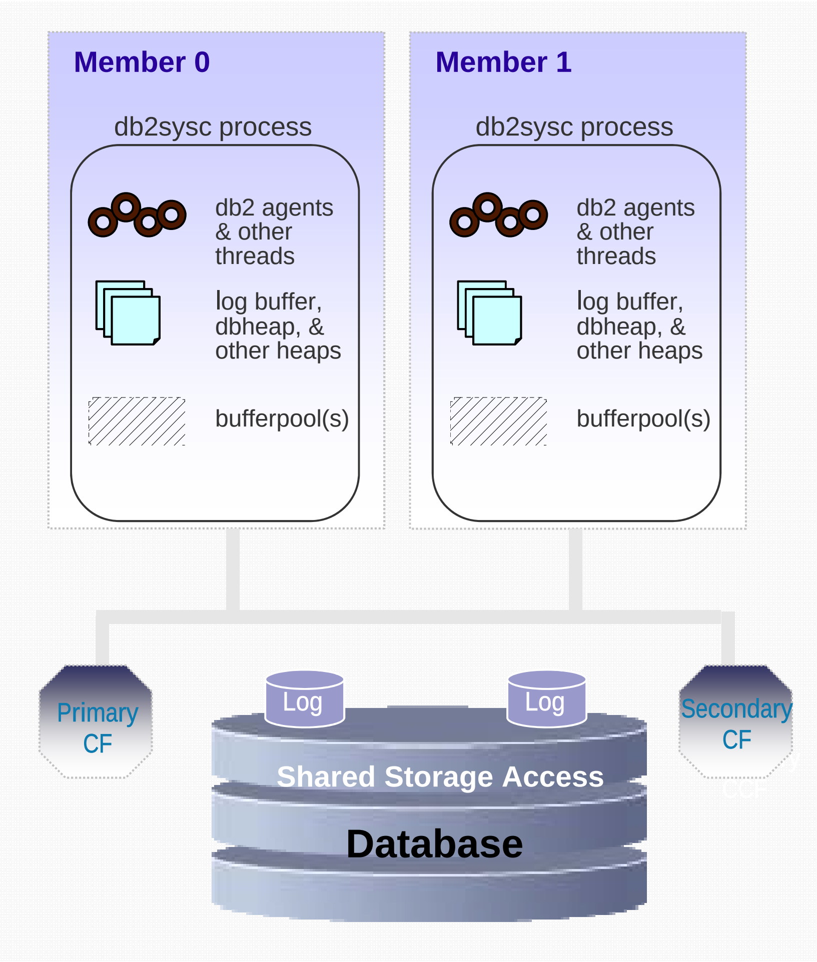
BladeCenter HS22

BladeCenter HS23

GPFS compatible storage

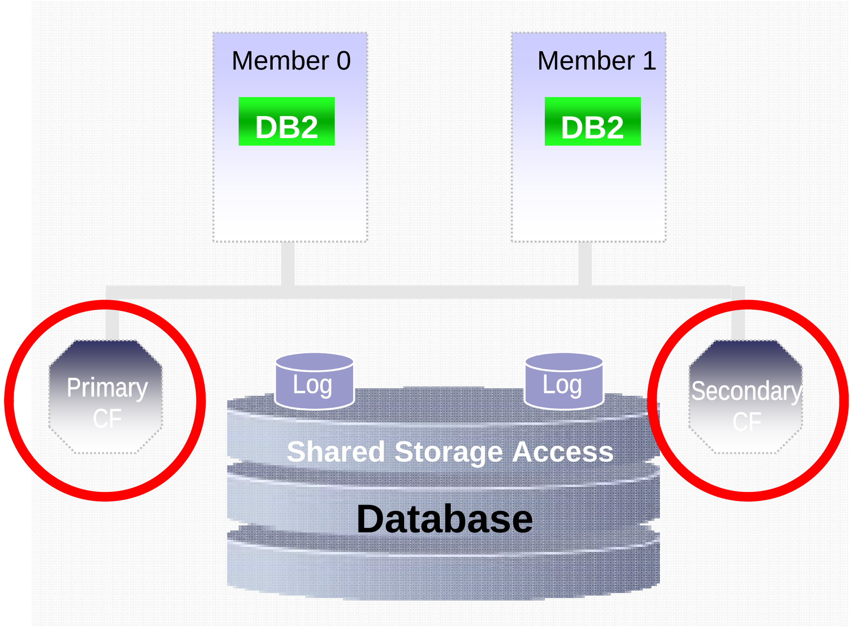
What is a Member ?

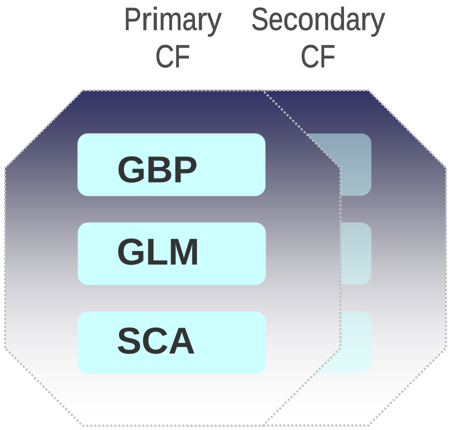
- **A DB2 engine address space**
  - i.e. a db2sysc process and its threads
- **Members Share Data**
  - All members access the same shared database
  - Aka “Data Sharing”
- **Each member has it’s own**
  - Buffer pools
  - Memory regions
  - Log files
- **Members are logical**
- **Can have**
  - **1 member per machine (recommended)**
  - 1+per machine



What is a Cluster Caching Facility (CF)?

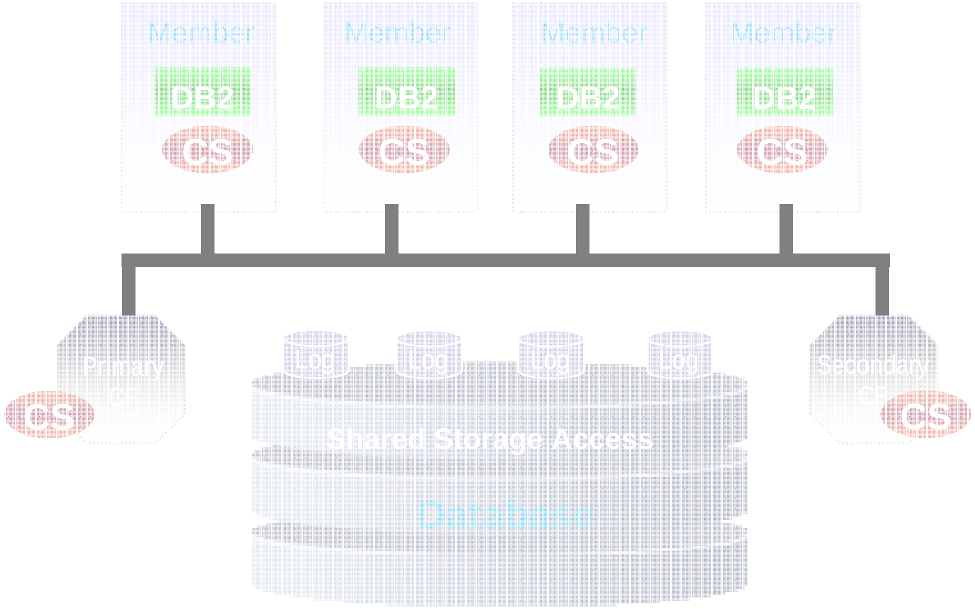
- **Software technology that assists in global buffer coherency management and global locking**
  - Shared lineage with System z Parallel Sysplex
  - Software based
- **Services provided include**
  - Group Buffer Pool (**GBP**)
  - Global Lock Manager (**GLM**)
  - Shared Communication Area (**SCA**)
- **Redundant CFs (recommended)**
  - Eliminates the single point of failure
  - Members automatically updates both CFs
  - Set up automatically





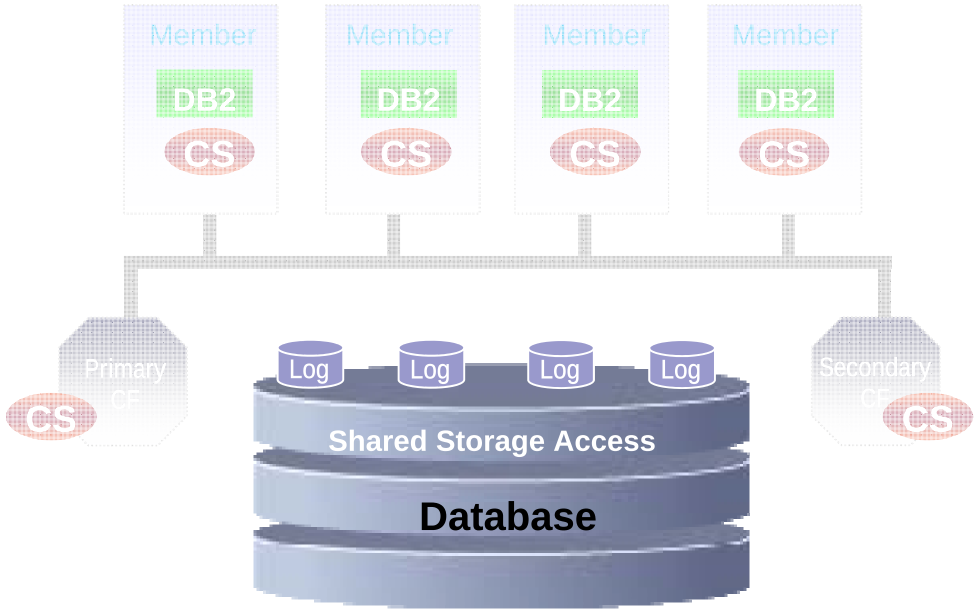
Cluster Interconnect

- Requirements
  - **Low latency, high speed interconnect** between members, and the primary and secondary CFs
  - **RDMA** capable fabric, to be able to make direct updates in memory without the need to interrupt the CPU
- Solutions
  - InfiniBand (IB) and uDAPL for performance
    - InfiniBand supports RDMA and is a low latency, high speed interconnect
    - uDAPL to reduce kernel time
  - RDMA on Converged Ethernet (RoCE)
  - TCP/IP on Ethernet for workloads that are not latency dependent.



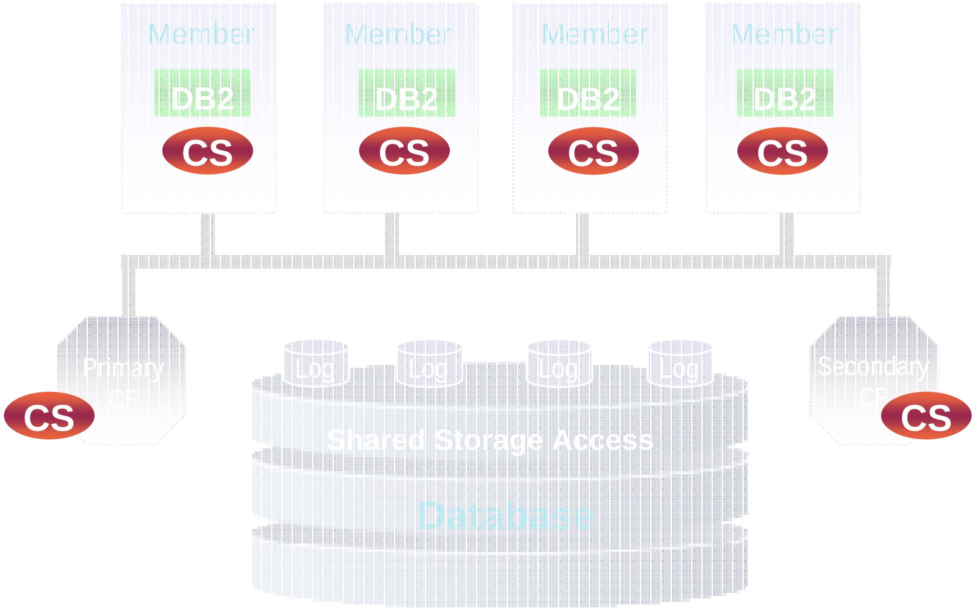
Cluster File System

- **Requirements**
  - Shared data requires shared disks and a cluster file system
  - Fencing of any failed members from the file system
- **Solution**
  - **General Parallel File System (GPFS)**
  - Shipped with, and installed and configured as part of DB2
  - We will also support a pre-existing user managed GPFS file system
    - Allows GPFS to be managed at the same level across the enterprise
    - DB2 will not manage this pre-existing file system, nor will it apply service updates to GPFS.
  - SCSI 3 Persistent Reserve recommended for rapid fencing



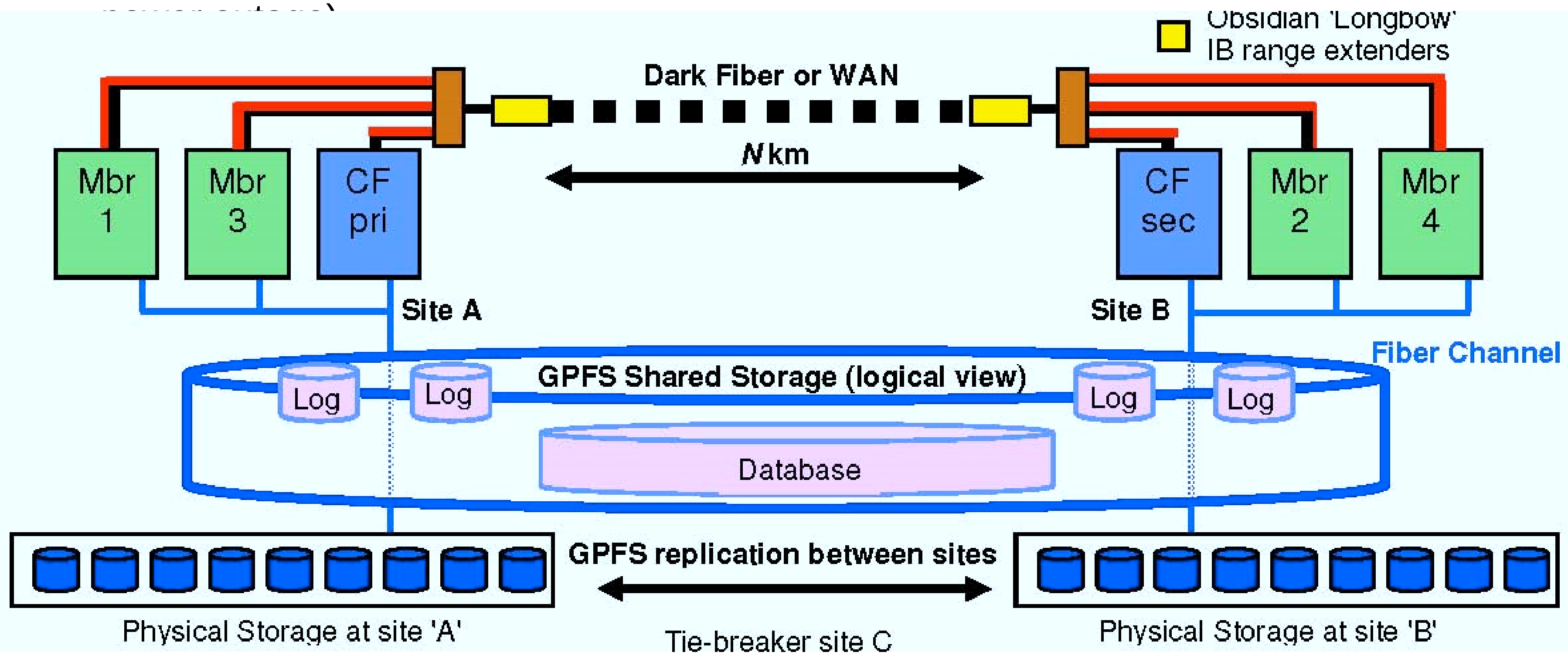
DB2 Cluster Services

- **Orchestrates**
  - – Unplanned event notifications to ensure seamless recovery and availability.
    - Member, CF, AIX, hardware, etc.
  - – Planned events
    - ‘Stealth’ maintenance (HW & SW)
- Integrates the following with DB2:
  - – **Cluster Management**
    - TSA (Tivoli System Automation)
  - – **Cluster File System**
    - GPFS (General Parallel File System)
  - – TSA and GPFS are shipped with, and installed and configured as part of the DB2 pureScale Feature



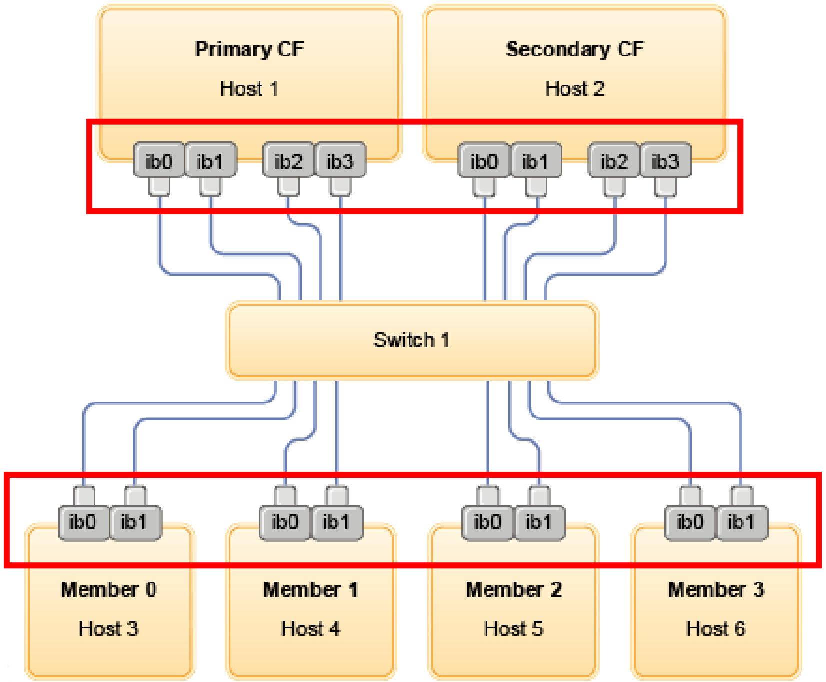
Geographically Dispersed pureScale Cluster

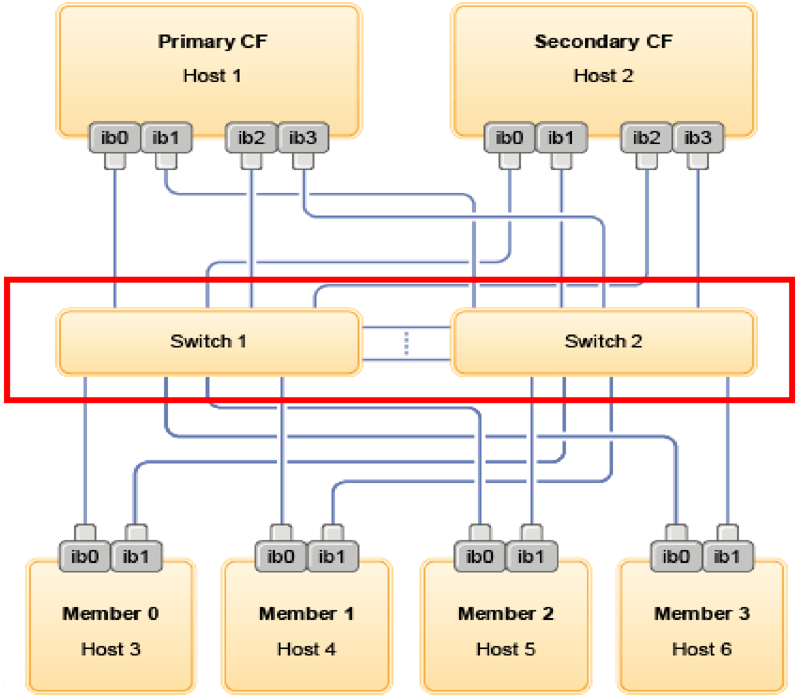
- Multiple site pureScale installation offers protection in case of disasters
  - Provides active/active access to one or more shared databases across the cluster
  - Enables a level of DR support suitable for many types of disasters (e.g. fire, localized power outage)


Obsidian 'Longbow' IB range extenders
Dark Fiber or WAN
N km
Mbr 1
Mbr 3
CF pri
CF sec
Mbr 2
Mbr 4
Site A
Site B
Fiber Channel
Log
Log
GPFS Shared Storage (logical view)
Log
Log
Database
GPFS replication between sites
Physical Storage at site 'A'
Tie-breaker site C
Physical Storage at site 'B'

Flexible Network Topology

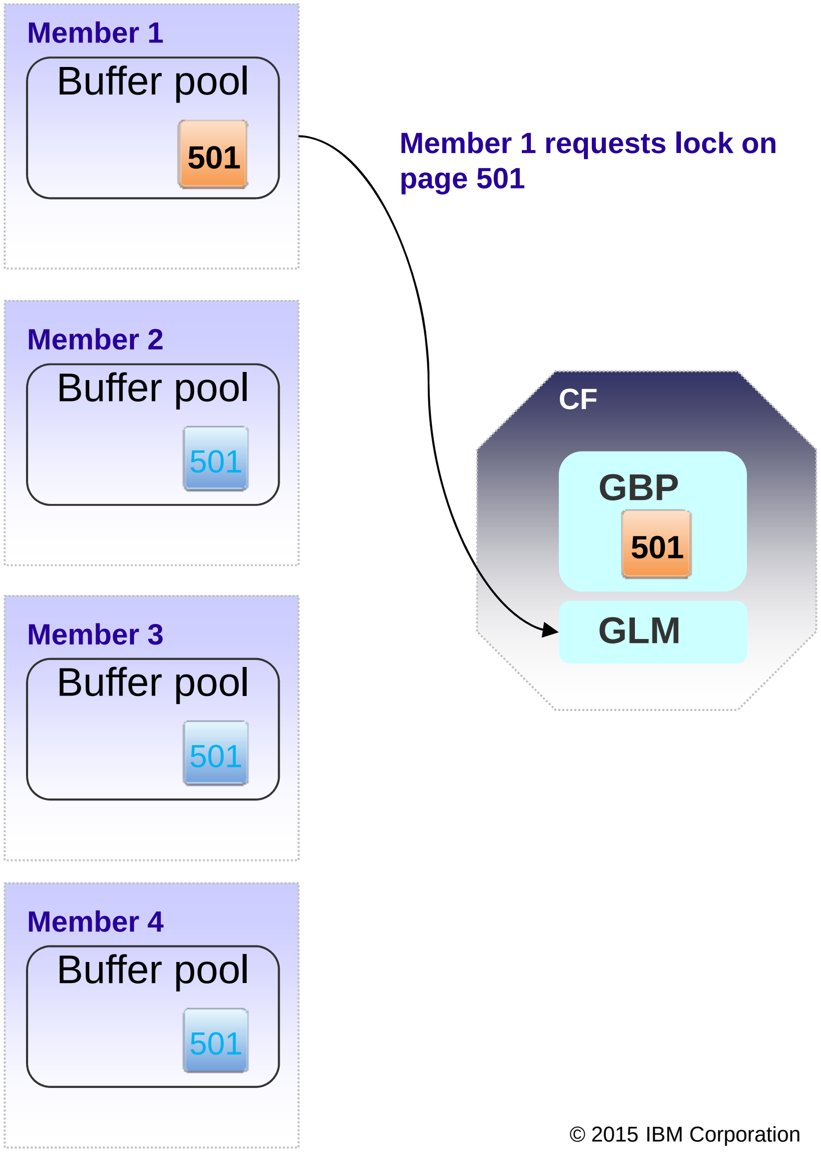
- **Multiple low-latency, high-speed cluster interconnects for the CFs and Members**
  - 1-switch configuration can increase the throughput of requests to CFs and Members
  - 2-switches configuration helps with increased throughput and high availability





Advantages of RDMA – An Example

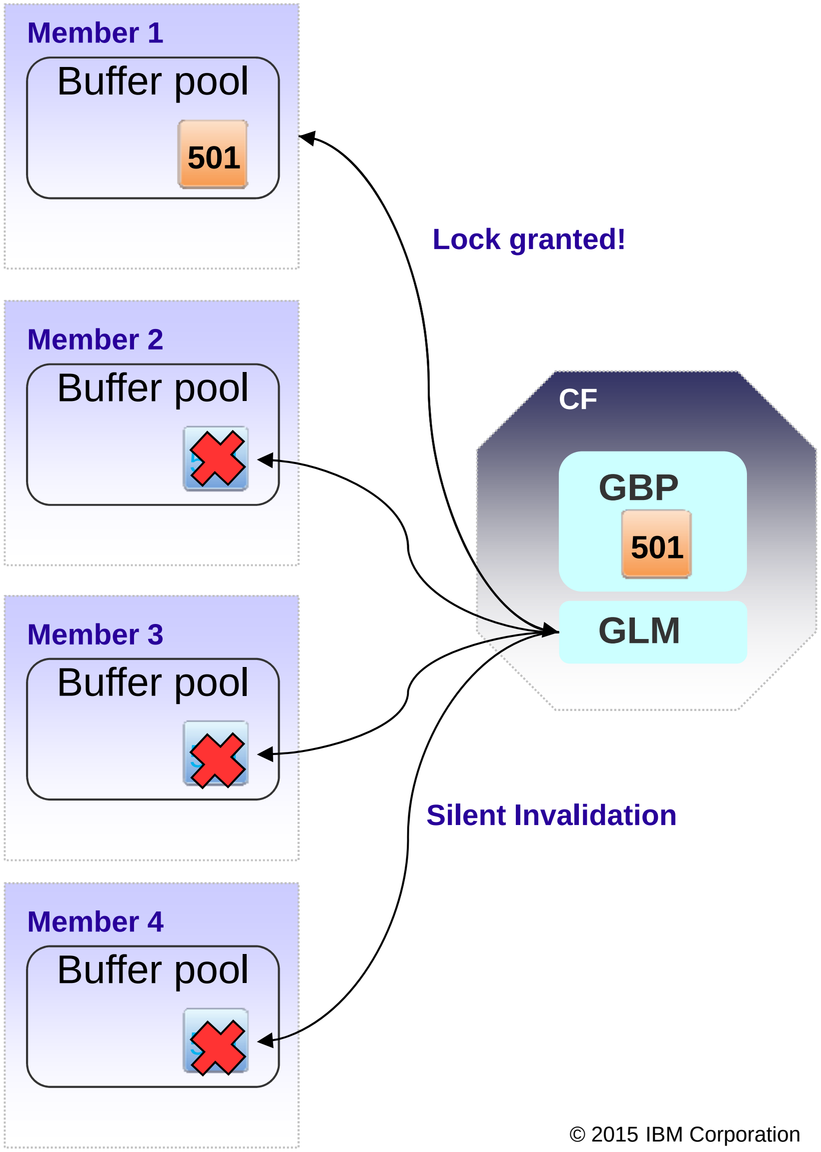
- **Deep RDMA exploitation over low latency fabric**
  - Direct memory access
  - Enables round-trip response time
    - **~10-15 microseconds**



Member 1 requests lock on page 501

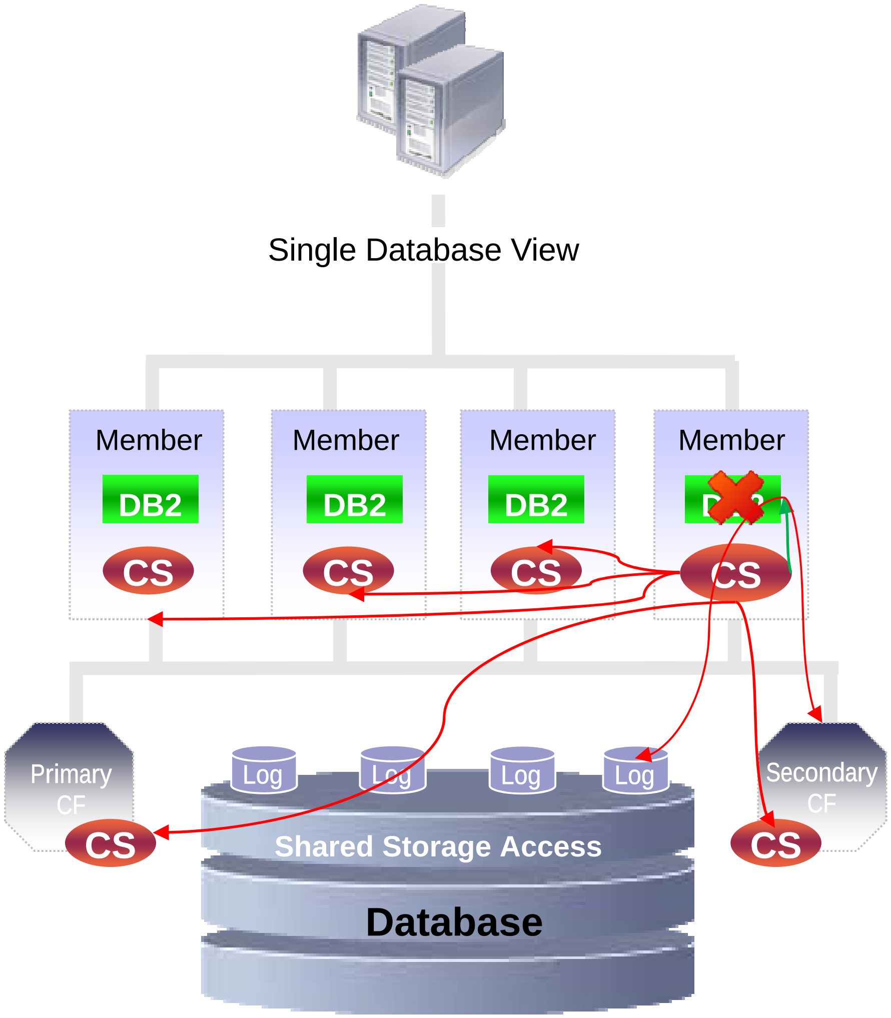
Advantages of RDMA – An Example

- **Silent Invalidation**
  - Informs members of page updates requires no CPU cycles on those members
  - No interrupt, No IP Socket Calls, No context switching, or other message processing required
  - Increasingly important as cluster grows
- **Hot pages available without disk I/O from GBP memory**
  - RDMA and dedicated threads enable read page operations in
    - **~10s of microseconds**



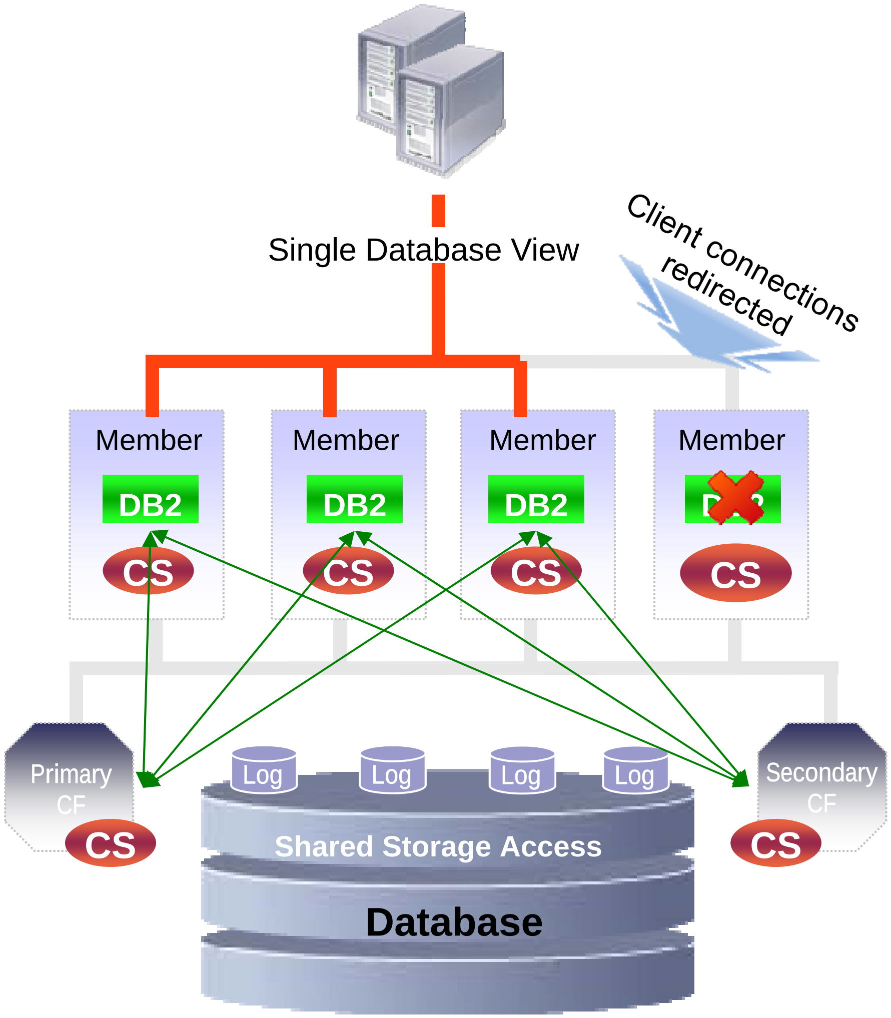
Member Software Failure Summary

- **Member Failure**
- **DB2 Cluster Services automatically detects member's death**
  - Inform other members, and CFs
  - Initiates automated member restart on same or remote host
  - Member restart is like crash recovery in a single system, but is much faster
    - Redo limited to in-flight transactions
    - Benefits from page cache in CF



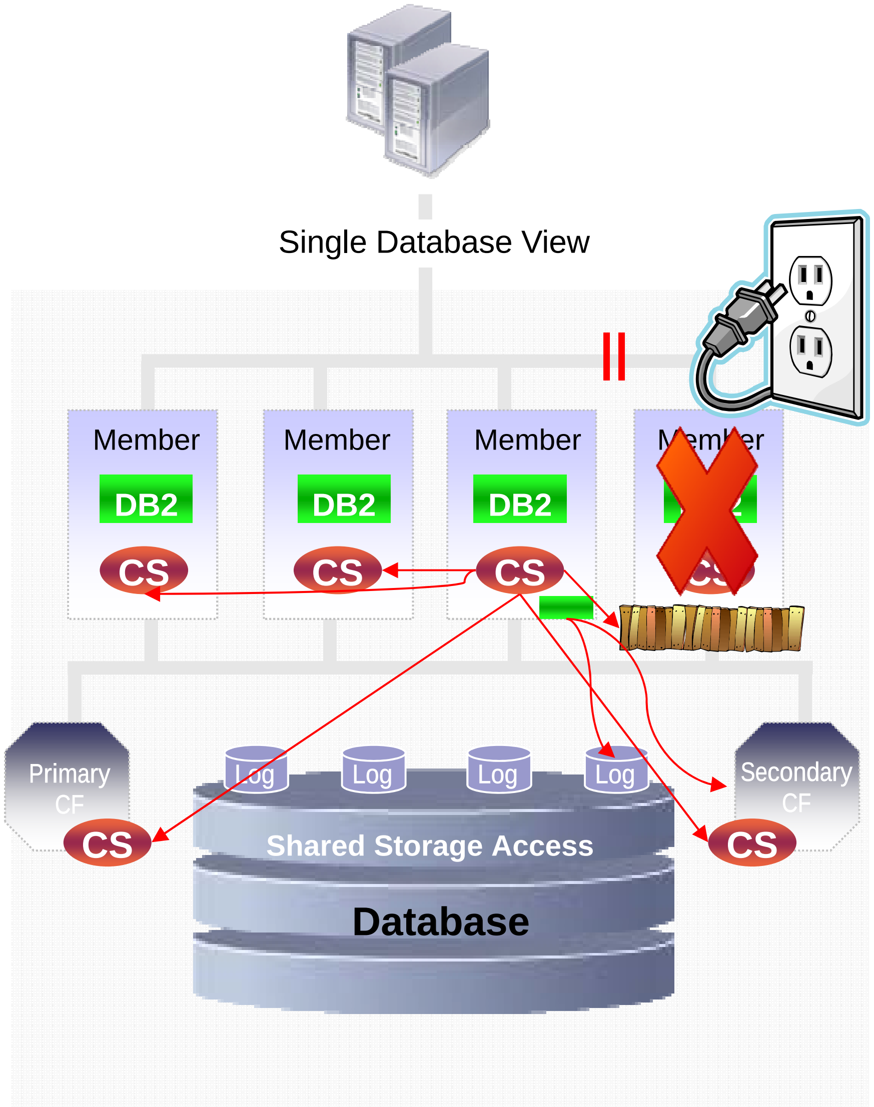
Member Software Failure Summary

- **Member Failure**
- **DB2 Cluster Services automatically detects member's death**
  - Inform other members, and CFs
  - Initiates automated member restart on same or remote host
  - Member restart is like crash recovery in a single system, but is much faster
    - Redo limited to in-flight transactions
    - Benefits from page cache in CF
- **Client <u>transparently</u> re-routed to healthy members**
- **Other members fully available at all times**
  "Online Failover"
  - CF holds update locks held by failed member
  - Other members can continue to read and update data not locked for update by failed member
- **Member restart completes**
  - Locks released and all data fully available



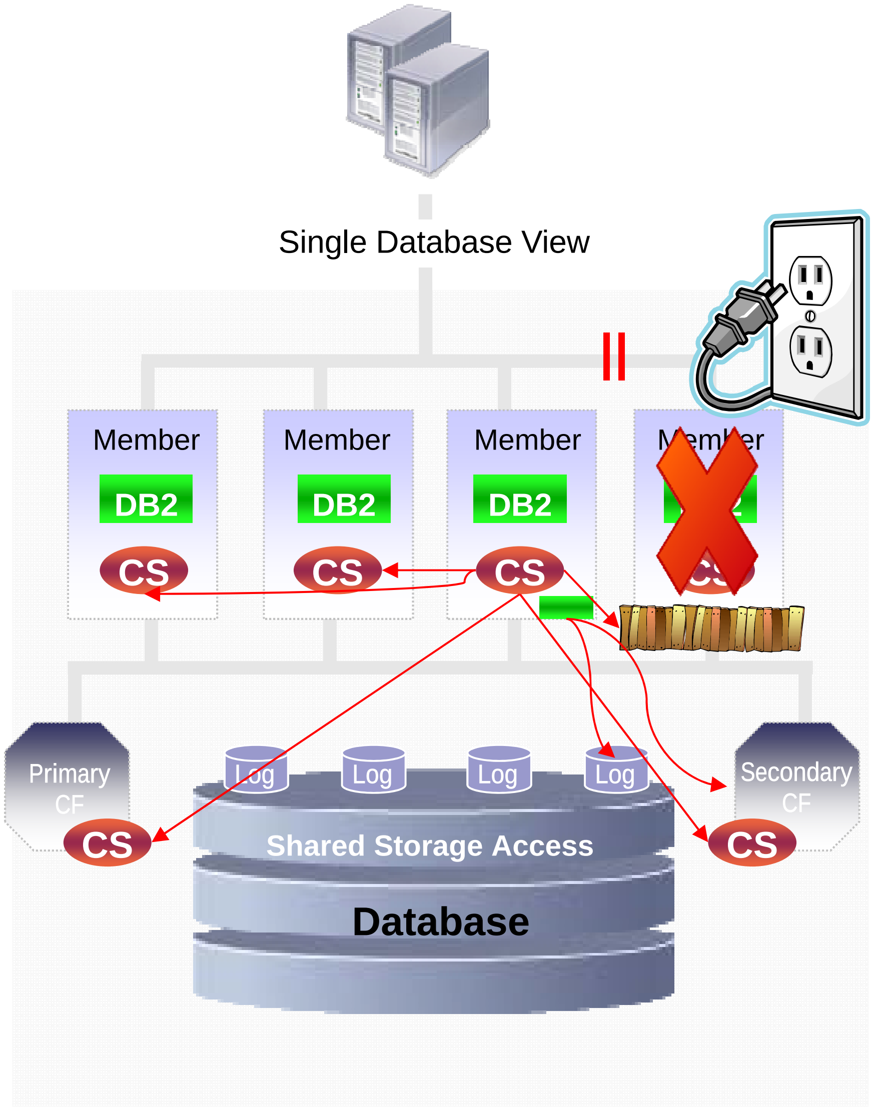
Member HW Failure – Member Restart on Guest Host

- Power cord tripped over accidentally
- DB2 Cluster Services loses heartbeat and declares member down
  - Informs other members & CFs
  - Fences member from logs and data
  - Initiates automated member restart on another (“guest”) host
    - Using reduced, and pre-allocated memory model
  - Member restart is like a database crash recovery in a single system database, but is much faster
    - Redo limited to in-flight transactions (due to FAC)
    - Benefits from page cache in CF



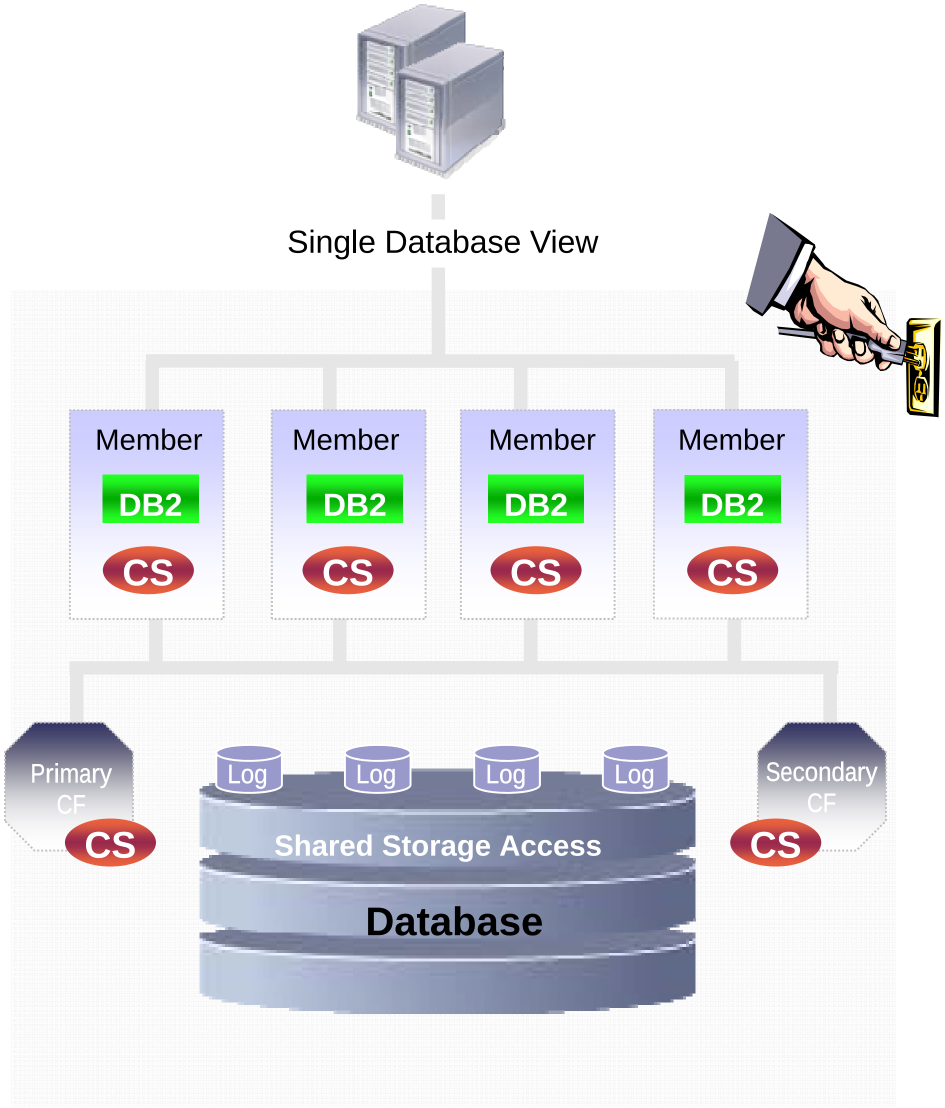
Member HW Failure – Member Restart on Guest Host

- In the mean-time, client connections are automatically re-routed to healthy members
  - – Based on least load (by default), or,
  - – Pre-designated failover member
- Other members remain fully available throughout – “Online Failover”
  - – Primary retains update locks held by member at the time of failure
  - – Other members can continue to read and update data not locked for write access by failed member
- Member restart on guest host completes
  - – Retained locks released and all data fully available



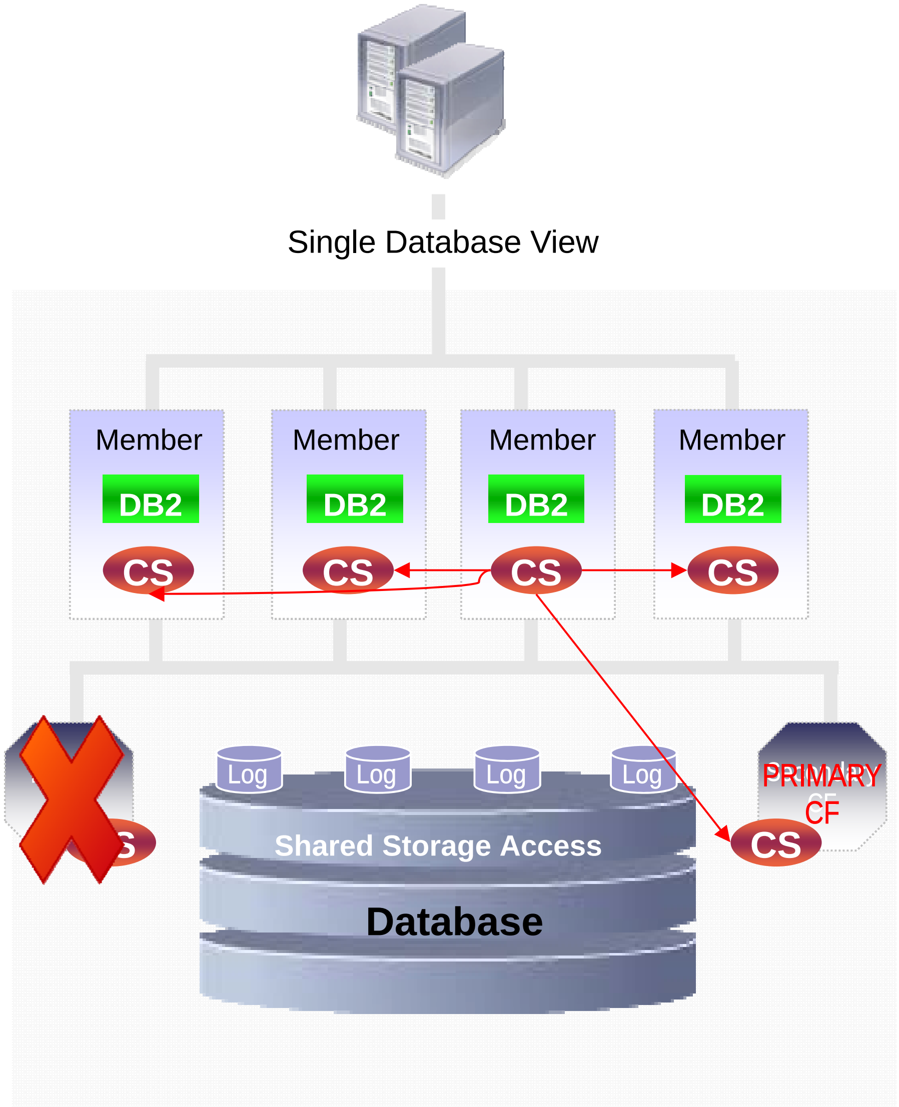
Failure Management – Member Failback

- Power restored and system re-booted
- DB2 Cluster Services automatically detects system availability
  - Informs other members and CFs
  - Removes fence
  - Brings up member on home host
- Client connections automatically re-routed back to member



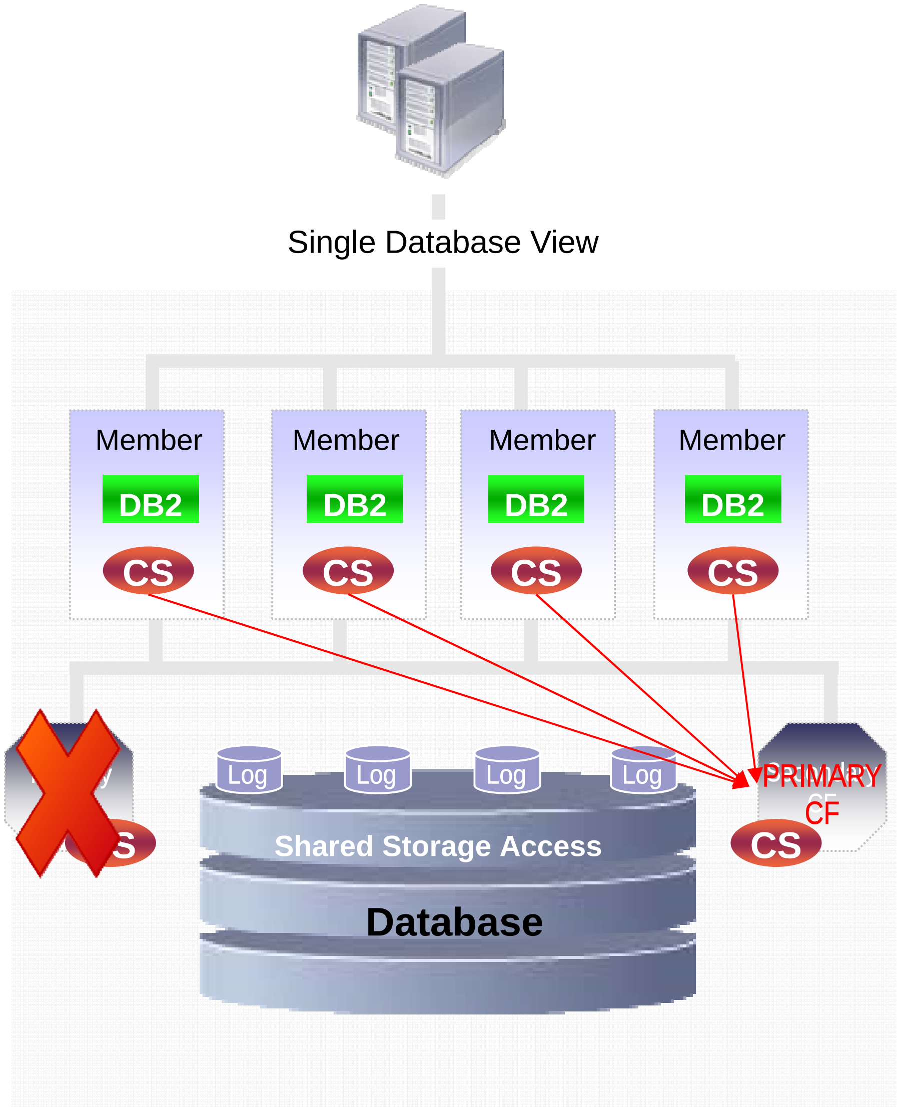
Failure Management – Primary CF Failure

- Power cord tripped over accidentally
- DB2 Cluster Services loses heartbeat and declares primary down
  - Informs members and secondary
  - CF service momentarily blocked
  - All other database activity that does not require a CF proceeds normally
    - E.g. accessing pages in local buffer pool, existing locks, sorting, aggregation, etc.



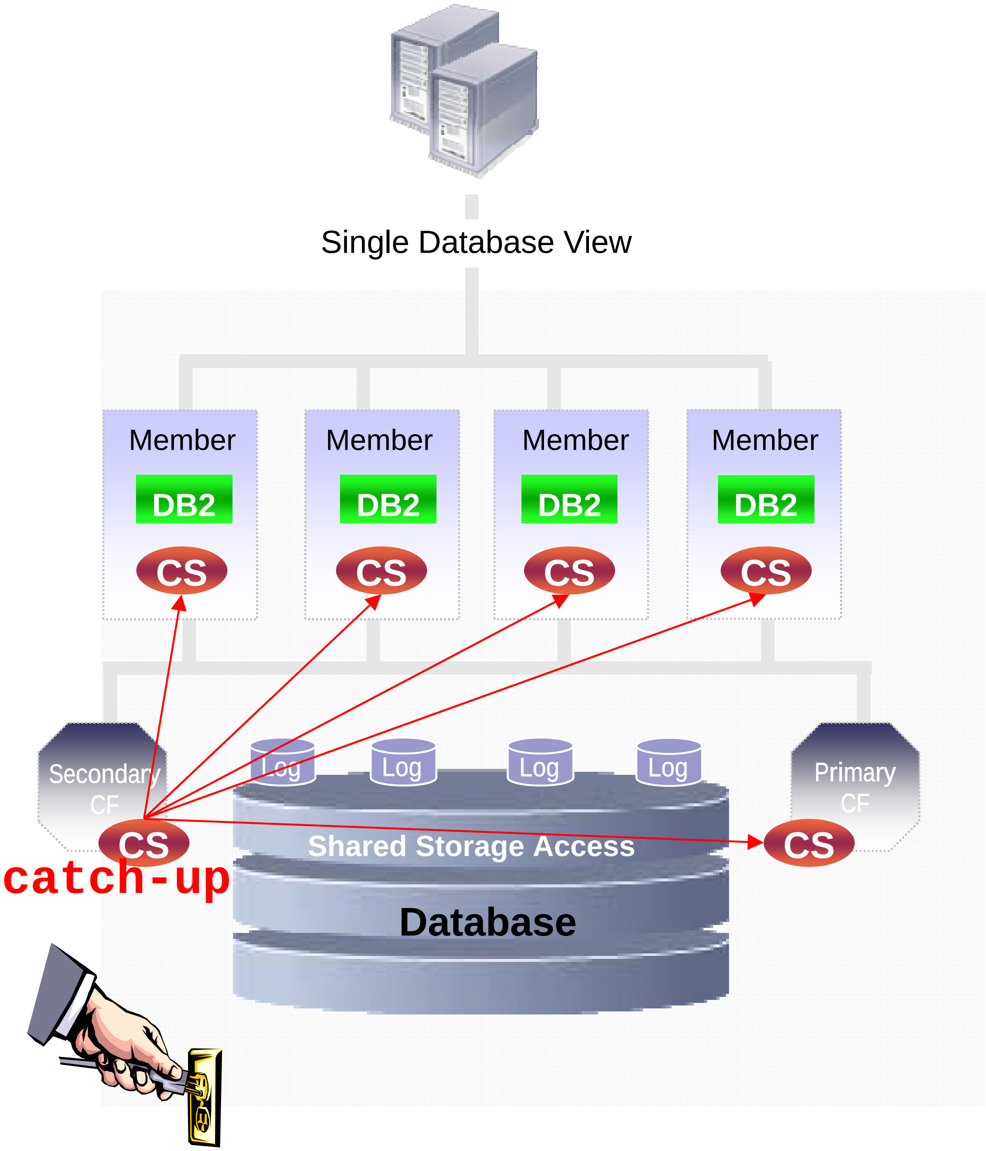
Failure Management – Primary CF Failure

- Members send missing data to secondary
  - E.g. read locks
- Secondary becomes primary
  - CF service continues where it left off
  - No errors are returned to DB2 members



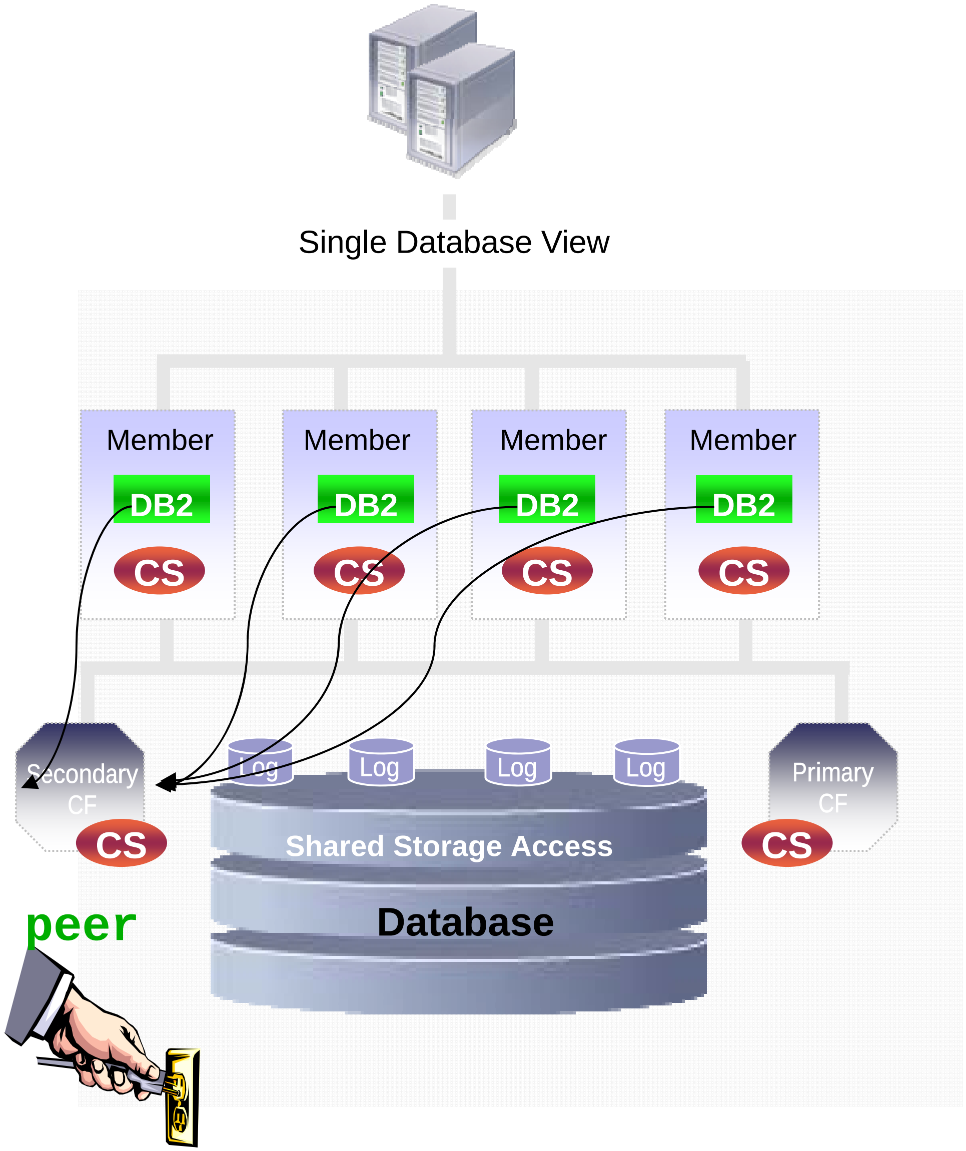
Failure Management – CF Re-integration

- Power restored and system re-booted
- DB2 Cluster Services automatically detects system availability
  - – Informs members and primary CF



Failure Management – CF Re-integration

- New system assumes secondary role in ‘catch-up’ state
  - Members resume duplexing
  - Members asynchronously send lock and other state information to secondary
- Catch-up complete
  - Secondary in peer state (contains same lock and page state as primary)



pureScale Deployment

- **pureScale on Power hardware**
  - Ideal for database workload consolidation
  - More powerful servers with high system utilization
  - Scaling by adding additional cores on existing servers or by adding more servers
  - Example Workloads: ERP, Forecasting & demand planning, Trading platforms

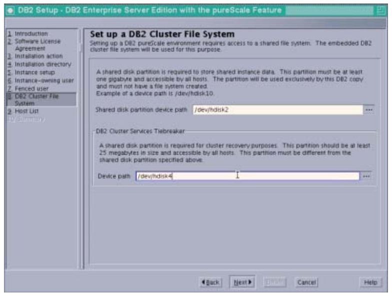
- **pureScale on Linux on x86 based hardware**
  - Ideal for active/active critical systems
  - Simple mass deployment of high availability for critical systems
  - Scaling in small units of additional servers
  - Example Workloads: Critical homegrown workloads, ISV applications, Risk Management

DB2 pureScale Installation Summary

- Start with pre-requisite setup, then follow the steps:

1. Copy image locally

2. Use db2setup to install pureScale on all nodes



DB2 Setup - DB2 Enterprise Server Edition with the pureScale Feature

Set up a DB2 Cluster File System

Setting up a DB2 pureScale environment requires access to a shared file system. The embedded DB2 cluster file system will be used for this purpose.

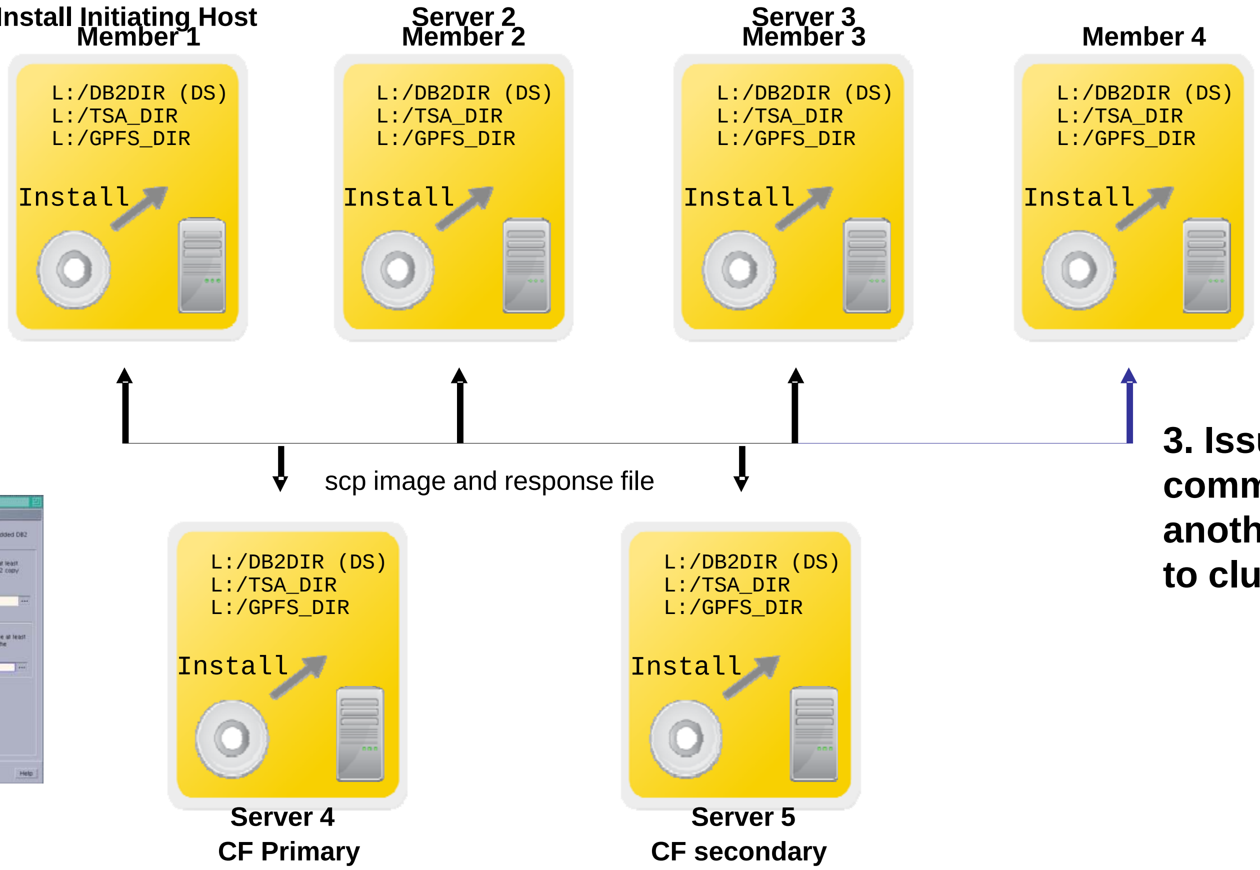
A shared disk partition is required to store shared instance data. This partition must be at least one gigabyte and accessible by all hosts. The partition will be used exclusively by this DB2 copy and must not have a file system created.
Example of a device path is /dev/hdisk10.

Shared disk partition device path: /dev/hdisk2

DB2 Cluster Services Tiebreaker

A shared disk partition is required for cluster recovery purposes. This partition should be at least 25 megabytes in size and accessible by all hosts. This partition must be different from the shared disk partition specified above.

Device path: /dev/hdisk4



All installation components copied over and installed

3. Issue db2iupdt command to add another member to cluster

Best practices for pureScale in a production environment

- **Duplex CFs (i.e., having a primary and a secondary):**
  - Synchronous duplexing of changes to the secondary keeps it in peer state, ready to take over if the primary fails
  - Using a single CF, you will have a single point of failure
- **One member per host:**
  - A host is either a physical system, or an LPAR (a host is a single instance of AIX)
  - Multiple members per host can be used in a development or QA environment
- **At least two machines:**
  - At a bare minimum, 2 machines should be used, each hosting a CF and a member
  - If only 2 physical AIX machines, ensure a CF on each physical box and members need to be split evenly between both boxes.
- **Use SCSI 3 Persistent Reserve if available:**
  - Provides rapid fencing of failed members (prevent I/O to shared disk of failed machine)
  - Reduces time of failover and fallback
- **If client affinity is needed, use it for:**
  - Help consolidate separate workloads/applications on same database infrastructure
  - Minimize total resource requirements for disjoint workloads

Enhancements in DB2 10

- **Added support to Range-partitioned tables**
  - **All roll-in and roll-out operations**
    - ADD/ATTACH/DETACH PARTITION

```
ALTER TABLE SALES ATTACH PARTITION pt1 on SALES …
```

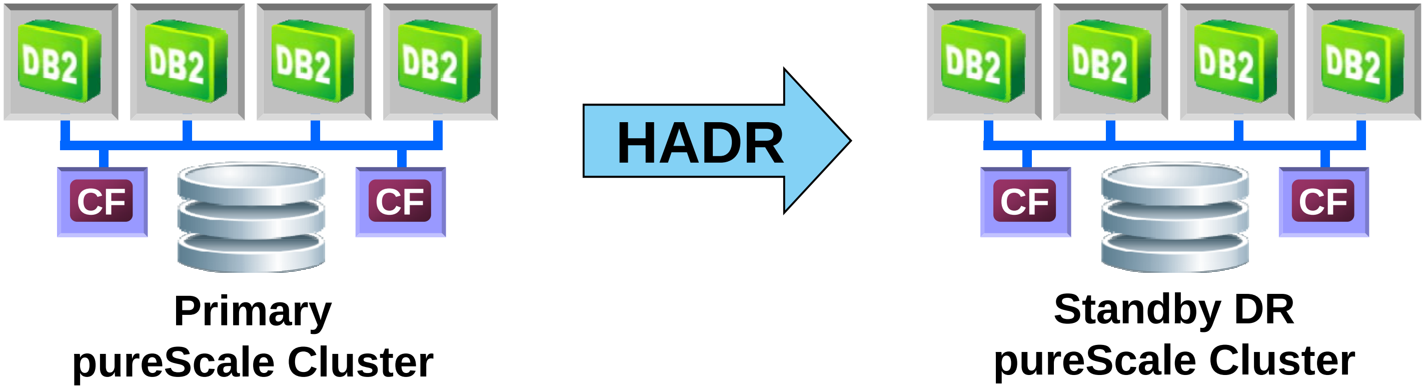
    - Asynchronous partition attach will start
    - This is only run on a single member
    - It may be different from the member that issued the attach
  - **Leverage Partitioned indexes and Partition REORGs**
- **DB2 Workload Manager** now available with DB2 pureScale
- **Using a split mirror as a backup image**
  - Added support to SET WRITE operations
- **New CURRENT MEMBER default value improves DB2 pureScale performance**
  - This member information can then be used to range partition a table or an index, and therefore reduce database contention.

DB2 10.5 pureScale Enhancements
Enhanced availability, optimized for OLTP Workloads

- **DB2 pureScale**
  - Robust infrastructure for OLTP workloads
  - Provides improved availability, performance and scalability
  - Application transparency
- **NEW pureScale enhancements**
  - Rich disaster recovery options
    - Integrated HADR support
    - QReplication and CDC
  - Improved administrative capabilities
    - Backup and restore between pureScale and non-pureScale environments
    - Snapshot backup scripts
    - Online fix pack updates
    - Add members online for additional capacity
  - Autotonomic improvements
    - Per member self tuning memory management
    - Member subsetting
    - Higher availability characteristics
  - POWER 8 optimizations
  - Random Key Indexes
  - Included in Advanced Workgroup and Advanced Enterprise editions

HADR in DB2 pureScale

- Integrated disaster recovery solution
  - – Very simple to setup, configure, and manage
- Support includes
  - – Asynchronous, super asynchronous modes
  - – Time delayed apply
  - – Log spooling
  - – Both non-forced (role switch) and forced (failover) takeovers
- Member topology must match between primary and standby clusters
  - – Different physical configuration allowed (less resources, sharing of LPAR, etc.)

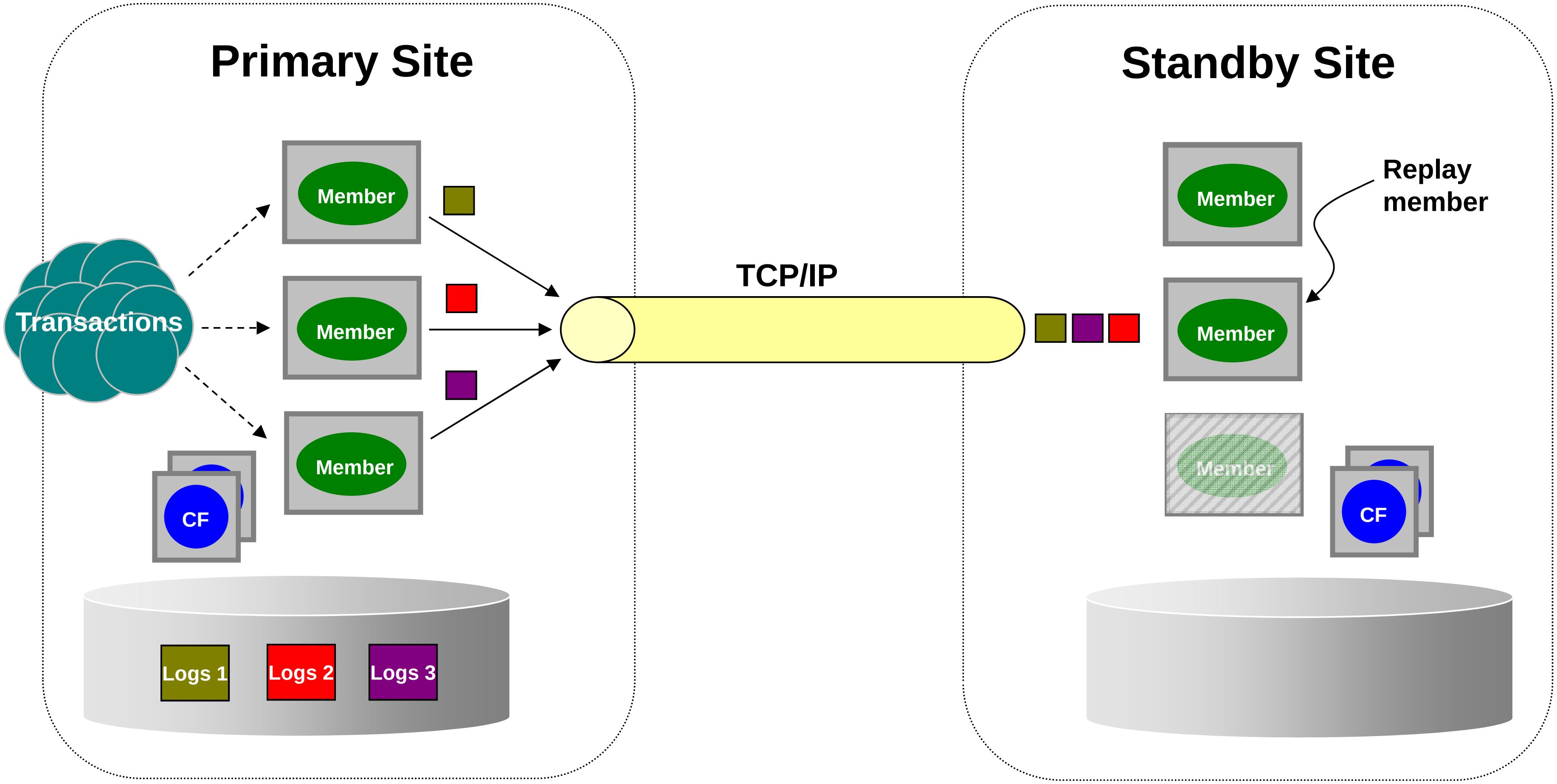


HADR in DB2 pureScale: New Concepts

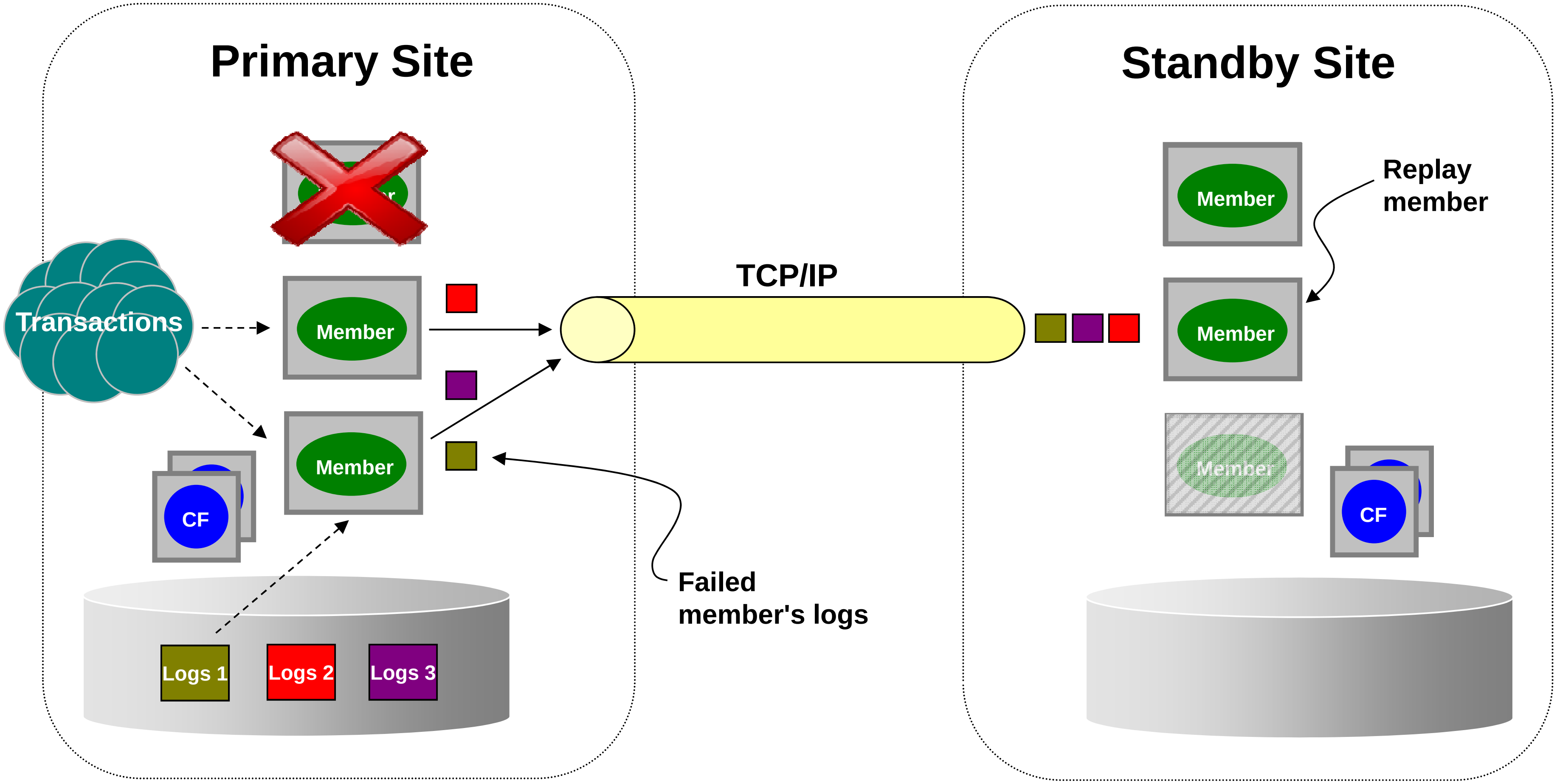
- Database only activated on one member in the standby cluster
  - Referred to as the replay member

- Can choose preferred replay member
  - May want to configure a member with more CPU power and memory
  - Member HADR started on is the preferred replay member
  - If replay member goes down normally or abnormally, DB2 will automatically migrate replay to another healthy member
- All primary members connect to replay member and send logs via TCP/IP
- Replay member on standby merges and replays the log streams
- If member in primary cluster fails or cannot connect to standby, logs for member shipped indirectly by another member to standby
  - Referred to as assisted remote catchup

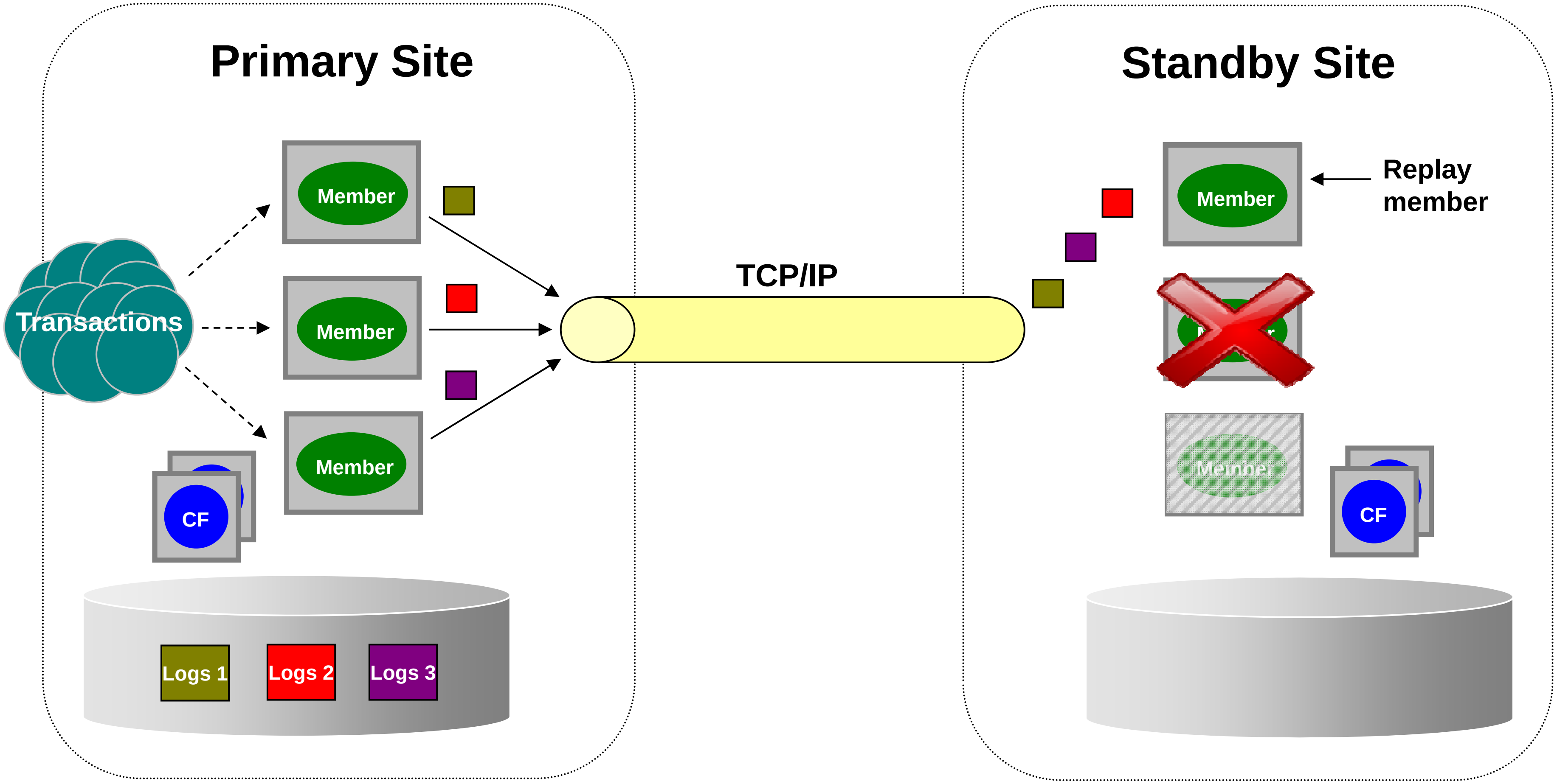
HADR in DB2 pureScale: Example



HADR in DB2 pureScale: Example



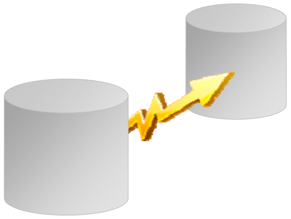
HADR in DB2 pureScale: Example



New Features in DB2 10.5 pureScale

- **Rolling Fix Pack Updates**
  - Transparently install pureScale fix packs or perform system maintenance in an online rolling fashion
  - No outage experienced by applications
  - Single `installFixPack` command run on each member/CF
  - Final `installFixPack` command to complete and commit updates
- **Online Add Member**
  - New members can be added to an instance while it is online
  - No change in add member command
    - `db2iupdt –add –m <newHost> -mnet <networkName> <instance>`
  - Offline backup no longer needed after adding new members
- **Topology-Changing Backup and Restore**
  - Backup and restore between topologies with differing numbers of members
  - Backup and restore from DB2 pureScale to non-DB2 pureScale (and vice-versa)
- **Multi Tenancy : Member Subsets**
  - Point applications to subsets of members which enables
    - Isolation of batch from transactional workloads
    - Multiple databases in a single instance to be isolated from each other

Snapshot Backup Scripts



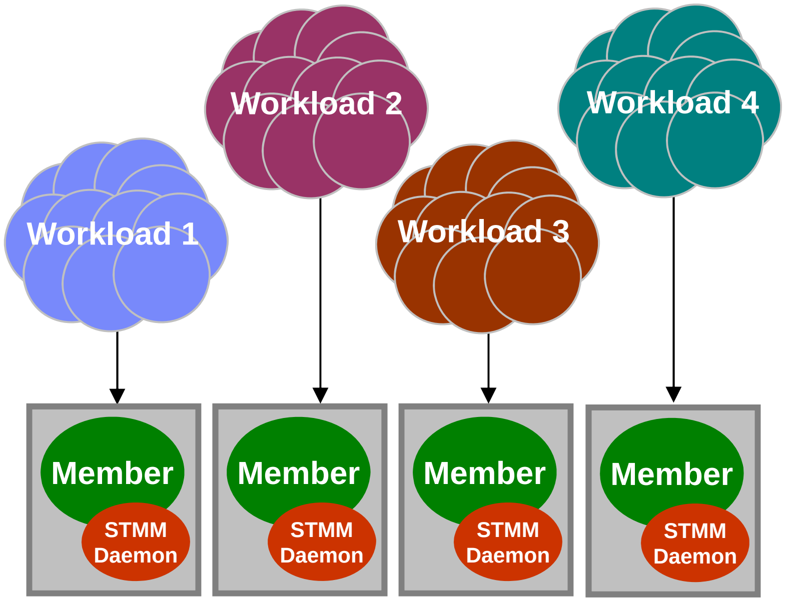
- Allows for integrated snapshot backup capabilities for those storage devices not supported by DB2 Advanced Copy Services (ACS)
  - Works with pureScale
- Custom script implements the DB2 ACS API
  - Users or storage vendors can write their own scripts
  - Write operations to the database are automatically suspended and resumed by DB2 during the backup process
- Benefits include
  - Wider storage support
  - Avoids need for manual snapshot backup process in pureScale
    - Manually running SET WRITE SUSPEND, SET WRITE RESUME, db2inidb, and storage vendor commands can be error prone
  - History file record is generated

```
BACKUP DATABASE PRODDATA USE SNAPSHOT SCRIPT '/scripts/snapshot.sh'
RESTORE DATABASE PRODDATA USE SNAPSHOT SCRIPT '/scripts/snapshot.sh'
  TAKEN AT 20130614120000
```

Multi-Tenancy: Self-Tuning Memory Management (STMM)

- Prior DB2 pureScale STMM design
  - Single tuning member makes local tuning decisions based on workload running on that member
    - Other member becomes tuning member in case of member failure
  - Broadcasts tuning decisions to other members
  - Works well in single homogeneous workload scenarios

- DB2 pureScale now allows per-member STMM tuning
  - Workload consolidation
  - Multi-tenancy
  - Batch workloads
  - Affinitized workloads



Random Key Indexes

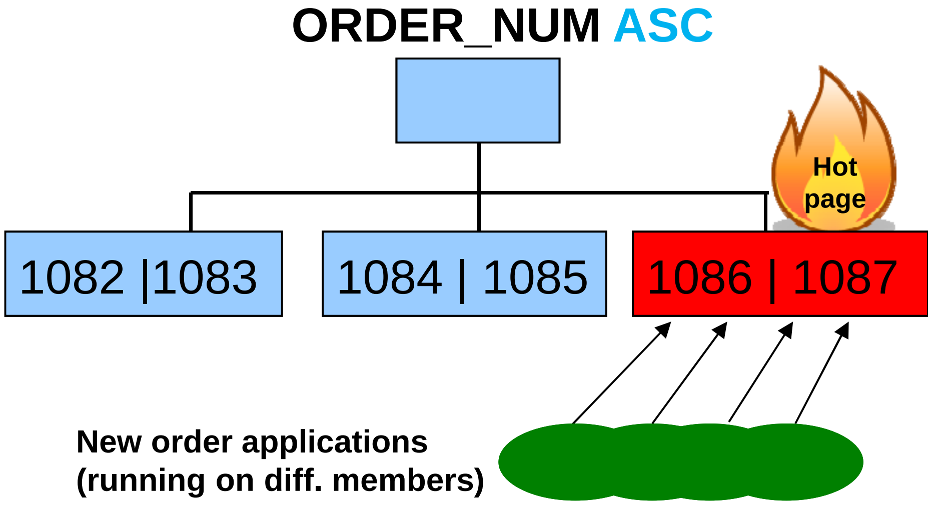
- Some workloads may experience page contention on frequently accessed index leaf pages
  - For example, an index key on a monotonically incrementing value
    - Such as a timestamp or identity column
  - Results in a "hot" index leaf page, which is the insert point for all new keys being generated
  - A typical example is an "order number" column in a retail industry schema
- Issue may be exacerbated in pureScale where pages are being shared across members
  - Hot index page gets reclaimed/negotiated over and over again between those members (routed through the CF)
- Random key indexes solve this issue

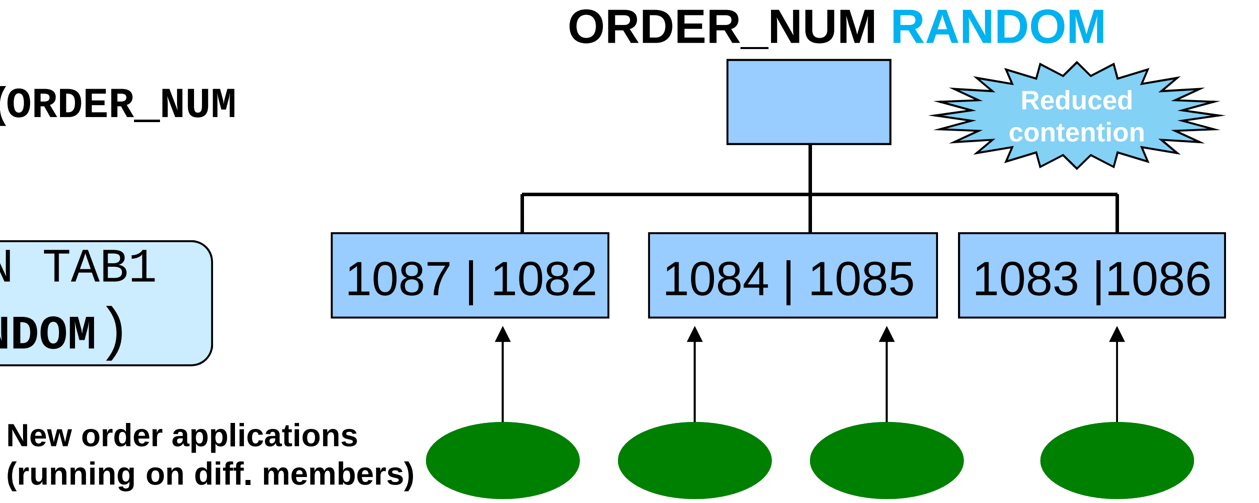
```
CREATE INDEX IX1 ON TAB1 (INT ORDER_NUM RANDOM)
```

Random Key Indexes (cont.)

- **Random key indexes allow you to randomize the placement of index key values**
- **Spreads out contention of the index high key**
- **Loss of order, so range queries become full index scans**
- **Allows for equality lookups (`ORDER_NUM = 1083`)**

```
CREATE INDEX IX1 ON TAB1
(INT ORDER_NUM RANDOM)
```





DB2 10.5 Cancun – pureScale
Simplified Deployment and Administration

- Simplified Deployment
  - TCP/IP interconnect (Sockets) with identical features to traditional RDMA-based (Infiniband/10GE) pureScale
  - VMWare and KVM support
  - Cluster Caching Facility with Self-Tuning Memory
  - Additional GDPC configurations & implementation services
  - Support for IBM POWER8 Hardware
- Administration
  - Online table re-orgs, Incremental Backup/Restore, Snapshot backups, DB2 Merge Backup Support
  - Additional OPM metrics for pureScale and HADR
  - Improved diagnostics, error detection, and upgrade all members & CF's in parallel
- Application Development
  - Federated Two phase commit and Spatial Extender Support
- Faster time to value with improved serviceability of installation, configuration, and updates
  - Parallelized DB2 instance upgrade of member and CFs

Summary

- Deliver higher levels of scalability and superior availability
- Continuous availability during regular maintenance or failures
- Improved SLA attainment
- Lower overall costs for applications that require high transactional performance and ultra high availability
- Single installation package for WSE, ESE and AESE
- DB2 10.5 pureScale has been enhanced to include things like
  - Split mirror technology
  - Multiple cluster-interconnects for the CFs and members
  - Multiple switches
  - Range-Partitioned tables
  - Workload Management
  - Online add member
  - Rolling fix pack support

The next steps…

The Next Steps…

- Complete the online quiz for this module
  - – Log onto SKI, go to “My Learning” page, and select the “In Progress” tab.
  - – Find the module and select the quiz
- Provide feedback on the module
  - – Log onto SKI, go to “My Learning” page
  - – Find the module and select the “Leave Feedback” button to leave your comments

Questions?
askdata@ca.ibm.com