blade.org

BLADE NETWORK TECHNOLOGIES

Chelsio Communications Accelerate

NetApp

A NETAPP COLLABORATIVE TECHNICAL REPORT

ORACLE 10*g* RAC: SEQUENTIAL WORKLOAD PERFORMANCE USING iSCSI AND NFS OVER 10GbE AND 4Gb FIBRE CHANNEL NETWORKS

Author: John Elliott, Network Appliance Inc.

Edited by: Blade Network Technologies, Chelsio Communications

November 2007 | TR-3628

EXECUTIVE SUMMARY

This paper demonstrates the increased throughput capabilities of an Oracle® 10*g*™ RAC database using a NetApp FAS6030 cluster storage system when accessed using today's faster transport technologies; specifically, 10 Gigabit Ethernet and 4 gigabit Fibre Channel.

TABLE OF CONTENTS

Table of Contents 2

INTRODUCTION 3

TEST LOAD DESCRIPTION 3

TEST CONFIGURATIONS AND SETTINGS 4

Hardware/Network Configuration............................ 4

Hardware and Software Configuration Details 5

Storage Provisioning Details for Fibre Channel and 10GbE/iSCSI Configurations 5

Storage Provisioning for NFS Configuration............ 7

NFS Mount Options 7

Oracle Parameter Settings (Nondefault Values) 8

Linux /etc/sysctl.conf Settings.................................. 9

/etc/security/limits.conf Settings.............................. 9

TEST RESULTS 9

Storage System Throughput.................................... 9

Database Throughput 10

Latency.. 11

CONCLUSION 12

ACKNOWLEDGEMENTS... 12

INTRODUCTION

The NetApp FAS6030 storage system was designed to handle very demanding enterprise applications. It utilizes the latest 64-bit CPU architecture and is compatible with the latest high-speed data transport technologies, including 10 Gigabit Ethernet (10GbE) and 4Gb Fibre Channel (FC). By utilizing these cutting-edge technologies, the FAS6030 provides the fast I/O features required by modern enterprise data warehouse operations.

To substantiate this claim, we utilized a three-node Oracle 10*g* RAC database running on IBM Blade Servers running Red Hat Enterprise Linux® Advanced Server 4. Our tests were performed with the following Oracle configurations:

- Oracle 10*g* RAC using NFS over 10GbE
- Oracle 10*g* RAC with Oracle Automatic Storage Management (ASM) using LUNs accessed over 10GbE with iSCSI
- Oracle 10*g* RAC with Oracle ASM using LUNs accessed over 4Gb FC

To fully demonstrate the impact of our high-speed data transport configurations, we utilized a workload that was very sequential in nature.

Note: Every effort was made to conform to Oracle and NetApp best practices as they were defined at the time the tests described in this paper were performed. However, best practices do occasionally change; therefore any discrepancy between this document and published best practices should be resolved in favor of the published best practices.

TEST LOAD DESCRIPTION

Our workload consisted of an Oracle query designed to simulate the analysis of parts inventories of suppliers across several different nations and to determine their suitability for promotional offers. Query execution resulted in sequential scanning of several tables along with joins and sort operations, which in turn generated a high volume of direct sequential read I/O from storage. The Oracle parallel query feature was used extensively to satisfy the queries' high data throughput requirements. To fully test network throughput with 10GbE and 4Gb FC, the database was sized with the goal of minimizing actual disk I/O. As a result, the storage system was able to cache in memory most of the data accessed by the test queries, thereby avoiding latencies associated with normal disk I/O. (Disk latencies are characteristic of all disk-based storage systems.) The data presented in this paper very closely reflects network throughput capabilities.

TEST CONFIGURATIONS AND SETTINGS

We used the following test configurations and settings.

HARDWARE/NETWORK CONFIGURATION

Figures 1 and 2 show block diagram representations of the two hardware configurations (4Gb/sec FC and 10GbE).

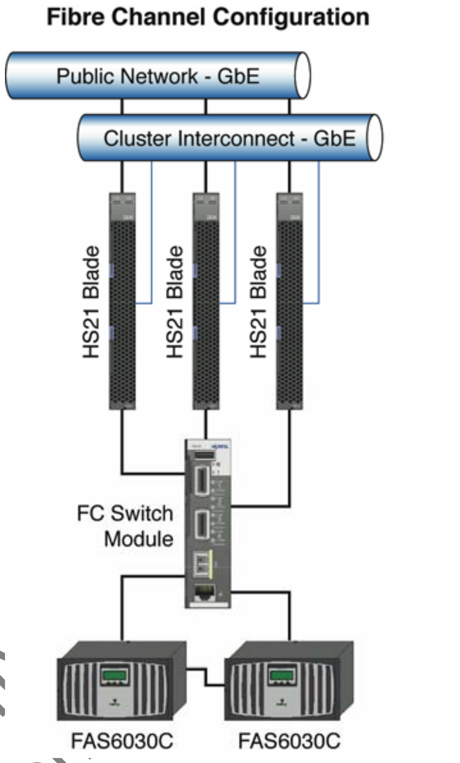

Figure 1) 4Gb/s FC configuration.

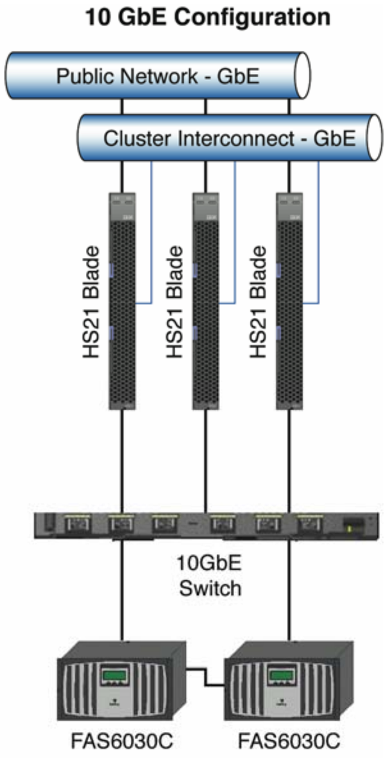

Figure 2) 10GbE configuration.

HARDWARE AND SOFTWARE CONFIGURATION DETAILS

Table 1) Configuration details.

	Item	Description
Software	Red Hat Enterprise Linux Release 4, Update 4	Operating System
	NetApp Data ONTAP® 7.2.3	Storage Controller Software
	Oracle 10*g* R2	RDBMS
Server	IBM BladeCenter H	Blade Chassis
	3-IBM HS-21	Blade Servers
	2-Dual Core 2.66 GHz Xeon®	Processors per blade
	8GB Physical Memory	Memory per blade
	Chelsio N310E-SR, 10GbE Server Adapters Driver Version: 1.0.113 Firmware Version: 4.6.0 Protocol SRAM Version: 1.1.0 Protocol EEPROM Version: 1.1.0	PCI Express Mezzanine Cards
	QLogic QMC2462S 4Gb/sec FC HBAs Driver Version: 8.01.07.15 BIOS Version: 1.04 Firmware Version: 4.00.26 Data Rate: 4Gb/sec Execution Throttle: 256	Fibre Channel HBAs per blade
Network	Blade Network Technologies 39Y9265 embedded switch Firmware Revision: 0100 Build ID: WMB01001	10GbE Layer 2/3 switch module blades, Jumbo Frames Enabled
	Embedded Brocade 32R1820 4G FC switch Firmware Revision: 504a Build ID: BREFSM	Fibre Channel switch module
Storage	FAS6030C	Active-active, highly scalable, enterprise class storage
	Chelsio S320e-SR NICs Driver Version: 1.0.113 Firmware Version: 4.6.0 Protocol SRAM Version: 1.1.0 Protocol EEPROM Version: 1.1.0	PCI-Express Network Interface Card
	Qlogic QMC2462S 4Gb/sec FC HBAs Driver Version: 8.01.07.15 BIOS Version: 1.04 Firmware Version: 4.00.26	Fibre Channel Target cards per controller

STORAGE PROVISIONING DETAILS FOR FIBRE

CHANNEL AND 10GbE/iSCSI CONFIGURATIONS

Storage layouts for the FCP and iSCSI configurations were identical. Figures 3 and 4 show the storage provisioning on both NetApp storage controllers. Note that a single data ASM disk group and a single log ASM disk group were created from the corresponding LUNs on both controllers of the storage cluster. In other words, a single disk group for data files was created from the three 100GB LUNs on controller 1 and the three 100GB LUNs on controller 2. In the same manner, a single disk group for log files and control files was created from the (four 10GB LUNs on both controllers. As shown in the figures, a single 3.1TB disk aggregate was created on each of the two storage systems. All of the volumes used by the database and clusterware were created in those aggregates.

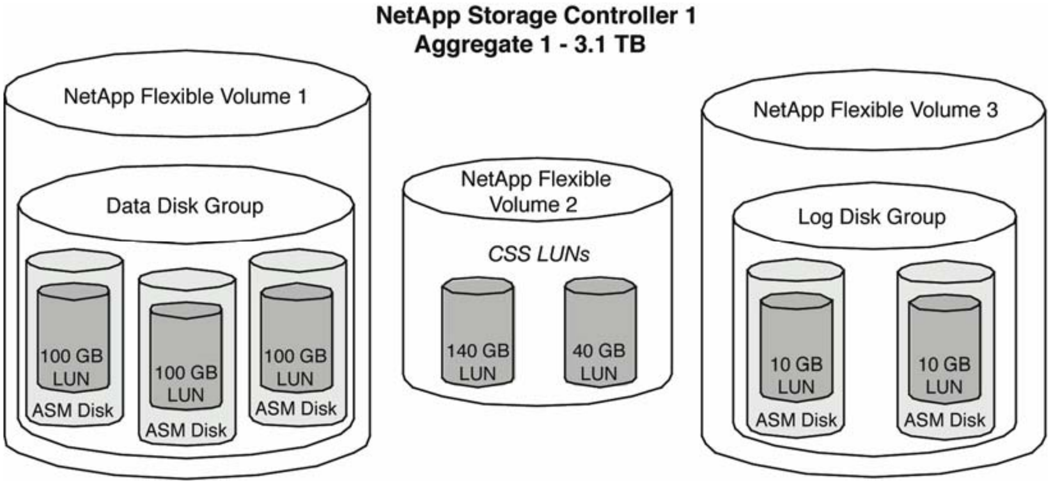

Figure 3) Storage provisioning on NetApp controller 1.

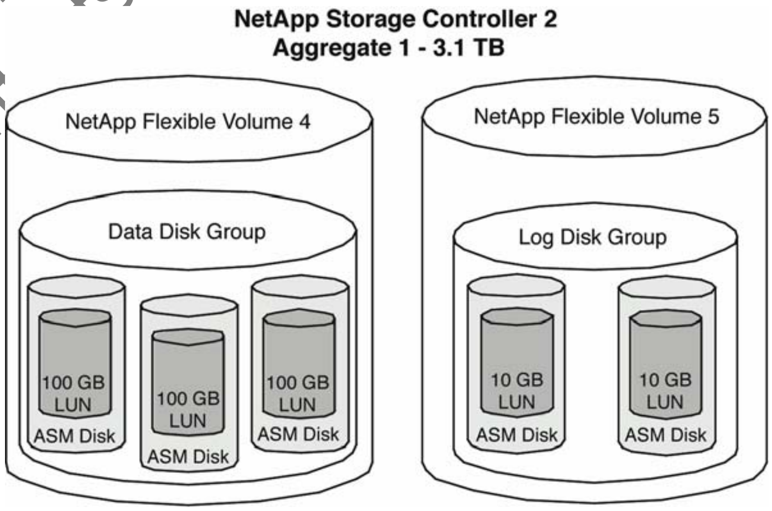

Figure 4) Storage provisioning on NetApp controller 2.

STORAGE PROVISIONING FOR NFS CONFIGURATION

As shown in Figures 5 and 6, volumes for the NFS configuration were laid out in much the same way as those used by the Fibre Channel and 10GbE/iSCSI configurations.

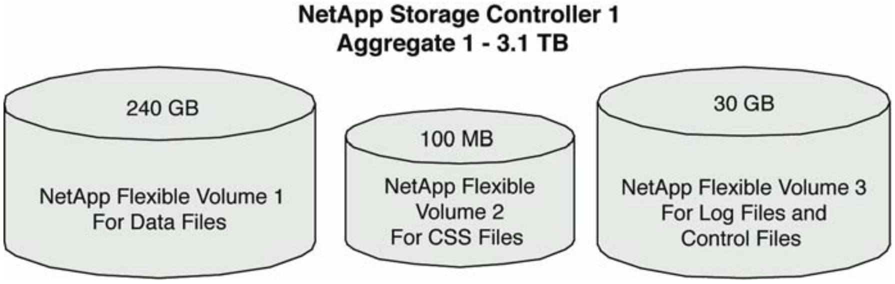

Figure 5) Storage provisioning for NFS configuration on NetApp controller 1.

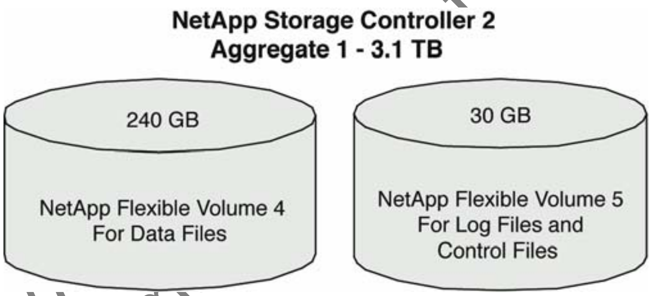

Figure 6) Storage provisioning for NFS configuration on NetApp controller 2.

NFS MOUNT OPTIONS

See NetApp kb7518 “Mount Options for Databases on NetApp NFS”:

http://now.netapp.com/Knowledgebase/solutionarea.asp?id=kb7518

ORACLE PARAMETER SETTINGS (NONDEFAULT VALUES)

Table 2) Oracle parameter settings (nondefault values only).

Oracle parameter settings (Nondefault Values)	
Parameter Name	**Value**
db_block_size	8192
db_cache_size	1008M
db_file_multiblock_read_count	128
db_files	200
db_writer_processes	2
filesystemio_options	setall
gcs_server_processes	4
hash_area_size	80000000
job_queue_processes	16
large_pool_size	160M
log_buffer	30519296
nls_date_format	YYYY-MM-DD
open_cursors	1024
parallel_execution_message_size	16384
parallel_max_servers	16
parallel_min_servers	8
parallel_threads_per_cpu	1
pga_aggregate_target	1300M
processes	900
session_max_open_files	300
sessions	900
sga_max_size	4608M
shared_pool_reserved_size	15099494
shared_pool_size	288M
sort_area_retained_size	2000000
sort_area_size	20000000
transactions	900

LINUX /ETC/SYSCTL.CONF SETTINGS

```
kernel.shmall=16000000000
kernel.shmmax=16000000000
kernel.shmmni=4096
kernel.sem=250 32000 100 128
net.ipv4.ip_local_port_range=1024 65000
net.core.wmem_default=16777216
net.core.rmem_default=16777216
net.core.wmem_max=16777216
net.core.rmem_max=16777216
fs.file-max=65536
fs.aio-max-nr=1048576
```

/ETC/SECURITY/LIMITS.CONF SETTINGS

```
oracle   soft      nproc   2047
oracle   hard      nproc   16384
oracle   soft      nofile   1024
oracle   hard      nofile   65536
```

TEST RESULTS

The table below shows the actual test results.

Table 3) Test results data.

Test Results - Comparison of 10GbE and 4Gb/s FC								
Storage Protocol	Database Throughput Avg (Gb/sec)	Database I/O Latency (ms)	Storage Throughput Avg (Gb/sec)	Peak Storage Throughput (Gb/sec)	Avg Host CPU Utilization			
					Idle	I/O Wait	System	User
10GbE NFS	14.18	5	7.74	8.95	19%	0%	48%	33%
10GbE iSCSI	12.50	13	6.98	7.21	1%	17%	60%	22%
4Gb/sec FC	5.94	29	3.22	3.36	9%	82%	3%	6%

STORAGE SYSTEM THROUGHPUT

Our test results clearly demonstrate the enhanced level of throughput to be expected from an enterprise-class storage system in an I/O transport-enhanced configuration. Figure 7 represents the average I/O per data transport cable from storage to the switch for the duration of each test run. This data was obtained using sysstat, a NetApp tool used to capture real-time data from NetApp storage systems. Because this data indicates network throughput, bits per second were converted to gigabits per second by using a decimal conversion factor (1000) instead of the binary conversion factor (1024), per standard practice. NFS on 10GbE clearly provided the highest throughput, with 4Gb/sec Fibre Channel providing the lowest throughput.

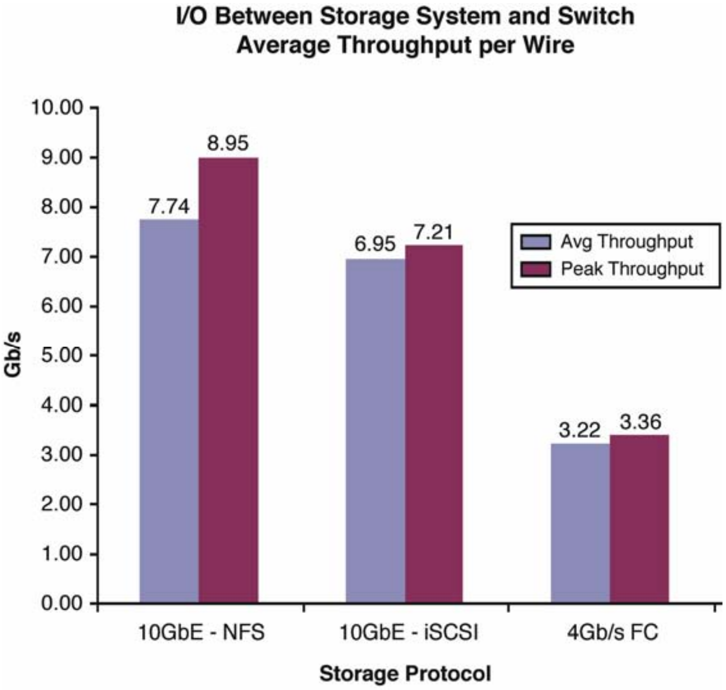

Figure 7) Summary of average storage system throughput (storage to switch).

DATABASE THROUGHPUT

Figure 8 presents the actual database throughput that corresponds to the data in Figure 7. This data was obtained from the Oracle Automatic Workload Repository (AWR) reports captured by each host during each test run. The database was basically fed by two wires from the storage system; because the data does not represent network throughput, the binary conversion factor (1024) was used to convert bits per second to gigabits per second. Again, NFS over 10GbE was the clear winner.

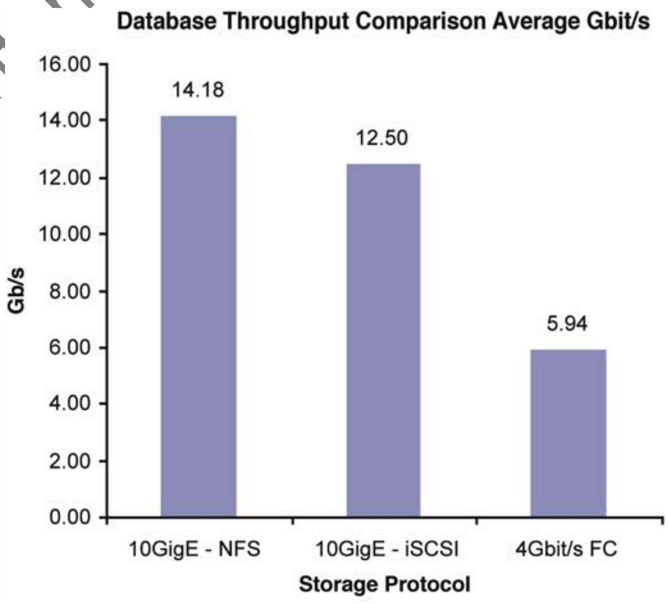

Figure 8) Database throughput comparison—three-node RAC.

CPU UTILIZATION

The database and the database workload were initially configured and tuned to maximize throughput of the 10GbE network with NFS, which served as a reference for the remaining tests. Although data transport utilization was maximized in all three sets of tests, CPU utilization on the database hosts was not optimized, resulting in excessive CPU I/O wait time for the non-NFS configurations, particularly with the much slower Fibre Channel configuration. Figure 9 summarizes the corresponding host-side CPU utilization as measured by the Linux vmstat utility. CPU calculations based on idle CPU time indicate a relatively high level of utilization for all three configurations, particularly the iSCSI configuration. For the two block-based protocols, a significant part of CPU utilization is CPU wait time, as previously explained. This indicates that processes are waiting for I/O. Although this is a measure of utilization, it does not indicate CPU saturation. Improvements in I/O bandwidth and processing capability will result in reductions in wait time statistics and increases in database throughput. Anticipated developments that will affect this include faster Fibre Channel transport (8Gb FC) and improved 10GbE tcp offload capabilities. Currently, 10GbE tcp offload is not supported by NetApp storage systems, and enabling the TCP Offload Engine (TOE) on the Linux hosts did not noticeably affect performance on the IBM blade side.

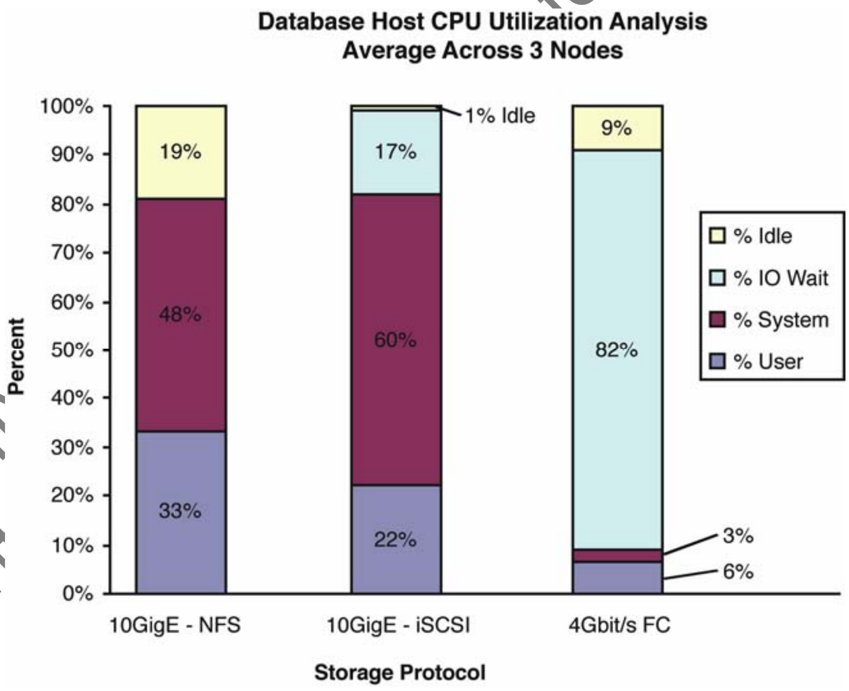

Figure 9) Analysis of host CPU utilization.

LATENCY

NFS over 10GbE also proved to be the low latency solution. In the DSS context, Oracle read latency can be defined as the average wait time for a required data block to be read from storage into the Oracle program global area (PGA). Figure 10 summarizes the average database read latency for each protocol. It basically follows the same trend as CPU I/O wait time, which was addressed previously in this document. Better tcp offloading on the iSCSI host side and faster Fibre Channel transport would improve that latency.

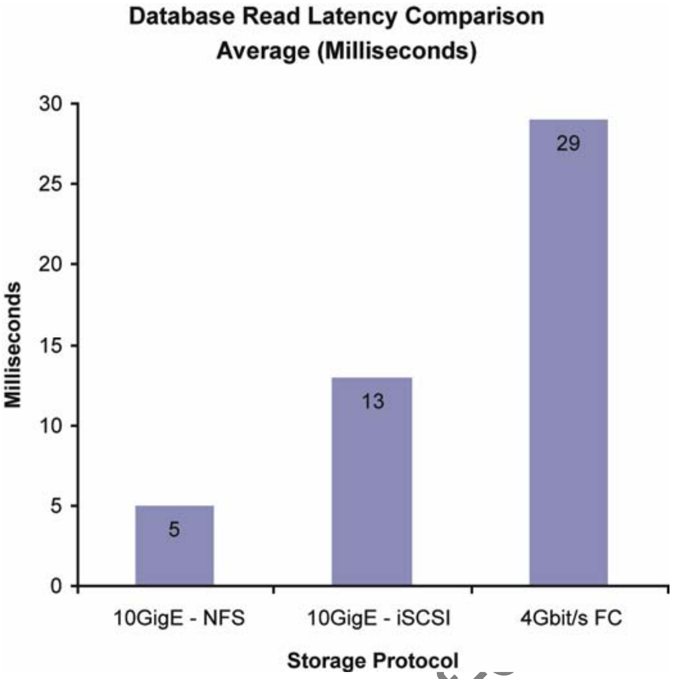

Figure 10) Average database read latency.

CONCLUSION

Network Appliance storage systems are compatible with the latest data transport technologies. The purpose of this paper is not to endorse one over the other, but there can be no doubt that NetApp storage with 10GbE is the clear winner in terms of database throughput capabilities, whether it's used with NFS or Oracle ASM. In addition, further technological developments will result in even better database performance. Enhanced 10GbE drivers are expected to result in even higher throughput for iSCSI as tcp offload capabilities are increased. Another development not mentioned in this paper is the new Oracle dNFS, a feature of Oracle Database 11g, which moves the NFS stack from the operating system to the Oracle software itself, resulting in improved performance and easier configuration. dNFS is also compatible with 10 Gigabit Ethernet. All of these new developments result in more and better options for breaking the speed barrier in the enterprise data center that existed with the older Gigabit Ethernet and 2Gb/sec Fibre Channel standards.

ACKNOWLEDGEMENTS

Special thanks to the following organizations for their contributions and support:

Blade.org

Blade Network Technologies, Inc.

Chelsio Communications, Inc.

Network Appliance, Inc.