InComparison

IBM pureScale Application System vs Oracle Exadata X2-2

An InComparison Paper by Bloor Research
Author : Philip Howard
Publish date : October 2010

When we commenced this exercise we expected to find that there were some areas in which IBM excelled and others in which Oracle did so. We have been surprised to find that that is not the case and that the IBM pureScale Application System out-competes Oracle Exadata in almost every area we have examined.

Philip Howard

Table of Contents

Executive summary 1
System descriptions 4
IBM DB2 pureScale 4
IBM pureScale Application System 4
Oracle Exadata X2-2 4
Scaling your system 6
Compression 7
InfiniBand 7
Performance 9
Flash storage 10
Database 10
SAP 11
Non-OLTP in OLTP environments 11
Managing growth 12
Administration and management 13
Implementation 13
High availability 13
Costs 14
Conclusion 15

Executive summary

This paper is organised into two parts: this section, together with the conclusion, is intended for executives who have no or limited technical knowledge or interest, and the remainder of the paper, which is intended for the more technically minded. Information in this section, in particular, is duplicated in the remainder of the paper, though we go into more detail there.

The basic theme of this paper and its companion piece (IBM Smart Analytics Systems vs Oracle Exadata X2-2) is to provide a comprehensive comparison of IBM and Oracle's offerings for on-line transaction processing (OLTP) and data warehousing respectively.

Oracle's view of these two sets of requirements is that a single solution, Oracle Exadata, is ideal to cover both of them; even though, in our view (and we don't think Oracle would disagree), the demands of the two environments are very different. IBM's attitude, by way of contrast, is that you need a different focus for each of these areas and thus it offers the IBM pureScale Application System for OLTP environments and IBM Smart Analytics Systems for data warehousing.

In practice IBM's approach is not quite as simple as this. In terms of OLTP there are, in fact, two possible approaches: to license DB2 pureScale or the IBM pureScale Application System where the latter contains the former but also (optionally) the WebSphere Application Server and where the whole package is built around the AIX operating system and IBM POWER7 servers. In effect, licensing DB2 pureScale is for those preferring a more DIY approach or for those wanting to run on x86 based hardware or on the Linux operating system, while the IBM pureScale Application System is for those that want a complete system ready to run. A similar concept applies to IBM's data warehousing offerings.

IBM refers to its approach as "workload optimised systems". That is, these offerings, and particularly the packages, have been designed and optimised for their specific environments.

The question we will address in this paper is which of these two approaches is best. Of course, one could make theoretical arguments in favour of either Oracle's or IBM's approach, in which case we could argue until the cows come home. Whether one concept is better or not from a theoretical point of view is beside the point; what counts is which is best in terms of performance, scalability, ease of administration and management, and cost.

Our main concern is that Exadata is over-specified for OLTP environments for all but the very largest organisations. While we know of users with 20Tb OLTP environments, there are very few enterprises that need more than gigabytes. Yet the minimum Exadata configuration has 21Tb of disk capacity that will provide 6Tb of usable capacity, and this before taking (optional) compression into account. So, Exadata is far larger than most users will require and, of course, you have to pay for that extra capacity. In addition, the storage in Exadata has been configured specifically to improve performance in data warehousing environments where you are typically reading a lot of data in sequential fashion, rather than randomly accessing the data, which is typical for OLTP. What is included within the Exadata configuration is so-called flash storage and this will certainly give OLTP performance a boost. However, we do not believe that this performance boost is enough to justify the extra costs involved except, as we have mentioned, for the very largest environments. If we are talking just about OLTP then a more pertinent comparison would perhaps be between DB2 pureScale and Oracle Real Application Clusters without Exadata. Now the Oracle environment is not over-specified.

Executive summary

It is also important to appreciate that transaction processing systems rarely, if ever, exist in isolation. In practice, there is invariably a significant level of reporting involved to support transaction processing. Here we would expect the Exadata sub-system to provide improved performance when compared to pureScale. In addition, it will often be the case that other applications are implemented on the same system as the OLTP-based applications (a consolidated approach) and these may also benefit from features of the Oracle environment that support data look-up, where these applications are not transaction intensive. In addition, if multiple applications are implemented on a single Exadata system then, of course, the over-specification issue goes away.

Aside from this capacity issue, how does IBM's clustering technology compare with Oracle's? Of course it is not as mature, though much of the technology involved has been derived from IBM's experience with mainframe database installations and no-one could argue that that is not mature. For example, IBM uses central lock management, which is a technique used within IBM's implementation of DB2 on the mainframe. This contrasts with the use of distributed lock management by Oracle. The result is that an IBM cluster has less traffic crossing the network and is faster to recover in the event of a failure. One can also argue that IBM has second mover advantage. As an example, Exadata runs using the InfiniBand interconnect. So does IBM. However, in order to support backward compatibility, Oracle uses the same transport protocol, across InfiniBand, as it has always used for real application clusters. This operates at a much higher level in the communications stack than IBM's protocol, which, again, means that there is much more traffic across the network in an Oracle environment. If you compare this with IBM's approach, this will mean reduced performance. One other feature that is worth commenting on at this level is that applications running against DB2 do not need to know about the cluster configuration and its details. By contrast, applications running against real application clusters need to be cluster-aware in order to optimise their performance.

In so far as OLTP performance is concerned, IBM currently holds the TPC-C (see note on page 7) benchmark record for both the best performance and best price-performance in the industry—records it took from Oracle. Vendors tend to leapfrog one another with these statistics but there were less than 9 months between Oracle's previous record and that of IBM's (the end of 2009 and mid-summer 2010) and, moreover, the record was taken by a significant margin. All of that said, benchmarks are artificial and do not represent your data or your environment. They do have advantages for vendors in that they can help them to determine extreme performance problems before any customers experience them but, in so far as comparisons between products and vendors are concerned, they should be treated as no more than indicative and with a large pinch of salt.

Another distinction is IBM's approach to upgrading the pureScale Application System. Oracle is prescriptive about Exadata: you can have a ¼ rack, a ½ rack or multiples of full racks. If you want to add new processing nodes you have to add one or more additional storage servers and, in addition to the hardware costs involved, you have to pay increased license fees to run the Exadata software, which comes at $10,000 per disk (and there are 12 disks per server) plus 22% maintenance; and all this because you need some processing power even if you don't need extra disk capacity! To be fair, Oracle has recognised this point and has announced Exadata X-2-8 to address this problem but a) it is not yet available and b) it is only available as a full rack, so it will have a large capacity to begin with.

Executive summary

Conversely, IBM does not mandate any particular storage. You can continue to use existing SAN-based storage if you wish. You can choose to use solid state (flash) disks if you want to but it is not mandated. Moreover, IBM offers what it calls "workload on demand": you can add an additional server into the cluster (which is a one step process apart from making the physical connections) and you can upgrade the number of cores (from 4 or 8 to 16, depending on the system) and amount of memory (from 32Gb to 48Gb or 64Gb) and use any of these on a temporary basis and then turn them off again or remove the node from the cluster (again, a one-step process). As far as license fees are concerned, you only pay for the additional capacity while it is enabled and once it is turned off your license fees go back to what they were before.

Going into pricing itself, IBM has targeted its pricing for pureScale Application Systems at Oracle environments that do not require any software licensing. That is, those customers with a universal licensing agreement (ULA). For these users, a small pureScale Application System, which has approximately the same processor performance and storage capacity as a ¼ rack Oracle Exadata system, costs around the same as its counterpart whereas the medium and large configurations are less expensive than their ½ and full rack competitors. Of course, if you don't have a ULA then these price comparisons are very much in IBM's favour. Bear in mind, however, that list prices are notoriously fickle and sometimes bear little resemblance to reality.

Perhaps the biggest differences between the two systems are in manageability and flexibility. We have already discussed the latter with respect to the configurations and disk capacities that are available, as well as workload on demand. In terms of manageability, the pureScale Application System is easier to install in the first place, it is much easier to grow and shrink, and its use of central lock management means that do you do not get a cluster freeze in the event of a node failing, which can be the case with Oracle Exadata. This last point can be critical in a bet-your-business OLTP environment.

The bottom line is that Oracle Exadata should offer performance benefits when compared to pureScale when it comes to query and reporting and this means that Oracle will be well-placed in consolidated environments. However, we are not convinced that its flash storage (which will give OLTP performance benefits) will outweigh the deficits that derive from using a distributed lock manager as opposed to a central one. In other respects the pureScale Application System appears to offer significant advantages over Oracle Exadata.

System descriptions

Before we begin to make any sort of comparisons we need to have a clear idea about the architectures of each product offering and what is and is not included within each.

IBM DB2 pureScale

DB2 pureScale is a licensable option for the DB2 database that runs on either IBM Power Systems or System x series, the former being based on IBM's POWER7 hardware running AIX, and the latter being based on x86 processors running Linux.

What pureScale adds to DB2 is the ability for the latter to run on clustered hardware. There is no difference in the functionality of DB2 except what is needed to support a clustered environment. Most notably this means that DB2 pureScale has a shared disk architecture rather than the shared nothing architecture that has traditionally been its métier in distributed environments. However, this is not to say that this is a completely new departure for DB2 since DB2 on the mainframe System z has always used a shared disk approach and, indeed, the IBM development team responsible for pureScale has borrowed heavily from the company's existing mainframe technology in order to introduce the pureScale product. Note that Oracle has always had a shared disk architecture so there is now no differentiation between the two companies in this respect.

The architecture of DB2 pureScale contains:

- Members, which are DB2 engine address spaces that may reside on their own server or in a logical partition. You can have multiple members on a server or in logical partition, and this may be useful for testing or development purposes, but it is not recommended for live installations. Each member has its own bufferpools, memory regions and writes to its own log files. The only time that a member can access the log file of another member is in the event of a failure of the latter.

- PowerHA pureScale instances (also known as the cluster caching facility, or CF), which represent software that has been designed to assist with global buffer coherency management and global lock management. While it is not mandatory, it is recommended that you have two of these, a primary and secondary, which are duplexed and act as back-up for one another.

- An InfiniBand interconnect. You can have two of these but they are not duplexed at present.

- Storage, which is shared, except for the independent log files that relate to each member, which are only accessible by another member in the event of a failure of the owner.

- Cluster services that are provided by IBM's Systems and Technology Group and Tivoli, with the former providing GPFS (General Parallel File System) and RSCT (Reliable Scalable Cluster Technology) and the latter providing Tivoli Systems Automation for multi-platforms (TSAMP).

IBM pureScale Application System

The difference between the IBM pureScale Application System and DB2 pureScale is that the former only runs on Power 770 hardware while the latter will run on all POWER7 servers except blades, as well as POWER6 550 and 595 servers, and on System x servers. In addition, while you can have different Power System or System x servers within the same configuration if you are running DB2 pureScale that is not true with respect to the Application System, though it may be possible in the future as more Power System servers are introduced.

What then is the advantage of the pureScale Application System? First, both DB2 and WebSphere Application Server (which is an option that can be licensed in conjunction with the pureScale Application System) have features that make specific use of AIX and the Power System architecture. However, you could argue that that's an advantage of using the Power System over the System x and the fact that there is tight integration between DB2 and WebSphere would apply regardless of the platform. The key advantage therefore (though this may change in future releases), is convenience: everything is pre-bundled and ready to go for those who want an IBM stack, therefore improving time to value.

Oracle Exadata X2-2

There are actually two Oracle Exadata Database Machine products: Exadata X2-2 and Exadata X2-8. The latter is a full-rack only system, primarily intended for the largest OLTP and consolidation environments. It is not yet available and we will therefore be focusing on Exadata X2-2. This consists of Oracle Database 11g Release 2, Oracle RAC (Real Application

System descriptions

Clusters) Database server grid, an InfiniBand interconnect, the Oracle Enterprise Linux operating system, and the Exadata Storage Server Grid using either High Performance (600Gb) or High Capacity (2Tb) disk storage, where the latter give more capacity but with lower performance.

The way that the system works is that data is stored in the Exadata Storage Server grid and the storage servers act as a sort of pre-processor for accessing data from disk in an optimised fashion, using what Oracle calls smart scans, before passing the results to the database itself. This significantly reduces the amount of data that the database has to process and is particularly efficient in data warehousing environments. In order to improve performance in OLTP environments Oracle Exadata also includes flash storage for caching hot data.

You can have multiple databases running within an Exadata environment and you can have multiple small databases on a single RAC node or you can have larger databases that span nodes. This means that you can have an OLTP system sharing an Exadata environment with a data warehousing implementation. You cannot similarly share pureScale with the Smart Analytics System. This is because Oracle uses a shared disk environment throughout whereas IBM uses shared disk for OLTP but a shared nothing architecture for data warehousing. On the downside, and admittedly it is a small point, you cannot re-purpose Exadata Storage Servers. If you decide in the future to move to some other vendor then you can re-use your RAC servers and you can re-use your pureScale servers but it will not be easy to do the same thing with the Exadata Storage Servers because of their particular functional design.

Scaling your system

The implementation options for Oracle Exadata X2-2 are illustrated in Table 1.

	¼ Rack	½ Rack	Full Rack	2-8 Racks
Database Servers	2	4	8	16–64
Exadata Storage Servers	3	7	14	28–112

Note that when upgrading these are the only options available: a ¼ rack can be upgraded to a ½ rack and a ½ rack to a full rack; you cannot have a ¾ rack, for example, and you can only have whole numbers of racks above a full rack. A ¼ rack configuration holds 21Tb of raw data and a full rack (using High Performance drives) contains almost 100Tb of raw disk capacity. If you are using High Capacity drives the capacity of a full rack is 336Tb. Each Exadata Storage Server also includes 4 flash cards, with a capacity of 96Gb each, scaling up to around 5Tb on a full rack. Note that you cannot scale upwards without adding extra disks: this means that you cannot simply add new processing capacity if you have, say, a CPU bottleneck: you have to have additional storage capacity even if you don't need it.

In practice, of course, actual disk capacity and usable disk capacity are very different things. To begin with there is disk mirroring, which is needed for resiliency, which halves your available capacity, and then there are considerations with regard to space needed for logs, temp space, indexes and so on. Oracle's own estimates are that 55% of disk capacity, before taking account of mirroring, is actually usable for storing data, which means that a ¼ rack actually provides around 6Tb of usable space, a ½ rack 14Tb and a full rack 28Tb. Of course, these figures do not take account of compression.

IBM similarly offers a choice of high performance and high capacity drives and can also expand to large numbers of servers (configurations with up to 128 servers are supported). Unlike Oracle it does not use direct attached storage but offers a variety of SAN-based approaches to storage that you can adopt, notably its XIV clustering architecture and its SONAS network attached storage. This also means that you can reuse your existing SAN-based storage. We will discuss IBM's use of solid state disks and its Easy Tier capability when we discuss the use of flash storage later.

Also different from Oracle is the fact that you can start very small with systems involving less than 1Tb of data. While we will discuss pricing in due course, this strongly suggests that Oracle Exadata does not scale down and will not be suitable, or will be overly expensive, for departmental environments, for small and mid-sized organisations, and even for some large companies, since there are many organisations that do not require 6Tb of (uncompressed) data capacity for OLTP purposes. The exception will be where a single Exadata environment is required to host both OLTP and warehousing capabilities but even then there are many companies that do not need anything like 6Tb, especially bearing in mind that this is raw capacity before compression is taken into account.

IBM pureScale also differs from Oracle in terms of upgrading servers. The pureScale Application System is available in small, medium and large configurations based on two Power 770 3.1GHz processors linked through an InfiniBand interconnect. Where the different options differ is that they have 4, 8 or 16 cores (out of a total of 16) activated on each

Scaling your system

of the two servers and with 32, 48 or 64Gb of memory (out of a total of 64Gb) activated. 16 cores and 64Gb are installed in all cases. Not only can you add new nodes, you can upgrade the number of cores and memory that is activated within a server.

Further, IBM's licensing model is such that you can upgrade cores and add servers on a temporary basis. Suppose that you need extra capacity on your system at year-end; then you can add another server to your pureScale implementation for the duration that you require that extra capability, and then remove it again. Of course, you will have to have a spare server available to do that but, as far as DB2 is concerned, additional license fees will only apply for that limited period. The same approach is adopted with respect to core upgrades: so, if you have licensed 4 cores per server but need 8 cores for one week in the year then you can turn these on temporarily and only pay additional license fees for that week. While we will discuss costs in a separate section it is not hard to see the potential advantages of this approach.

Compression

We should add here a comment about compression. Oracle uses two different types of compression: one which is designed specifically for data warehousing and archival environments, which will be discussed in more detail in the companion paper to this, and what is known as 'advanced compression', which is used by Oracle for OLTP purposes. Oracle quotes 2 to 4 times compression for transaction data. In practice, the best compression rates will be achieved when the data has been pre-sorted, which is unlikely to be (or remain) the case with live transactional data. Oracle also compresses indexes, though the compression ratios here are less impressive.

IBM not only uses a different technique for compression (a form of tokenisation) but also compresses temporary data. Without going into every nuance of the different compression technologies used by the two companies it will be useful to give an example: suppose that you are compressing a customer index then every time that 'Bloor' appears Oracle will store 'Bloor' plus a row ID. The first thing to understand is how indexing works in the first place. In the case of Oracle, if there are 250 entries for 'Bloor' there will be 250 separate Bloor-Row ID pairs. IBM however, would only store Bloor once, followed by a string of 250 Row IDs. Thus IBM's approach is more efficient to begin with. When it comes to compression Oracle compresses each Bloor-Row ID pair whereas IBM compresses 'Bloor' and Row IDs separately and therefore more efficiently (because you can use different algorithms depending on the datatype) so we would expect better index compression ratios from IBM. Further, IBM also compresses temporary data so we would expect IBM's compression rates to exceed those of Oracle overall.

InfiniBand

However, scaling up isn't just about what capacity you can add and the disk space you can expand to but also about how you add capacity and the efficiency with which you can scale. In order to understand how this works, and explain one of the key differences between the two rival offerings, it will be necessary to understand how the two companies make use of their InfiniBand interconnect and how they implement locking.

As we have noted, both systems employ InfiniBand (because InfiniBand has a greater bandwidth capacity than Ethernet and significantly lower latency). However, IBM pureScale was designed specifically to work with InfiniBand while Oracle RAC (which is the important element here) was designed back in the last century before InfiniBand had been developed. As a result, the two companies use very different approaches towards leveraging InfiniBand: Oracle employs RDS (Reliable Datagram Sockets), which is a protocol similar to TCP/IP and which relies on message passing; IBM, on the other hand, uses RDMA (Remote Data Memory Access), which is a much lower level protocol designed specifically for InfiniBand use which, as its name suggests, allows direct access to the main and cache memories residing on the various servers within the network. The advantage of this latter approach is that it eliminates context switching, there is no requirement for interrupt or message processing, and there are no CPU cycles involved when members need to be informed of page updates (because memory is used instead). As a result, typical round-trip response times are of the order of 10–15 microseconds where you would expect Oracle's to be in the range of hundreds of microseconds or milliseconds. Furthermore, when a transaction on one member updates and commits new data, pureScale is able to communicate the existence of this newly

Scaling your system

committed data to all other members of the cluster, without involving any host cycles on those other members. This is true no matter how large the cluster. pureScale does this through a technique called 'silent invalidation', which depends on pureScale's exploitation of RDMA. As the cluster grows tens or hundreds of members, this technique is key to application transparent scalability. Even with 100 members, when one member commits new data, all other members are made aware that this occurred, without invoking any host cycles. With Oracle, which does not have a comparable technique to silent invalidation, the host cycles that are required grow as the cluster grows.

A further consequence of the way that the interconnect is used is reflected in the different approaches to locking taken by the two companies. Put simply, Oracle uses a distributed lock manager while IBM uses a global lock manager, where the latter is located on the PowerHA pureScale node or nodes. In other words, in an Oracle RAC environment every node is responsible for some of the locks held by the system whereas IBM pureScale holds all locks centrally.

There are two consequences to this approach. The first (which we include in this section for completeness) is that you can get problems with Oracle Exadata when a node goes down because you may get a cluster freeze while the remaining nodes work to rebuild the lock list, unraveling the locks held by the failed node. When a DB2 pureScale node fails, on the other hand, the only thing that happens is that data in-flight on the failed member remains locked during the recovery process (which is automated) but otherwise processing proceeds as normal.

The second consequence is that you get a huge amount of message passing in Oracle environments. While this is by no means the only factor involved, and we do not have figures to substantiate this claim, based on their respective uses of InfiniBand and how they implement log management, one would expect pureScale to scale better (in terms of the extra performance you get by adding extra servers) than Oracle Exadata: as nodes are added the amount of traffic across the interconnect will increase exponentially and this can only mean that performance (and therefore scalability) will suffer. It should be noted, however, that this issue will not apply to warehousing environments where access is read only and locking is not an issue.

Finally, within the context of scaling your system there is the question of what happens when you add a node to your cluster. IBM offers an automated load balancing capability that will automatically recognise when nodes are added to or removed from the cluster and it will automatically re-direct transactions as appropriate: there is no tuning required. This is not true with respect to Oracle Exadata. There are similar load balancing facilities provided but there is typically an initial tuning requirement when a new node is added to a cluster or one removed. We will discuss the ease with which you can add a new node in a later section.

Performance

Oracle has conducted a benchmark internally on a ¼ rack system running Siebel software, simulating a call centre environment with heavy traffic. The benchmark simulated more than 30,000 users who performed over 400,000 transactions per hour with an average response time of 0.12 seconds, with 75% of those transactions being served out of flash storage (see later). This is impressive, but perhaps most interesting is the fact that CPU utilisation (per database server node) was only 22%. In other words, the system tested was over-specified for its environment if this is all it is expected to do.

More generally (and while we have doubts about the utility of benchmarks that are not performed specifically for customers with their own data and their own workload) it is worth noting the current TPC-C* benchmarks. Towards the end of 2009 Oracle posted what was then a record performance at 7,646,486 tpmC (transactions per minute type C) at a cost of $2.36 per tpmC. This represented an approximate 25% performance improvement compared to the previous record (held by IBM) and a 16% improvement in cost. More recently, in August 2010, IBM took the record with a performance rating of 10,366,254 tpmC at a cost of $1.38 per tpmC. This represents an improvement of Oracle's figures of roughly 35% and 42% respectively. This is an unusually large jump, especially in price/performance terms.

However, these figures should be taken as no more than indicative. Moreover, any estimate of performance is not merely the sum of the parts but is the software, the operating system and the hardware, all working together, which together provide the best performance. From this perspective it is important to recognise that while Oracle 11g Release 2 has features in it that exploit Exadata, it was not, essentially, designed for that purpose. In particular, and as an example, the legacy of Real Application Clusters means that Oracle cannot take full advantage of InfiniBand. At first sight one might make the same argument for IBM and DB2. However, DB2 on the mainframe has always been tightly coupled with the z series operating system and hardware, with centralised locking and centralised buffer pools: what IBM has done with pureScale (in a project that started some 6 years ago) is to adopt the same principles on its distributed systems.

We will now go on to discuss particular elements of each system that militate in favour of good performance. Some of these features we have already discussed, notably the advantages that IBM has in terms of InfiniBand and lock management. One further feature of the hardware architecture (as opposed to DB2 compared to Oracle 11g, which we will come to shortly) that we need to discuss is the use of flash storage.

*__IBM POWER7 Benchmark Result__:
IBM Power 780: 10,366,254 tpmC at $1.38USD/tpmC available October 13, 2010, running on 3 nodes with a total of 24 processors, 192 cores and 768 threads.

__Oracle Sun Benchmark Result__:
Sun SPARC Enterprise T5440: 7,646,486 tpmC at $2.36USD/tpmC, available March 19, 2010, running on 12 nodes with a total of 48 processors, 384 cores and 3,072 threads.
Results current as of August 17, 2010.

TPC, TPC Benchmark, TPC-C and tpmC are trademarks of the Transaction Processing Performance Council. TPC-C results available at www.tpc.org.

Performance

Flash storage

The IBM pureScale Application System does not come with solid state disks though you can use them if you choose to. With the DS8700 storage system (and, later in 2010, on mid-range storage servers) it is used in conjunction with IBM's Easy Tier technology. The idea here is that some (hot) data is held on solid state disks (SSD arrays) and the remainder on conventional hard drives, with data migrating up to SSD arrays or down to hard disks as appropriate, with the relocation of data being automatically handled by the software.

There are two major differences between IBM (with Easy Tier) and Oracle in their use of flash. The first is that Oracle uses a PCIe flash card as opposed to solid state disks. The advantage of this is that you don't have a disk controller between the flash storage and processor, which can potentially slow the environment down if the disk controller has not been designed to operate at flash speeds. The other difference between Oracle and IBM is in the way that the two companies use flash. Oracle, which refers to its technology as the Exadata Smart Flash Cache, actually uses this as a read cache. That is, it copies hot data from storage into the cache as opposed to IBM's Easy Tier in which hot data is actually stored on the SSD. Deciding what data should be held in the flash cache is handled automatically though users can define directives at the database table, index or segment level to ensure that specific application data is held in flash, subject to the proviso that the software is smart enough to know when data will not fit into the cache. The downside of the cache-based approach is that all Oracle locking is in the database, which means that when a page is updated it is no longer valid in the cache. Taken in isolation this would mean that you will have to wait until the cache is refreshed from disk or you would need to read directly from disk, but, on the assumption that we are talking about a recent update then this would be in the database buffer in memory, so there would be no need to access that data from either flash or disk. Nevertheless, defining appropriate directives will be an important consideration, because it is data that doesn't change much that will give you most benefit from being located in the cache. Of course, IBM's approach also has a downside in that data will be moving from SSD arrays to hard disk and vice versa but this can be done as a background task without interfering with normal operations.

Having made all of the preceding comments it must be borne in mind that the pureScale Application System does not come with solid state disks as standard. Thus the comparison in most situations will be between flash cache on the one hand and conventional bufferpools on the other. For the right applications, especially where the environment is very large, the use of flash disks should provide a significant advantage for Oracle. How applicable this will be for smaller environments is another question entirely.

Database

Bloor Research has regularly conducted performance comparisons between the DB2 and Oracle database systems, in 2003, 2005 and 2007 prior to the current paper. We do not intend to go into the sort of detail that those reports have done, otherwise we would double the length of this paper! In general there are individual features of each product that we like but these roughly balance out. The same tends to be true as the two vendors leapfrog one another as new versions of their respective database systems are released. Of those elements that pertain to OLTP, as opposed to data warehousing, we have historically seen the two engines as more or less comparable: we have preferred Oracle's indexing and clustering, and preferred IBM's support for XML and its tuning and management capabilities. With the exception of cluster support, where pureScale is the discussion point in hand, we see no need to change these opinions. However, Oracle has made significant strides in reducing its administrative requirements so that it has narrowed the gap, but we still rate it as lagging behind IBM. The one area where there is a significant difference between the two products, and where this is likely to remain the case, is with respect to support for XML. Both companies claim to support XML natively but they mean different things by this assertion: Oracle supports a native XML datatype while IBM not only does this but also stores XML natively. This means that XML documents do not need to be shredded in order to store the data or re-combined when it is retrieved, which, in turn, means that DB2 should out-perform Oracle when it comes to reading and writing XML data. Of course, this will not be relevant for all users but it may be significant if XML is important to your organisation.

Performance

SAP

While this will not be applicable to all potential users of either Exadata or pureScale it is worth pointing out that DB2 has a number of performance optimisations built into it in order to support SAP application environments. In particular, DB2 understands the SAP environment within which it is working so that, for example, it can recognise relevant details of the SAP configuration in use and set defaults against these when the system is initially installed. You can also install DB2 as a part of the SAP installation process. DB2 is also aware of SAP workloads and the database's built-in tuning capabilities can use this fact when it makes recommendations; and the same applies to troubleshooting, whereby diagnostics also understand the SAP environment.

Non-OLTP in OLTP environments

We have never run across a database environment that only processes transactions. Invariably there are various reports that have to be produced (aged debt analysis, for instance) on a regular basis. In addition, it may well be that additional applications leverage the same environment. For example, a manufacturer with an ERP implementation will not just be doing things like sales order processing but also tasks such as capacity planning. Where this is the case, Oracle Exadata should offer advantages over the pureScale environment under certain conditions, specifically where (multiple) whole table scans are required within a query. These will be more common within a data warehouse but will also occur sometimes within an operational environment and, in this case, Oracle will have a significant performance advantage when compared to IBM. A detailed discussion of how Oracle's technology works in this respect is included in the companion paper to this.

Managing growth

We have already discussed the physical aspects of adding nodes to a cluster and the general approach to growth adopted by the two vendors, with Oracle offering rack-based growth and IBM taking a more modular stance, allowing you to add cores and/or additional nodes, as required. However, we should also consider what adding nodes means in software terms.

In a pureScale environment adding a node means that the operating system has been installed, that the node is physically attached to the network and that access to shared disks has been enabled. Once that is done you enter the command "`db2iupdt -add -m <MemHostName:MemIBHostName> InstName`" to add a new member and DB2 does everything else (copies the image and response file, runs install, sets up access to the cluster file system, and so on) for you. A similar process is used for dropping a member or adding or dropping a PowerHA pureScale Server. However, in this release (this will change) extending or shrinking an instance is an offline process.

Associated with this process it needs to be understood that pureScale supports application transparency. That is, applications running against DB2 do not need to know anything about the hardware environment which is supporting them, because DB2 handles that automatically. This means that there are no coding changes required when you add or drop a node and, similarly, there is no requirement for application testing or for tuning of the infrastructure.

Adding a new node to an Oracle RAC implementation, on the other hand, means provisioning the new node, installing CRS, installing the RAC software, adding LISTENER to the new node, adding the database software, manually adding an ASM instance and manually adding a database test instance, which seems rather more complex. In addition, with Oracle RAC, applications need to be cluster-aware in order to optimise the performance benefits associated with the cluster, which means that applications may need to be tuned as the environment scales up (or down for that matter).

Administration and management

We have already commented on the fact that although Oracle has made significant strides in reducing its administrative requirements in recent releases, we still do not believe that its autonomic and self-tuning capabilities match those of IBM. Apart from this perhaps the biggest differences are in implementation and high availability.

Implementation

While we do not have figures for installing an Oracle Exadata system, we do have such figures for an Oracle Database 11g Release 2 implementation on a 4 node cluster: according to independent research conducted by the Winter Corporation there are 208 steps involved in installing such a system. By way of contrast, an IBM pureScale Application System is built, tested and the software installed on IBM premises before being delivered to you. It is taken apart during the physical shipment process and then the engineer will cable it up for you. At that point you can start loading data. In other words 208 steps for RAC versus none for pureScale, and probably more than that for Exadata. Similar (though not so extreme) differences apply when it comes to upgrades and fixes, which is a single installation across all software components in the case of IBM.

High availability

Oracle uses direct attached storage while IBM uses a SAN-based approach. As a result, Oracle relies on software-based disk mirroring: when a disk fails this is automatically detected by the software and this recreates (and rebalances) new mirrors on other good disks. In addition, Oracle offers a high redundancy option whereby you can set up three copies of the data so that double failures cannot cause a problem. More generally, both vendors claim that they have no single source of failure. In neither case is this true. IBM only has one interconnect (you can have two but they do not act as back-up to one another); conversely Oracle does have two interconnects, each with their own link and which are bonded so that one can act as the back-up for the other. On the other hand, there is only a single disk controller in each Exadata Storage Server, which means that you have to failover to another Storage Server if a disk controller fails: technically this is not a single point of failure but it's an expensive back-up option. IBM, on the other hand, ensures redundancy throughout, with dual adapters, dual controllers, dual cables and so on.

In addition, there is the possibility of a cluster freeze within an Oracle environment in the event of a node failure, because of the locking issues previously discussed.

A major feature of DB2 is its support for Oracle environments.

You can import Oracle schemas directly into a DB2 database and DB2 has native (not emulated) support for Oracle concurrency control (but DB2 does it in a different way in order to avoid the locking issues that cause performance degradation in Oracle environments), SQL, PL/SQL, packages, built-in packages, OCI (Oracle call interface), JDBC, online schema changes and SQL*Plus scripts, amongst other features.

What this all means is that the vast majority of applications, stored procedures and other constructs written to run against an Oracle database will run unchanged, possibly with better performance because of the improved locking, against a DB2 database.

According to IBM it has tested more than 750,000 lines of PL/SQL and it has achieved an average compatibility of 98.43%.

This is truly impressive.

Costs

Oracle has an unbundled pricing structure for Exadata while IBM has a bundled approach for the pureScale Application System. Thus, in the case of the former you have to separately license the database itself, RAC, partitioning, advanced compression and the tuning and diagnostic packs. Table 2 illustrates the list price for different Exadata configurations, excluding these additional components (even though some of these are included as standard by IBM) as well as first year maintenance and support, against comparable figures for a pureScale Application Server configuration with similar server performance characteristics and storage capacity, although the IBM figures do not include the use of solid sate disks whereas Oracle's figures do include flash disk. Note that the disk capacities quoted are for usable, uncompressed capacities.

Oracle system	Oracle list price	Oracle list no s/w	IBM system	IBM list price
¼ rack (6Tb)	$1.53m	$0.73m	Small	$0.74m
½ rack (14Tb)	$3.11m	$1.51m	Medium	$1.22m
Full rack (28Tb)	$6.10m	$2.90m	Large	$2.21m

Note that we have also included Oracle prices without any ULA-based software costs (that is, excluding database licenses but including Storage Server licenses). This is because customers having an existing universal license agreement with Oracle will not have to pay some or any of these additional fees (depending on their agreement) so, for these organisations, this element of the Exadata pricing equation will be partly or wholly irrelevant. What is noticeable is that IBM has deliberately priced the pureScale Application System to be equivalent to, or more attractive, than the comparable Exadata configuration even if all database software license fees are excluded.

Of course the proviso must be made that these are only list prices and are subject to potentially substantial discounts.

The other major cost factor that we have already alluded to is IBM's flexible approach (known as 'capacity on demand') to temporary additional requirements where you only pay license fees for what you use. Further, IBM does not charge for idle standby DB2 systems in a high availability environment, while in a VMware virtualised environment it only charges for doing DB2 work on the server. None of this is the case with Oracle. Moreover, as we have also discussed, with Oracle you can only upgrade from a ¼ rack to a ½ rack to a full rack and then by adding additional racks. You cannot, for example, add additional database servers separately from storage servers and with storage server licensing costing $10,000 per disk drive (plus 22% maintenance) this is an expensive option if you only need extra compute power.

Finally, if we are correct in our assertion that DB2 is more easily manageable than Oracle and the pureScale environment than Exadata, then we would expect the latter to require additional administration over and above that needed by IBM. This in itself represents an expense.

Conclusion

When we commenced this exercise we expected to find that there were some areas in which IBM excelled and others in which Oracle did so. We have been surprised to find that that is not the case and that the IBM pureScale Application System out-competes Oracle Exadata in almost every area we have examined. The exception is performance. It is certainly true that flash storage will give you a performance boost in OLTP environments and, though IBM can offer this, it is not standard. On the other hand, IBM offers more efficient locking (and locking is very important in OLTP environments) as well as less messaging across the interconnect. Which of these is most important within an OLTP environment is debatable and will depend on your particular circumstances. Otherwise we would certainly expect Oracle Exadata to provide performance advantages for ancillary capabilities such as queries and reporting, as well as in supporting non-OLTP consolidated applications running on the same platform.

In all other respects, from scalability to flexibility, through ease of use and high availability, to cost (at least at list prices), IBM appears to offer significant advantages. Oracle may be able to out-perform IBM for certain consolidated functions (especially complex queries) but this is poor recompense for IBM's advantages for transaction processing, which is typically the key purchasing criteria for such environments.

Further Information

Further information about this subject is available from http://www.BloorResearch.com/update/2063

Bloor Research overview

Bloor Research is one of Europe's leading IT research, analysis and consultancy organisations. We explain how to bring greater Agility to corporate IT systems through the effective governance, management and leverage of Information. We have built a reputation for 'telling the right story' with independent, intelligent, well-articulated communications content and publications on all aspects of the ICT industry. We believe the objective of telling the right story is to:

- Describe the technology in context to its business value and the other systems and processes it interacts with.
- Understand how new and innovative technologies fit in with existing ICT investments.
- Look at the whole market and explain all the solutions available and how they can be more effectively evaluated.
- Filter "noise" and make it easier to find the additional information or news that supports both investment and implementation.
- Ensure all our content is available through the most appropriate channel.

Founded in 1989, we have spent over two decades distributing research and analysis to IT user and vendor organisations throughout the world via online subscriptions, tailored research services, events and consultancy projects. We are committed to turning our knowledge into business value for you.

About the author

Philip Howard
Research Director - Data

Philip started in the computer industry way back in 1973 and has variously worked as a systems analyst, programmer and salesperson, as well as in marketing and product management, for a variety of companies including GEC Marconi, GPT, Philips Data Systems, Raytheon and NCR.

After a quarter of a century of not being his own boss Philip set up what is now P3ST (Wordsmiths) Ltd in 1992 and his first client was Bloor Research (then ButlerBloor), with Philip working for the company as an associate analyst. His relationship with Bloor Research has continued since that time and he is now Research Director. His practice area encompasses anything to do with data and content and he has five further analysts working with him in this area. While maintaining an overview of the whole space Philip himself specialises in databases, data management, data integration, data quality, data federation, master data management, data governance and data warehousing. He also has an interest in event stream/complex event processing.

In addition to the numerous reports Philip has written on behalf of Bloor Research, Philip also contributes regularly to www.IT-Director.com and www.IT-Analysis.com and was previously the editor of both "Application Development News" and "Operating System News" on behalf of Cambridge Market Intelligence (CMI). He has also contributed to various magazines and published a number of reports published by companies such as CMI and The Financial Times.

Away from work, Philip's primary leisure activities are canal boats, skiing, playing Bridge (at which he is a Life Master) and walking the dog.

Copyright & disclaimer

2nd Floor,
145–157 St John Street
LONDON,
EC1V 4PY, United Kingdom

Tel: +44 (0)207 043 9750
Fax: +44 (0)207 043 9748
Web: www.BloorResearch.com
email: info@BloorResearch.com

POL03075-USEN-00