DB2 pureScale
By Jaqui Lynch

DB2 pureScale is a tightly integrated, scalable, clustered database solution that uses IBM DB2 as its foundation. It runs on Linux, UNIX and Windows on IBM POWER or System X servers. In this article we will focus on IBM's POWER servers with AIX or Linux.

Historically, data clusters were really partitioned LPARs where the data was partitioned and each host/LPAR owned a separate partition – this is sometimes referred to as a shared nothing approach. DB2 pureScale is designed to provide a shared storage environment – it is a clustered solution where all the hosts/LPARs can access the same data while still providing data integrity and performance. This provides for continuous availability, increased capacity and scalability, and ease of design and access. Shared nothing environments are best used for data mining environments that are high on parallelism and low on updates, ideally environments such as data warehouses. Shared storage environments are designed for OLTP environments, where there are many updates. pureScale provides a shared storage environment that allows DB2 to act as a single clustered database system.

Key to the success of pureScale is the CF node (Cluster Caching Facility). This node acts as a master cache manager responsible for the integrity and synchronization of the database locks and buffers. There is a failover node to provide redundancy and techniques such as parallel logs are in place to ensure logging is consistent and does not suffer from bottlenecks.

DB2 pureScale is supported in either the DB2 Advanced Workgroup or the DB2 Advanced Enterprise editions of DB2. A typical pureScale cluster consists of a CF node and its failover node, member nodes, the global bufferpool, global lock manager, DB2 cluster services, a cluster interconnect, the cluster filesystem which is based on GPFS (General parallel filesystem) and the storage that the database goes onto. The components listed below provide the core of the pureScale architecture. At a minimum you need two members and two CFs.

Cluster Caching Facility (CF)
This is a critical component within the pureScale cluster. It manages the central resources that are shared amongst the members, specifically the global bufferpool and the global lock manager. The CF is duplexed so there is no single point of failure and provides the centralized global locking and page cache management to ensure high levels of availability and scalability. CF hosts do not need pureScale licenses as you only license the cores on which the member nodes are running. CFs should, however, be in separate LPARs from members. Each CF has a specific role to play – typically the primary CF holds all the lock information while the other CF is a peer and only has the lock information that is necessary for it to take over should the primary fail. At least one CF node must

be online for a database to be available. For performance it is recommended that CF nodes have dedicated cores – member nodes can use micropartitioning if so desired.

Cluster Members

Cluster members are the nodes that accept client requests and perform work on behalf of those clients. Each member has its own local memory with bufferpools, caches, locklists, etc and they also have full access to the database and keep logs on a shared filesystem. Transactions can run on any member of the cluster irrespective of the data being accessed. This allows for automatic workload balancing amongst the member nodes. It should be noted that the member cluster is designed to maximize recovery from failures – this means that only in-flight data remains locked until member recovery completes. Additionally, pureScale members can be added while the cluster is online which allows you to scale up very easily to a maximum of 128 member nodes with no downtime.

Cluster Interconnect

Cluster members are connected via a high speed low latency connection. This is an RDMA (remote direct memory access) capable interconnect and requires specialized network cards, specifically an Infiniband card or a 10 gigabit Ethernet RoCE (RDMA over Converged Ethernet) card. RDMA provides a low latency method of remotely changing the state of memory pages on another host without interrupting the kernel at the other host. This provides for a very fast method to move dirty memory pages between members and CFs.

For production it is recommended that the 10 gigabit Ethernet card be used as small message response time is better there. For performance, the interconnect should be dedicated to the LPAR, not virtualized.

Global Bufferpool (GBP)

This is sometimes referred to as the group bufferpool. Memory in a pureScale instance is a two-tiered environment. There are local bufferpools on each member with copies of the pages needed by that member and there are global bufferpools at the CF. The GBP holds a copy of every dirty page in the cluster as well as references to all the pages in the local bufferpools across the cluster. It also maintains a list of which members have a copy of those pages. Dirty pages are pages that have rows that have been updated, inserted or deleted in the instance. Member nodes will refresh pages from this pool when another member updates a page that they are interested in.

Global Lock Manager (GLM)

Before a row on any page can be updated the member needs to request locks from the GLM. As this occurs, the pages on all the members can be invalidated if needs be. This means a member can make an update and ensure that other members are required to get a new copy of that page before they act on it.

DB2 cluster services

DB2 cluster services provides integrated failure detection, recovery automation and the clustered filesystem. RSCT (reliable scalable clustering technology) is used to provide heart beating and domain management and is automatically installed when pureScale is installed assuming it is not already on the system. RSCT monitors all hardware components in the system including the network adapters and it provides the cluster management. GPFS is installed to provide the clustered filesystem. TSA (Tivoli System Automation) is installed to assist with identifying and recovering from failures automatically, monitoring the members and the cluster caching facilities (CFs).

Cluster Filesystem

The cluster filesystem is based on the low latency non-blocking GPFS filesystem. GPFS allows multiple hosts to share a distributed filesystem. All hosts can write to it and all hosts can see the changes immediately.

Maintenance

Because pureScale is a clustered technology IBM paid significant attention to how maintenance could be performed. It is designed to be done in a rolling fashion by applying fixpacks one member node at a time. The member gets quiesced, the fixpack is installed and the instance is updated. Then the member gets unquiesced. The member continues to behave as if it is running on the previous fixpack. Once all of the member nodes are updated then a cluster-wide command is run to complete and commit the updates. At this point all the member nodes will switch to running at the new fixpack level. pureScale is supported on POWER6 and above at specific minimum firmware and AIX levels. Information on these levels is available at the Installation prerequisites link below. Because of the clustered nature of the environment, supported release levels are very important although the installation of pureScale will update GPFS and TA if they are not at the correct level.

Summary

DB2 pureScale is designed to provide the highest levels of service combined with maximum scalability for OLTP environments. It takes advantage of well proven technologies such as GPFS, RSCT and DB2 itself to ensure a reliable high-performing environment that includes workload balancing as well as the ability for all member nodes to access all data. As with any clustered environment, pureScale requires that attention be given to proper design and architecture but the end result is well worth the effort. There is a great deal more to pureScale than we have touched on here, but hopefully this is enough information to get you started on your pureScale experience.

References

DB2 pureScale Clustered Database Solution: Part 1

http://www.ibmbigdatahub.com/blog/db2-purescale-clustered-database-solution-part-1

Plan your DB2 pureScale Feature Installation on AIX
https://www-01.ibm.com/support/knowledgecenter/SSEPGG_10.5.0/com.ibm.db2.luw.qb.server.doc/doc/r0061534.html

What is DB2 pureScale?
http://www.ibm.com/developerworks/data/library/dmmag/DBMag_2010_Issue1/DBMag_Issue109_pureScale/

Db2 pureScale
http://www-01.ibm.com/software/data/db2/linux-unix-windows/purescale/

Compare the distributed DB2 10.5 database Servers
https://www.ibm.com/developerworks/data/library/techarticle/dm-1311db2compare/index.html?ca=drs

Installation prerequisites for DB2 10.5 pureScale
https://www-01.ibm.com/support/knowledgecenter/api/content/SSEPGG_10.5.0/com.ibm.db2.luw.qb.server.doc/doc/r0054850.html