IDUG The Worldwide DB2 User Community

North America Conference 2011

Achieve Breakthrough Performance and Availability with DB2 pureScale

Philip K. Gunning

Gunning Technology Solutions, LLC

Session Code: D01

May 3, 2011 12:45 – 1:45 PM | Platform: LUW

Outline

- Building the case for
  - Setting the stage for DB2 pureScale
- DB2 prior to pureScale
- DB2 pureScale defined
- DB2 pureScale Prerequisites
- Installation and Configuration
- Application Transparency
- Cluster Caching Facility Components
- Monitoring
- Unlimited Scalability and High Availability

The CASE for DB2 LUW

- DB2 for LUW has been providing superior performance, scalability and high availability for many years
- 2003 – Large Student Loan Organization, Mainframe application moved to DB2 for Windows V7.2!
- 2003 – Large State Department of Education Agency, School District Demographic processing, DB2 v7.2 ESE, AIX 5.1, 300 trans per sec
- 2004 – Large Financial Services Co, Oracle to DB2 v7.1 migration, PeopleSoft EDW, AIX 5.1
- 2005 – Major Bus Manufacturer, Oracle to DB2 v8.1 migration, BAAN MRP Application, AIX 5.3

The CASE for DB2 LUW

- 2006 – Large Oil Additive Company, Mainframe application migrated to DB2 for Windows WSE, v8.1 -- Mainframe replaced!
- 2006 – eCommerce Company, DB2 for Windows, WSE, DB2 V8.2 up for 6 months without an outage -- 1550 trans per sec
- 2007 – Credit Report processing application, Interbase to DB2 for LINUX v8.2 migration, RedHat Linux
- 2008 – Major yearbook publication company, migration from MySQL to DB2 9.5 WSE/ESE on SUSE Linux -- 2,000 trans per sec/Disk snapshot technology

The CASE for DB2 LUW

- 2008 – Large School District, DB2 v8.2, PeopleSoft HR and Financials, AIX 5.3, 28,000 employees!
- 2009 – Large School District, Enterprise Data Warehouse Migration from DB2 for z/OS v8 to DB2 9.5 ESE on AIX 6.1, Mainframe application replaced, achieved 10x performance improvement! Disk snapshot technology
- 2010 – Large Financial Services Co, Check scanning application, DB2 for Windows V9.7, WSE – Microsoft Clustering and HADR
- 2010 – Financial Adviser Company, Mutual Fund Pricing Application, DB2 9.5 ESE, SOLARIS

What we had prior to DB2 pureScale

- Superior Performance
  - Inter and Intra-partition parallelism
  - Ability to exploit multiple CPUs/Cores
  - MQT/MDC
  - Range Partitioning – partition elimination
- Almost unlimited scalability
  - Database Partitioning Feature
  - Inter and Intra-partition parallelism
  - Shared nothing
  - Multiple servers/machines

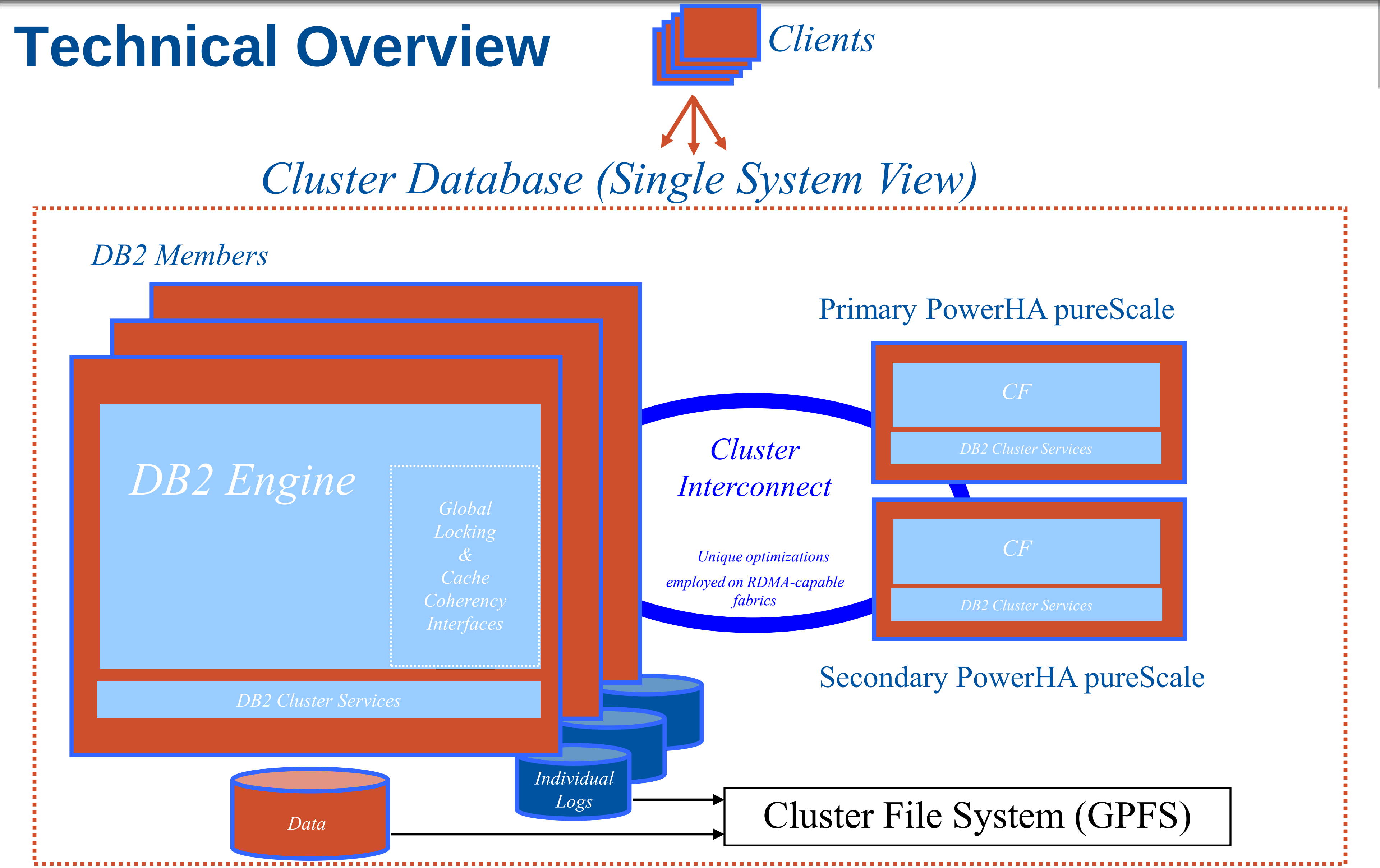
What we had prior to DB2 pureScale

- High Availability
  - Clustering Technology
  - High Availability Disaster Recovery (HADR)
  - Database Backup Snapshot
  - Disk Snapshot Technology
  - Online REORG
  - Schema Evolution (online object changes)
  - Read-only on the STANDBY

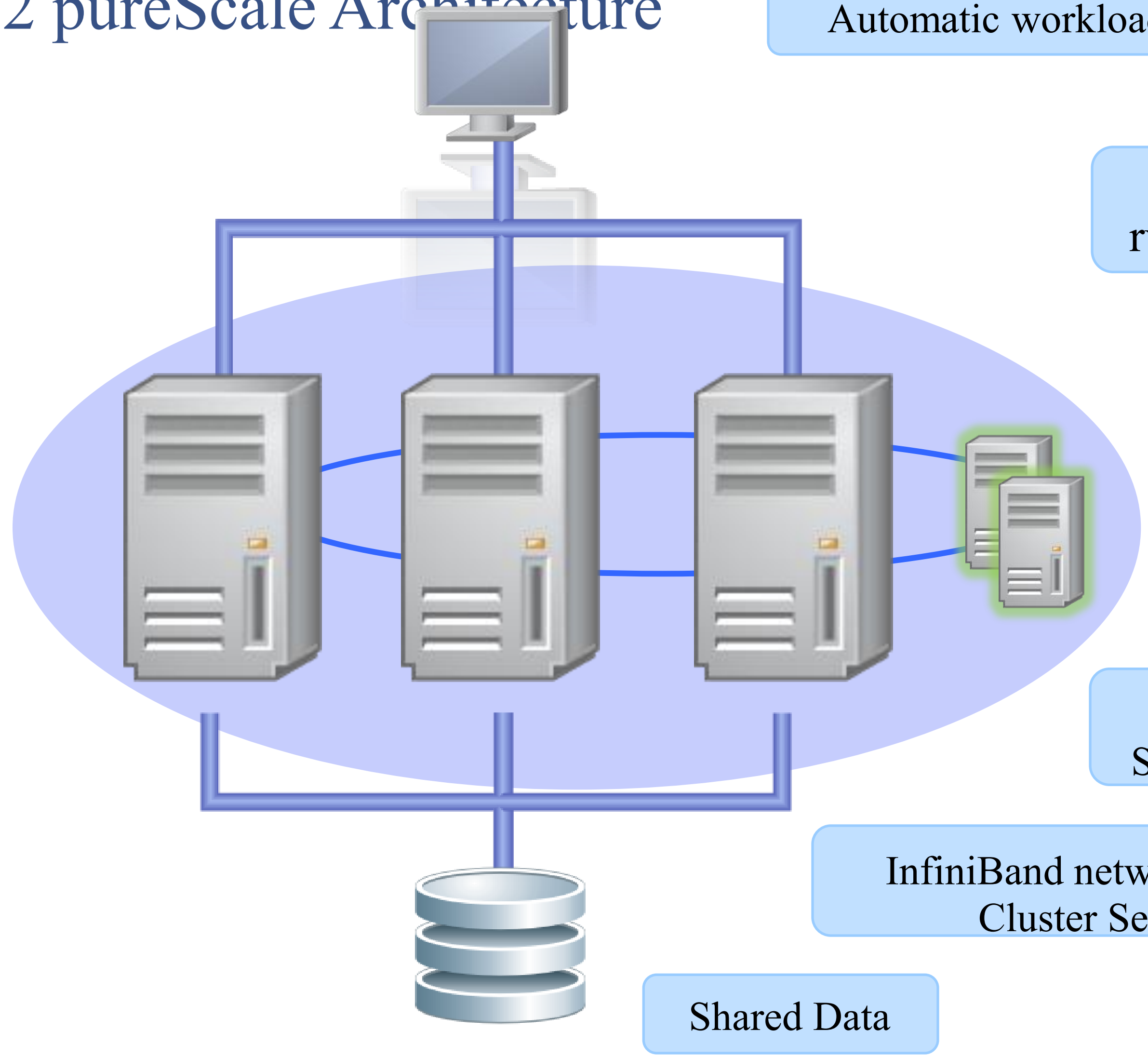
DB2 pureScale Defined

- DB2 pureScale Feature installed on multiple servers referred to as “members”
  - Installed using DB2 Setup GUI or db2_install script
- Members use the following:
  - General Parallel File System for data storage and sharing between members
  - Tie-breaker shared disk
  - Cluster interconnect – InfiniBand fabric
  - Cluster Caching Facility (CF)
  - Specialized server hardware/firmware
  - PowerHA pureScale which is underlying technology behind DB2 pureScale

Technical Overview


Clients
Cluster Database (Single System View)
DB2 Members
DB2 Engine
Global Locking & Cache Coherency Interfaces
DB2 Cluster Services
Cluster Interconnect
Unique optimizations employed on RDMA-capable fabrics
Primary PowerHA pureScale
CF
DB2 Cluster Services
CF
DB2 Cluster Services
Secondary PowerHA pureScale
Individual Logs
Data
Cluster File System (GPFS)

DB2 pureScale Architecture



Automatic workload balancing

Cluster of DB2 nodes running on Power servers

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

Integrated Tivoli System Automation

InfiniBand network & DB2 Cluster Services

Shared Data

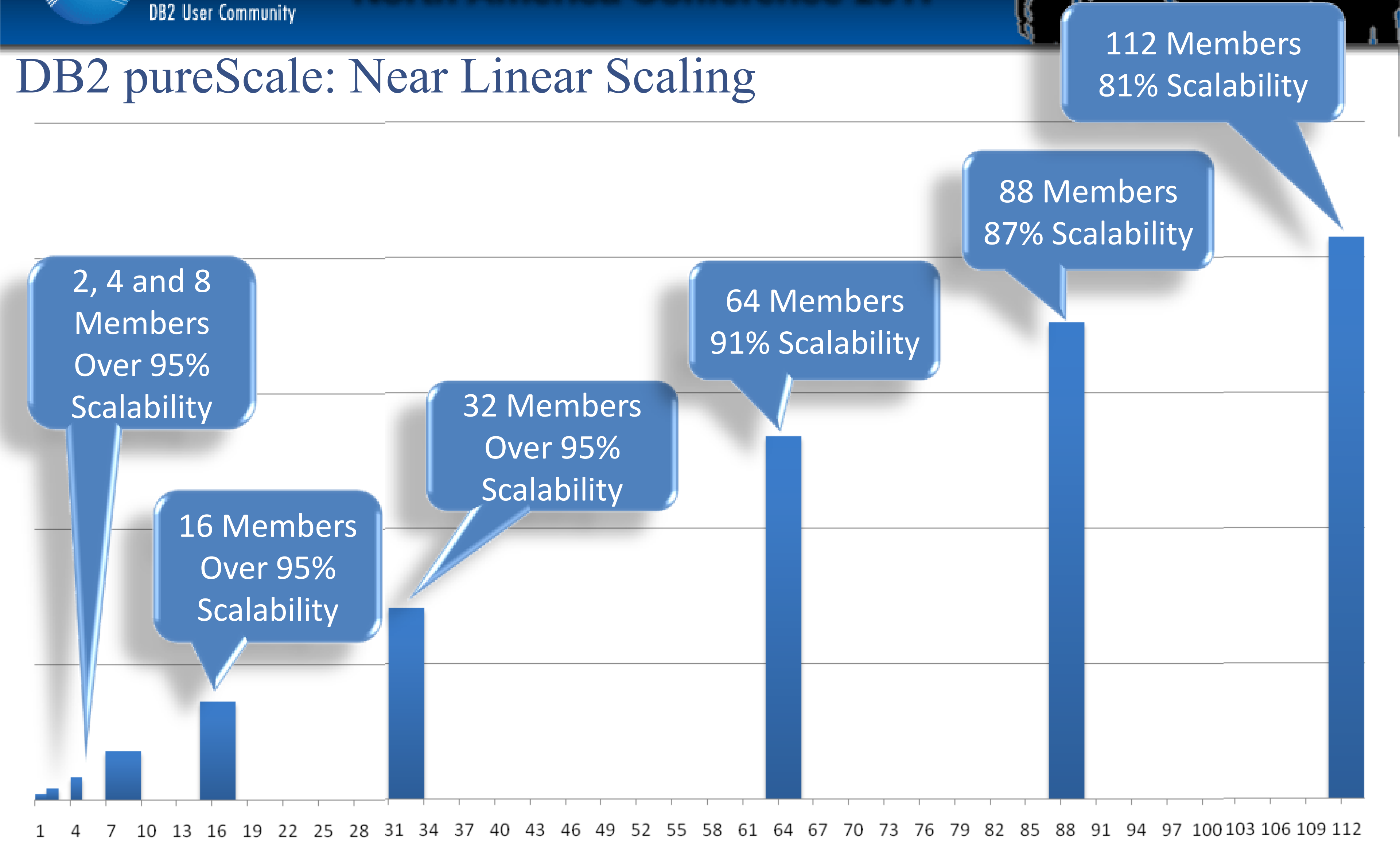
Application Transparency

- No need for application to track or know what member it is connected to
- Connection and routing done automatically
- Clients are aware of member status through frequent updates

Unlimited Scalability

- Add capacity on demand
  - Month End
  - Quarter End
  - Year End
- Simply shutdown member when not needed
  - Only pay for capacity while you need it!
- Can add up to 128 members!

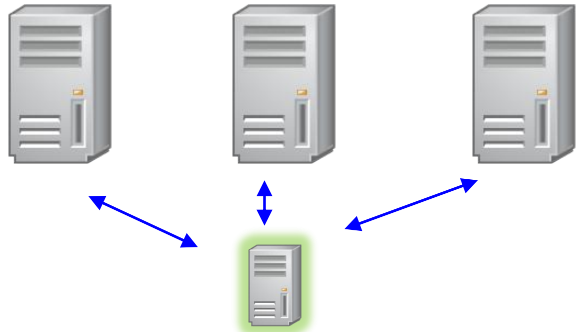
DB2 pureScale: Near Linear Scaling



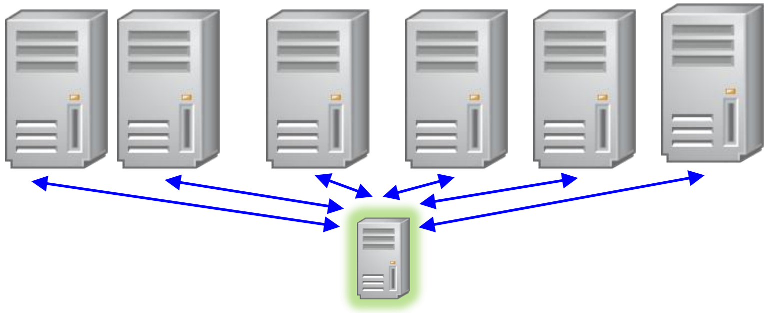
Number of Members in the Cluster

Reduce System Overhead by Minimizing Inter-node Communication

DB2 pureScale's central locking and memory manager minimizes communication traffic



DB2 pureScale grows efficiently as servers are added



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 for only a single day

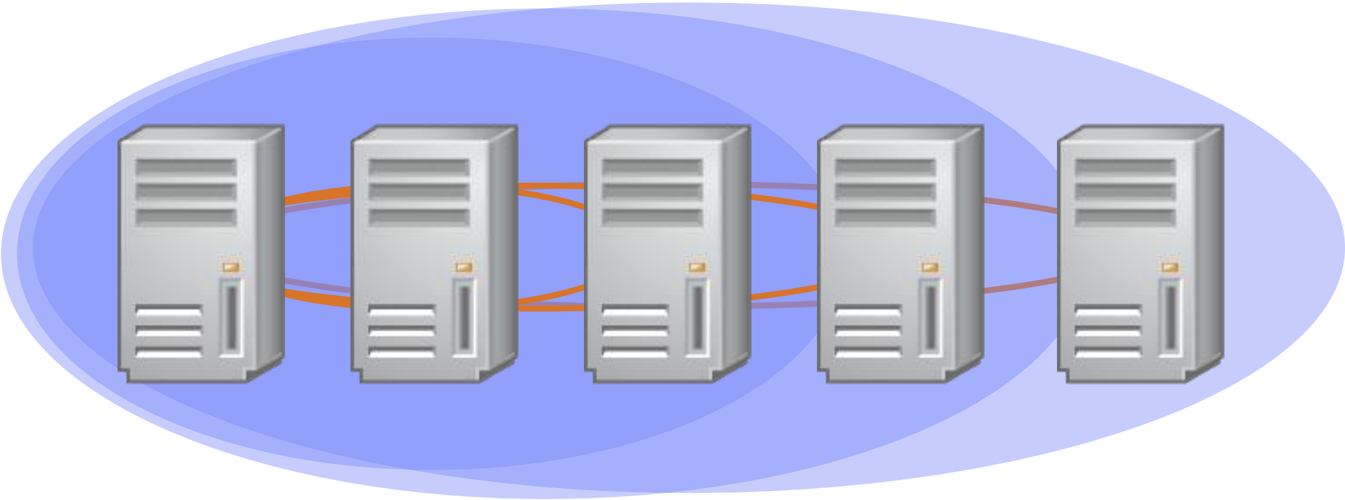
Issue:

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

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

Solution:

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

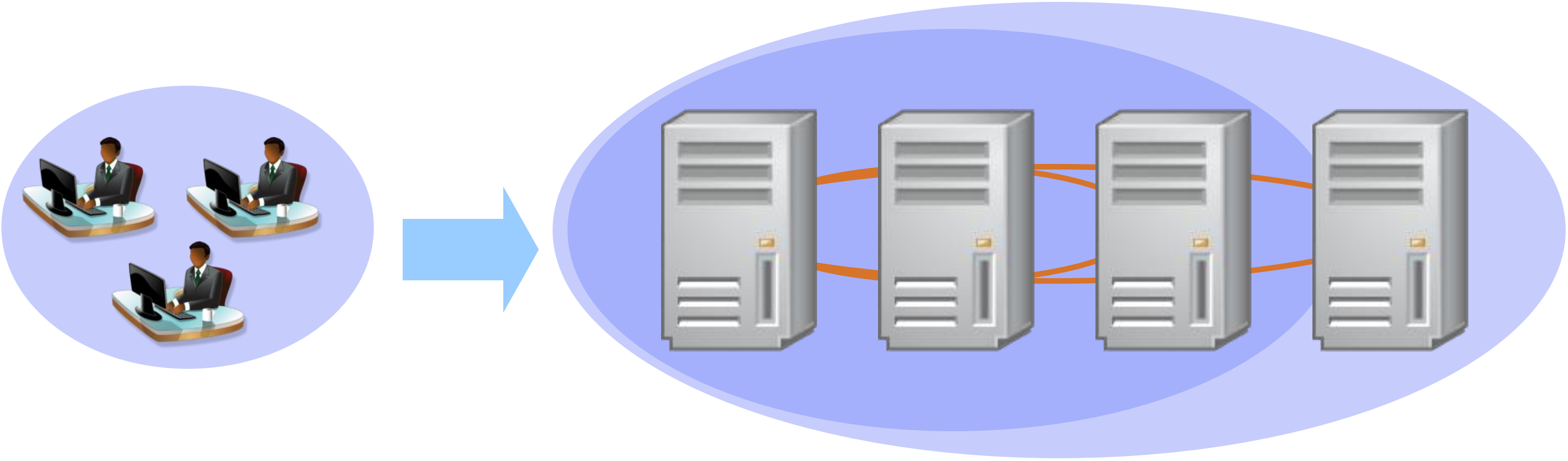


Over 100+ node architecture validation has been run by IBM

Application Transparency

Take advantage of extra capacity instantly

- No need to modify your application code
- No need to tune your database infrastructure



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

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

DB2 Data Server and DB2 Client Requirements

- DB2 9.8 DB2 pureScale Feature
- DB2 9.7 FP1 or later
  - Transaction and connection level workload balancing
  - Automatic Client Reroute based on workload
  - Client Affinities
- DB2 9.1, 9.5, and 9.7 (before FP1)
  - Only connection level workload balancing (transaction level not available)
  - Automatic Client Reroute based on workload
  - No Client Affinities

Hardware Requirements

- IBM Power 6
  - 550
  - 595
    - Firmware level 3.5.3 or higher
    - HMC 3.5.0 or higher
    - InfiniBand Network Adapter Feature Code 5609
    - InfiniBand Channel Conversion Cables (12x to 4x, FC 1854)

Hardware Requirements

- IBM Power 7
  - 710,720,730,740,750,770,780,795
    - Firmware level 7.1.0 or higher
    - HMC Release 7.1.0 Modification 0 or higher
    - InfiniBand Network Adapter Feature Code 5266*
    - InfiniBand Channel Conversion Cables (12x to 4x*, FC 1854*)

- See latest Release Notes for all Feature Codes and Cables

Operating Systems Supported/Prerequisites

- AIX
- IBM X Series Servers
- RedHat and SUSE LINUX

DB2 pureScale is Easy to Deploy

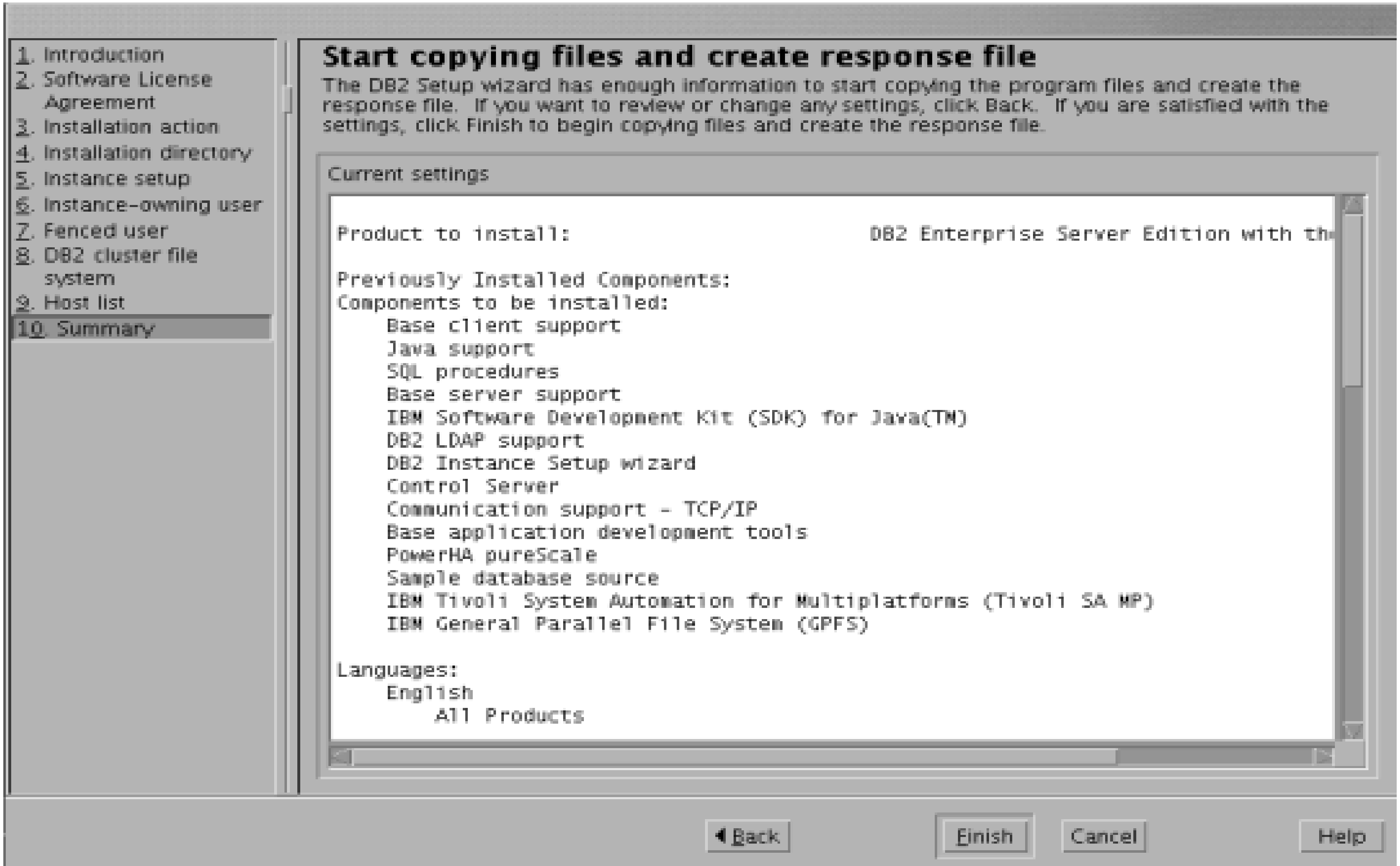
Single installation for all components

Monitoring integrated into
Optim tools

Single installation for fixpaks
& updates

Simple commands to add
and remove members

DB2 Setup – Last Step



Start copying files and create response file

The DB2 Setup wizard has enough information to start copying the program files and create the response file. If you want to review or change any settings, click Back. If you are satisfied with the settings, click Finish to begin copying files and create the response file.

Current settings

```
Product to install:                      DB2 Enterprise Server Edition with th

Previously Installed Components:
Components to be installed:
    Base client support
    Java support
    SQL procedures
    Base server support
    IBM Software Development Kit (SDK) for Java(TM)
    DB2 LDAP support
    DB2 Instance Setup wizard
    Control Server
    Communication support - TCP/IP
    Base application development tools
    PowerHA pureScale
    Sample database source
    IBM Tivoli System Automation for Multiplatforms (Tivoli SA MP)
    IBM General Parallel File System (GPFS)

Languages:
    English
        All Products
```

Install Notes

- db2icrt can only be used to create an instance on two hosts (one member or one CF at a time)
- Differences between db2icrt and
  - db2isetup can be used to create an instance with multiple hosts
- Only DB2 pureScale instance hosts are supported in a DB2 pureScale environment, non-partitioned ESE only for upgrade purposes, no other instance types supported

Lock Management

- Global Lock Manager runs in CF
- Most locks are global
  - Locks protect data that is shared by members
  - Lock request from all members must be considered before granting such locks
- Local Lock Manager runs on each member
- Local locks to protect data that is not shared
  - Internal data structures
  - Catalog cache
- Snapshot Lock monitor locking elements
- Elements without the "_global" suffix track all lock waits in the system
  - Within and between members

Lock Monitoring

- Most locks are global
  - Locks protect data that is shared by members
  - Lock request from all members must be considered before granting such locks
- Local locks to protect data that is not shared
  - Internal data structures
  - Catalog cache
- Lock monitoring elements
- Elements without the “_global” suffix track all lock waits in the system
  - Within and between members

Lock Monitoring Elements

- lock_timeouts_global
- lock_wait_time_global
- lock_wait_time_global_top
- lock_waits_global
- lock_escals_global
- lock_escals_locklist
- lock_escals_maxlocks

CF Lock Monitoring Elements

- current_cf_lock_size
- configured_cf_lock_size
- target_cf_lock_size

db2pd CF Monitoring Commands

- db2pd enhanced to support DB2 pureScale
- db2pd -cfinfo [cf_num|primary|secondary] [perf | gbp | sca | list | lock | gcl]
- Example :
  - db2pd –db sample –cfinfo gbp

Lock-related Configuration Parameters

- Memory is needed on the CF to store lock information for the various members. This memory is controlled by the cf_lock_sz database configuration parameter.
- More memory is needed to support the use of locks for internal concurrency control across members. This has an impact on the total amount of memory that is used to support locking, set LOCKLIST to (# members x original LOCKLIST setting) or AUTOMATIC (the default)

Lock-related Configuration Parameters

- More memory is needed for each table being accessed for locking purposes.
  - Requires more database heap
  - More memory needed to manage communication between local and global lock manager
- Use default AUTOMATIC DBHEAP DB CFG setting

Deadlock Detection

- db2dlock EDU on each member
  - Detect and break deadlocks among applications running on that member
- db2glock EDU is used to detect deadlocks running among applications that are running on different members
- One db2glock EDU is started on each member and one db2glock is assigned roles of “acting db2glock”
  - If member running this fails, a different member is chosen to host the “acting db2glock”
  - No need to wait for failed member to recover

CF Group Buffer Pool (GPB)

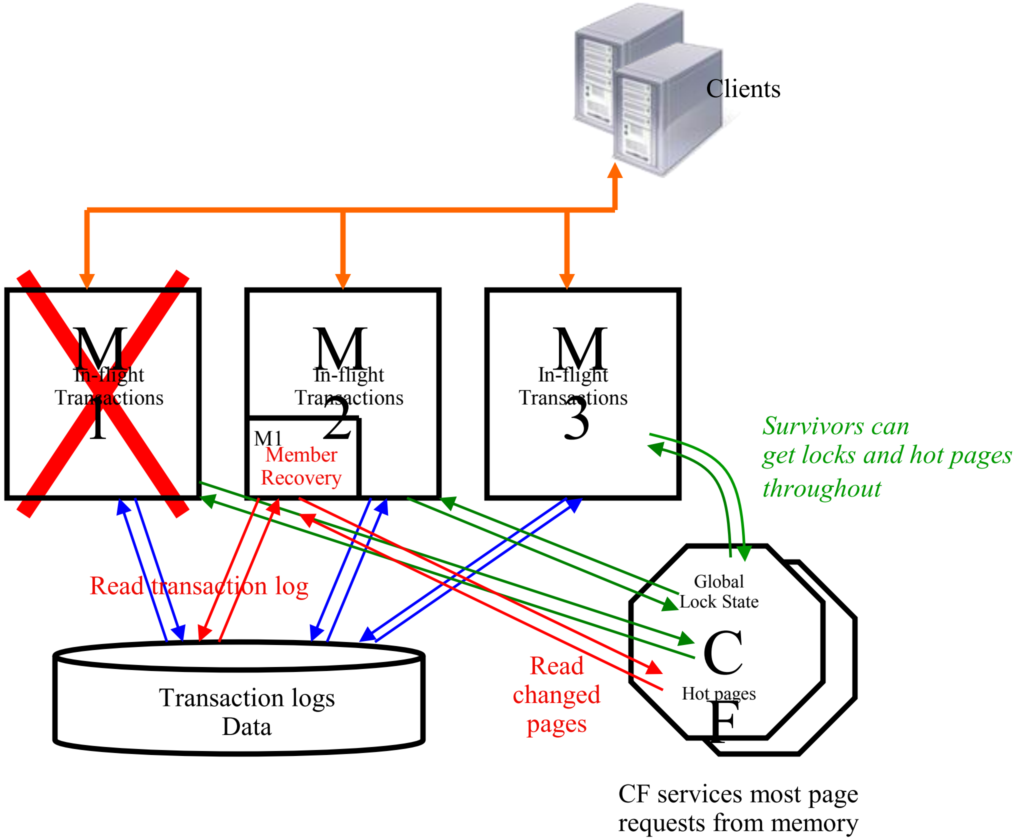
- The **cf_gbp_sz - Group buffer pool DB CFG parameter controls the size of the CF group buffer pool**
- Created upon first database connection or activation on any member
- Set to AUTOMATIC by default
- GPB is used to hold directory entries and data elements
  - Directory entry stores metadata information pertaining to a page
  - A data element stores the actual page data
- Members continue to cache pages in their own local buffer pools

Group Bufferpool Monitoring

- MON_GET_BUFFERPOOL table function
  - Returns information at the member level
  - No aggregation but can be done via SQL
  - Look for GBP in monitoring element output
    - POOL_DATA_GBP_L_READS
  - Look for LBP in monitoring element output
    - POOL_DATA_LBP_PAGES_FOUND

Member Recovery – Machine / OS / LPAR Failure

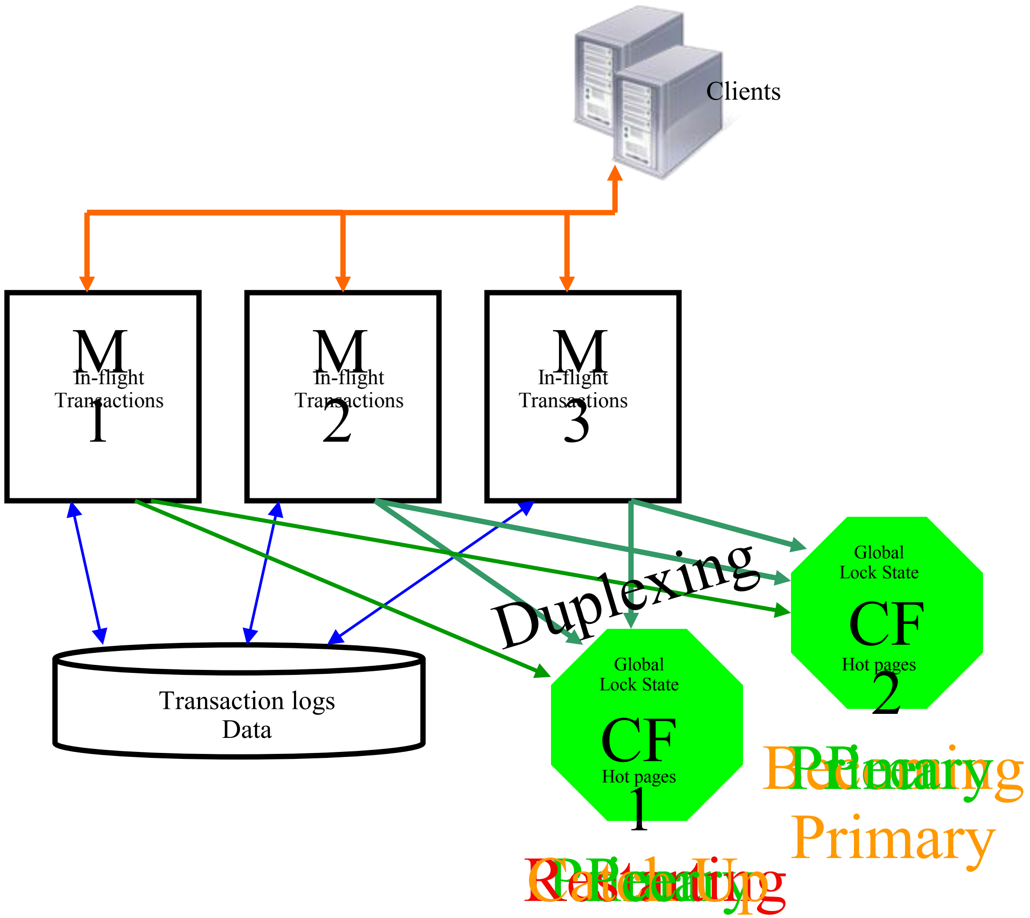
- All members available
- Host the member runs on has AIX or hardware failure
- Other members continue processing
  - Only in-flight data on failed member is unavailable
  - Transaction work rerouted to surviving members
- Member restarted on guest host – restart light
  - Member recovery completed
  - All data available
- AIX or hardware failure is corrected
- Member is restarted on home host
- All members available
  - Transaction work routed to recovered member



Primary PowerHA pureScale Failure

5 seconds

- Normal: CF1 is primary, CF2 in peer state
- CF1, the primary, fails
  - Failure is detected; notify members to stop duplexing
- CF2 becoming primary
  - Construct missing data on the CF2 (peer)
  - Momentary blip in CF response time
- CF2 assumes primary role
  - Notify members of new primary
- CF1 is automatically restarted
  - Becomes the secondary CF
  - Starts catch up
- Completes catch up
  - Notify members of new peer
  - Start duplexing
- Normal: CF2 is primary, CF1 in peer state



There is no or minimal impact on application response time.

Summary

- DB2 pureScale offers the following industry best business core functions:
- Application Transparency
- Automatic Detection, Failover, Recovery, with No OUTAGE!
- Near linear SCALABILITY
- Continuous Availability!
- Capacity on Demand as NEEDED!
  - Monthly, Quarterly, Yearly
- Low Latency
  - InfiniBand and RDMA

Moving Your Applications to DB2 is Easy

Proven Results

- Customers are no longer locked into Oracle RAC
- Integrated, cross-platform tools supporting Oracle database as well
- Customers and partners have moved in only days

References

- IBM DB2 pureScale Feature Installation and Configuration Guide, http://publib.boulder.ibm.com/infocenter/db2luw/v9r8/topic/com.ibm.db2.luw.sd.doc/doc/db2dsi.pdf
- DB2 pureScale Home Page -- http://www-01.ibm.com/software/data/db2/linux-unix-windows/editions-features-purescale.html
- Power Systems eBook -- ftp://public.dhe.ibm.com/common/ssi/pm/bk/n/imm14058usen/IMM14058USEN.PDF

Philip K. Gunning

Gunning Technology Solutions, LLC

pgunning@gts1consulting.com

D01

Achieve Breakthrough Performance and Availability with DB2 pureScale