Database Clusters with IBM eServer xSeries 366 and Oracle Database 10*g* Real Application Clusters (RAC)

Phil Horwitz and Martha Centeno
IBM eServer xSeries Performance Development & Analysis
Research Triangle Park, NC USA

Abstract

Customers running an Oracle database understand the need for building a flexible infrastructure that will allow them to scale according to the demands of their business. The latest addition to the IBM® eServer™ xSeries® product line, the IBM eServer xSeries 366 featuring the 64-bit Intel® Xeon™ Processor MP, is a versatile server that can help customers meet the demands of their Oracle implementations.

With the introduction of the x366, customers now have an opportunity to deploy and take advantage of both 32-bit and 64-bit versions of applications like Oracle Database 10g with Real Application Clusters (RAC). The x366 is designed to achieve unprecedented levels of performance and availability through its ground-breaking, IBM eServer X3 Architecture, the third-generation IBM Enterprise X-Architecture XA-64e™ chipset. With eServer X3 Architecture, this compact 3U 4-way server proves 64-bit solutions are available today for small to large customers.

This paper will describe a proof-point with a 2-node Oracle Database 10*g* RAC cluster running on Red Hat Enterprise Linux 3 Advanced Server. The workloads used in this proof-point were queries from a decision support system database. Results clearly show overall system scalability in CPU-intensive queries as processing power is added, and improved performance[1] from the cluster interconnect with the introduction of InfiniBand® as the interconnect technology.

---

[1] Performance is an internal throughput rate based on measurements and projections using standard IBM benchmarks in a controlled environment. The actual throughput that any user will experience will vary depending upon considerations such as the amount of multiprogramming in the user's job stream, the I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve throughput improvements equivalent to the performance ratios stated here.

Introduction

The x366 provides breakthrough, four-socket performance with 64-bit memory addressability through IBM eServer X3 Architecture, the third generation of mainframe-inspired IBM Enterprise X-Architecture™. It is the first-to-market, industry-standard x86 server powered by the latest 64-bit Intel Xeon Processor MP and can simultaneously run 32- and 64-bit applications.
The IBM x366 offers the latest performance enhancements that reduce latency and increase data transfer bandwidth for commercial x86 applications, including XceL4v Dynamic Server Cache, Active PCI-X 2.0 up to 266 MHz, DDR2-based Active Memory™ and highly reliable Serial Attached SCSI (SAS) hard drives.

Oracle Database 10*g* RAC is designed to help build flexible, high-performance, highly available, clustered database solutions on Linux®.

The audience for this paper is customers interested in implementing a database cluster based on the latest in server technology from IBM and Oracle who want to evaluate the far lower total cost of ownership (TCO) (compared to older UNIX® based servers) that a Linux-based IT architecture may provide.

This paper is organized in three major sections as follows:

- Architecture and Concepts

  It is important to understand the basic architecture of each of the components included in a cluster. More importantly is how each of the features enables higher performance of the total solution. This section Includes descriptions of the IBM eServer xSeries 366 Architecture and Oracle Database 10*g* Cache Fusion with new functions such as Automated Storage Management (ASM) and InfiniBand support

- System Test Configuration

  This section includes the configuration for the system under test. It includes setup details on Oracle Database 10*g* running on the x366, ASM running on the DS4500, and the interconnect configuration

- Workload Measurements and Analysis

  The workload consists of unique but representative queries from a decision support type workload. The queries have different characteristics and therefore each stresses the system in different ways. The measurements show the capability of the subsystems when running processor-, disk- and interconnect-intensive workloads.

Architecture and Concepts

IBM eServer xSeries 366 Architecture

It was important to choose a hardware platform that showcases performance, flexibility and manageability. The IBM eServer xSeries 366 was designed from the ground up to emphasize those attributes. It utilizes the IBM XA-64e chipset, which is a newly redesigned version of its predecessor chipset first introduced in the IBM eServer xSeries 440 in 2002.

Major features of the XA-64e chipset include:

- Support for latest Intel Xeon Processor MP with Extended Memory 64 Technology (EM64T)
- North Bridge and L4 Controller combined into a single high-performance controller
- Twin Front-Side buses.
- High-performance DDR2 memory technology
- PCI-X support up to PCI-X-266MHz speeds

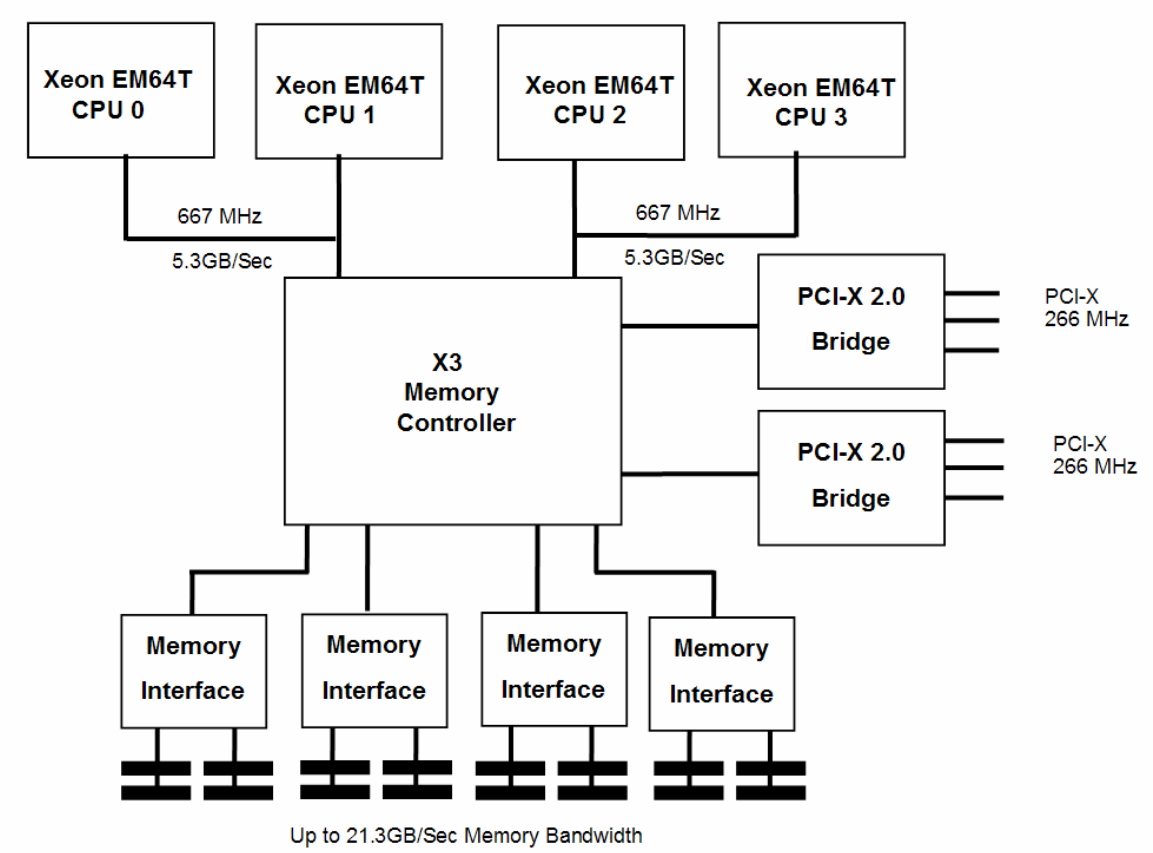

Figure 1: IBM eServer xSeries 366 Architecture

The x366 chassis accommodates up to four 64-bit Intel® Xeon™ Processors MP and includes these major features:

- Enhanced microarchitecture
- 90 nm process technology
- HyperThreading Technology
- Support up to four Intel Xeon Processors MP at 3.66GHz with 1MB L2 cache
- Front-Side Bus (FSB) enhancements that include a 667 MHz FSB on 133 MHz clock and speed that is 67% faster than current 4-way buses

The x366 is designed to accommodate dual-core 64 bit Intel® Xeon™ Processors MP when available. This enables investment protection for future versions of Intel processors.

This Database Cluster proof-of-concept required that large amounts of highly available disk storage be connected to the cluster. The storage subsystem was designed around the IBM TotalStorage® DS4500 Storage Server. This is a RAID storage subsystem that contains Fibre Channel interfaces to connect both the host systems and the disk drive enclosures. With its 2Gbps controllers and high-availability design, the DS4500 delivers the throughput to support this proof-of-concept.

For more information about the IBM eServer xSeries 366, visit IBM's Web site:

www-1.ibm.com/servers/eserver/xseries/

Oracle Database 10*g* RAC and Cache Fusion

Oracle Database 10*g* RAC can be used in a large, flexible cluster or in the consolidation of multiple Oracle workloads (or clusters) to a single cluster. Oracle10*g* RAC capabilities include:

- Availability— Oracle Database 10*g* RAC is fault-resilient and allows nodes to join an application in the event of a down server.
- Scalability—Applications scale well due in part to Oracle's Cache Fusion technology.
- Flexibility—Multiple Oracle database applications can share a SAN from within a single cluster, reducing administrative overhead, and nodes can be provisioned from one application to another.

Cache Fusion allows the database buffers that are cached in each Oracle instance to behave as one large global cache. This is enabled by an interconnect which quickly transfers buffers to a requesting cluster node transparently to the application.

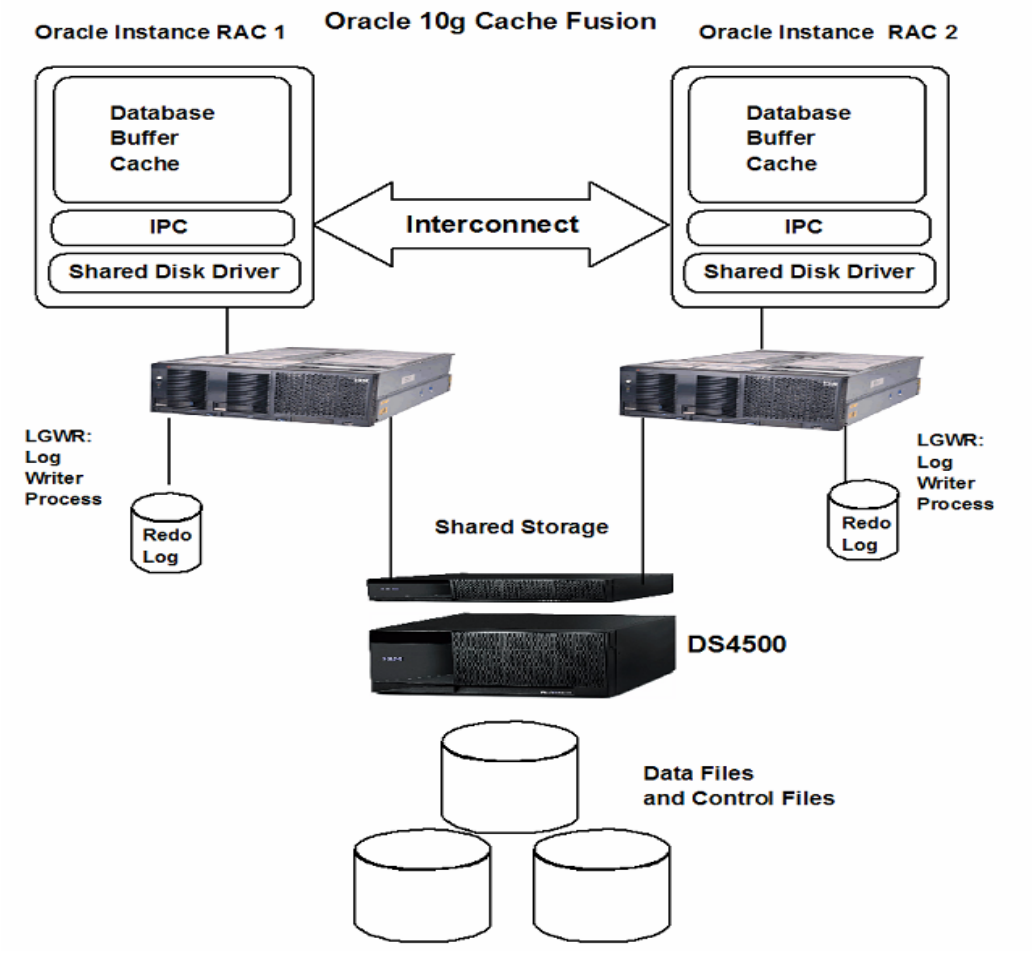

Figure 2: Oracle Cache Fusion

IBM xSeries continuously improves system design to enhance performance in a database environment. The following table shows a comparison of the previous generation server, the x365, and the recently announced x366.

Feature	xSeries 365	xSeries 366	Benefits of x366 to Oracle
Processor	32-bit	EM64T (64-bit)	Addresses a larger System Global Area (SGA)
Front-Side Bus	Single 400MHz	Twin 667MHz	Separate buses reduce interruption of Oracle processes running on other CPUs and improve efficiency.
Chipset	EXA-2	X3 Architecture	Isolates running Oracle processes from I/O traffic to reduce contention. Allows more efficient transfers of buffers from database drives to the SGA in memory while allowing Oracle processes to run with minimal interruption.
Memory	DDR PC2100	DDR2 PC3200	Reduces time to access Oracle database cache buffers in memory
I/O	PCI-X (4 slots at 133MHz)	PCI-X 2.0 (6 slots at 266MHz)	Increases bandwidth for RAC interconnects and disk subsystem.

Table 1: Comparison of x365 to x366 and Benefits of x366 to Oracle Database

For more information on Oracle RAC technologies, visit:

www.oracle.com/technology/products/database/clustering/index.html

Oracle Database 10*g* RAC and Interconnect

Oracle Database 10*g* supports advanced technologies such as InfiniBand for RAC interconnect. As the number of nodes per cluster rises, Oracle Database 10*g* RAC implementations will benefit from a high bandwidth, low latency interconnect such as InfiniBand. InfiniBand performance attributes include:

- High bandwidth – A 4X (X=2.5Gbps) InfiniBand Host Channel Adapter (HCA) has a theoretical capability of 10Gbps.
- Low latency – Low-level measurements have shown latencies of less than 10 microseconds (versus 100 microseconds for Gigabit Ethernet).
- Scalability – InfiniBand has capabilities to support 12X (30Gbps) throughput for future implementations.

In any Oracle RAC implementation, the transfer of blocks between instances is managed by the Global Cache Service Processes also known as Lock Management Service (LMS processes).

These LMS processes handle all update activity to the blocks and ensure that only one instance can make changes at a time. In Oracle 9*i* the LMS processes were limited to 10, in Oracle 10*g* it has been increased to 20. LMS processes are critical to performance and scalability because of their functions, which include:

- Coordinating block accesses – Ensure consistency of a particular block even if it appears in more than one RAC instance.
- Maintaining status – Update the status flag to maintain record (read/write access) of a particular image block.
- Forwarding blocks to the requesting instance – Handles and coordinates shipping blocks between database buffer caches of different instances.

Oracle Database 10*g* RAC and ASM

One of the new features in Oracle Database 10*g* is the Automatic Storage Management (ASM). ASM integrates the file system and volume manager into one instance to simplify management and optimize performance. Oracle Database 10*g* ASM Performance attributes include:

- Distribution of I/O

  ASM will distribute I/O across a disk subsystem to improve utilization and performance. The goal is to reduce the possible contention for any one disk. It will also reduce the amount of manual I/O tuning required to optimize performance.

- Redistribution of I/O

  ASM will redistribute I/O dynamically without user intervention. When disks are added or removed ASM will rebalance the data across the diskgroup. This rebalancing is transparent and adjustable. It can be configured to consume more or less bandwidth so as not to affect overall system performance.

- Fault Tolerance

  ASM supports several types of fault tolerance, including mirroring and triple-mirroring. If failure groups are created, ASM can intelligently distribute data between the groups to further protect against failure of a set of resources.

- Integration with Hardware Arrays

  It is possible to enable RAID protection in either ASM or in hardware. Enabling hardware protection allows greater flexibility in choosing RAID implementations. It also offloads RAID processing to the array controller and therefore reduces CPU utilization. Since overhead is reduced, the processor is able to take on other tasks.

- Stripe Size and Allocation Units

  ASM uses 1MB to stripe database file extents. It can also implement a smaller stripe size (128K) for devices that may benefit from finer granularity such as log files. These allocation units can work in tandem with stripe size implemented in the hardware storage arrays to further distribute I/O's to reduce contention.

For more information on Oracle ASM technologies, visit:

www.oracle.com/technology/products/database/asm/index.html

Red Hat Enterprise Linux

As one of the world's fastest-growing operating systems, Linux has been embraced by IBM and is now recognized as an operating system suitable for enterprise-level applications like Oracle Database 10*g* Real Application Clusters. Red Hat Enterprise Linux 3 Advanced Server is designed to support the high-end, mission-critical systems, and with the release of the Update 2 version, it supports EM64T architectures. Besides additional kernel packages for support of

EM64T, the 2.4 kernel was modified to include some of the most useful 2.6 kernel features, including:

- Native Posix Threads Library (NPTL) resulting in better scalability and faster multithreading

- Asynchronous I/O (AIO) permits applications to continue processing and not have to wait for I/Os to complete. This helps improve performance in IO intensive applications like Oracle 10*g* RAC

- The Translation Lookaside Buffers (TLB) is a small cache that stores the virtual to physical address mappings of currently accessed memory. TLB misses are expensive and the HugeTLB support provides a mechanism to manage memory in larger segments. This becomes critical to database server performance as these systems are configured with large amounts of RAM.

- Scheduler support for Hyperthreaded CPUs, which allows for correct identification of physical rather than logical CPUs and creates more efficient work queues within processors

- ReverseMap Virtual Memory, which is a kernel memory management subsystem that improves performance for memory constrained systems, NUMA systems and systems with large aggregate virtual address spaces

For more information on RedHat Enterprise Linux, visit:

www.redhat.com/software/rhel/

System Test Configuration

A two-node cluster was created to demonstrate the performance characteristics while running several database queries. The queries are unique and each stresses the system differently.

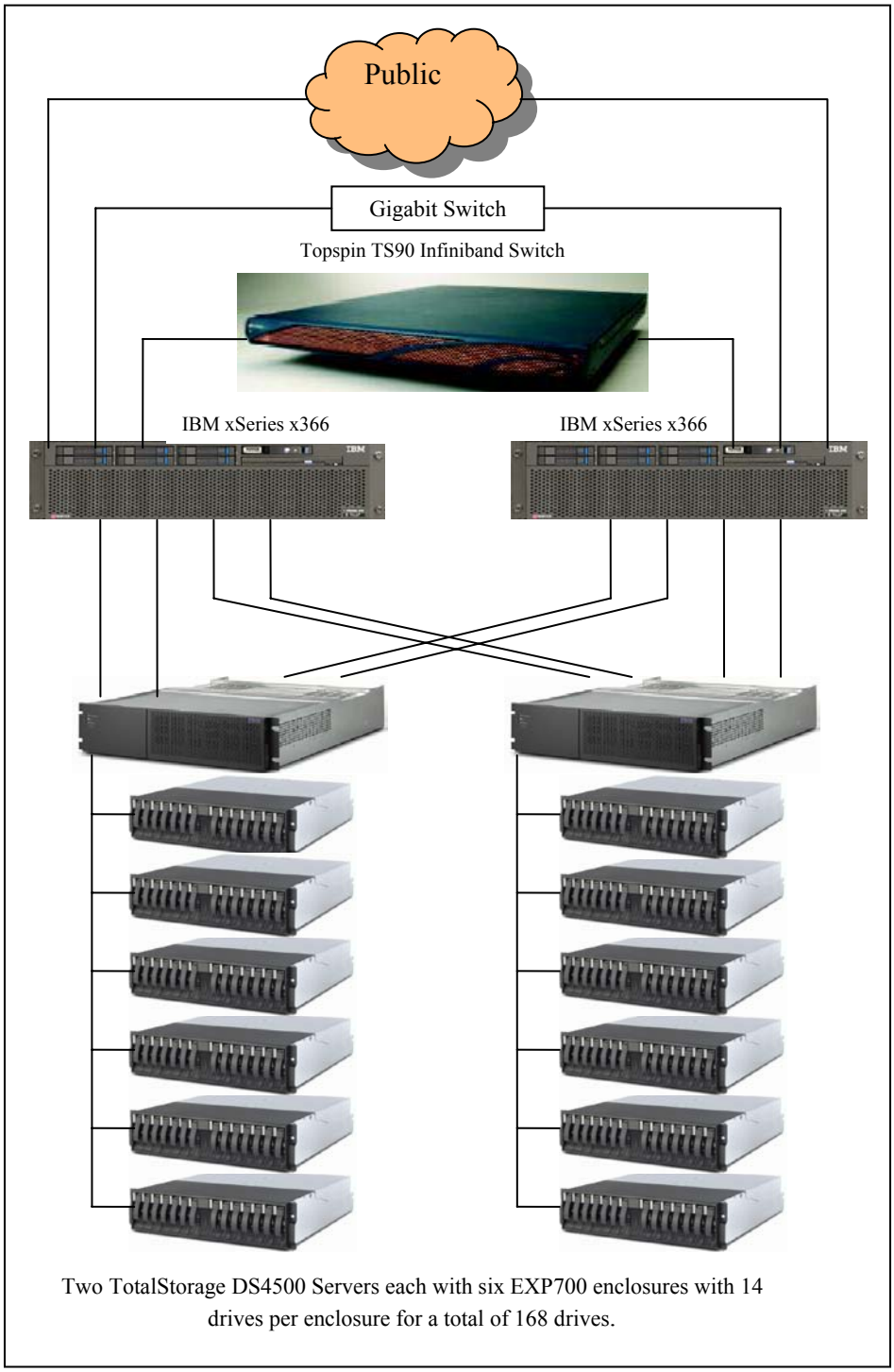

Figure 3: Components of the Database Cluster System – Two-Node x366 Cluster

Node Configuration

The xSeries x366 server is a 3U rack-optimized chassis with 4-socket 64-bit Intel Xeon Processors MP and designed to support future dual-core processors. Its design is centered around the IBM XA-64e third-generation chipset. The x366 can have up to four 4-DIMM memory cards of DDR2-based Active Memory for a total of 16 DIMM memory slots allowing up to 64GB of RAM to be configured within the system. It can accommodate up to six 2.5" SAS hot swap hard disks and can implement RAID-5 with the addition of an optional ServeRAID adapter. Internally, there are six 64-bit PCI-X 2.0 (266 MHz) slots.

For this proof point, each of the server nodes was configured with:

- Four 64-bit Intel Xeon Processors MP at 3.66GHz with 1MB L2 cache
- 32GB RAM
- One internal SAS drive for booting OS
- Four IBM TotalStorage DS4000 FC2-133 Host Bus Adapters
- One Topspin InfiniHost PCI-X HCA

Storage Configuration

The storage subsystem consisted of two IBM TotalStorage DS4500 Storage Servers, twelve DS4000 EXP700 drive enclosures, eighty-four 18.2GB HDDs, and eighty-four 36.4GB Fibre Channel hard disks. The 168 disks provided approximately 4.5TB of storage. The DS4500 with its dual active 2GB RAID controllers and 2GB battery-backed cache is an enterprise-class storage server designed for performance and flexibility in data-intensive computing environments. When combined with the EXP700 enclosures, the DS4500 can be used to help build a complete 2Gb SAN environment.

Each DS4500 server was configured with six fully populated EXP700 enclosures. The DS4000 TotalStorage Manager software was used to set up individual Logical Unit Numbers (LUNs). Each LUN appears as a device to the Red Hat operating system as /dev/sd<n>, where <n> is the device identifier. LUNs for the Oracle database storage were configured as:

- Twenty drives
- RAID-0[2]
- Segment Size 64K
- Read cache enabled
- Write cache without batteries
- Cache Read Ahead Multiplier = 1

[2] RAID-0 was used in this test environment in order to take advantage of highest I/O throughput. It is not recommended to run non-redundant storage subsystems in production environments.

Figure 4 depicts how "Array 1" was distributed across the 20 drives. Once configured, this LUN was available to the Red Hat operating system as the device "/dev/sdb".

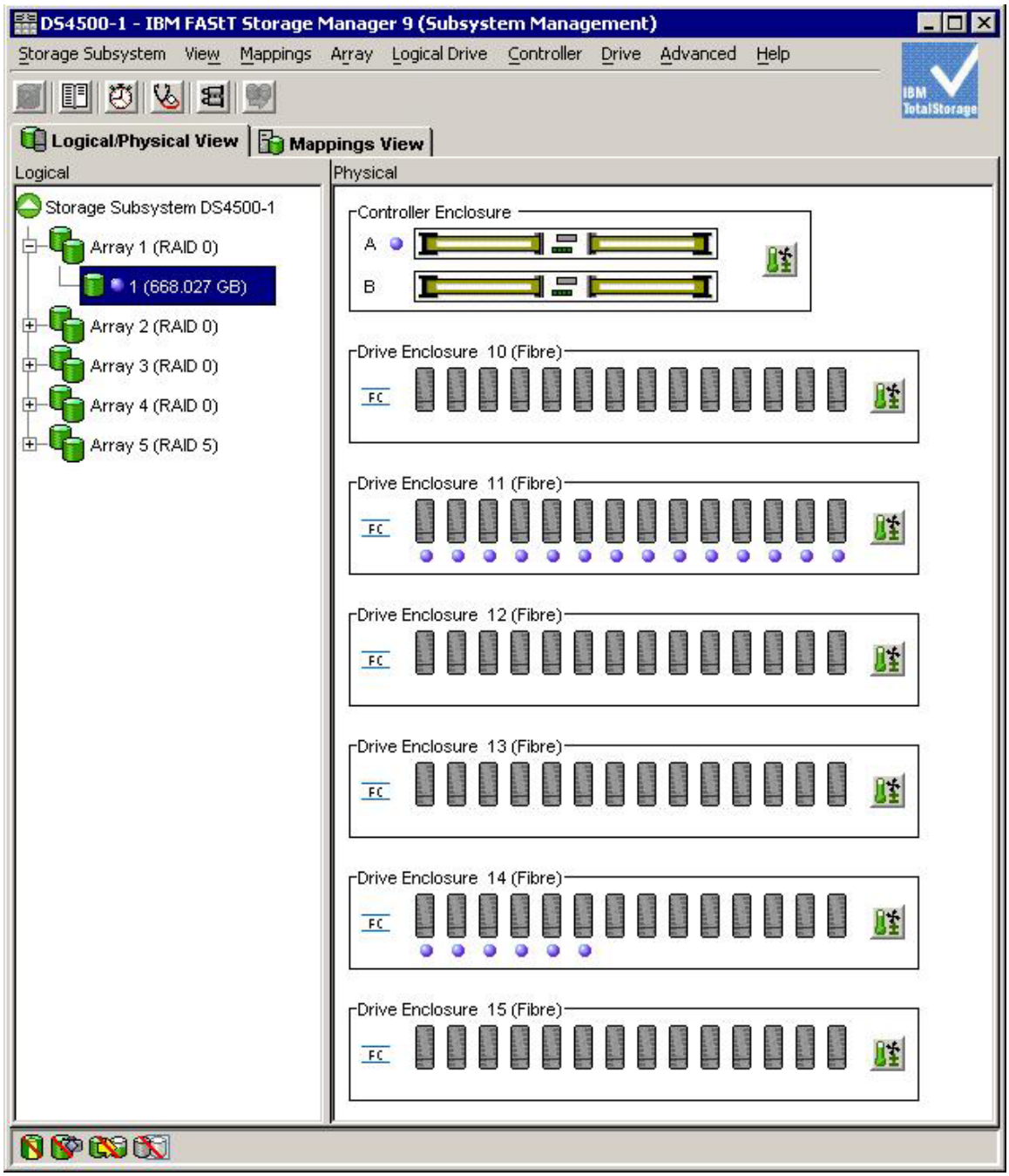

Figure 4: DS4500 Configuration of a LUN

Oracle's Automatic Storage Management (ASM) was used to simplify the configuration of the storage subsystem. Introduced as a new feature in Oracle Database 10*g*, ASM allows database administrators (DBAs) to define the disk groups that are managed internally by the Oracle kernel. ASM reduces the complexity involved with managing Oracle data storage by automatically managing the placement and naming of files, dynamically reallocating space and improving I/O throughput by distributing it across all disks in a diskgroup. Figure 5 shows how the diskgroups were defined as the repository for the database files. When creating the diskgroups, the "external

redundancy" option was chosen so that the RAID implementation was handled at the hardware layer. This frees up processor cycles to handle user requests as opposed to handling the mirroring functions.

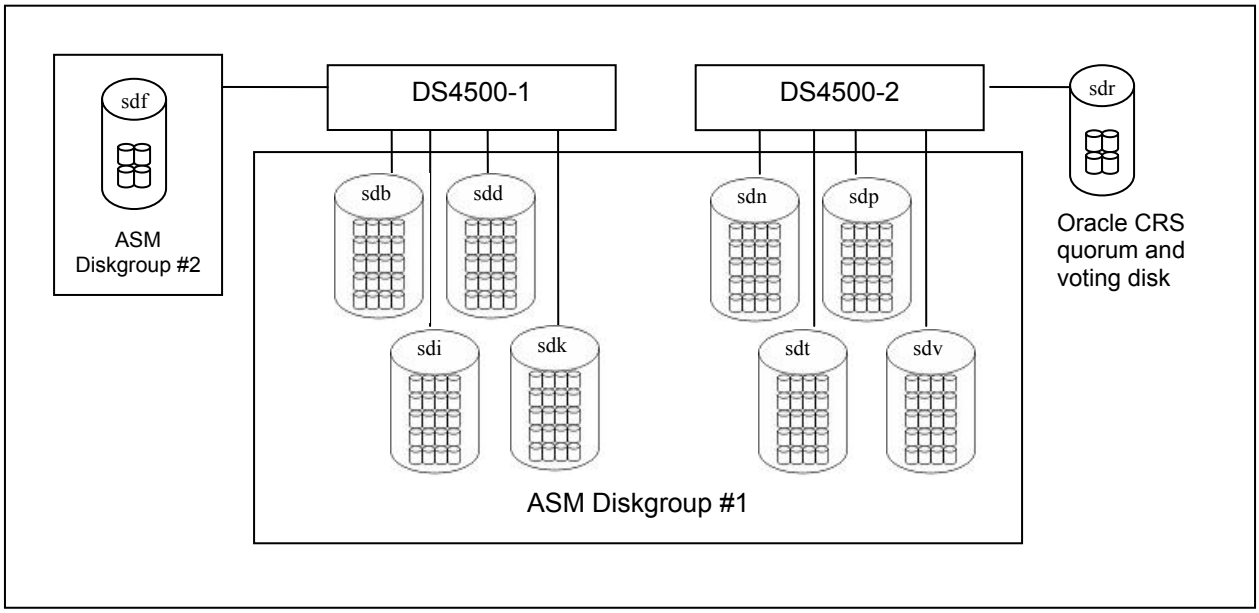

Figure 5: ASM Diskgroups

Network Configuration

Oracle documentation states a requirement that servers in a RAC configuration must have at least two network adapters; one dedicated public access for users connecting to the system and one for private access for Cluster Ready Services (CRS) heartbeat, Cache Fusion traffic and Inter-Process Communication (IPC). The x366's integrated dual Gigabit Ethernet interfaces were used to set up the initial private and public networks to complete the Oracle installation.

In addition to the onboard Gigabit interfaces, each server was configured with a Topspin InfiniHost PCI-X HCA in order to baseline and compare the performance of Gigabit vs. InfiniBand for the Oracle interconnects. The InfiniBand measurements used the IP over InfiniBand protocol (IPoIB).

Workload Measurements and Analysis

The goal of this proof-of-concept was to validate an all 64-bit operating system and Oracle Database 10*g* RAC installation on the IBM eServer xSeries x366 server based on the 64-bit Intel Xeon Processor MP. Some key concepts that were learned from this proof point include:

- CPU-bound queries scale when adding processor power and when going from a non-cluster Oracle configuration to a RAC database.
- The use of InfiniBand as an interconnect technology can improve performance when executing IPC-bound queries.

Scalability in a DSS Environment with CPU-Intensive Queries

The Decision Support System (DSS) database schema used in this proof point was based on a business intelligence workload and provided a way to execute analytical queries about customer purchases. The schema contained an order, line-item, customer, part, and supplier tables. Since this database simulates a DSS environment, the queries were executed as ad hoc queries with their elapsed times used as the major metric. The characteristics of DSS queries with heavy access to data through indexing, scanning, joining, sorting and aggregation makes the workload suitable to show an example of CPU scalability on the x366.

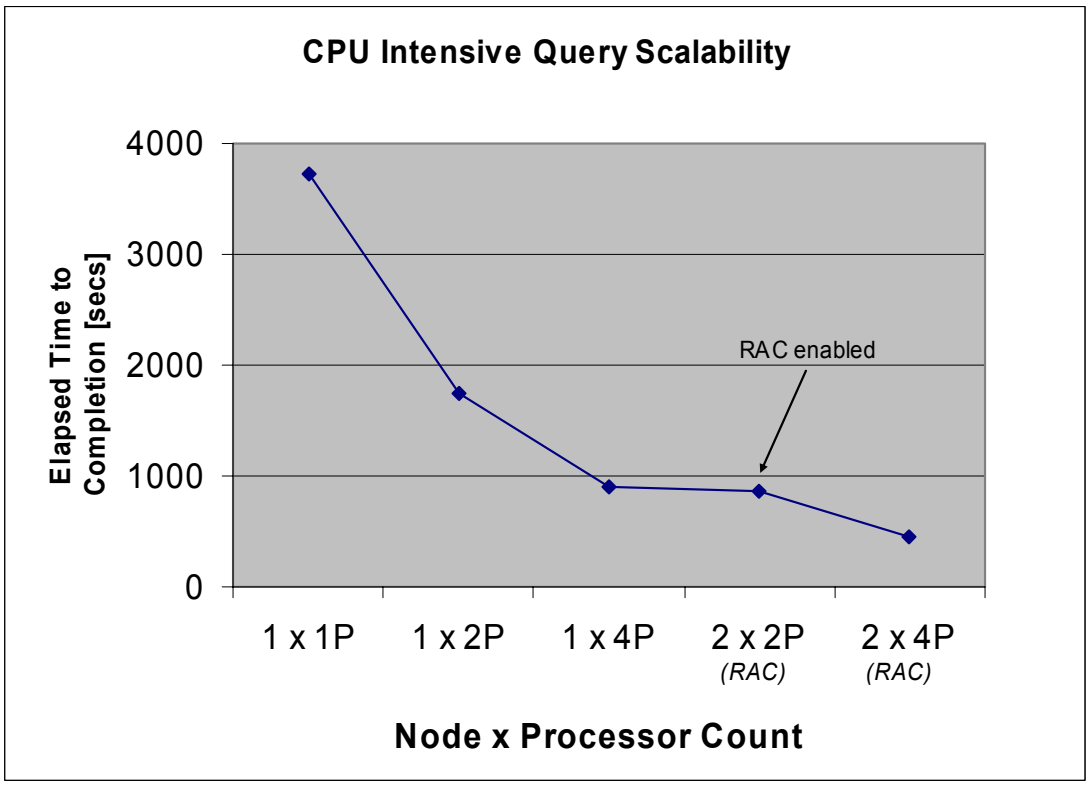

Figure 6: Scalability: CPU - Reduction in Completion Time

As expected, Oracle Database 10*g* RAC on the IBM xSeries 366 showed excellent scalability when running a CPU-intensive query. Scalability from 2P to 4P was 1.95, and from 4P to 8P scalability was 1.91 (2.0 would have been perfect scalability when doubling processors). The measurements shown in Figure 6 also show the impact when enabling RAC. This is shown in the

1 x 4P (four processors on one node) compared to 2 x 2P (two processors on each of two nodes) measurement, which experiences little performance difference. There is minimal overhead associated with running Oracle Database 10*g* RAC for this particular query.

Scalability in a DSS Environment with I/O Bound Queries

Table scans are the basis for many queries in a DSS environment leading to a demand for a robust storage subsystem.

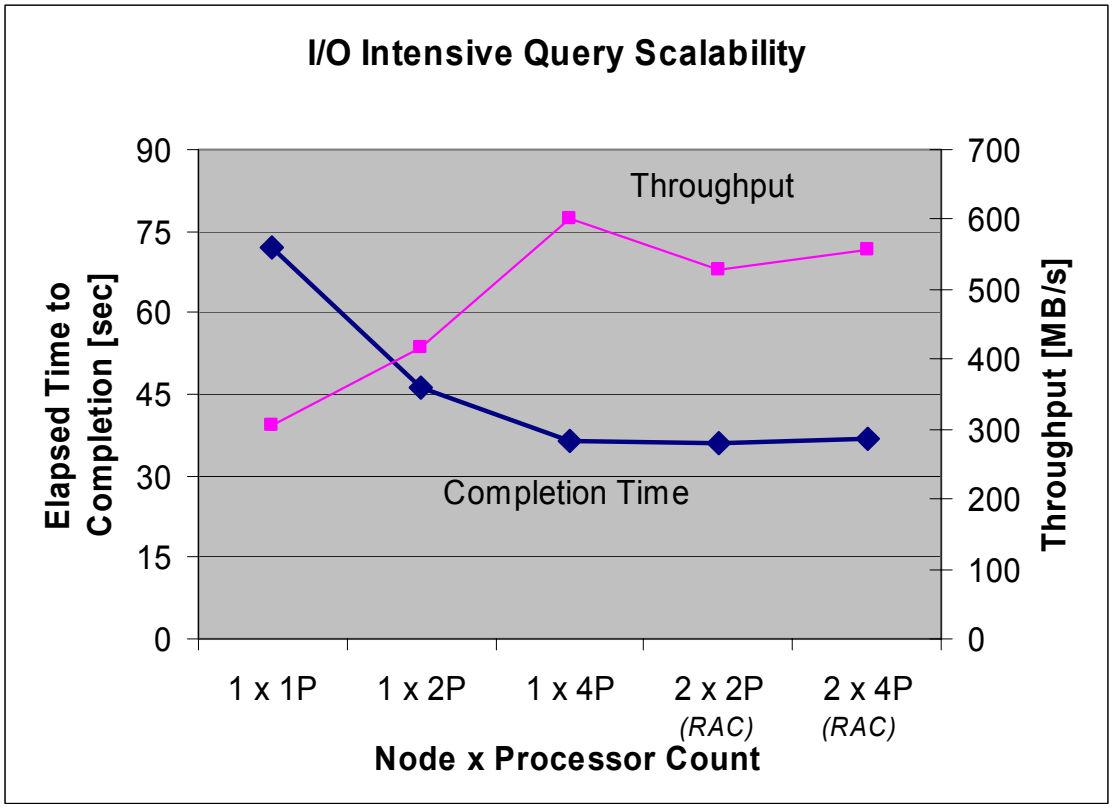

Figure 7: I/O-Intensive Query

The measurements show good scalability from 1P to 2P and some improvement from 2P to 4P; however, there is little improvement beyond four processors for this query. Compare this to the previous measurements, which showed excellent scalability. The throughput for the disk subsystem was included on the chart to explain the performance result. This query created a disk drive bottleneck (notice the essentially flat throughput line from 4P to 8P); therefore, the completion time remained flat from 4 to 8 processors as well. This configuration requires a faster disk subsystem (i.e., more drive spindles) to keep up with the true potential scalability of the x366.

An obvious but often overlooked conclusion can be made from comparing the two previous charts. Both measurements used the same hardware, operating system, and database software but yielded different scalability measurements. It was the unique workloads that stressed different bottlenecks and therefore produced different performance results. These results show how critical it is to understand the exact production workloads when making comparisons of hardware/software platforms. It is not enough to ask how well a particular system scales; the question should be how well does this system scale running a specific software stack and a specific workload.

Interconnect Analysis

In the Oracle Database 10*g* RAC "shared everything" architecture, the cluster interconnect becomes a very important component as users embrace the scale-out deployment of solutions to achieve high-end performance from industry-standard servers. Oracle's Cache Fusion is the method by which multiple servers in a cluster can share the global cache and reduce the need to access the I/O subsystem in order to guarantee data consistency. Oracle documentation recommends Gigabit Ethernet as a standard solution. However, as node count in cluster deployments increases, and the potential for performance bottlenecks increases, there is an opportunity to evaluate new technology for the cluster interconnect.

InfiniBand with its low-latency and higher bandwidth is a solution that can be used to reduce potential IPC bottlenecks as node counts increase. Support for InfiniBand fabrics as the cluster interconnect is straightforward and does not need additional software from Oracle. Once InfiniBand drivers are installed on the server, creating the interface is completed with the regular Linux network tools.

The test setup had a Topspin dual-port InfiniHost PCI-X HCA installed in each x366 node. Once the drivers[3] were installed, there were two additional network interfaces available for configuration: ib0 and ib1. The **ifconfig** command was used to assign network addresses to these new interfaces. A Topspin TS90 12-port InfiniBand switch was used to connect the nodes. Once IP over InfiniBand (IPoIB) communication was established by successfully pinging the servers, the init.ora parameter files for each of the nodes instances were modified to include the cluster_interconnects=xx.xx.xx.xx to override the default interconnect interface from the initial Oracle product installation.

A test scenario was designed so that in a 2-node cluster, one node would act as the "master" node holding a lock on table row data, and the second node would act as a "client" and request the blocks from the first node. Since the query was created to force a remote cache access, this would force all blocks to be transferred across the interconnect forcing all Cache Fusion traffic. The number of client processes was increased in order to stress the technologies.

[3] Topspin drivers used for this proof point were topspin-ib-rhel3-3.0.0-160 and topspin-ib-mod-rhel3-2.4.21-27.EL-3.0.0-160.

The measurements in Figure 8 show that there is a performance advantage to using IPoIB with regards to throughput of global cache consistent read (database buffer cache) blocks and the response time on those requests. By using IPoIB, the client node was able to process approximately 75% more database buffer cache blocks than Gigabit Ethernet.

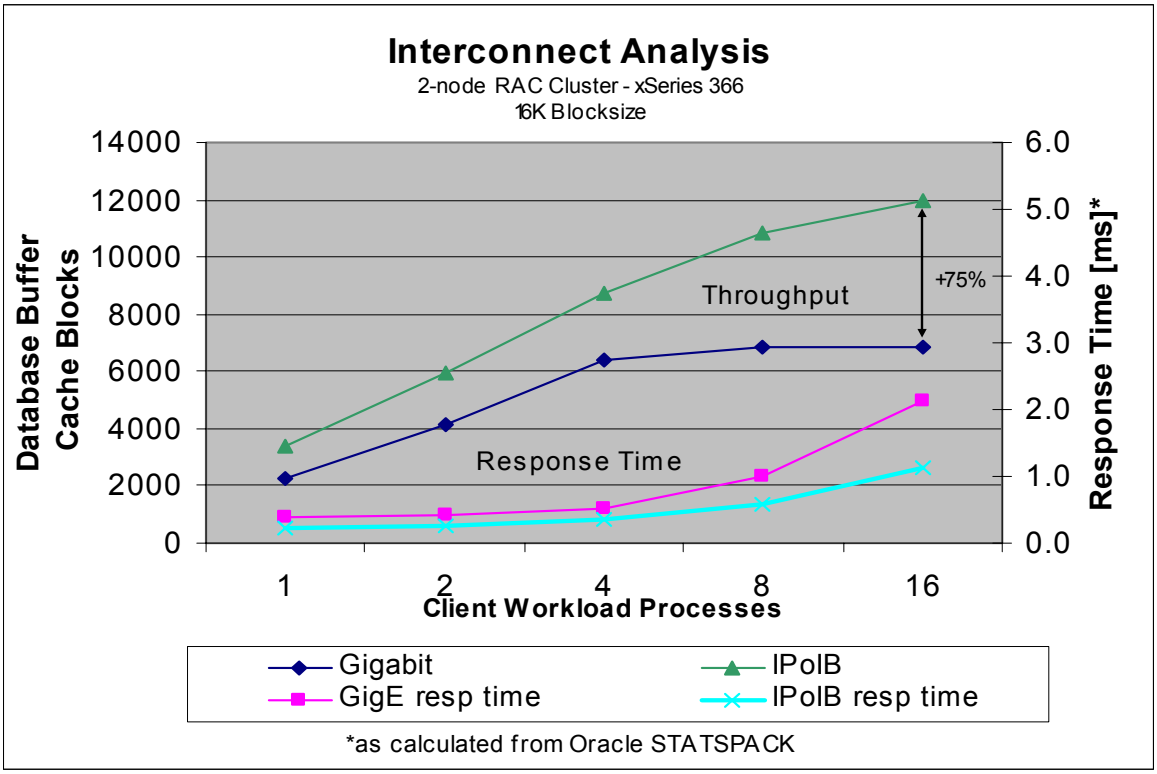

Figure 8: Interconnect Throughput

The measurement shown here was intended to show the capabilities of the interconnect under an intense workload. A typical 2-node cluster would not be able to drive the interconnect to saturation in a production environment. This measurement proves there are interconnect technologies in place that will enable larger scale-out environments in the future.

Summary

The goal of this proof-of-concept was to validate the IBM xSeries x366 as a 64-bit hardware platform as capable of running 64-bit applications today. The analysis presented in this paper confirms that:

- The new performance attributes of the x366 make it an ideal Oracle Database 10*g* cluster node. These attributes include faster 64-bit enabled processors, greater memory addressability and greater I/O bandwidth.
- The x366 supports Red Hat Enterprise Linux 3 – Update 4 and Oracle Database 10*g* RAC (10.1.0.3) for x86-64.
- InfiniBand as the interconnect technology has the performance bandwidth to enable larger scale-out environments. This technology is available and supported on the IBM x366 and Oracle Database 10*g* RAC.
- IBM partners with vendors such as Oracle, Red Hat and Topspin to provide total solutions with optimized performance.

For more performance-related information, visit the IBM eServer xSeries Benchmarks Web site:

www.pc.ibm.com/ww/eserver/xseries/benchmarks/

IBM Systems and Technology Group

Department 23U

Research Triangle Park, NC 27709

Produced in the USA.

4-05

Visit *www.ibm.com/pc/safecomputing* periodically for the latest information on safe and effective computing. Warranty Information: For a copy of applicable product warranties, write to: Warranty Information, P.O. Box 12195, RTP, NC 27709, Attn: Dept. JDJA/B203. IBM makes no representation or warranty regarding third-party products or services including those designated as ServerProven or ClusterProven.

KIND, EITHER EXPRESS OR IMPLIED, INCLUDING THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE. Some jurisdictions do not allow disclaimer of express or implied warranties in certain transactions; therefore, this statement may not apply to you.

This publication may contain links to third party sites that are not under the control of or maintained by IBM. Access to any such third party site is at the user's own risk and IBM is not responsible for the accuracy or reliability of any information, data, opinions, advice or statements made on these sites. IBM provides these links merely as a convenience and the inclusion of such links does not imply an endorsement.