NETAPP TECHNICAL REPORT

PERFORMANCE COMPARISON OF NETAPP FAS6070C AND FAS980C USING ORACLE10*G*™ RAC

John Elliott, Network Appliance, Inc.

February 2007 | TR-3538-0207

EXECUTIVE SUMMARY

This paper demonstrates the enhanced capabilities of the NetApp FAS6070C storage system by comparing its performance to that of the FAS980C, using the load generated by an Oracle10*g* Real Application Clusters (RAC) database.

TABLE OF CONTENTS

1 INTRODUCTION 3

2 TEST LOAD DESCRIPTION 3

3 TEST RESULTS 3

4 CONCLUSIONS 6

APPENDIX A: HARDWARE CONFIGURATION – DATABASE SERVERS (NODES) 6

APPENDIX B: HARDWARE CONFIGURATION – NETAPP STORAGE SYSTEMS 6

APPENDIX C: FCP NETWORK CONFIGURATION 7

APPENDIX D: FCP ADAPTERS AND SWITCH – HARDWARE AND SOFTWARE CONFIGURATION 8

APPENDIX E: STORAGE PROVISIONING DETAILS 8

APPENDIX F: NETAPP/LINUX/OCFS2 STORAGE DEVICE MAPPINGS 9

APPENDIX G: HOST SOFTWARE CONFIGURATION 9

APPENDIX H: ORACLE CONFIGURATION (PFILE) 9

APPENDIX I: LINUX /ETC/RC.LOCAL FILE 11

APPENDIX J: LINUX /ETC/SYSCTL.CONF FILE 11

APPENDIX K: TIPS AND LESSONS LEARNED 1

ACKNOWLEDGEMENTS 1

1 INTRODUCTION

The NetApp FAS6070 storage system was designed to handle the most demanding enterprise applications. It utilizes the latest technology, 64-bit CPU architecture, and four times as much memory as the workhorse FAS980. In order to provide a known point of comparison, we recently ran a series of performance tests comparing the FAS6070C to the FAS980C.

An Oracle10*g* RAC database running on Red Hat Enterprise Linux® AS 4 was used to generate an OLTP test load. OCFS2 file systems were used for Oracle® data files, log files, and control files. We used NetApp LUNs provisioned over FCP for OCFS2 file systems, the Oracle Cluster Registry (OCR) disk, and the Oracle Cluster Synchronization Services (CSS) voting disk. The database was roughly 350GB in size.

In terms of throughput, the FAS980C performed well against the FAS6070C for smaller RAC configurations, but the throughput gap between the two storage systems increased as the node count increased. Finally, with a 12-node RAC database, the FAS6070C provided about 40% higher throughput than the FAS980C, with much lower latency. Latency results were even more dramatic. Read latencies in terms of average Oracle db file sequential read time ranged from 5 to 24 milliseconds for the FAS980C as the RAC node count increased from 2 to 12, while the read latency for the FAS607C increased from 3 to 6. The results are discussed in detail in the following sections.

2 TEST LOAD DESCRIPTION

In order to interpret the test results, it is necessary to understand how the load was generated. This section discusses the details of the workload used for these tests.

The OLTP test load was generated by a series of scripts that simulate an online ordering system, executing a continuous set of transactions against an Oracle10*g* RAC database. We performed a single round of tests on each of several Oracle RAC configurations, beginning with two RAC nodes and incrementing the node count by 1 with successive test rounds until the node count reached 12. As the node count increased, the load on each node remained unchanged, resulting in an incremental load increase with the addition of each node. This set of tests was performed on a FAS6070 cluster, followed by identical tests on a FAS980 cluster. The FAS980C and FAS6070C environments were identical in every way (disk drives, Data ONTAP® version, controller settings, FCP adapters, etc.), with the only difference being the storage controller used (FAS6070 vs. FAS980). Database throughput was measured in terms of “OETs per minute,” with an OET being defined as a single order entry transaction. Latency was measured in milliseconds.

3 TEST RESULTS

Although the FAS980C was a strong performer, the FAS6070C came out on top in terms of throughput, load scalability, and latency. Figure 1 summarizes the throughput results.

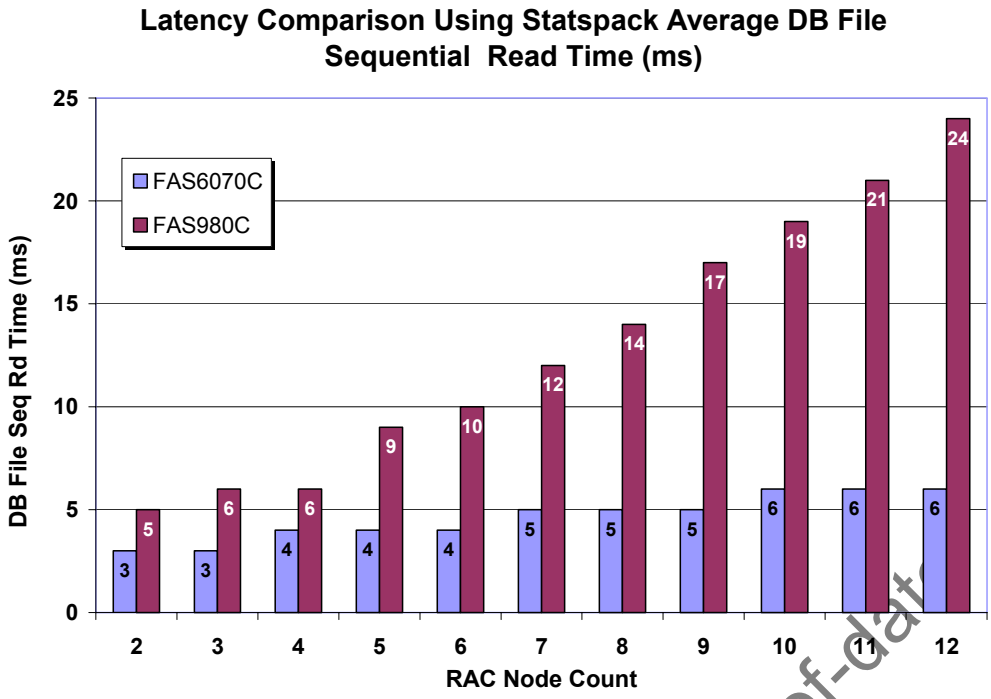

Figure 1) Throughput measurements for 2- to 12-node configurations.

You can see that throughput for the two systems is very close for node counts up to four, demonstrating that the two, for the most part, handled the load equally well. Beyond four nodes, however, the FAS6070C really begins to demonstrate its superior scalability. Even so, the most dramatic results can be seen in terms of database latency and FCP latency. Figure 2 compares database read latency measured by Oracle statspack and average db file sequential read time. ("Average db file sequential read time" can be defined as the average amount of time in milliseconds required for Oracle to read a single block from storage. It is considered to be an Oracle-centric measurement of read latency. In other words, it represents the Oracle application's view of read latency.)

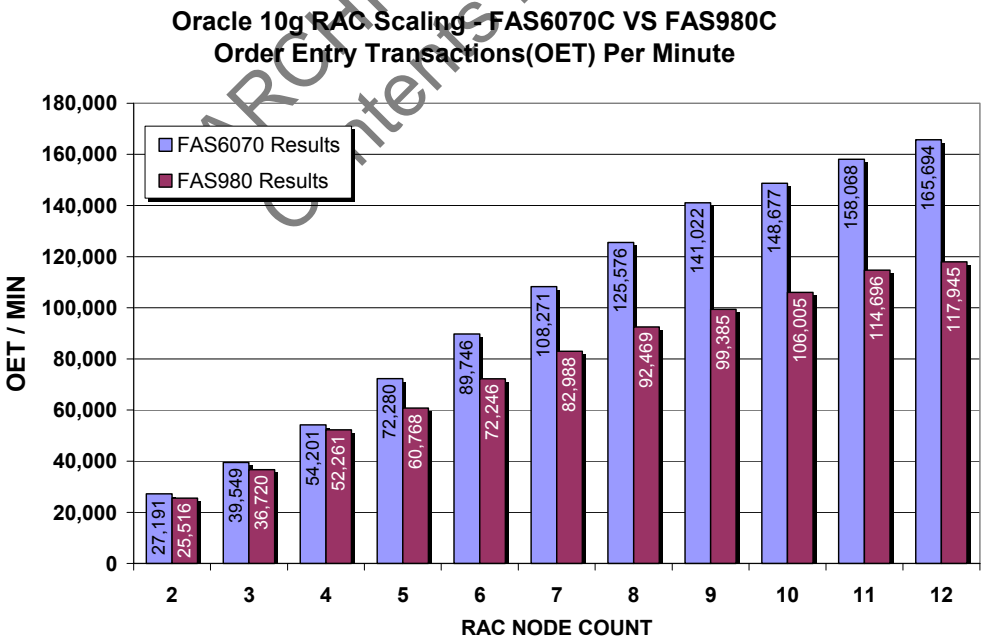

Figure 2) DB file sequential read time from Oracle statspack.

Average db file sequential read time is usually closely related to latency on the data storage system, which was certainly the case in this study. Figures 3 and 4 represent the storage system views of read and write latency. Note the close correlation between the application (Oracle) read latency (figure 1) and the storage system FCP read latency. FCP latency on the storage system is the key to good application latency.

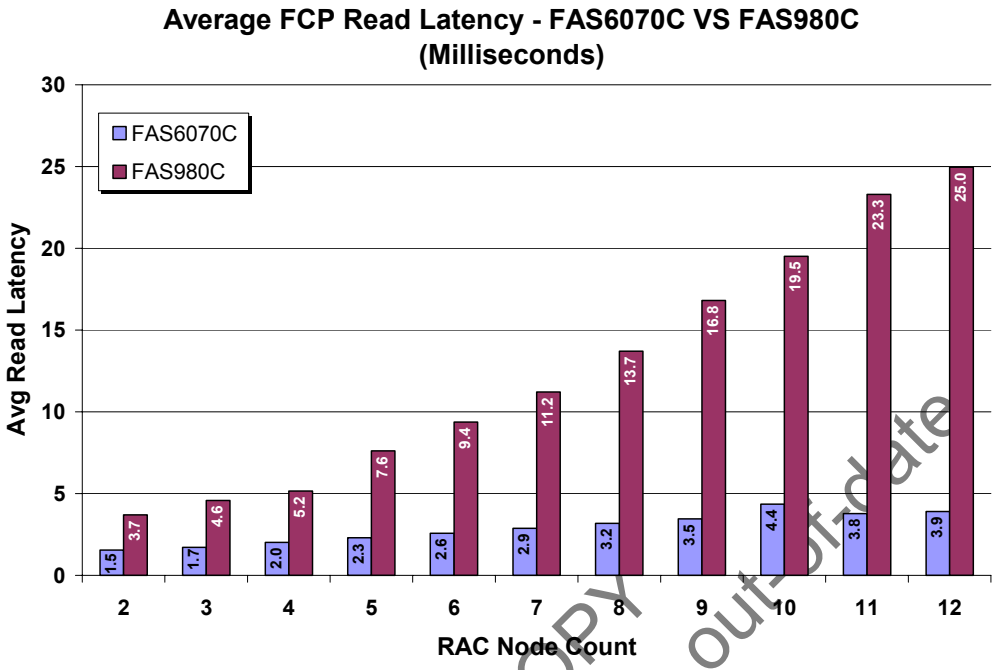

Figure 3) Average FCP read latency measured from the storage system.

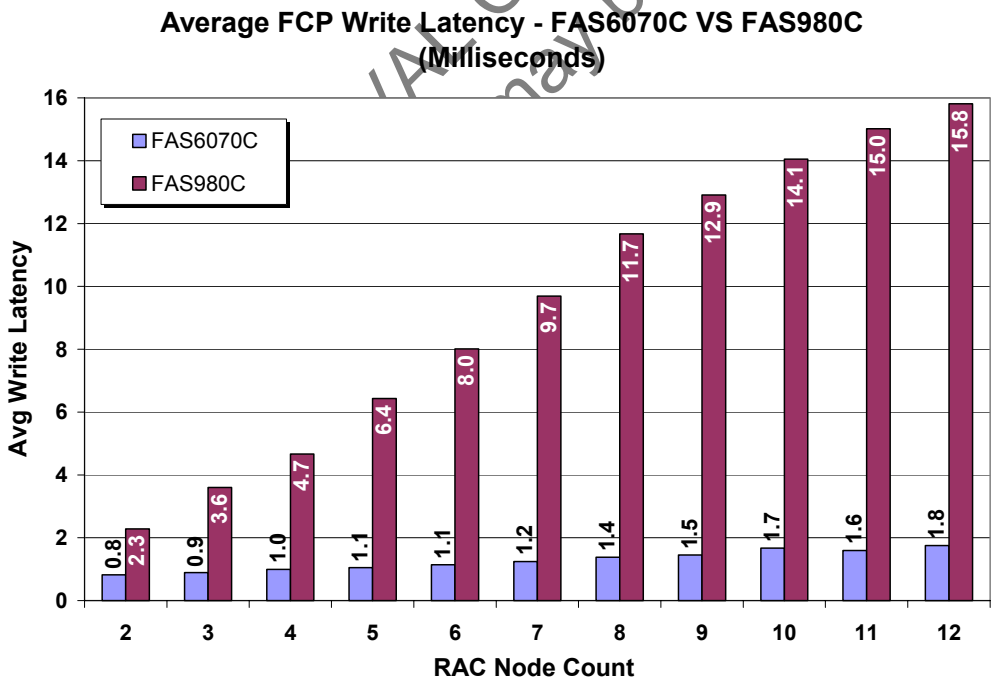

Figure 4) Average FCP write latency measured from the storage system.

As you can see from the two figures above, FCP latency for reads and writes is consistent with the Oracle db file sequential read time.

4 CONCLUSIONS

Performance tests using an Oracle10*g* RAC database on Red Hat Enterprise Linux AS 4 clearly demonstrated the superior performance of the NetApp FAS6070C storage system as compared to the FAS980C. We found that database throughput measurements for the two storage systems were comparable for a 2-node RAC, with the FAS6070C generating about 6.6% better overall throughput; however, that gap grew significantly as the node count increased. At 12 nodes the gap in throughput increased to 40%, with the FAS6070C clearly outperforming the FAS980C.

Differences in latencies between the two configurations were far more dramatic on both the database host and the storage system. With the addition of each node in the RAC cluster, the latencies measured on the FAS6070C remained relatively flat, in a range of approximately 3 to 6 milliseconds. However, latencies on the FAS980C increased an additional 1 to 3 milliseconds with the addition of each Oracle RAC node. Specifically, the Oracle read latency for a 2-node RAC was 67% higher with the FAS980C than with the FAS6070C. For a 12-node RAC, Oracle latency on the FAS980C was 300% higher!

The FAS6070C leverages advances in both hardware and software to meet the demands of the large enterprise data center, which is clearly evident in our test results. The improvements in throughput and latency with the FAS6070C can be largely attributed to the following factors, to list a few:

- Increased physical memory in the FAS6070 (four times that of the FAS980)
- Faster CPUs in the FAS6070
- Enhancements to Data ONTAP to make effective use of storage systems with four or more CPUs

APPENDIX A: HARDWARE CONFIGURATION – DATABASE SERVERS (NODES)

Twelve Fujitsu Primergy RX300 S2 servers, each equipped with the following:

- (4) Gigabytes physical memory
- (2) Intel® Xeon™ 3.60 GHz CPUs
- (2) QLogic QLA2342 FCP Host Adapters
- (2) Broadcom BCM5721 GbE NICs

APPENDIX B: HARDWARE CONFIGURATION – NETAPP STORAGE SYSTEMS

Two FAS6070 storage systems configured in a cluster, each equipped with the following:

- (4) QLogic QLA2432 FCP target adapters
- (112) 144GB FC disks (15K RPM)

Two FAS980 storage systems configured in a cluster, each equipped with the following:

- (4) QLA2432 FCP target adapters
- (112) 144GB FC disks (15K RPM)

APPENDIX C:
FCP NETWORK CONFIGURATION

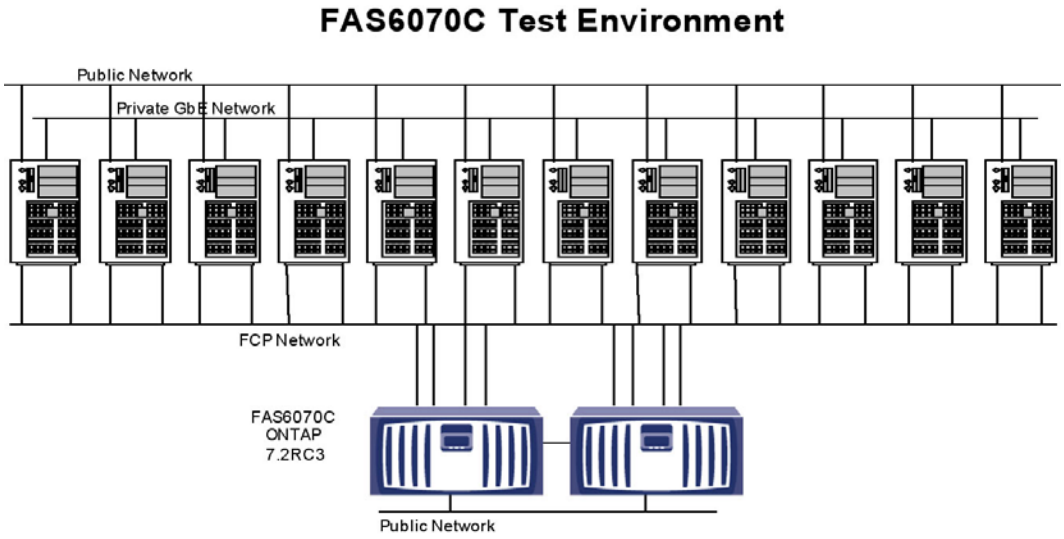

Figure 5) FAS6070C network configuration.

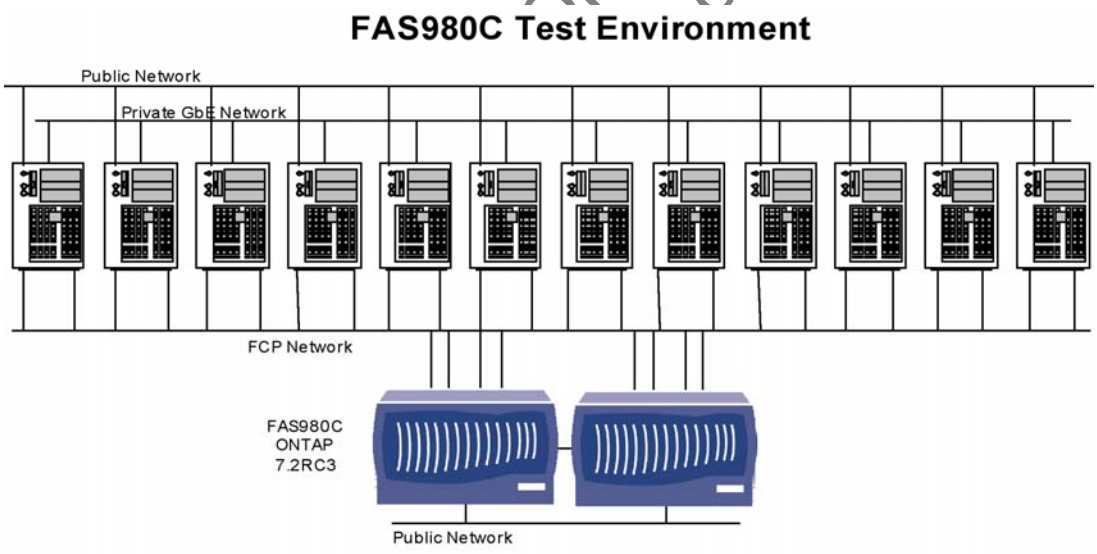

Figure 6) FAS980C network configuration.

Note that the network configurations for the FAS980C and FAS6070C are identical. Each database node is connected to the FCP network using two QLA2342 HBAs. Each system is connected to the FCP network using four FCP targets. Also note the private GbE network, which serves as the cluster interconnect.

APPENDIX D: FCP ADAPTERS AND SWITCH – HARDWARE AND SOFTWARE CONFIGURATION

QLogic QL2342 FCP HBA

- Driver Version: 8.01.06
- Bios Version: 1.21
- Firmware Version: 3.03.19
- Settings:
  - Data Rate: 2 Gbps
  - Frame Size: 2048
  - LUNs Per Target: 8
  - Execution Throttle: 256

Switch Model: QLogic SANbox2-64

- Firmware Version: V4.2.0.18-0

APPENDIX E: STORAGE PROVISIONING DETAILS

All RAID groups were 16 disks in size and configured to use double parity.

Storage Controller	Aggregate Name	Volume Name	LUN	Size	Description
Controller 1	Aggr0			3 Disks	
		/vol/vol0		95147788 KB	Root Volume
	Aggr1			96 Disks	Database Aggregate
		/vol/oradata1		560GB	
			/vol/oradata1/lun1	300GB	Oracle Data Files
		/vol/oradata2		560GB	
			/vol/oradata2/lun2	300GB	Oracle Data Files
		/vol/oralog1		170GB	
			/vol/oralog1/lun3	70GB	Oracle Log and Control Files
Controller 2	Aggr0			3 Disks	
		/vol/vol0			Root Volume
	Aggr1			96 Disks	Database Aggregate
		/vol/oradata3		560GB	
			/vol/oradata3/lun1	300GB	Oracle Data Files
		/vol/oradata4		560GB	
			/vol/oradata4/lun2	300GB	Oracle Data Files
		/vol/oralog2		170GB	
			/vol/oralog2/lun3	70GB	Oracle Log and Control Files
		/vol/css_vol		640MB	
			/vol/css_vol/lun4	140MB	OCR Disk
			/vol/css_vol/lun5	80MB	CSS Voting Disk

APPENDIX F: NETAPP/LINUX/OCFS2 STORAGE DEVICE MAPPINGS

LUN	Linux Device	Raw Device	OCFS2 Mount	Oracle Usage
/vol/oradata1/lun1	/dev/sdb		/mnt/oradata1	Data File
/vol/oradata2/lun2	/dev/sdc		/mnt/oradata2	Data File
/vol/oradata3/lun1	/dev/sde		/mnt/oradata3	Data File
/vol/oradata4/lun2	/dev/sdf		/mnt/oradata4	Data File
/vol/oralog1/lun3	/dev/sdd		/mnt/oralog1	Log & Control Files
/vol/oralog2/lun3	/dev/sdg		/mnt/oralog2	Log & Control Files
/vol/css_vol/lun4	/dev/sdh	/dev/raw/raw1		OCR Disk
/vol/css_vol/lun5	/dev/sdi	/dev/raw/raw2		CSS Voting Disk

APPENDIX G: HOST SOFTWARE CONFIGURATION

- Operating System: Red Hat Enterprise Linux Advanced Server 4 Update 3
  - Kernel 2.6.9-34.ELsmp
- Oracle: Oracle 10.2.0.2.0
- OCFS2 (Oracle Cluster File System): OCFS2 1.2.3
  - Mount Options: rw,_netdev,nointr,heartbeat=local

APPENDIX H: ORACLE CONFIGURATION (PFILE)

```
*._db_file_noncontig_mblock_read_count=1
*._log_parallelism=4
*.audit_file_dest='/home/oracle/admin/tpcc/adump'
*.background_dump_dest='/home/oracle/admin/tpcc/bdump'
*.cluster_database_instances=10
*.cluster_database=true
*.compatible='10.2.0.1.0'
*.control_files='/mnt/oralog1/control01.ctl','/mnt/oralog2/control02.ctl'
*.core_dump_dest='/home/oracle/admin/tpcc/cdump'
*.cpu_count=4
*.db_16k_cache_size=201326592
*.db_8k_cache_size=1024
*.db_block_size=4096
*.db_cache_size=1157627904
*.db_domain=''
*.db_file_multiblock_read_count=8
*.db_files=300
*.db_name='tpcc'
*.db_writer_processes=4
*.dispatchers='(PROTOCOL=TCP) (SERVICE=tpccXDB)'
*.filesystemio_options='SETALL'
```

```
tpcc12.instance_number=12
tpcc11.instance_number=11
tpcc10.instance_number=10
tpcc9.instance_number=9
tpcc8.instance_number=8
tpcc7.instance_number=7
tpcc6.instance_number=6
tpcc5.instance_number=5
tpcc4.instance_number=4
tpcc3.instance_number=3
tpcc2.instance_number=2
tpcc1.instance_number=1
*.java_pool_size=16777216
*.job_queue_processes=100
*.open_cursors=600
*.parallel_execution_message_size=4096
*.parallel_max_servers=0
*.parallel_threads_per_cpu=2
*.pga_aggregate_target=41943040
*.processes=300
*.remote_listener='LISTENERS_TPCC'
*.remote_login_passwordfile='EXCLUSIVE'
*.session_max_open_files=300
*.sessions=600
*.sga_max_size=2097152000
*.sga_target=2097152000
*.shared_pool_size=352321536
*.streams_pool_size=0
tpcc12.thread=12
tpcc11.thread=11
tpcc10.thread=10
tpcc9.thread=9
tpcc8.thread=8
tpcc7.thread=7
tpcc6.thread=6
tpcc5.thread=5
tpcc4.thread=4
tpcc3.thread=3
tpcc2.thread=2
tpcc1.thread=1
*.trace_enabled=FALSE
*.undo_management='AUTO'
tpcc1.undo_tablespace='UNDOTBS1'
tpcc2.undo_tablespace='UNDOTBS2'
tpcc3.undo_tablespace='UNDOTBS3'
tpcc4.undo_tablespace='UNDOTBS4'
tpcc5.undo_tablespace='UNDOTBS5'
tpcc6.undo_tablespace='UNDOTBS6'
tpcc7.undo_tablespace='UNDOTBS7'
```

```
tpcc8.undo_tablespace='UNDOTBS8'
tpcc9.undo_tablespace='UNDOTBS9'
tpcc10.undo_tablespace='UNDOTBS10'
tpcc11.undo_tablespace='UNDOTBS11'
tpcc12.undo_tablespace='UNDOTBS12'
*.user_dump_dest='/home/oracle/admin/tpcc/udump'
```

APPENDIX I: LINUX /ETC/RC.LOCAL FILE

```
touch /var/lock/subsys/local
/sbin/modprobe hangcheck-timer
/bin/chown root:dba /dev/raw/raw1
/bin/chown oracle:dba /dev/raw/raw2
/bin/chmod 640 /dev/raw/raw1
/bin/chmod 644 /dev/raw/raw2
```

APPENDIX J: LINUX /ETC/SYSCTL.CONF FILE

```
# Controls IP packet forwarding
net.ipv4.ip_forward = 0
# Controls source route verification
net.ipv4.conf.default.rp_filter = 1
# Do not accept source routing
net.ipv4.conf.default.accept_source_route = 0
# Controls the System Request debugging functionality of the kernel
kernel.sysrq = 0
# Controls whether core dumps will append the PID to the core filename.
# Useful for debugging multi-threaded applications.
kernel.core_uses_pid = 1
# Settings for Oracle
kernel.shmall=4000000000
kernel.shmmax=4000000000
kernel.shmmni=4096
kernel.sem=250 32000 100 128
fs.file-max=65536
net.ipv4.ip_local_port_range=1024 65000
net.core.rmem_default=262144
net.core.rmem_max=262144
net.core.wmem_default=262144
net.core.wmem_max=262144
```

APPENDIX K: TIPS AND LESSONS LEARNED

Oracle Real Application Clusters (RAC) environments can become quite complex as the node count is increased. Following are some suggestions for configuring a stable environment with optimum performance:

- Incorrect configuration of FCP zones can result in significant performance degradation and OCFS2 instability, particularly when clustered storage systems are being used. When zoning an FCP switch, make sure that the LUNs are being accessed via direct FCP connection and not via the storage system partner. One way to verify this is to run the "fcp stats" on the storage system and check the virtual target interconnect (VTIC) statistics. "read ops" and "write ops" output should both be 0. (VTIC is the clustering interface between two storage systems configured in a cluster.)
- Gigabit Ethernet should always be used for the Oracle cluster interconnect and the OCFS2 cluster interconnect for cluster stability and performance. Also, performance with jumbo frames should be tested. In many environments, the use of jumbo frames results in higher throughput.
- In many cases the default queue depth settings on HBAs are too low, resulting in poor performance. They should be checked and set high enough to handle the database load. Guidelines listed in the HBA documentation should be followed. In our test environment we used a queue depth (execution throttle) setting of 256.
- OCFS2 is relatively new, compared to OCFS. Be sure to use the latest version.
- Directio should be used with FCP for best results.
- Asynchronous I/O is good for performance and should be used if the host operating system and file system support it. Fortunately, in our case, both directio and asynch io were available.

ACKNOWLEDGEMENTS

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

Network Appliance, Inc.

Phil Larson, Keith Griffin, Ricky Stout, Lee Dorrier, Steve Daniel