Technical Report

AIX Performance with NFS, iSCSI, and FCP Using an Oracle Database on NetApp Storage

Sanjay Gulabani, NetApp
May 2009 | TR-3408

ARCHIVAL COPY
Contents may be out-of-date

ABSTRACT

This technical report covers tuning recommendations that can increase the performance of AIX 5L environments running Oracle10g Databases over NFS, iSCSI and FCP and provides a performance comparison among these protocols. The focus of this paper is technical, and the reader should be familiar with AIX system administration, Oracle10g Database administration, network connectivity, Fiber Channel administration and NetApp storage system administration.

TABLE OF CONTENTS

ARCHIVAL COPY
Contents may be out-of-date

1 INTRODUCTION AND SUMMARY

This technical report provides performance-tuning recommendations and performance comparisons for running Oracle® Databases over NFS, FCP and iSCSI on NetApp® storage systems in an AIX® 5L 5.2 and a forthcoming AIX 5.3 ML-03 environment (please see the Appendix on a new ML3 bug). An OLTP work load using an Oracle 10.1.0.2.0 Database was used for this comparison.

A primary result of this study indicates that AIX 5.3 ML-03 with NFS provides OLTP performance comparable to iSCSI and FCP. This result was achieved using a new AIX 5.3 NFS mount option "`cio`" (concurrent I/O). This option allows concurrent I/O to a single file without access synchronization in the kernel. This relieves a significant performance bottleneck. Alternatively, in an AIX 5.2 NFS environment, the OLTP workload could not drive a 2 CPU AIX box beyond 50% CPU utilization primarily due to lock contention in the kernel.

This report details the specific environment configurations, tuning recommendations and performance results of the OLTP workload using NFS, iSCSI and FCP.

2 HARDWARE AND SOFTWARE ENVIRONMENT

2.1 SERVER

For purposes of comparison an AIX 5L pSeries 650 class server was used. The two actual configurations are shown in Table 1 and Table 2.

Table 1) Server configuration.

Component	Details	
Operating System	IBM AIX 5L	
Version	5.2 ML-005	
System Type	PowerPC pSeries 650	
Database Server	Oracle 10.1.0.2.0	
Swap Space	8GB	
Total Physical RAM	4GB	
Processor	2 * 1.2 GHz PowerPC	
Storage Network	1Gb Ethernet for NFS/iSCSI	2Gb FC-AL for FCP

Table 2) Server configuration.

Component	Details	
Operating System	IBM AIX 5L	
Version	5.3 Maintenance Level 03 (beta release)	
System Type	PowerPC pSeries 650	
Database Server	Oracle 10.1.0.2.0	
Swap Space	8GB	
Total Physical RAM	4GB	
Processor	2 * 1.2 GHz PowerPC	
Storage Network	1Gb Ethernet for NFS/iSCSI	2Gb FC-AL for FCP

ARCHIVAL COPY
Contents may be out-of-date

2.2 STORAGE

The NetApp storage system configuration is described below in Table 3.

Table 3) Storage system configuration.

Component	Details	
Operating System	Data ONTAP® 6.5.1R1	
Storage Interconnect	1GbE for NFS and iSCSI	2Gb FC-AL for FCP
Disks	4 DS14s of 144GB, 10K RPM disks	
Storage System Model	FAS940c	
DS to Storage System	2 backside FCAL	
Storage Switches	Server and storage system were direct connected with crossover cables	

2.3 AIX 64-BIT KERNEL

The AIX kernel is available in 32-bit or 64-bit mode. Though 32-bit AIX kernel can support 64-bit applications, IBM recommends using 64-bit kernels if the hardware supports it. The following commands can help you switch to 64-bit kernel, provided the environment has the "`bos.mp64`" file set installed.

```
#ln –sf /usr/lib/boot/unix_64 /unix

#ln –sf /usr/lib/boot/unix_64 /usr/lib/boot/unix

#bosboot –ak /usr/lib/boot/unix_64

#shutdown –Fr
```

ARCHIVAL COPY
Contents may be out-of-date

2.4 GIGABIT ETHERNET

High performance storage connections require a dedicated Gigabit Ethernet network to connect storage system and host running a database application.

Enabling jumbo frames on each network link is also recommended. The steps to accomplish AIX host and a NetApp storage system are provided below.

Steps on the AIX host:

1. Enable jumbo frames on the NIC:

   ```
   Aixhost>chdev -l 'ent2' -a jumbo_frames='yes'
   ```

2. Update the `mtu` size on the NIC:

   ```
   Aixhost>chdev -l 'en2' -a mtu='9000'

   Aixhost>lsattr -El en2 -a mtu

   mtu 9000 Maximum IP Packet Size for This Device True
   ```

Steps on the NetApp storage system:

```
nafiler>ifconfig e10 mtusize 9000 up
```

In the above example, `chdev` first enables jumbo frames on the network card and then changes the MTU (maximum transfer unit) size from its standard 1,500 bytes to the new jumbo frame size of 9,000 bytes. Make sure to include the storage system's ifconfig statement in its `/etc/rc` file or setting will be lost at the next storage system reboot. The following example shows the ifconfig line from the `/etc/rc` file on the NetApp storage system:

```
ifconfig e5 `hostname`-e5 netmask 255.255.255.0 mtusize 9000 flowcontrol full
```

If the connection topology involves a switch, make sure the switch also has support for jumbo frames enabled.

2.5 KERNEL PARAMETERS

Most system parameters such as shared memory, semaphores and message queues are automatically tuned by AIX. However, when running Oracle workloads of the type in this report, several other system wide limits need to be increased from the defaults.

1. Change system wide limits in `/etc/security/limits`:

   ```
   fsize=-1

   data=-1

   rss=-1

   stack=-1

   nofiles=2000
   ```

2. Change maximum processes per user (default of 128 is too low for most workloads):

   `chdev –l sys0 –a maxuproc='512'`

   Ideally `maxuproc` will be set based upon the total user processes expected in the workload.

There are several ways to set kernel tuning parameters. The preferred method is via the `smit` utility. AIX also offers command-line tools such as `vmo`, `ioo`, `no` and `chdev` to change various settings.

2.6 BUFFER CACHE HANDLING

The OS provides a filesystem buffer cache that operates between the host filesystem and the storage. This buffer cache can have a large performance impact with both NFS and JFS2 filesystems. Setting the correct buffer cache value is important. Note that when the 'cio' mount option is used, the filesystem buffer cache is bypassed for I/O associated with that mountpoint.

The parameter `maxclient%` specifies the maximum percentage of RAM that can be used for caching client I/O. When using direct I/O or concurrent I/O options, it is important not to allocate too much memory for the filesystem buffer cache, since the cache does not play a role in I/O to mount points using these options. Specifically, if direct I/O or concurrent I/O is set, the recommended `maxclient%` setting is `20` (representing 20% of memory). This is the recommended setting when using AIX 5.3 ML- 03, NFS and the 'cio' mount option.

When using JFS2 or NFS with AIX 5.2 the recommended setting for `maxclient%` is 80 (representing 80% of memory).

The setting can be changed with the following command:

```
#vmo –o maxclient%=80
```

The parameter `maxclient%` specifies the point above which the page-stealing algorithm steals only file pages. The recommended value of `maxclient%` is the same as `maxclient%` for database workloads.

ARCHIVAL COPY
Contents may be out-of-date

2.7 ASYNCHRONOUS I/O

AIX offers two implementations of async I/O: legacy AIO and POSIX AIO. Legacy async I/O is an API created and implemented on AIX before the industry standard POSIX async I/O API was defined.

Oracle recommends using legacy AIO. This is enabled via `smit` by the following commands:

```
#smit aio
-> Change/Show characteristics of Async I/O
STATE to be configured at system startup = Available
Click OK
```

AIX has a default value of `maxaioservers = 10`; this value should be moved up to at least 250 with the following command:

```
#chdev -l aio0 -a maxaioservers='250'
```

In Oracle10g and AIX the recommended database settings are a single db writer process and async I/O turned on. The following init parameters are required to enable async I/O:

```
disk_asynch_io=true
filesystemio_options=setall or async
```

Additionally, `filesystemio_options` should be set to '`setall`' when using async I/O as well as direct I/O. Setting `filesystemio_options` to "`directIO`" disables async I/O and is not recommended.

3 AIX NFS CONFIGURATION

The following mount options are used for the volumes in the AIX 5.3 ML-03 NFS tests:
`cio,rw,bg,hard,intr,proto=tcp,vers=3,rsize=32768,wsize=32768,timeo=600`

Note that the following two options are new in AIX 5.3 ML-03 NFS filesystems:

- Concurrent I/O (cio)
- Direct I/O (dio)

Concurrent I/O provides the best performance for Oracle Databases since it:

- Bypasses Virtual Memory Manager (VMM) module code
- Avoids caching of file data in the kernel
- Avoids contention on the per file write lock that blocks readers, therefore relying on the applications to do file access synchronization

Concurrent I/O is enabled with the "cio" mount option. More information on "cio" is available in the IBM paper *Improving Database Performance with AIX Concurrent I/O*.

While that paper describes concurrent I/O in relation to the JFS2 filesystem, the concepts are applicable to NFS starting with AIX release 5.3 ML-03.

For AIX 5.2 the following mount options are recommended:
`rw,bg,hard,intr,proto=tcp,vers=3,rsize=32768,wsize=32768,timeo=600`

ARCHIVAL COPY
Contents may be out-of-date

4 AIX ISCSI CONFIGURATION

The NetApp AIX host attach kit can be downloaded from the NetApp NOW™ site.

The attach kit offers a "sanlun" utility for monitoring and mapping rhdisk devices on AIX to LUNs on the storage system.

The AIX iSCSI driver has a default queue_depth setting of 1. This value is low for database workloads and needs to be increased to 128 to improve database performance.

The queue_depth setting can be checked on each individual device using this command:

```
#lsattr -El hdisk5 -a queue_depth
```

To change this setting use the following command:

```
#chdev -l hdisk5 -a queue_depth=128
```

5 AIX FCP CONFIGURATION

All of the previous configuration recommendations for iSCSI also apply to FCP devices. Additionally the particular FC-AL card has a parameter setting for num_cmd_elems. This setting defaults to 200 and the maximum setting is 2048. For database workloads, this parameter should be set to some high value, for example, 2000.

6 STORAGE SYSTEM TUNING

The only tuning option needed on the storage system is to set no_atime_update = on and to set minra = on for all volumes used for the Oracle Database.

Note: Historically, NetApp had recommended disabling aggressive storage readahead for OLTP (Online Transaction Processing) database workloads by setting the Data ONTAP parameter "minra" to "on". Data ONTAP 6.5.1, however, introduced significant changes to the readahead algorithm, making it more intelligent and efficient. Hence, disabling readahead for database workloads is no longer recommended. Recent experience indicates that in fact, enabling minra may lower the overall database performance. As a result, NetApp now recommends that the minra setting be left in the default "off" state unless explicit guidance to do otherwise is given by the NetApp Global Support Organization.

For the tests used to generate this report, all Oracle data files were placed on one volume. The volume had eight disks in each RAID group with a total of 48 disks.

ARCHIVAL COPY
Contents may be out-of-date

7 ORACLE TUNING

Below is a list of Oracle initialization parameters that were used for these OLTP tests:

```
*._in_memory_undo=false
*._undo_autotune=false
*.compatible='10.1.0.0.0'
*.control_files='/oradata/control_001','/oradata/control_002'
*.cursor_space_for_time=TRUE
*.db_16k_cache_size=320M
*.db_4k_cache_size=20M
*.db_8k_cache_size=16M
*.db_block_size=2048
*.db_cache_size=2400M
*.db_files=119
*.db_name='NTAP'
*.db_writer_processes=1
*.disk_asynch_io=true
*.dml_locks=500
*.enqueue_resources=2000
*.filesystemio_options='setall'
*.lock_sga=true
*.log_buffer=2097152
*.parallel_max_servers=0
*.plsql_optimize_level=2
*.processes=300
*.recovery_parallelism=40
*.sessions=300
*.shared_pool_size=256M
*.statistics_level='basic'
*.transactions=300
*.transactions_per_rollback_segment=1
*.undo_management='auto'
*.undo_retention=1
*.undo_tablespace='undo_1'
```

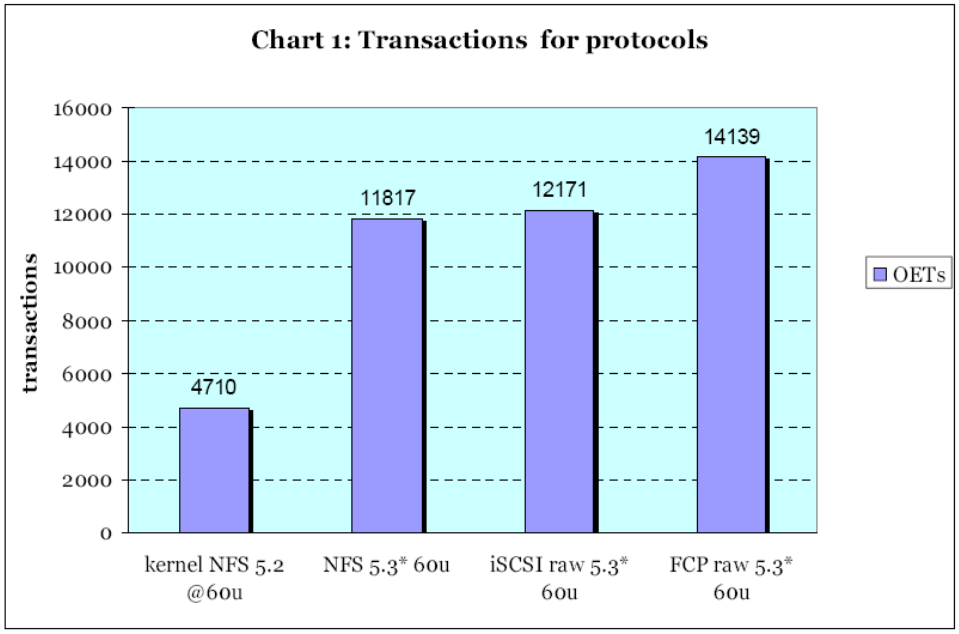
Note that since AIX 5.2 doesn't support direct I/O with NFS, `filesystemio_options = async` was set for AIX 5.2 tests.

ARCHIVAL COPY
Contents may be out-of-date

8 ORACLE PERFORMANCE COMPARISON BETWEEN NFS 5.2, NFS 5.3 ML-03, ISCSI AND FCP

The OLTP workload "Order Entry Benchmark" is a series of small (typically 2KB or 4KB in size) I/Os that are a mixture of reads and writes (approximately 2:1 reads to writes). The data set is a small number of large size files.

The benchmark was run in *server-only* mode. This means that all the users and the database engine were running on the pSeries 650.



ARCHIVE
Contents m

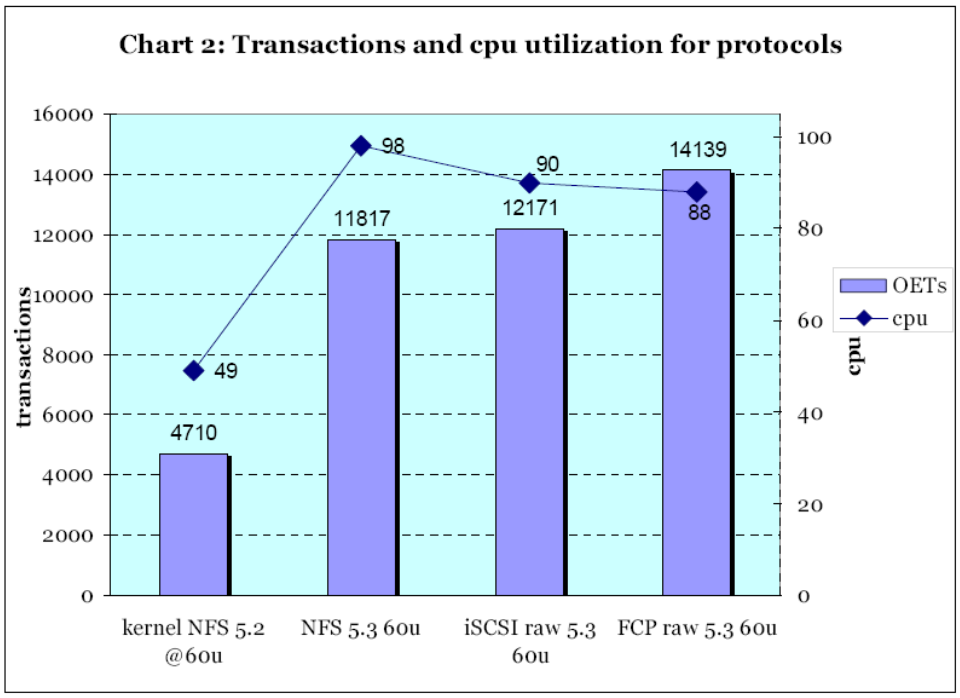
Server-only mode also implies that the users are running without any think times. This means that the users continually submit transactions without simulating any delay between transactions. Typical OLTP users have some amount of think time between keystrokes as they process transactions.

The following key metrics were defined for comparisons: OETs: Order Entry Transactions. This is a measure of OETs per specified constant set of time (in this case one minute).

Chart 1 shows OETs for three protocols (NFS, iSCSI and FCP). The chart contains NFS results for both AIX 5.2 and AIX 5.3 ML-03. Please note that AIX 5.3* in the chart above and other charts that follow is the beta release of AIX 5.3 ML-03. The results for iSCSI and FCP on 5.2 iSCSI numbers are not charted; however, the 5.3 ML-03 results are very similar to 5.2 results for FCP and iSCSI. Note that NFS performance improves significantly from AIX 5.2 to 5.3 ML-03 due to the new cio mount option and other enhancements. Results from NFS 5.3 ML-03 without the cio mount option are comparable to the AIX 5.2 results.

The NFS performance on 5.3 ML-03 and cio is comparable to iSCSI on raw devices. FCP shows a 15% to 20% performance improvement. Note that most modern database deployments do not use raw devices on block storage. Typically some type of filesystem is deployed to increase manageability. The presence of a filesystem decreases performance.

Chart 2 depicts the percentage CPU utilization in relation to the OETs completed.

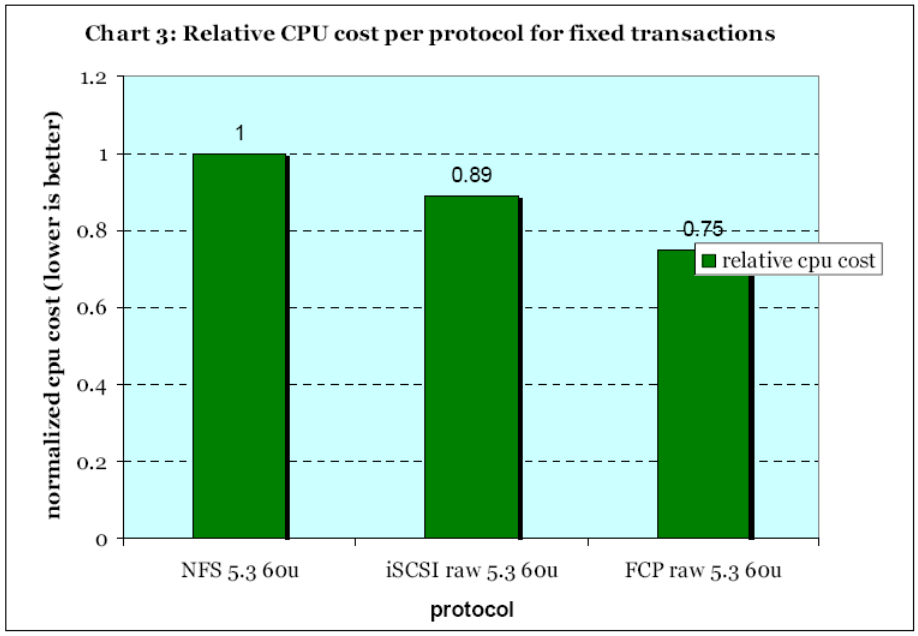


Notice that CPU utilization for NFS 5.2 is shown at 49%. This value does not increase regardless of how many additional users are added. This is a direct result of AIX 5.2 NFS lack of support of concurrent I/O. Clearly AIX 5.3 ML-03 improves this situation.

Note that both iSCSI and FCP have slightly lower CPU costs. Again these differences will likely decrease when a filesystem is applied in the iSCSI and FCP environments.

ARCHIVAL COPY
Contents may be ou

For purposes of efficiency evaluation, Chart 3 shows CPU cost per fixed number of transactions. Chart 3: Relative CPU cost per protocol for fixed transactions.

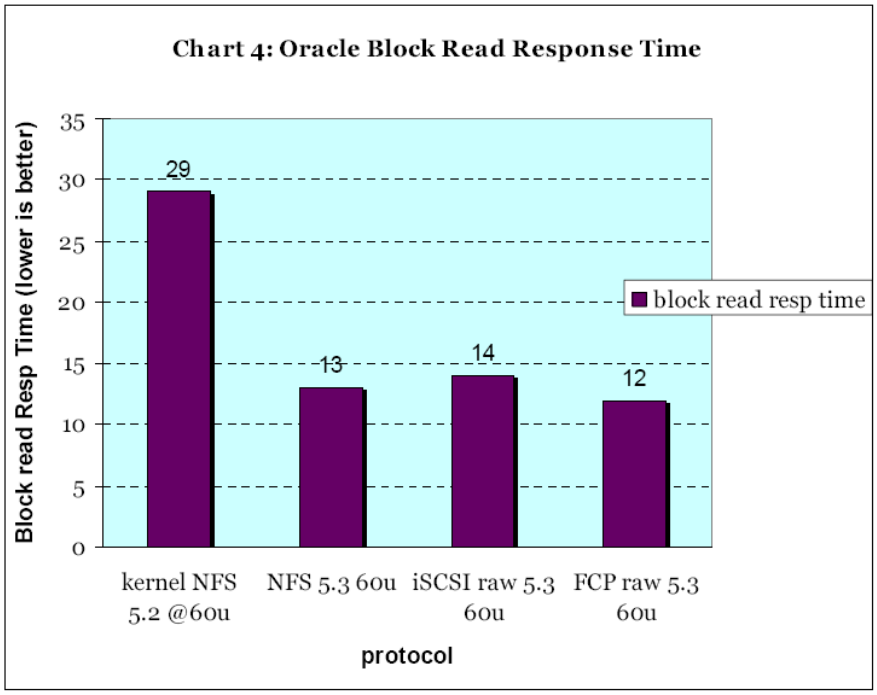


Notice that CPU costs shown are calculated using (%CPU used/ OETs) * K, where K is a constant such that NFS CPU cost per K transactions is 1.0 and relative cost for iSCSI and FCP is computed using the same constant K transactions.

The above chart shows about 11% CPU efficiency going from NFS on 5.3 ML-03 to iSCSI raw and about 19% efficiency gains by going to FCP in comparison with iSCSI. Please note that iSCSI and FCP numbers obtained are with raw device access and in real-world workloads one would want to add JFS2 or another filesystem. Also note that in the case of FCP a hardware card is used that offloads processing and there is a dollar cost associated with that, which has been left out from the comparison above.

ARCHIVAL COPY
Contents may be out

The block read response times for various protocols obtained from Oracle Statspack are shown in chart 4 below.



Note that the response times or random read latencies for all protocols in AIX 5.3 ML-03 are very comparable.

ARCHIVAL CO
Contents may be

9 CONCLUSIONS

This paper demonstrates that the AIX 5.3 ML-03 (see Appendix) with NFS and the new concurrent I/O 'cio' mount option delivers an OLTP performance environment comparable to that of block-based protocols (iSCSI and FCP). Additionally, this paper outlines clear configuration guidelines for Oracle Database workloads on all the protocols.

As with any environment, tuning a particular workload is an art. These are merely suggestions used in a lab environment and should give you good results. Individual results will vary depending upon type of the workload.

Please contact sanjay.gulabani@netapp.com with any questions or comments on this document.

10 APPENDIX

There has been a bug observed in AIX 5.3 ML-03 with how retransmits by the NFS client are handled. A fix is available in the maintenance level now referred to as AIX 5.3 TL-04. We recommend for Oracle Databases over NFS, AIX 5.3 TL-04 level is used along with other recommendations indicated in this report.

11 ACKNOWLEDGEMENTS

NetApp

Glenn Colaco, Steve Daniel, Dan Morgan, Jeff Kimmel, Darrell Suggs

IBM Corporation

Diane Flemming, Augie Mena

12 REVISION HISTORY

Date	Author	Comments
October 2005	Sanjay Gulabani	Original draft
March 2006	Sanjay Gulabani	
May 2009	Esther Smitha	Updated changes to the readahead setting recommendation.

ARCHIVAL COPY
Contents may be out-of-date

NetApp
www.netapp.com

© 2009 NetApp. All rights reserved. Specifications are subject to change without notice. NetApp, the NetApp logo, Go Further, Faster, and Data ONTAP are trademarks or registered trademarks of NetApp, Inc. in the United States and/or other countries. Oracle and Oracle 10g are registered trademarks of Oracle Corporation. AIX is a trademark of IBM Corp. All other brands or products are trademarks or registered trademarks of their respective holders and should be treated as such.