Unleashing DB2 pureScale

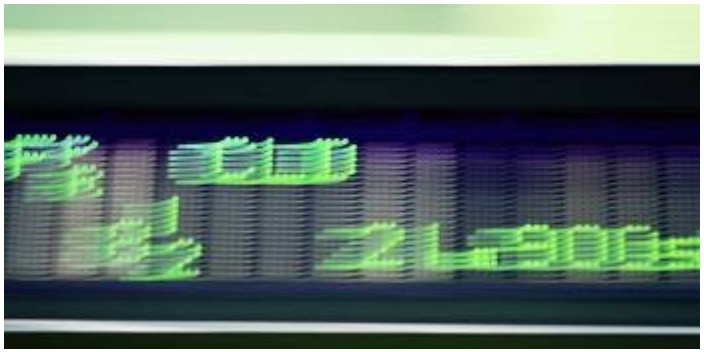
Drew Bradstock,
Program Director, DB2 Product Management
Aug 31, 2010

ibm.com/db2/labchats

Critical IT Applications Need Reliability and Scalability

- **Local Databases are Becoming Global**
  - Successful global businesses must deal with exploding data and server needs
  - Competitive IT organizations need to handle rapid change

Customers need a highly scalable, flexible solution for the growth of their information with the ability to easily grow existing applications



- **Down-time is Not Acceptable**
  - Any outage means lost revenue and permanent customer loss
  - Today's distributed systems need reliability

IT Needs to Adapt in Hours…Not Months

- **Handling Change is a Competitive Advantage**
- **Dynamic Capacity is not the Exception**
  - Over-provisioning to handle critical business spikes is inefficient
  - IT must respond to changing capacity demand in days, not months

Businesses need to be able grow their infrastructure without adding risk

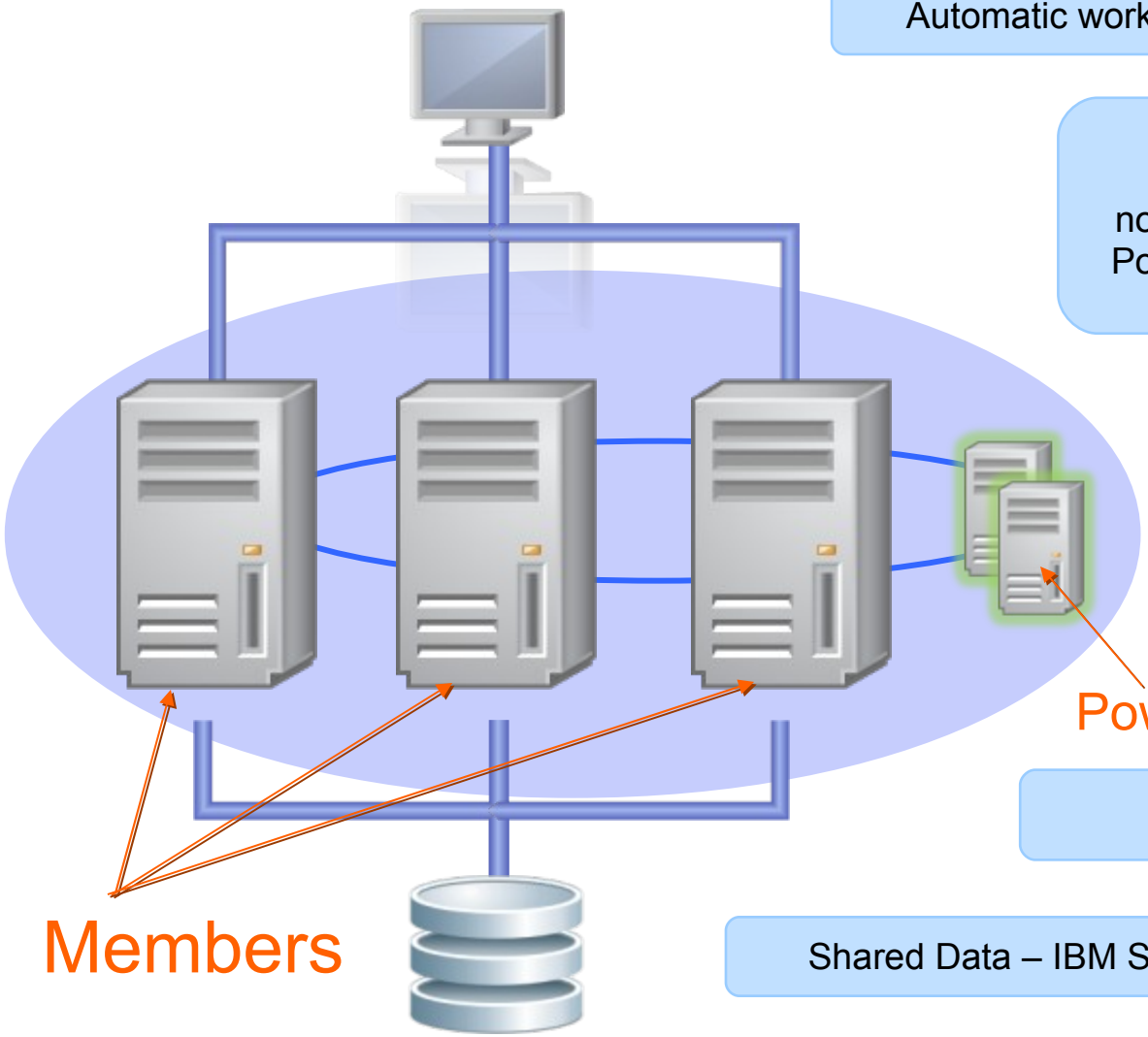
- **Application Changes are Expensive**
  - Changes to handle more workload volume can be costly and risky
  - Developers rarely design with scaling in mind
  - Adding capacity should be stress free

DB2 pureScale

- **Unlimited Capacity**
  - Buy only what you need, add capacity as your needs grow
- **Application Transparency**
  - Avoid the risk and cost of application changes
- **Continuous Availability**
  - Deliver uninterrupted access to your data with consistent performance

Learning from the undisputed Gold Standard... System z

DB2 pureScale Architecture



Automatic workload balancing

Cluster of DB2 nodes/members running on Power or System X servers

PowerHA pureScale technology drives the clustering,
Runs on an LPAR or a stand alone server

PowerHA pureScale Servers

InfiniBand network

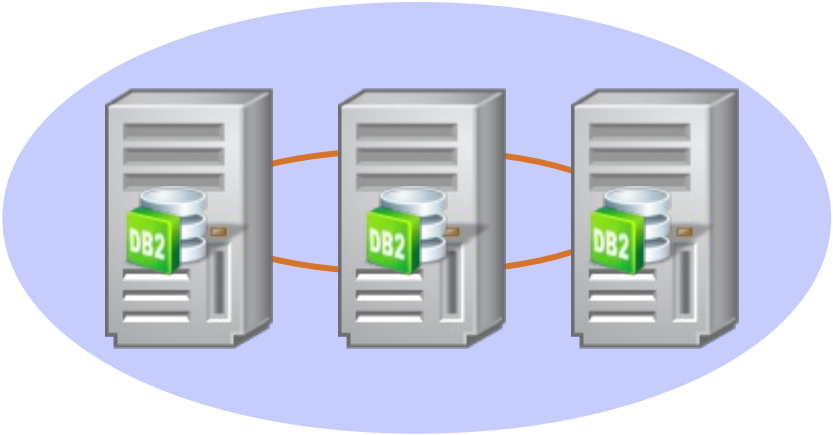
Members

Shared Data – IBM Storage supported

Unlimited Capacity

- DB2 pureScale has been designed to grow to whatever capacity your business requires
- Flexible licensing designed for minimizing costs of peak times
- Only pay for additional capacity when you use it – even if only for a single day

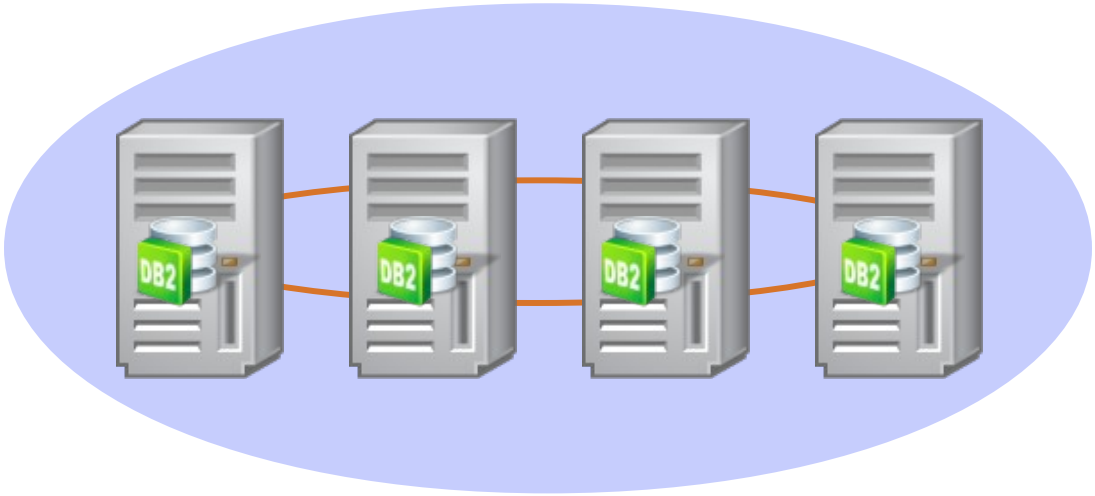
Issue:
All year, except for two days, the system requires 3 servers of capacity.



Unlimited Capacity

- DB2 pureScale has been designed to grow to whatever capacity your system requires
- Flexible licensing designed for minimizing costs of peak times
- Only pay for additional capacity when you use it – even for a single day

Solution:
Use DB2 pureScale and add another server for those two days, and only pay software license fees for the days you use it.

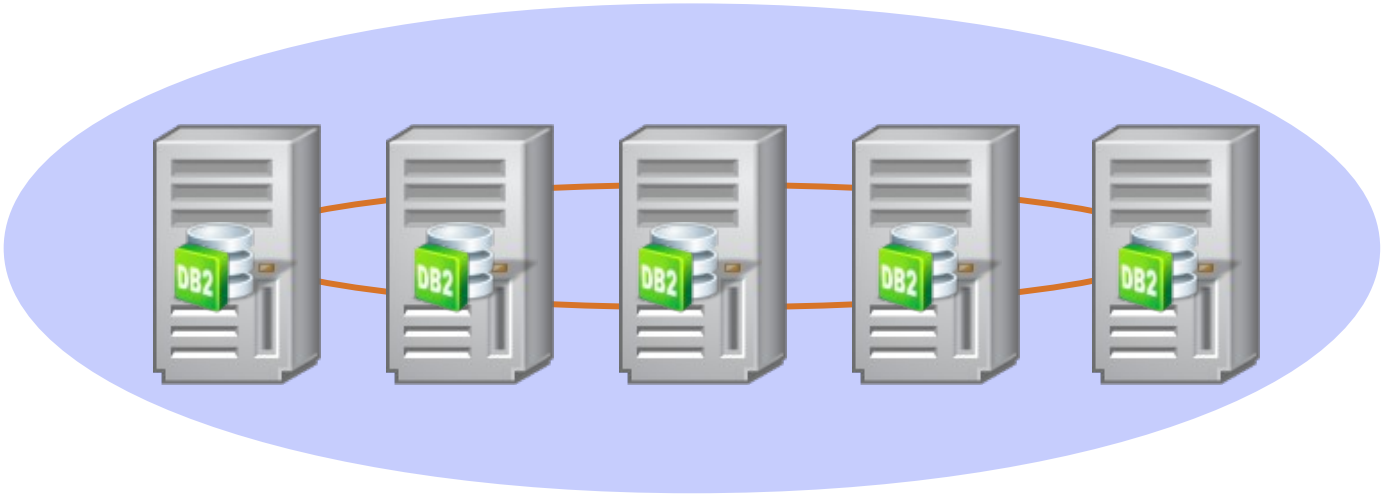


DB2 pureScale helps CIOs handle business critical peak periods & save costs

Unlimited Capacity

- DB2 pureScale has been designed to grow to whatever capacity your system requires
- Flexible licensing designed for minimizing costs of peak times
- Only pay for additional capacity when you use it – even for a single day

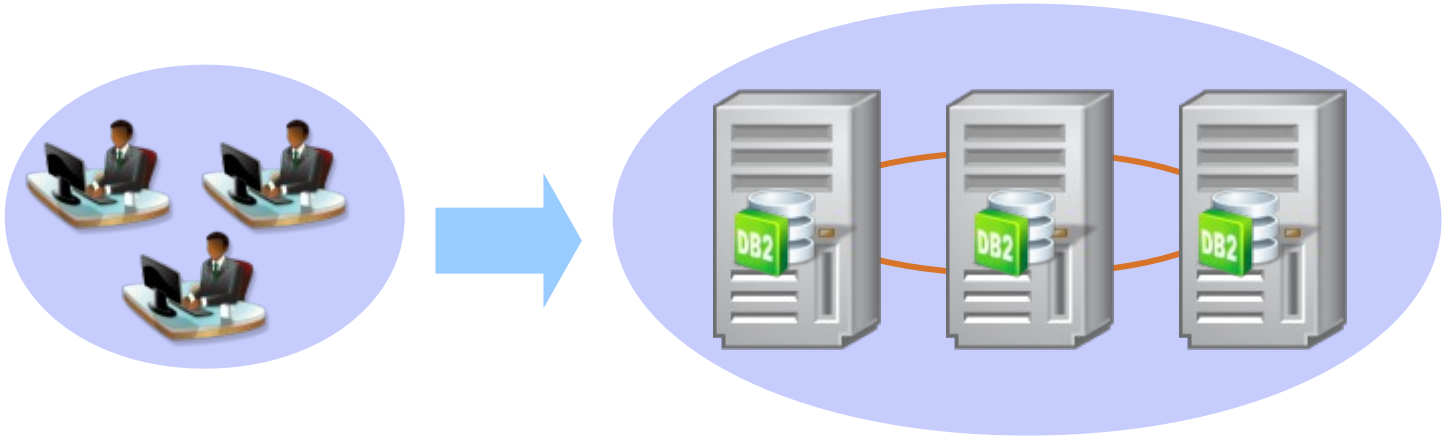
Need more?
Just deploy another server and then turn off DB2 when you're done.



Over 100+ node architecture validation has been run by IBM

Application Transparency

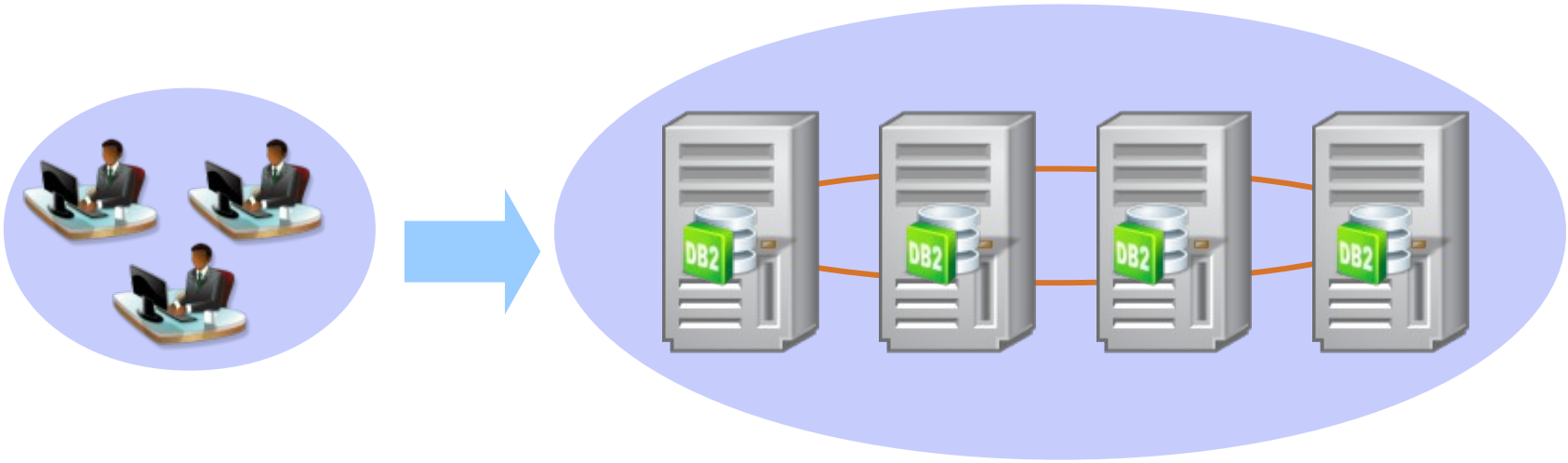
- Avoid the risk and cost of application changes
- Take advantage of extra capacity instantly
  - No need to modify your application code
  - No need to re-tune your infrastructure



Your DBAs can add capacity without re-tuning or re-testing

Application Transparency

- Avoid the risk and cost of application changes
- Take advantage of extra capacity instantly
  - No need to modify your application code
  - No need to re-tune your infrastructure



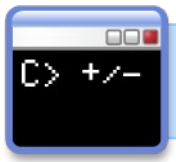
Your developers don't even need to know more nodes are being added

DB2 pureScale is Easy to Deploy

Single installation for all components

Monitoring integrated into Optim tools

Single installation for fixpaks & updates



Simple commands to add and remove members

Continuous Availability

- Protect from infrastructure outages
  - Architected for no single point of failure

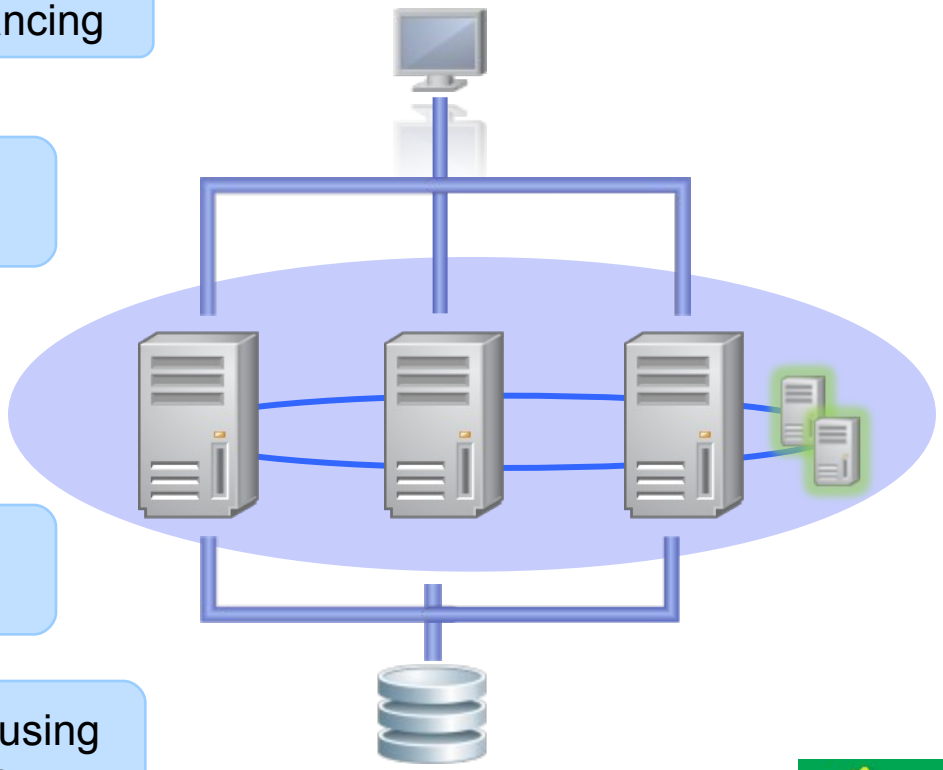
Automatic workload balancing

Duplexed secondary global lock and memory manager

Tivoli System Automation automatically handles all component failures

DB2 pureScale stays up even with multiple node failures

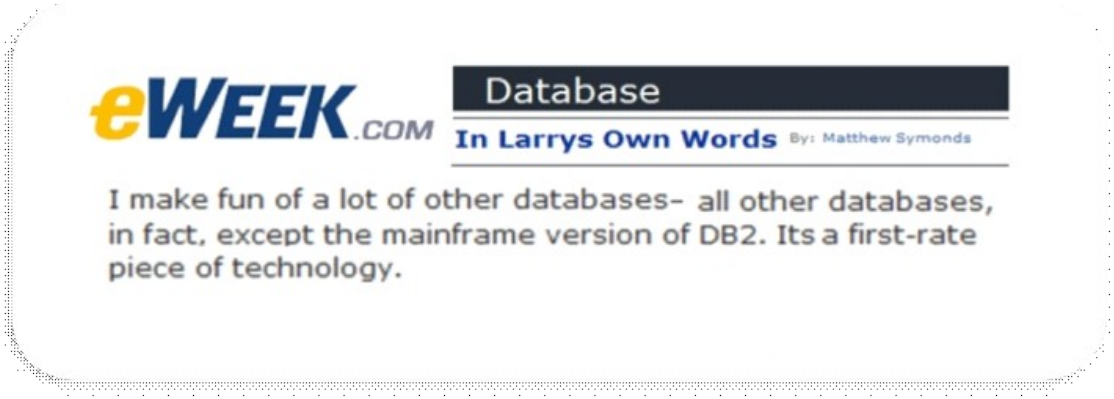
Shared disk failure handled using disk replication technology



DB2 for z/OS Data Sharing is the Gold Standard

- **Everyone recognizes DB2 for z/OS as the "Gold" standard for scalability and high availability**

- **Even Oracle agrees:**

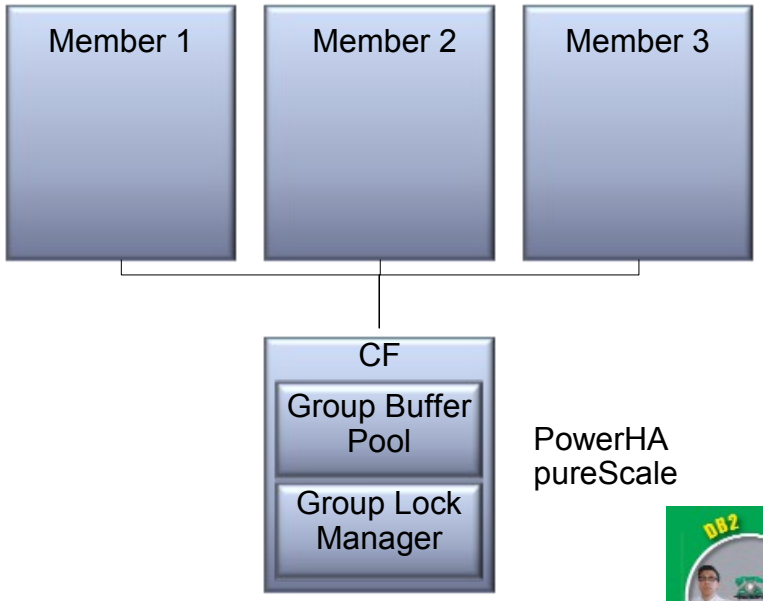


- **Why?**
  - **The Coupling Facility!!**
    - Centralized locking, centralized buffer pool deliver superior scalability and superior availability
  - **The entire environment on z/OS uses the Coupling Facility**
    - CICS, MQ, IMS, Workload Management, and more

PowerHA pureScale with RDMA – the secret sauce!

- **Efficient Centralized Locking and Caching**
  - As the cluster grows, DB2 maintains one place to go for locking information and shared pages
  - Optimized for very high speed access
    - DB2 pureScale uses **Remote Direct Memory Access (RDMA)** to communicate with the powerHA pureScale server
    - No IP socket calls, no interrupts, no context switching

- **Results**
  - **Near Linear Scalability to large numbers of servers**
  - Constant awareness of what each member is doing
    - If one member fails, no need to block I/O from other members
    - Recovery runs at memory speeds

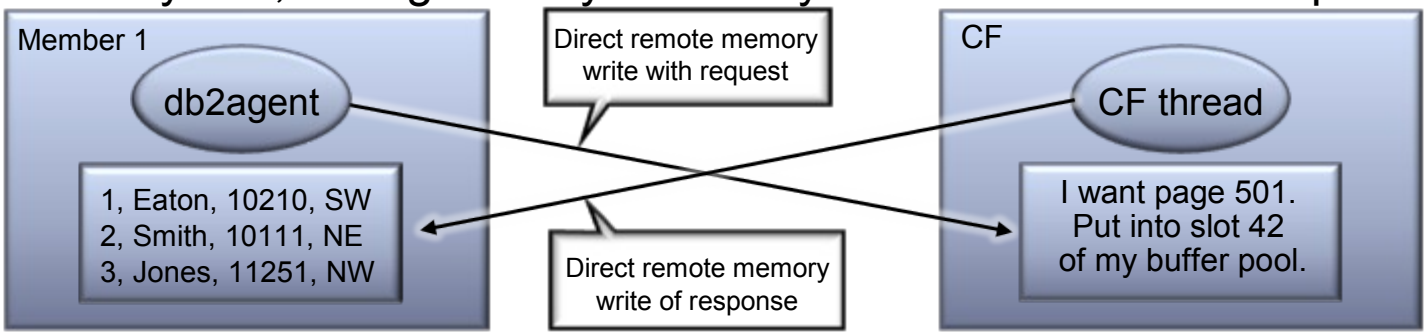


The Advantage of DB2 Read and Register with RDMA

1. DB2 agent on Member 1 writes directly into CF memory with:
   - Page number it wants to read
   - Buffer pool slot that it wants the page to go into

1. CF either responds by writing directly into memory on Member 1:
   - That it does not have the page **or**
   - With the requested page of data

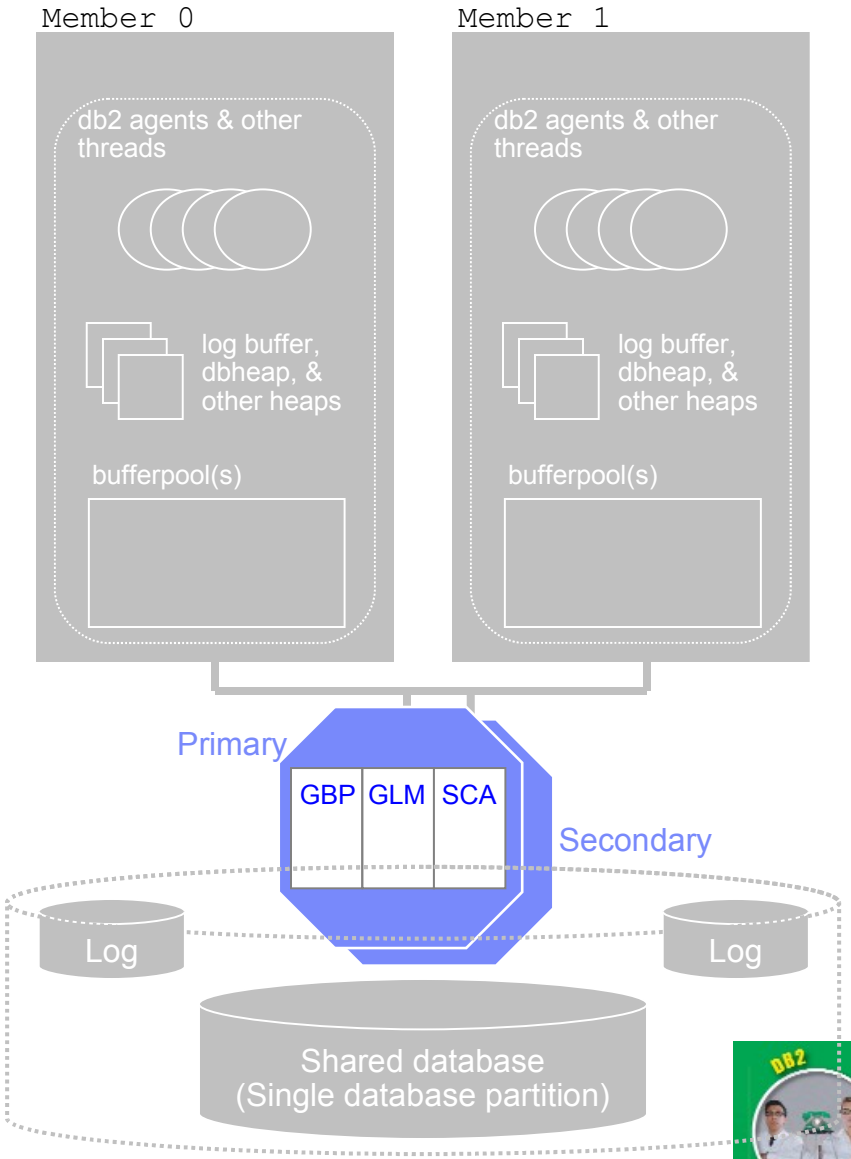
- Total end to end time for RAR is measured in microseconds
- Calls are very fast, the agent may even stay on the CPU for the response



Much more scalable, does not require locality of data

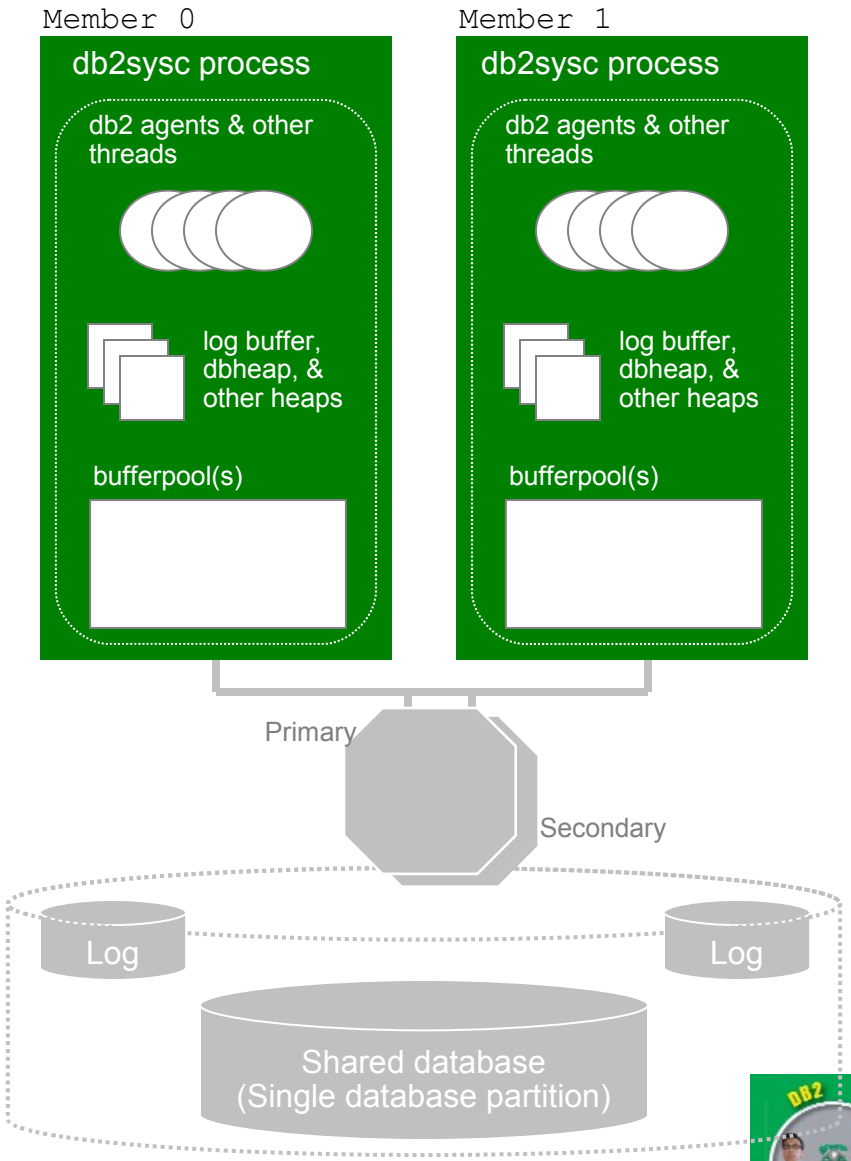
What is a PowerHA pureScale Server?

- **Software technology that assists in global buffer coherency management and global locking**
  - Derived from System z Parallel Sysplex & Coupling Facility technology
  - Software based
- **Services provided include**
  - Group Bufferpool (GBP)
  - Global Lock Management (GLM)
  - Shared Communication Area (SCA)
- **Members duplex GBP, GLM, SCA state to both a primary and secondary**
  - Done synchronously
  - Duplexing is optional (but recommended)
  - Set up automatically, by default



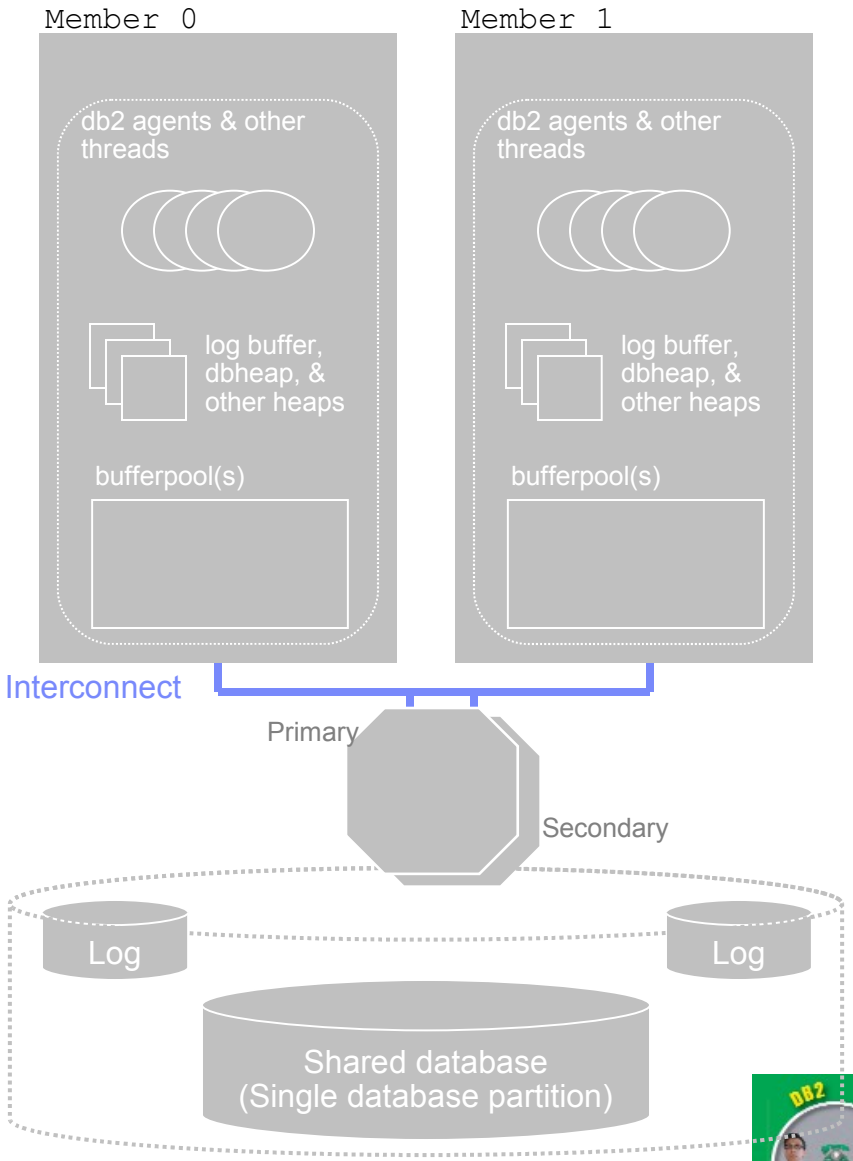
What is a Member ?

- A DB2 engine address space
- Members Share Data
  - All members access the same shared database
  - Aka "*Data Sharing*"
- Each member has it's own …
  - Bufferpools
  - Memory regions
  - Log files
- Members are logical.
  Can have …
  - 1 per machine or LPAR (recommended)
  - >1 per machine or LPAR (not recommended for production)
- Member != Database Partition
  - *Member* = db2sysc process
  - *Database Partition* = a partition of the database



Cluster interconnect

- **Requirements**
  1. Low latency, high speed interconnect between members, and the primary and secondary PowerHA pure scale servers
  2. RDMA capable fabric
     - To make direct updates in memory without the need to interrupt the CPU
- **Solution**
  - InfiniBand (IB) and uDAPL for performance
    - **InfiniBand** supports RDMA and is a low latency, high speed interconnect
    - **uDAPL** to reduce kernel time in AIX



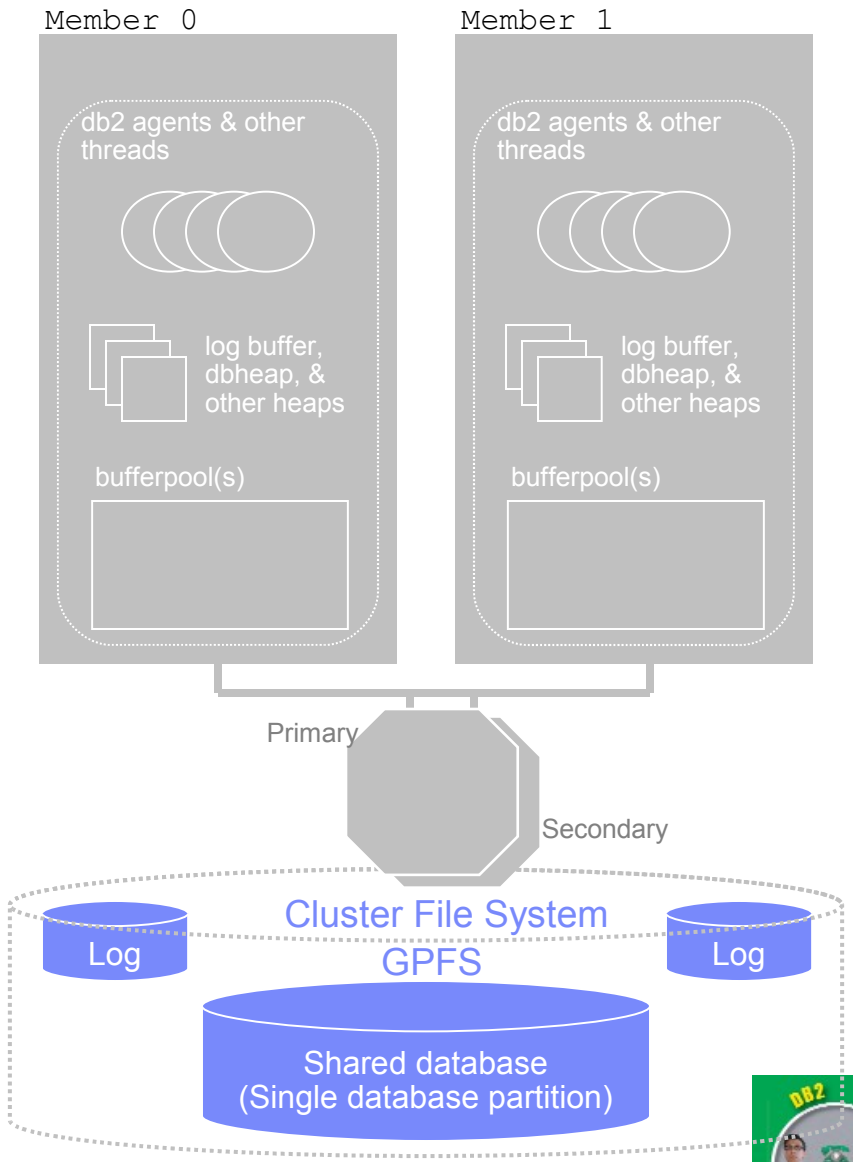
Cluster file system

- **Requirements**
  1. Shared data requires shared disks and a cluster file system
  2. 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

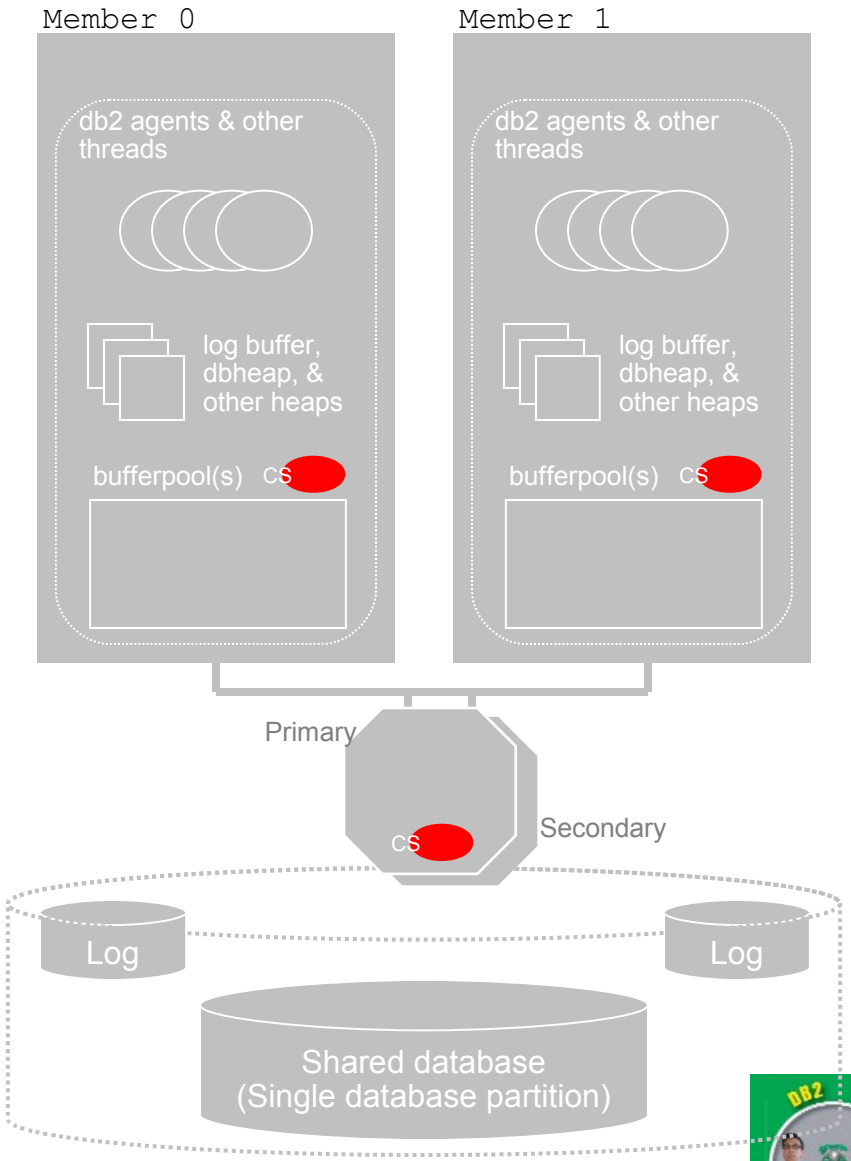
      Allows GPFS to be managed at the same level across the enterprise

      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

- **Orchestrate**
  - Unplanned event notifications to ensure seamless recovery and availability.
    - Member, PowerHA pureScale, AIX, hardware, etc. unplanned events.
  - Planned events
    - 'Stealth' maintenance
- **Integrates with:**
  - RSCT, Tivoli SA MP and GPFS
    - Packaged, shipped and installed with DB2 pureScale



IBM pureScale Application System

Scale Up, Scale Out, Scale Within

Scale seamlessly, effortlessly from 8 cores up to 8,192 cores [1]

IBM Power 770 IBM Power 770

Scale Up

Expand each member
up to 64 cores

Scale Out

Add additional members,
up to 128 total [2]

1 – Assumes 64 cores x 128 members (see note 2)

2 - Architecture proven in lab testing to 128 nodes using pre-release code

IBM pureScale Application System

Scale Up, Scale Out, Scale Within

Scale seamlessly, effortlessly from 8 cores up to 8,192 cores [1]

IBM Power 770

IBM Power 770

Scale Up

Expand each member
up to 64 cores

Scale Out

Add additional members,
up to 128 total [2]

IBM Power 770

1 – Assumes 64 cores x 128 members (see note 2)

2 - Architecture proven in lab testing to 128 nodes using pre-release code

IBM pureScale Application System

Scale Up, Scale Out, Scale Within

Scale seamlessly, effortlessly from 8 cores up to 8,192 cores [1]

IBM Power 770

IBM Power 770

Scale Up
Expand each member
up to 64 cores

Scale Out
Add additional members,
up to 128 total [2]

IBM Power 770

Scale Within
Add additional cores as needed with
Capacity OnDemand

1 – Assumes 64 cores x 128 members (see note 2)

2 - Architecture proven in lab testing to 128 nodes using pre-release code

DB2 pureScale on System X & Linux

- **Initial Choice of System X Servers**
  - **x3850 X5**
    - A 1-4 socket server with 4/6/8-core Nehalem EX processors
    - 64 DDR3 DIMMs

      Can be further expanded with an additional 32 DIMMs with the MAX5 memory drawer
  - **x3690 X5**
    - A 1-2 socket server with 4/6/8-core Nehalem EX processors
    - 32 DDR3 DIMMs

      Can be further expanded with an additional 32 DIMMs with the MAX5 memory drawer
  - **x3650 servers** with PCI-E Gen2 are also supported (currently available)
- **Infiniband**
  - Any Cluster350-qualified IB Switch and Mellanox ConnectX2 IB card
- **Operating System**
  - Novell Suse SLES10 SP
    - SLES11 and RHEL5 to follow in Q4

System x X5 servers with DB2 pureScale

- **Minimal Configuration**
  - A pair of machines, with at least 1 socket populated
    - DB2 member and CF are concurrently active on each machine

      Can dedicate certain cores to DB2 or CF, only DB2 cores count for PVU

- **Typical Configuration**
  - A cluster of 2-4 machines with 2-4 sockets populated
    - 2 machines share the primary and secondary CF
    - 2 additional machines can be used for DB2

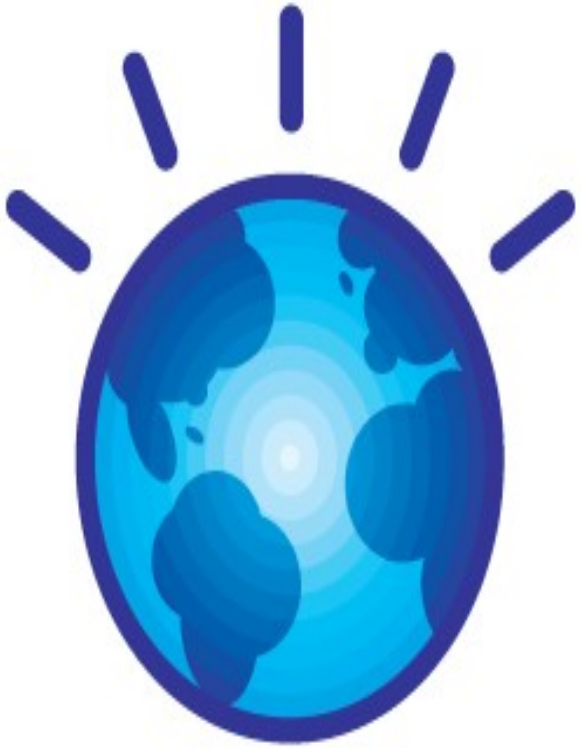
- **Maximum Configuration**
  - A cluster of 10 machines with 4 8-core sockets
    - 2 machines are for the primary and secondary CF
    - 8 additional machines for DB2

      Total of 8 x 4 x 8 = 256 cores of DB2 processing power

Thank You!

ibm.com/db2/pureScale

Contact me at drewkb@ca.ibm.com



Power your planet… Smarter systems for a Smarter Planet

> Questions

Thank You!

ibm.com/db2/labchats

Thank you for attending!