Demartek

May 2012

High Performance Oracle RAC Architecture with 16 Gbps Fibre Channel

Evaluation report prepared under contract with Brocade

Executive Summary

Today's datacenters face a myriad of challenges brought by mounting data volumes, larger user populations, and increasingly more complex data analyses. Many bandwidth-intensive applications such as data warehousing, data mining, high-end video production, seismic data interpretation and high-performance computing applications need faster throughput to satisfy complex, large-scale challenges.

In the majority of cases, enterprises are constrained by the available bandwidth of the connections between the servers and the storage.

Recent innovations in networking and storage technologies can help alleviate bandwidth issues for high traffic applications. With the recent introduction of networking solutions based on the new 16 Gbps Fibre Channel standard, bandwidth constraints are now loosened. In addition to lower power consumption, these next-generation SAN products provide consistent performance with deterministic latency which is critical for many of these environments.

Solid-state storage (SSD) is another innovative, emerging technology within data centers that is proving its ability to provide extremely good performance for applications that demand high performance and low latency. SSD-based storage systems deliver storage speeds that are far greater than are even theoretically possible or economically feasible with conventional, magnetic storage devices. To fully make use of this speed, SSDs typically connect to servers or networks through multiple high-speed channels.

When combined, these two innovative technologies provide excellent performance and low latency and are a compelling solution for bandwidth-intensive applications such as Oracle Real Application Clusters (RAC).

Demartek was commissioned to review a new Brocade/Texas Memory Systems reference architecture and validate its performance. This test scenario was designed to simulate a typical environment deployed in thousands of mid-to large enterprises. In this testing evaluation, the reference architecture showed significant bandwidth of more than 7,200 MB/sec with a three-server Oracle Real Application Cluster (RAC) configuration using five host ports of 16 Gbps FC connected to a high-performance SSD storage solution. In reviewing benchmark results from the Storage Performance Council, this same level of performance requires at least 256 15,000 RPM hard disk drives.

Oracle RAC Background

Oracle RAC is a clustering option within the Oracle relational database management system (RDBMS) that allows multiple computers to run Oracle RDBMS software simultaneously while accessing a single database. In an Oracle RAC environment, two or more computers (each with its own Oracle RDBMS instance) concurrently access a single database. This allows an application or user to connect to either computer and have access to a single coordinated set of data. This improves the ability of an organization's users, such as finance and human resources, to collaborate, enhance productivity and make more intelligent business decisions.

For most enterprises, Oracle RAC environments present considerable challenges to storage area networks (SANs), especially when competing for bandwidth resources during peak periods, such as preparing for quarterly financial reporting. As the number of users and the size of the database increase, along with the need of the business to perform more complex queries and analysis, performance typically diminishes.

Traditional methods to improve performance, such as adding memory to servers, increasing the number of servers tied to the database, or fine-tuning the Oracle database operations, usually produce diminishing returns, resulting in additional complaints from users.

Recently, Brocade and Texas Memory Systems developed a reference architecture designed to improve the performance of database clusters, including Oracle RAC. Components of the reference architecture include:

- Brocade 6510 Data Center Switch
- Brocade 1860 Fabric Adapter
- Texas Memory Systems RamSan-630 SSD Storage Array

Application Environment and Limitations of Current Solutions

Oracle RAC is a widely deployed, classic example of an application that places significant loads on SANs and as a result, creates major bandwidth headaches for IT staff.

When system administrators look to storage they frequently try different approaches to resolve performance problems:

- **Increase the number of disks –** By increasing the number of disks, the I/O from a database can be spread across more physical devices. As most applications do not have the high levels of concurrency that drive more parallel requests to the storage, this generally has a trivial impact on decreasing the bottleneck.
- **Adding controller cache in hopes to avoid access to disk –** As rotating disks are not currently available over 15,000 RPM, few databases and other applications can be moved to a faster tier of rotating disks. This leaves storage administrators with few choices when staying with the legacy technology.
- **Increase processing power –** Often, additional processing power alone will do little or nothing to improve Oracle performance. This is because the processor, no matter how fast, finds itself constantly waiting on mechanical storage devices for its data. While every other component in the "data chain" moves in terms of computation times and the raw speed of electricity through a circuit, hard drives move mechanically, relying on physical movement around a magnetic platter to access information.

In the last twenty years, processor speeds have increased at a geometric rate. At the same time, however, conventional storage access times have only improved marginally. The result is a massive performance gap, felt most painfully by database servers, which typically carry out far more I/O transactions than other systems. Extremely fast processors and massive amounts of bandwidth are often wasted as storage devices take several milliseconds just to access the requested data.

When servers wait on storage, users wait on servers. This is I/O wait time, also known as latency. Solid state disks are designed to solve the problem of I/O wait time by offering as much as 25x faster access times (0.2 milliseconds instead of 5) and up to 200x more I/O transactions per second (1,000,000 instead of 5,000) than a hard disk RAID.

Benefits of the High Performance Oracle RAC Reference Architecture

The new reference architecture developed by Brocade and Texas Memory Systems deploys next-generation technology in order to significantly increase performance and scalability. Components of the reference architecture include

- New six-core servers that have one rack unit (1U) form factor
- 16 Gbps Fibre Channel switches and adapters from Brocade
- All-flash/SSD storage arrays from Texas Memory Systems

Brocade 6510 Switch

The Brocade® 6510 Switch provides exceptional price/performance value, combining flexibility, simplicity, and enterprise-class functionality for virtualized data centers and private cloud architectures. Designed to enable maximum flexibility and investment protection, the Brocade 6510 is configurable in 24, 36, or 48 ports and supports 2, 4, 8, 10, or 16 Gbps speeds in an efficiently designed 1U package. It also provides a simplified deployment process and a point-and-click user interface — making it both powerful and easy to use.

Brocade 1860 Fabric Adapter

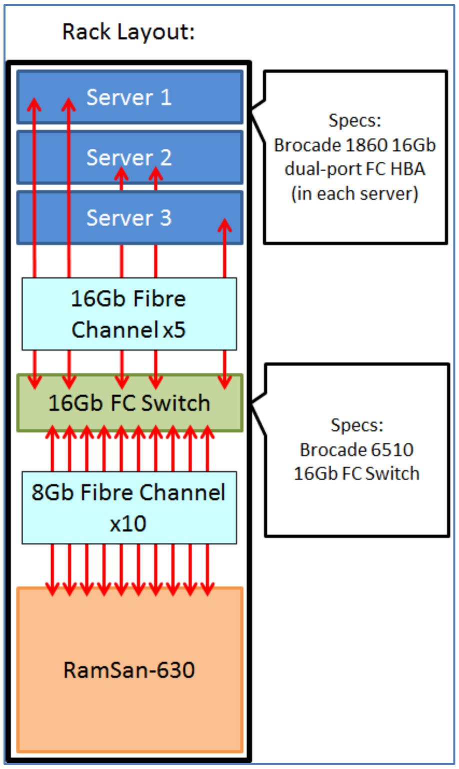
Brocade 1860 Fabric Adapters are a new class of “stand-up” network adapter cards for servers with standard PCIe-based multi-function expansion slots. They feature Brocade’s industry-unique AnyIO™ technology that allows any individual adapter port to be configured “on-demand” by software command as either a 16 Gbps Fibre Channel (FC) Host Bus Adapter (HBA), 10GbE Converged Network Adapter (CNA) or 10GbE Network Interface Card (NIC). With a dual-port Brocade1860 Fabric Adapter, FC and DCB FCoE, TCP/IP and iSCSI I/O protocols can all be running simultaneously on the same adapter card. The Brocade 1860 provides line-rate 16 Gbps FC performance data streaming bandwidth and over 1 million low-latency transaction IOPS per dual-port adapter. Most importantly, the Brocade 1860 can consolidate and dramatically reduce the number of network adapters required in a server while maintaining high-availability hardware redundancy such that cost/space-efficient 1U servers with as few as two PCIe slots can be deployed in the reference architecture solutions.

Texas Memory Systems RamSan-630

The RamSan-630 rack-mount solid state storage system provides 10 TB of shareable, high performance storage for IT organizations that need to respond to the growing storage and performance needs of their users and applications. One RamSan-630 system can replace the performance of an entire rack of high-end hard drives. SLC Flash and innovative controller designs give the RamSan-630 enterprise reliability and data protection features.

With a usable capacity of 10 TB in a 3U enclosure that uses only 500 Watts, the RamSan-630 can handle data growth very efficiently. It installs quickly and easily, integrating seamlessly into almost any SAN environment using up to 10 Fibre Channel or InfiniBand interfaces. It is certified by IBM for use with IBM SAN Volume Controller, providing additional management and reliability features. Capable of over 800,000 IOPS and 8 GB/s bandwidth with Fibre Channel, the versatile RamSan-630 is designed for high performance computing, data warehousing, and batch processing applications.

Test Environment and Performance Results



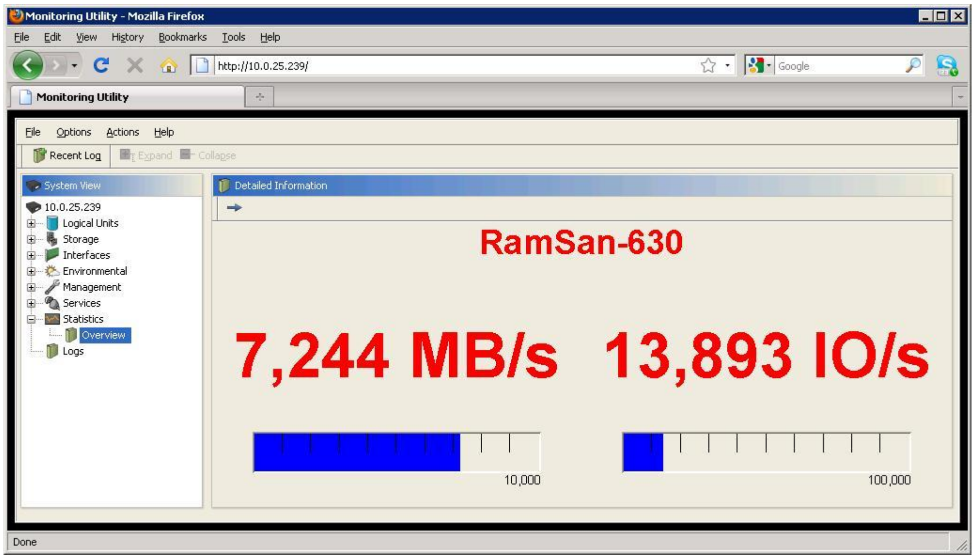
Demartek audited the tests performed at Texas Memory Systems using a combination of 16 Gbps Fibre Channel infrastructure from Brocade and a Texas Memory Systems RamSan-630 rackmount solid state storage system. The RamSan-630 provides 10TB of shareable, high-performance SLC Flash storage in a three rack unit (3U) form factor and consumes only 500 watts.

We configured an Oracle RAC environment using three servers, each having a dual-port Brocade 1860 16 Gbps Fabric adapter. These adapters support both 16 Gbps Fibre Channel and 10GbE DCB for TCP/IP, iSCSI and FCoE. For these tests, we used Fibre Channel connectivity only. We connected five of the six total 16 Gbps FC host ports to a Brocade 6510 Fibre Channel switch. A RamSan-630 storage system was connected to that switch with ten (10) 8Gb Fibre Channel connections. This provided 80 Gb/sec total bandwidth from the servers, and 80 Gb/sec total bandwidth from the storage.

The test consisted of a simple SQL Select statement that accessed all the rows from the database. The database contained 10.2 billion records.

We ran this set of tests multiple times and observed very consistent performance across the multiple runs. We achieved a sustained rate of approximately 7200 MB/sec during each of the test periods. The RamSan-630 provides a management interface that provides real-time performance statistics, shown below.

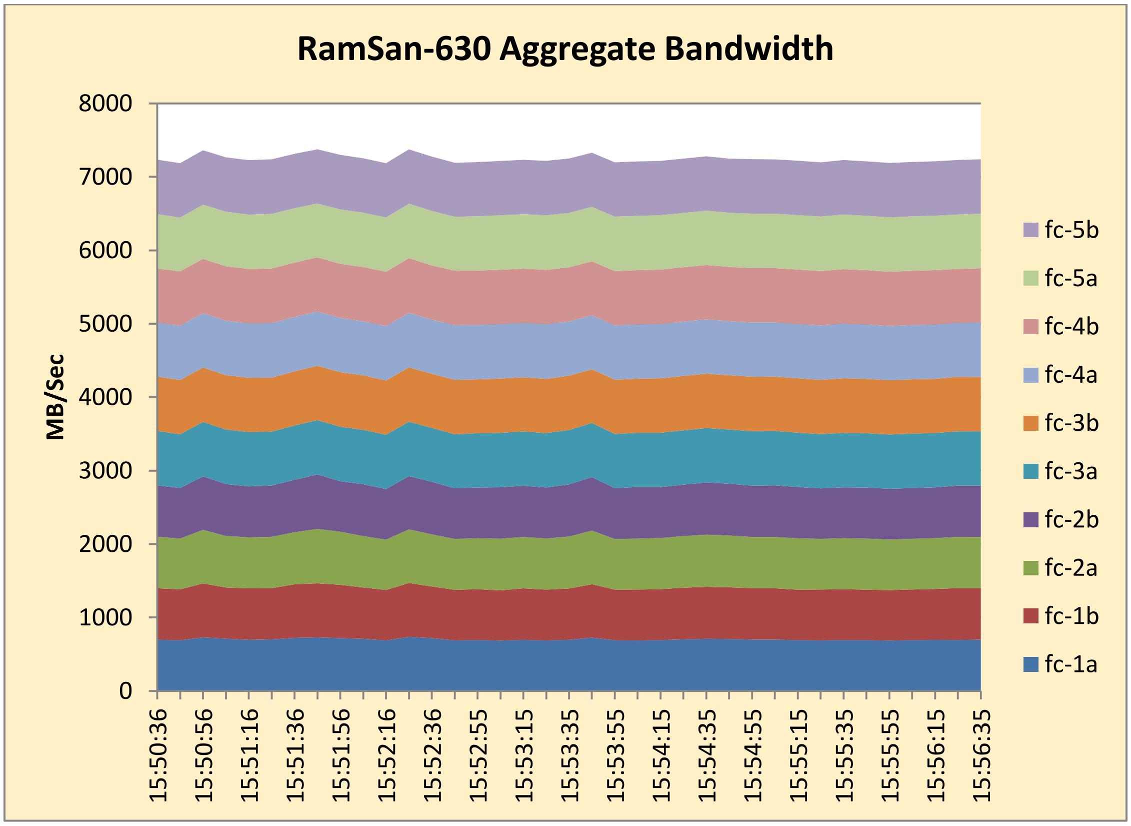
Improving Oracle RAC performance at the network and storage layers is a better option than thre traditional approach of deploying more servers. Often, additional processing power alone does little or nothing to improve database and application performance and typically results in increased Oracle licensing costs. This is due to the processor. Because of the performance discrepancy against mechanical disks, the processor finds itself constantly waiting for the data to present to the user or to analyze. Systems running at 20 percent CPU utilization will only see a 10 percent recovery when doubling CPU performance as the 80 percent of processor waits remain the bottleneck.



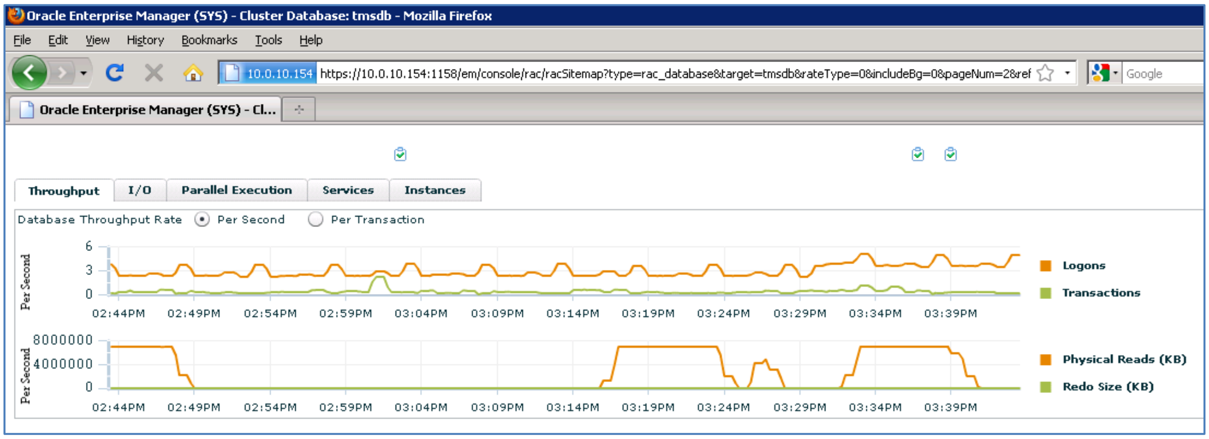
RamSan-630

7,244 MB/s 13,893 IO/s

The RamSan-630 also provides individual port statistics that can be exported. This is shown below.



The Oracle database system also provides a large number of statistics regarding the performance of the system from the host perspective. The Oracle throughput statistics graphs are shown below. Oracle reported approximately the same throughput rate as was reported by the RamSan-630.



Database Throughput Rate ● Per Second ○ Per Transaction

What to Move to SSD Drives?

Database performance is very important to company profitability. There is some subset of databases that help companies make more money, lose less money, or improve customer satisfaction if they process faster. Solid state disks can help make these companies more profitable.

Following deployment of the Brocade/TMS reference architecture to address I/O subsystem problems, the next step is to determine which components of your Oracle database are experiencing the highest I/O and in turn causing I/O wait time. The following database components should be looked at:

Entire Database

There are some databases that should have all of their files moved to solid state disk. These databases tend to have at least one of the following characteristics:

- **High concurrent access** – Databases that are being hit by a large number of concurrent users should consider storing all of their data on solid state disk since, as we know from the previous section, each user process in Oracle does its own disk reads. This will make sure that storage is not a bottleneck for the application and maximize the utilization of servers and networks. I/O wait time will be minimized and servers and bandwidth will be fully utilized.
- **Frequent random accesses to all tables** – For some databases, it is impossible to identify a subset of files that are frequently accessed. Many times these databases are effectively large indices themselves.

Large databases needing shorter reporting/analytics times

Given the fixed costs associated with architecting a RAID system for performance (buying a large cache, buying a lot of spindles for striping), it is economical and much faster to buy a RamSan Flash solution in order to accelerate large databases. RamSan Flash systems offer Performance/Capacity ratios much higher than disks. If the performance requirement of the database exceeds the performance of disks for the same capacity, excessive disk striping will quite often lead to a cheaper solution of RamSan Flash systems.

Redo Logs

Redo logs are one of the most important factors in the write performance for Oracle databases. Whenever a database write occurs, Oracle creates a redo entry. Redo logs are used in sequence with the best practice configuration using mirrored redo log groups, a minimum of two groups is required. Each redo entry is written to the two or more mirrored redo logs. Oracle strongly encourages the use of mirrored redo logs so that a backup redo log is available in the event of a failure. The operation is considered committed once the write to the redo logs is complete. Redo logs are used with linear output, if desired; the administrator can also configure redo logs to automatically archive. Archiving makes a copy of a filled log to another location before it can be reused. Archiving can be another source of waits in a slow disk based system.

The redo logs are a source of constant I/O during database operation. It is important that the redo logs are stored on the fastest possible disk. Writing a redo log to a solid state disk is a natural way to improve overall database performance.

Indices

An index is a data structure that speeds up access to database records. An index is usually created for each table in a database. These indices are updated whenever records are added and when the identifying data for a record is modified. When a read occurs an index is consulted so that Oracle can quickly get to the correct record. Furthermore, many concurrent users may read any index simultaneously. The activity to the disk drive is characterized by frequent, small, and random transactions. Under these conditions, disk drives are unable to keep up with demand and I/O wait time results. As data changes over time, indexes themselves will become fragmented and expose the performance limitations of spinning disks.

By storing indices on a solid state disk, performance of the entire application can be increased. For on-line transaction processing (OLTP) systems with a high number of concurrent users this can result in faster database access. Because indices can be recreated from the existing data, they have historically been a common Oracle component to be moved to solid state disk.

Temporary Tablespace

Temporary segments are used to support temporary data during certain Oracle operations. The temporary tablespace segments support complex sort, hash, global temporary table, and bitmap index operations. Because temporary tablespaces support many kinds of operations they can quickly become fragmented. In internal tests at Texas Memory Systems, the company found that Oracle database performance degrades quickly as data becomes fragmented.

When complex operations occur they will complete more quickly if the temporary tablespace is moved to solid state disk. Because the I/O to the temporary tablespaces can be frequent, disk drives cannot easily handle them.

Rollback Data

In databases with a high number of concurrent users, the rollback segments (undo tablespace in newer versions) can be a cause of contention. Undo data is created any time an Oracle transaction changes a record. In other words, if a delete command is issued, all of the original data is stored in the undo tablespace until the operation commits. If the transaction is rolled-back, then the data is moved from the undo tablespace back to the table(s) it was removed from.

Because the undo tablespace is hit with every change operation, it is useful to have the undo tablespace stored on solid state disk. This will provide fast writes when the update transaction is created and will make undo tablespace available more quickly for the next operation.

Frequently Accessed Tables

It is estimated that only 10 percent of data stored in OLTP systems is frequently accessed. These tables typically account for a large percentage of all database activity and thus I/O to storage. When a large number of users hit a table, they are likely going after different records and different attributes. As a result, the activity on that table is random. Disk drives are notoriously bad at servicing random requests for data. In fact, the peak performance of a disk drive drops as much as 95 percent when servicing random transactions. When a table experiences frequent access, transaction queues develop where other transactions are literally waiting on the disk to service the next request. These queues are another sign that the system is experiencing I/O wait time.

It makes sense to move the frequently accessed tables to solid state disk. SSD performance is not impacted if performance is random. Additionally, solid state disks by definition have faster access times than disk drives. Therefore, application performance can be improved up to 25x if frequently accessed tables are moved to SSD.

Conclusion

Obtaining performance efficiencies in bandwidth constrained environments such as Oracle RAC is a major pain point for IT and storage administrators. The combination of 16 Gbps Fibre Channel and SSD technology unlocks new levels of performance for bandwidth-intensive applications. As we have shown in this example, these levels of bandwidth can be achieved in a relatively small amount of rack space and power consumption.

Demartek observed a sustained database cluster performance of greater than 7200 MB/sec throughput using this new Brocade/Texas Memory Systems architecture using only three servers and a total of five 16 Gbps host bus adapter ports communicating with ten 8 Gbps storage target ports. This performance can be scaled linearly up by deploying additional 16 Gbps host ports and additional storage.

Appendix – Test Environment

Server Specifications (each of 3 servers)

- 1x Intel Xeon X3430, 2.4 GHz, 4 total cores (no hyper-threading)
- 8GB RAM
- RedHat Enterprise Linux 5.6, kernel 2.6.32-100.26.2.el5
- Oracle RAC 11.2.0.1.0
-

The original version of this report is available at www.demartek.com/Demartek_Brocade_Texas_Memory_Systems_16GFC_Oracle_RAC_Evaluation_2012-05.html.