Компания **Netwell** - российский дистрибьютор высокотехнологичного оборудования. Основные направления деятельности – сетевые технологии, системы хранения данных, сетевая и информационная безопасность. **Netwell** является **официальным дистрибьютором компании NetApp.**

NETAPP TECHNICAL REPORT

Руководство по наилучшим способам использования систем NetApp с Oracle®

NetApp, Inc.
TR-3369
Редакция: Май 2008.

Eric Barrett
Technical Global Advisor
Bikash R. Choudhury
Technical Global Advisor
Bruce Clarke
Consulting Systems Engineer
Blaine McFadden
Technical Marketing
Tushar Patel
Technical Marketing
Ed Hsu
Systems Engineer
Christopher Slater
Database Consulting Systems Engineer
Michael Tatum
Database Consulting Systems Engineer

Оглавление

1. Введение 5

2. Конфигурирование NetApp 5

2.1. Сетевые настройки 5

2.1.1. Ethernet — Gigabit Ethernet, Autonegotiation, и Full Duplex 6

2.2. Настройки и опции Volume (том) и Aggregate (аггрегейт) 6

2.2.1. Базы данных 6

2.2.2. Aggregates и FlexVol Volumes или Traditional Volumes 6

2.2.3 Размеры тома 7

2.2.4 Рекомендации по типам томов для баз данных и логов Oracle 7

2.2.5. Oracle Optimal Flexible Architecture (OFA) на NetApp 7

2.2.6. Размещение Oracle Home 8

2.2.7. Наилучшие решения для Control и Log файлов 9

2.3. RAID Group Size 10

2.3.1 Традиционный RAID (RAID-4) 10

2.3.2 RAID-DP 11

2.4. Snapshot and SnapRestore® 11

2.5. Резервирование места под снэпшоты (Snap Reserve) 11

2.6. Настройки системы хранения 12

2.6.1. Опция minra 12

2.6.2. Обновление значения File Access Time 12

2.6.3. Настройки NFS 13

3. Операционные системы 13

3.1. Linux 13

3.1.1. Linux — рекомендованные версии 13

3.1.2. Linux — патчи ядра 14

3.1.3. Linux — настройки OS 14

3.1.4. Сеть в Linux — Full Duplex и Autonegotiation 15

3.1.5. Сеть в Linux — адаптеры Gigabit Ethernet 15
3.1.6. Сеть в Linux — Jumbo Frames в GbE 16
3.1.7. Протокол NFS в Linux — опции монтирования 16
3.1.8. iSCSI Initiators для Linux 16
3.1.9. FCP SAN Initiators для Linux 16
3.2. Sun™ Solaris Operating Systems 17
3.2.1. Solaris — рекомендованные версии 17
3.2.2. Solaris — патчи ядра 17
3.2.3. Solaris — настройки OS 18
3.2.4. Сеть в Solaris — Full Duplex и Autonegotiation 18
3.2.5. Сеть в Solaris — адаптеры Gigabit Ethernet 19
3.2.6. Сеть в Solaris — Jumbo Frames в GbE 20
3.2.7. Сеть в Solaris — Увеличение производительности сети 20
3.2.8. Solaris IP Multipathing (IPMP) 21
3.2.9. Протокол NFS в Solaris — опции монтирования 22
3.2.10. iSCSI Initiators для Solaris 24
3.2.11. Fibre Channel SAN для Solaris 24
3.3. Microsoft® Windows Operating Systems 25
3.3.1. OS Windows — Рекомендованные версии 25
3.3.2. OS Windows — Сервиспаки 25
3.3.3. OS Windows — Настройки реестра 25
3.3.4. Сеть в Windows — Autonegotiation и Full Duplex 26
3.3.5. Сеть в Windows — Gigabit Ethernet Network Adapters 26
3.3.6. Сеть в Windows — Jumbo Frames и GbE 26
3.3.7. iSCSI Initiators для Windows 26
3.3.8. FCP SAN Initiators для Windows 27
4. Настройки Oracle 27
4.1. DISK_ASYNCH_IO 27
4.2. DB_FILE_MULTIBLOCK_READ_COUNT 27

4.3. DB_BLOCK_SIZE 28
4.4. DBWR_IO_SLAVES и DB_WRITER_PROCESSES 28
4.5. DB_BLOCK_LRU_LATCHES 28
5. Резервное копирование, восстановление, катастрофоустойчивость 28
5.1. Как создать резервную копию данных с системы хранения NetApp 28
5.2. Создание онлайн-копий с помощью Snapshot 29
5.3. Восстановление отдельных файлов из Snapshot 30
5.4. Восстановление данных с помощью SnapRestore 30
5.5. Консолидирование резервных копий с помощью SnapMirror 30
5.6. Создание катастрофоустойчивой системы с помощью SnapMirror 31
5.7. Создание оперативных резервных копий с помощью SnapVault 31
5.8. NDMP и резервное копирование-восстановление на лентах 31
5.9. Использование Tape Devices с системами NetApp 32
5.10. Поддерживаемые инструменты резервного копирования других компаний 32
5.11. Наилучшие методы резервного копирования и восстановления 33
5.11.1. SnapVault и резервное копирование базы 33
5.12. SnapManager for Oracle – Практики резервного копирования и восстановления 36
5.12.1 SnapManager for Oracle – копирование и восстановление с использованием ASM 36
5.12.2 SnapManager for Oracle – копирование и восстановление с использованием RMAN 37
5.12.3 SnapManager for Oracle – клонирование 37
Ссылки 38
История изменений 38

1. Введение

Тысячи пользователей систем хранения NetApp успешно установили и используют СУБД Oracle на системах хранения NetApp для своих критически важных задач и приложений. NetApp и Oracle совместно работают несколько лет, для того, чтобы проверить и подтвердить корректную работу продуктов Oracle при использовании их на системах NetApp и широком спектре серверных платформ. NetApp и техподдержка Oracle создали совместную команду по отработке пользовательских задач, и проблем совместного использования наших продуктов. В процессе работы над такими проблемами удалось установить, что большинство их вызвано отклонениями от рекомендаций Best Practices при использовании Oracle на NetApp

Этот документ описывает наилучшие методы и решения задач по запуску СУБД Oracle на системах хранения NetApp, при использовании на серверных платформах OS Solaris™, HP/UX, AIX, Linux®, и Windows®. Этот документ отражает работу, проделанную NetApp, Oracle, и инженерами NetApp на различных пользовательских инсталляциях и задачах. Этот документ должен рассматриваться как стартовая точка и предлагает минимум требований, которые должны быть удовлетворены при развертывании Oracle на NetApp.

Это руководство подразумевает наличие базовых знаний, понимания технологий и действий для продуктов NetApp, и содержит рекомендации по планированию, развертыванию и использованию систем хранения NetApp, для максимально эффективного их использования

2. Конфигурирование NetApp

2.1. Сетевые настройки

При конфигурировании сетевых интерфейсов на новой системе, наилучшим решением будет запустить команду `setup`, чтобы автоматически настроить интерфейсы и обновить файлы `/etc/rc` и `/etc/hosts`. Команда `setup` потребует перезагрузки, для применения сделанных настроек.

Однако если система уже работает, и перезагрузка нежелательна, то интерфейсы могут быть сконфигурированы с помощью команды `ifconfig`. Если NIC уже включены и в онлайне, и требуют переконфигурирования, вы сначала должны перевести их в оффлайн. Для минимизации даунтайма интерфейса, группу последовательно выполняемых команд можно объединить в одну строку с помощи символа «точка с запятой» (*semicolon*, ' ; ').

Пример:

```
filer>ifconfig e0 down;ifconfig e0 'hostname'-e0 mediatype auto netmask

255.255.255.0 partner e0
```

При конфигурировании и реконфигурировании NIC или VIF в кластере, очень важно включить соответствующий `partner <interface> name` **или** `VIF name` **в конфигурацию NIC или VIF партнера в кластере, чтобы обеспечить отказоустойчивость в случае кластерного takeover. Пожалуйста, проконсультируйтесь со специалистом в поддержке NetApp, чтобы получить необходимую помощь. NIC или VIF, которые используются базой данных не должны переконфигурироваться, когда база данных работает. Это может вызвать серьезное повреждение базы.**

2.1.1. Ethernet — Gigabit Ethernet, Autonegotiation, и Full Duplex

Любая база данных, использующая систему хранения NetApp, должна использовать Gigabit Ethernet как на стороне системы хранения, так и на стороне сервера базы данных.

Адаптеры NetApp Gigabit II, III, и IV разработаны для автоматического определения конфигурации интерфейса, и имеют возможности интеллектуально самоконфигурироваться, если процесс autonegotiation не удался. По этой причине, **NetApp рекомендует, чтобы линки Gigabit Ethernet на клиентах, коммутаторах, и системах NetApp, оставались в их состоянии по умолчанию, то есть «autonegotiation»**, если линк не поднимается, производительность низка, или существуют иные проблемы соединения. Это позволит минимизировать путь поиска источника проблем.

Значение flow control должно быть установлено в «full» на системе хранения, в файле `/etc/rc`, записью вида (предположим, что интерфейс Ethernet у нас e5):

```
ifconfig e5 flowcontrol full
```

Если вывод команды `ifstat -a` не показывает flow control типа full, то тогда порт коммутатора также должен быть настроен, для поддержки этого значения. (Команда `ifconfig` на системе хранения всегда покажет установленные в настройках значения; `ifstat`, напротив, покажет, как flow control был на самом деле распознан на коммутаторе.)

2.2. Настройки и опции Volume (том) и Aggregate

2.2.1. Базы данных

В настоящий момент нет эмпирических данных того, насколько разделение базы на несколько физических томов увеличивает или уменьшает ее производительность. Поэтому, решение о том, какую структуру томов выбрать, следует принимать на основе требований резервного копирования, восстановления и зеркалирования.

Отдельный инстанс (экземпляр) базы данных не должен размещаться на нескольких некластеризованных системах хранения, так как база, с разделением на несколько систем хранения требует обслуживания с необходимостью выключения, что, как правило, трудно спланировать, и это повышает общее время недоступности базы. Если файлы одного экземпляра базы данных должна быть распределены на несколько отдельных систем хранения для повышения производительности, позаботьтесь о правильном планировании, чтобы влияние процессов обслуживания и резервного копирования было минимальным. Рекомендуется, если это возможно, проводить деление базы данных таким образом, чтобы ее части на различных системах хранения могли периодически выводиться в оффлайн.

2.2.2. Aggregates и FlexVol Volumes или Traditional Volumes

Начиная с Data ONTAP™ 7G, системы хранения NetApp поддерживают объединение большого числа

дисков в структуру «aggregate», и построение виртуальных томов (так называемых томов типа FlexVol) поверх этих дисков. Все это имеет множество преимуществ для баз данных Oracle, см подробности в [1].

Для баз данных Oracle рекомендуется помещать все ваши диски в один большой aggregate и использовать тома FlexVol для датафайлов и логфайлов, как будет описано ниже. Это обеспечивает преимущества более простого администрирования, особенно для растущих или уменьшающихся томов без влияния на производительность. Для деталей относительно точных рекомендаций по разбивке, смотрите [2].

2.2.3 Размеры тома

NetApp рекомендует пользователям выбирать размеры тома в соответствии с требованиями резервного копирования и восстановления, а также иных аспектов дизайна системы хранения.

2.2.4 Рекомендации по типам томов для баз данных и логов Oracle

В ходе нашего тестирования, мы выяснили, что предлагаемые схемы адекватны для большинства сценариев применения. Общая рекомендация – использовать один aggregate, содержащий все flexvol, на которых размещены компоненты базы данных.

В случае Flexible Volumes и Aggregates

Database binaries	Выделенный FlexVol volume	
Database config files	Выделенный FlexVol volume	Multiplex with transaction logs
Transaction log files	Выделенный FlexVol volume	Multiplex with config files
Archive logs	Выделенный FlexVol volume	Используйте SnapMirror
Data files	Выделенный FlexVol volume	
Temporary datafiles	Выделенный FlexVol volume	Выключите на нем снэпшоты
Cluster related files	Выделенный FlexVol volume	

В случае Traditional Volumes

Для traditional volumes, мы в общем случае рекомендуем вам создавать один отдельный том для каждой базы данных и логов. Если ORACLE_HOME будет размещаться на системе хранения, то сделайте для него дополнительный том.

2.2.5. Oracle Optimal Flexible Architecture (OFA) на NetApp

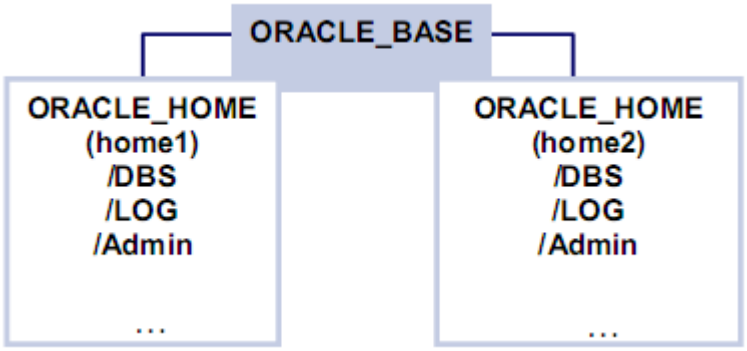
Распределите файлы по различным томам на различных физических дисках, чтобы обеспечить балансировку нагрузки ввода-вывода:

- Отделите файлы с высоким уровнем ввода-вывода от системных файлов, для лучшего времени отклика
- Упростите резервное копирование и восстановление data- и log-файлов, поместив их в отдельные логические тома.
- Убедитесь в возможности быстрого восстановления, для минимизации простоя после сбоя
- Обеспечьте логическое разделение компонентов Oracle, для упрощения обслуживания и администрирования
- Архитектура OFA хорошо работает со структурой multiple Oracle home (MOH)

Для дополнительных сведений про Oracle OFA для RAC или Non-RAC и для сведений об Oracle9i в сравнении с Oracle10g, посетите следующие ссылки:

- OFA для Non-RAC: http://download-west.oracle.com/docs/html/B14399_01/app_ofa.htm#i633126

- Для RAC, OFA для ORACLE_HOME:
  http://download-west.oracle.com/docs/cd/B19306_01/install.102/b14203/apa.htm#CHDCDGFE



Тип файлов	Описание	Точка монтирования OFA	Размещение
ORACLE_HOME	Oracle libraries and binaries	/u01/app/oracle/product/9.2.0/ /u01/app/oracle/product/10.1.0/db_unique_name	Локальная файловая система или система хранения
Database files	Oracle database files	/u02/oradata	Директория NFS на системе хранения
Log Files	Oracle redo archive logs	/u03/oradata	Директория NFS на системе хранения
CRS_HOME (For 10.1.x.x RAC)	Oracle CRS HOME	/u01/app/oracle/product/10.1.0/crs_1	Директория NFS на системе хранения
CRS_HOME (For 10.2.x.x RAC)	Oracle CRS HOME	/u04/crs/product/10.2.0/app/ (Oracle 10g™ R2)	Директория NFS на системе хранения

2.2.6. Размещение Oracle Home

Структура OFA достаточно гибка, чтобы размещать ORACLE_HOME на локальной файловой системе, или на смонтированном томе NFS. Для Oracle 10g, ORACLE_HOME может быть совместно использован конфигурацией RAC, когда один набор программ (binaries) и библиотек (libraries) совместно используются различными экземплярами (instances) той же базы

Некоторые детали совместного использования ORACLE_HOME рассматриваются ниже.

Что такое совместно используемый (Shared) ORACLE_HOME?

- Совместно используемый (shared) ORACLE_HOME это директория ORACLE_HOME, совместно используемая двумя или более хостами. Это директория установки ПО, и, обычно, содержит программы, библиотеки, сетевые файлы (listener, tnsnames, и т.д....), oraInventory, dbs, и т.д.
- Совместно используемый ORACLE_HOME, это директория ПО Oracle, которая смонтирована с сервера NFS, и имеет доступ сразу с 2 или более хостов по одному и тому же пути.
- Директория ORACLE_HOME будет выглядеть, в соответствии с OFA, примерно как (/u01/app/oracle/product/10.2.0/db_1).

Что поддерживает Oracle при использовании Oracle 10g?

- Отдельный экземпляр (инстанс) Oracle 10g поддерживает использование смонтированного в NFS ORACLE_HOME на один хост.

- Экземпляр (инстанс) Oracle RAC (Oracle 10g) поддерживает использование смонтированного в NFS ORACLE_HOME на 1 или более хостов.

Каковы преимущества совместно используемого ORACLE_HOME в Oracle 10g?

- Не нужно использовать избыточные копи для разных хостов. Это особенно эффективно в тестовой системе, где необходим быстрый доступ к бинарным файлам с похожих хост-систем.
- Экономия дискового пространства.
- Применение патчей для множества систем может быть выполнено более быстро.
- Проще добавлять ноды.

Каковы недостатки совместно используемого ORACLE_HOME в Oracle 10g?

- При наложении патчей на общий ORACLE_HOME, все базы данных, использующие этот home, должны быть перезапущены.
- В системе высокой степени доступности, использование shared ORACLE_HOME может вызвать перерыв в работе для большего количества серверов, если что-то случается.

Какие варианты совместного использования ORACLE_HOME поддерживает NetApp?

- Мы ПОДДЕРЖИВАЕМ совместно используемый ORACLE_HOME для систем RAC.
- Мы ПОДДЕРЖИВАЕМ совместно используемый ORACLE_HOME для одного экземпляра базы, когда он смонтирован на одну хост-систему.
- Мы НЕ ПОДДЕРЖИВАЕМ совместно используемый ORACLE_HOME в продакшне, который требует высокой доступности для какого-либо из использующих его экземпляров базы. Другими словами, несколько баз не должны совместно использовать один NFS-раздел с ORACLE_HOME если хотя бы одна из баз работает в продакшне.

2.2.7. Наилучшие решения для Control и Log файлов

Online Redo Log Files

Распределите ваши лог-файлы по нескольким местам. Чтобы сделать это, следуйте рекомендациям:

- Создайте как минимум две группы online redo log, каждую с тремя участниками (members). Поместите первую группу online redo log в один том, а следующую в другой том. Экземпляр процесса LGWR (Log Writer) сбрасывает REDO Log Buffer, который содержит как проведенные (committed), так и непроведенные (uncommitted) транзакции, для всех участников текущей группы online redo log, и когда группа заполнена, то переключает лог на следующую группу, при этом LGWR пишет во всех участников группы, до тех пор, пока они не заполнятся, и так далее. Чекпойнты не вызывают переключение логов, но на практике, много чекпойнтов возникают пока группа логов заполняется, также чекпойннт возникает при переключении логов.

- Предлагаемый вариант: [*]
  Redo Grp 1: $ORACLE_HOME/Redo_Grp1 (на томе /vol/oralog)
  Redo Grp 2: $ORACLE_HOME/Redo_Grp2 (на томе /vol/oralog)

Файлы архивных логов (Archived Log Files)

- Установите ваш (init) параметр, ARCHIVE_LOG_DEST, на директорию в томе логов, такую, как $ORACLE_HOME/log/ArchiveLog (на томе /vol/oralog).

Control Files

Распределите ваши control files. Для этого:

- Установите параметр CONTROL_FILE_DEST, чтобы он указывал, по меньшей мере на два разных тома:
  Dest 1: $ORACLE_HOME/Control_File1 (на локальной файловой системе или на томе системы хранения /vol/oralog)
  Dest 2: $ORACLE_HOME/log/Control_File2 (на томе системы хранения /vol/oradata)

2.3. RAID Group Size

Если время реконструкции (reconstruction rate) RAID-группы после сбоя это важный фактор, то должны использоваться RAID-группы из небольшого количества дисков. Далее даются рекомендации по наилучшему выбору размера RAID-группы при использовании как традиционного NetApp RAID тип 4 или RAID-DP™.

2.3.1 Традиционный RAID (RAID-4)

NetApp рекомендует использовать размер RAID-группы по умолчанию, размером в 8 дисков, для большинства приложений.

RAID-группа большего размера увеличивает влияние при дисковой реконструкции по следующим причинам:

- Требуется большее количество чтений
- Требуется большее количество ресурсов RAID
- Удлиняется период, во время которого производительность ввода-вывода снижена (реконструкция RAID-группы большого размера занимает большее время, по этой причине, на время реконструкции большой RAID-группы, производительность ввода-вывода снижается на более длительный срок)

Эти факторы приводят, в результате, к большему влиянию на производительность при типичной пользовательской нагрузке и/или более медленному процессу реконструкции. RAID-группа большого размера также увеличивает вероятность двойной дисковой ошибки, ведущей к потере данных. (В большой RAID-группе выше шансы того, что два диска выйдут из строя одновременно в одной и той же группе.)

---

[*] Прим. перев.: По видимому ошибка в оригинальном тексте, предлагаемый вариант противоречит тексту с описанием выше

2.3.2 RAID-DP

С выходом версии Data ONTAP 6.5, появилась возможность использовать «RAID с двойной четностью» или RAID-DP. В RAID-DP, в каждую RAID-группу добавляется диск дополнительной четности. При использовании этой дополнительной защиты, вероятность потери данных в результате двойной ошибки дисков практически устраняется, поэтому появляется возможность использовать RAID-группы большего размера.

В Data ONTAP 6.5 и позднее, RAID-группа размером более 16 дисков может быть безопасно создана при использовании RAID-DP. Однако мы рекомендуем использовать размер RAID-группы по умолчанию, в 16 дисков, при использовании RAID-DP.

2.4. Snapshot and SnapRestore®

NetApp настоятельно рекомендует использовать Snapshot и SnapRestore для резервного копирования баз Oracle. Snapshot предоставляет возможность моментальной копии всей базы данных, без ущерба для производительности системы хранения, а SnapRestore восстанавливает базу данных целиком, из любого предыдущего снэпшота.

Для того чтобы снэпшоты могли эффективно использоваться с базами данных Oracle, они должны быть скоординированы со средствами Oracle hot backup. По этой причине, NetApp рекомендует выключить автоматически создаваемые снэпшоты на томах, хранящих данные баз данных Oracle.

Для выключения автоматических снэпшотов для тома, используйте следующую команду:

```
vol options <volname> nosnap on
```

Если вы хотите сделать директорию `.snapshot` невидимой для клиентов, выполните следующую команду:

```
vol options <volname> nosnapdir on
```

При выключенном автоматическом создании снэпшотов, снэпшоты создаются как часть бэкап-процесса Oracle, когда база находится в консистентном состоянии.

Для дополнительных сведений о использовании Snapshot и SnapRestore для резервного копирования и восстановления Oracle Database, смотрите [3].

2.5. Резервирование места под снэпшоты (Snap Reserve)

Установка величины snap reserve на томе, выделяет часть пространства на томе под использование для хранения данных снэпшотов. Заметьте: снэпшоты могут использовать больше места на томе, чем задано в значении snap reserve, но пользователь не может использовать для своих данных место из зарезервированного под снэпшоты.

Чтобы посмотреть рамер резерва на томе, введите команду:

```
snap reserve
```

Для установки размера snap reserve для тома (значение по умолчанию 20%), введите команду:

```
snap reserve <volume> <percentage>
```

Не пишите знак процента (%) когда задаете величину в процентах.

Резервирование snap reserve должно быть настроено по размеру на величину, слегка превышающую максимальный объем снэпшотов для данного тома. Максимум объема снэпшотов можно установить

наблюдая за системой в течении нескольких дней, при наличии в это время высокой рабочей нагрузки.

Размеры snap reserve могут быть изменены в любое время. Не превышайте уровнем snap reserve количества свободного места на томе, в противном случае подключенный этому тому сервер может перестать работать с сообщением о нехватке дискового места.

NetApp рекомендует регулярно наблюдать за размером использованного снэпшотами пространства в snap reserve. Не позволяйте им превышать выделенного в резерве места. Если snap reserve исчерпан, то увеличьте процент места, выделенного под snap reserve или удаляйте какие-то из снэпшотов, пока количество использованного на томе места не опустится ниже 100%. Программа NetApp DataFabric® Manager (DFM) может помочь в мониторинге.

2.6. Настройки системы хранения

2.6.1. Опция minra

Когда включена опция `minra` (minimize read-ahead), то в процессе чтения минимизируется количество блоков, которые считываются в кэш с упреждением (read-ahead). По умолчанию, `minra` выключена, и система хранения производит активное упреждающее чтение в кэш для каждого тома. Эффект от упреждающего чтения зависит от характера ввода-вывода приложения. Если данные считываются последовательно, например, когда база данных проводит полный просмотр таблиц и индексов (full scan), упреждающее чтение увеличит производительность ввода-вывода. Если доступ к данным происходит с полностью случайным характером, то упреждающее чтение должно быть выключено, так как оно снижает производительность за счет лишнего чтения дисковых блоков, и непроизводительно загружает системные ресурсы.

Следующая команда используется для *включения* опции `minra` для тома, и *вылючения* упреждающего чтения:

```
vol options <volname> minra on
```

В общем случае упреждающее чтение выгодно для баз данных, и включать `minra` не следует. Однако NetApp рекомендует поэкспериментировать с опцией minra, и пронаблюдать за влиянием на производительность, так как заранее невозможно строго определить то, какое приложение использует более случайный, нежели последовательный доступ к данным. Эта опция прозрачна в плане доступа клиентов к данным, и может переключаться без прерывания процесса ввода-вывода данных. Убедитесь, что вы выждали две или три минуты, прежде чем оценивать изменения значения производительности.

2.6.2. Обновление значения File Access Time

Еще одна опция, которая может улучшить время доступа, это отключение обновления времени последнего доступа к файлу. Если приложение не зависит от этого атрибута, и не использует его в работе, эта опция может быть отключена. Используйте этот метод, только если приложение генерирует большой трафик чтения. Следующая команда выключает обновления времени последнего доступа к файлам для тома:

```
vol options <volname> no_atime_update on
```

2.6.3. Настройки NFS

Для файлов баз данных и mountpoints, NetApp поддерживает использование TCP в качестве механизма передачи данных в текущем клиенте NFS V3.0. Не поддерживается UDP для файлов баз данных и mountpoints.

3. Операционные системы

Для полного, актуального списка платформ, сертифицированных для Oracle, смотрите: http://www.netapp.com/partners/oracle/tech.html

3.1. Linux

Для дополнительной информации об использовании Linux и технологий NetApp, см. [4].

3.1.1. Linux — рекомендованные версии

Различные дистрибутивы OS Linux основаны на том или ином ядре (kernel). Для любого дистрибутива важно сосредоточить внимание на его ядре, чтобы понять возможные особенности применения и совместимости.

Kernel 2.4

Клиент NFS в этом ядре имеет множество улучшений, по сравнению с клиентом для ядра 2.2, большинство из которых относится к производительности и стабильности. Клиент NFS ядра после 2.4.16 имеет значительные улучшения производительности и стабильности.

Некоторые спорные изменения были внесены в ветку 2.4, что было препятствием для разработчиков дистрибутивов использовать поздние версии ядер этой ветки. Хотя некоторые заметные улучшения NFS были сделаны в версии 2.4.15, Торвальдс изменил часть подсистемы VM, сделав версии ядра 2.4.15, 2.4.16, и 2.4.17 нестабильными на большой нагрузке.

При использовании ядер 2.4 на оборудовании с более чем 896MB памяти, в них нужно включать при компиляции опцию CONFIG_HIGHMEM, которая необходима для доступа к памяти выше 896MB. Клиент NFS в Linux ядра 2.4 имеет известную проблему в этой конфигурации, выражающуюся в случайном подвисании приложения, или всей клиентской системы в целом. Эта проблема была устранена в ядре 2.4.20, но по прежнему может проявляться в дистрибутивах Red Hat и SUSE, использующих более ранние ядра.

Рекомендации по выбору ядра Linux

В NetApp протестирован множество вариантов дистрибутивов, и основанные на ядрах ветви 2.6 в настоящий момент рекомендуются к применению.

Рекомендованные дистрибутивы включают в себя Red Hat Enterprise Linux Advanced Server 3.0 и 4.0, а также SUSE Enterprise Linux 9.0 (SLES9).

Этот раздел будет обновляться в будущем, по мере проведения дополнительных тестирований.

Manufacturer	Version	Tested	Recommended
Red Hat	Advanced Server 2.1	Yes	No
Red Hat	Advanced Server 3.0	Yes	Yes
Red Hat	Advanced Server 4.0	Yes	Yes
SUSE	7.2	Yes	No
SUSE	SLES 8	Yes	No

SUSE	SLES 9	Yes	Yes

3.1.2. Linux — патчи ядра

В общем случае, в первую очередь должны быть применены патчи ядра, рекомендованные Oracle для конкретной версии СУБД. В общем случае эти рекомендации не конфликтуют с приведенными нами здесь, но если конфликт возник, свяжитесь с поддержкой Oracle или NetApp, для разрешения проблемы, до того, как начнете применение патчей.

Патч для использования режима uncached I/O (некэшированного ввода-вывода) впервые появился в Red Hat Advanced Server 2.1, update 3, с kernel errata e35 и выше. Необходимо в обязательно применять uncached I/O при использовании Oracle9i™ RAC с системами хранения NetApp в NAS-режиме. Uncached I/O не кэширует данные в буферах файловой системы Linux, во время проведения операций ввода-вывода для тома, смонтированного с опциями монтирования noac. Для включения uncached I/O, добавьте следующие записи в файл `/etc/modules.conf`, и перезагрузите узлы кластера:

```
options nfs nfs_uncached_io=1
```

Тома, используемые для хранения файлов баз данных, по-прежнему требуют использования опции монтирования noac для баз Oracle9i RAC.

Патч uncached I/O был разработан Red Hat и протестирован Oracle, NetApp, и Red Hat.

3.1.3. Linux — настройки OS

3.1.3.1. Увеличение буферов Transport Socket Buffers на клиенте NFS

Увеличение буферов транспортных сокетов (transport socket buffers), которые Linux использует для трафика NFS, помогает снизить использование ресурсов на клиенте, снизить колебания в производительности, и увеличить максимальные показатели пропускной способности для данных и операций. В будущих релизах клиентов, следующие процедуры не будут необходимы, так как клиент сможет выбрать самостоятельно оптимальные размеры буферов сокета.

Будучи пользователем root на клиенте, выполните следующие команды:

```
cd /proc/sys/net/core

echo 1048576 > rmem_max

echo 262143 > wmem_max

echo 1048576 > rmem_default

echo 262143 > wmem_default
```

Перемонтируйте файловую систему NFS на клиенте.

Дистрибутив Red Hat после 7.2 содержит файл с именем `/etc/sysctl.conf` куда могут быть добавлены эти изменения, так, что их не нужно будет вводить после каждой перезагрузки. Добавьте следующие строки в файл `/etc/sysctl.conf` на системе под Red Hat:

```
net.core.rmem_max = 1048576

net.core.wmem_max = 262143
```

```
net.core.rmem_default = 1048576

net.core.wmem_default = 262143
```

3.1.3.2. Прочие улучшения TCP

Следующие настройки могут помочь снизить объемы работы клиентов и системы хранения, когда вы используете NFS по TCP:

```
echo 0 > /proc/sys/net/ipv4/tcp_sack

echo 0 > /proc/sys/net/ipv4/tcp_timestamps
```

Эти действия отключают некоторые дополнительные возможности TCP, сберегая немного процессорных ресурсов и сетевой полосы пропускания.

Когда вы компилируете ядро, то убедитесь, что опция CONFIG_SYNCOOKIES отключена. SYN cookies замедляют соединение TCP, добавляя небольшое количество операций на обоих концах сокета. Некоторые дистрибутивы Linux поставляют ядро с включенными SYN cookies.

Ядра Linux 2.2 и 2.4 поддерживают большие окна TCP (large TCP windows RFC 1323) по умолчанию. Изменения по включению поддержки больших окон не требуются.

3.1.4. Сеть в Linux — Full Duplex и Autonegotiation

Большинство сетевых карт используют автоопределение, для оптимальной настройки, доступной карте и порту коммутатора, к которой она подключена. Иногда встречающиеся несовместимости приводят к постоянным процессам переопределения настроек (renegotiation), ошибочному включению half duplex или низкой скорости. Когда вы ищете причины сетевых проблем, убедитесь, что настройки Ethernet соответствуют ожидаемым, прежде чем искать глубже. Избегайте использовать принудительные установки, вместо решения проблемы автоопределения, так как они могу только маскировать более глубинную проблему. Производители карт и коммутаторов должны помочь вам в решении таких проблем.

3.1.5. Сеть в Linux — адаптеры Gigabit Ethernet

Если сервера Linux используют высокопроизводительную сеть (gigabit или быстрее), обеспечьте достаточно ресурсов CPU и полосы пропускания памяти, чтобы обработать прерывания и поток данных. ПО клиента NFS и драйвер гигабитного адаптера уменьшает количество доступных приложению ресурсов, так что позаботьтесь, чтобы запас был адекватен. Большинство карт Gigabit Ethernet поддерживают 64-bit PCI или новее, и должны показывать хорошую производительность.

Все базы данных, использующие систему хранения NetApp, должны использовать Gigabit Ethernet на обоих концах соединения, как на системе хранения, так и на сервере, для достижения оптимальной производительности.

NetApp считает, что следующие сетевые карты Gigabit Ethernet хорошо работают под Linux:

- **SysKonnect**. Карты серии SysKonnect SK-98XX работают очень хорошо с Linux и поддерживают как single- так и dual-fiber, а также «медный» интерфейс для лучшей производительности и доступности. Стабильный и отлаженный драйвер для этих карт входит в ядро 2.4 и позднее.

- **Broadcom.** Многие карты и коммутаторы используют этот чипсет, включая вездесущий 3Com. Это предоставляет высокий уровень совместимости между сетевыми коммутаторами и

клиентами Linux. Драйвер для этого чипсета появился в ядре 2.4.19 и включался в дистрибутивы Red Hat с ранними ядрами 2.4. Убедитесь, что firmware чипсета обновлено.

- **AceNIC Tigon II**. Некоторые карты, такие как Netgear GA620T, используют этот чипсет, но больше они не производятся. Стабильный и активно поддерживаемый драйвер для этого чипсета включен в дистрибутив ядра.

- **Intel® EEPro/1000**. Это, возможно, быстрейший гигабитный сетевой адаптер, но драйвер включен только в наиболее свежие дистрибутивы ядра (2.4.20 и позднее) и может быть иногда нестабильным. Сообщают, что jumbo frame MTU для карт Intel равен только 8998 байт, а не стандартные 9000 байт.

3.1.6. Сеть в Linux — Jumbo Frames в GbE

Все карты, описанные выше, поддерживают опцию jumbo frames для Gigabit Ethernet. Использование jumbo frames может повысить производительность системы, использующей Linux NFS clients и системы хранения NetApp в немаршрутизируемой сети. Удостоверьтесь, что проверили в документации на каждый используемый коммутатор его возможности по работе с jumbo frames. Существует несколько известных проблем в драйверах Linux при использовании максимального размера фрейма (9000 байт). Если вы столкнулись с неожиданными замедлениями при использовании jumbo frames, то попробуйте уменьшить размер MTU до 8960 байт.

3.1.7. Протокол NFS в Linux — опции монтирования

Для базовых знаний о NFS и краткому описанию того, что делают разные опции монтирования в клиенте NFS, прочтите документ NetApp TR: Using the Linux NFS Client with NetApp [4]

Таблица по следующей ссылке суммирует в наиболее свежем виде список рекомендованных опций монтирования для клиентов NFS, для различных версий Oracle и платформ OS.

http://now.netapp.com/Knowledgebase/solutionarea.asp?id=kb7518 (требуется логин на NOW)

3.1.8. iSCSI Initiators для Linux

Поддержка iSCSI для Linux недавно стала доступной во множестве различных форм. Начали появляться аппаратные и программные инициаторы, но они пока не достигли уровня, пригодного для безоговорочного принятия. Объем тестирования пока недостаточен, чтобы рекомендовать какие-то безусловно наилучшие решения в этой области. Этот раздел будет переработан в будущем, для включения в него рекомендаций и практических решений по запуску баз данных Oracle на Linux с iSCSI initiators.

3.1.9. FCP SAN Initiators для Linux

NetApp поддерживает протокол доступа Fibre Channel для баз данных Oracle, работающих на Linux. Подключение к системе хранения NetApp может быть сделано через коммутатор Fibre Channel (SAN) или напрямую (direct-attached). NetApp в настоящий момент поддерживает Red Hat Enterprise Linux 2.1 и 3.0 а также SUSE Enterprise Server 8, работающие с системой хранения NetApp с OS Data ONTAP 6.4.1 и выше.

Для подробностей о требованиях к системе и инсталляции, смотрите [4].

NetApp рекомендует использовать Fibre Channel SAN для Oracle Databases на Linux, там, где имеются существующие капиталовложения в инфраструктуру Fibre Channel, или когда постоянная загрузка канала передачи данных превышает 1Gbit в секунду (~110 мегабайт в секунду).

3.2. Sun™ Solaris Operating Systems

3.2.1. Solaris — рекомендованные версии

Manufacturer	Version	Tested	Recommended
(Sun) Solaris	2.6	устарела	No
	7	Yes	No
	8	Yes	No
	9	Yes	Yes
	10	Yes	Yes

NetApp рекомендует использовать Solaris 9 Update 5 и выше, для оптимальной производительности сервера.

3.2.2. Solaris — патчи ядра

Патчи для Solaris часто обновляются, так что любой список будет неполным и немедленно устареет. Указанный список патчей является минимально необходимым; более поздние ревизии могут содержать дополнительные исправления, но могут вызывать и неожиданные проблемы.

NetApp рекомендует устанавливать наиболее свежие ревизии каждого патча Sun.

Эти рекомендации дополняют, но не заменяют рекомендации о патчах Solaris, включенные в инсталляцию Oracle или release notes.

Список желательных патчей Solaris 8 на 16 марта 2006:

Solaris 8

117000-05 SunOS 5.8: kernel patch (obsoletes 108813-17)

108806-20 SunOS 5.8: Sun Quad FastEthernet qfe driver

108528-29 SunOS 5.8: kernel update patch

116959-13 SunOS 5.8: nfs and rpcmod patch

(116959-05 относится к багу Solaris NFS client caching [wcc] bug 4407669: ОЧЕНЬ важный патч для производительности)

111883-34 SunOS 5.8: Sun GigaSwift Ethernet 1.0 driver patch

Список желательных патчей Solaris 9 на 16 марта 2006:

Solaris 9

112817-27 SunOS 5.9: Sun GigaSwift Ethernet 1.0 driver patch

113318-21 SunOS 5.9: nfs patch

(addresses Solaris NFS client caching [wcc[bug 4407669: также относится к багу Solaris

4960336 fdio: ОЧЕНЬ важный патч для производительности)

113459-03 SunOS 5.9: udp patch

112233-12 SunOS 5.9: kernel patch

112854-02 SunOS 5.9: icmp patch

117171-17 SunOS 5.9: patch /kernel/sys/kaio

112764-08 SunOS 5.9: Sun Quad FastEthernet qfe driver

Список желательных патчей Solaris 10 на 16 марта 2006:

Solaris 10

120030-01 SunOS 5.10: mountd patch

118375-06 SunOS 5.10: nfs patch

118822-30 SunOS 5.10: kernel patch

Неустановка перечисленных выше патчей может вызывать отказы в работе базы данных, «падения» и замедление работы. Они должны обязательно быть установлены. Пожалуйста, отметьте, что «Sun EAGAIN bug» — SUN Alert 41862, на который ссылается патч 108727—может вызывать аварийное завершение Oracle Database с сообщением ошибки:

```
SVR4 Error 11: Resource temporarily unavailable
```

Перечисленные патчи могут иметь определенные зависимости, не названные выше. Прочтите все инструкции по инсталляции к каждому патчу, чтобы быть уверенными, что все зависимости, а также относящиеся к процессу патчи также установлены.

3.2.3. Solaris — настройки OS

Вы можете использовать специальные настройки некоторых параметров в Solaris, чтобы добиться максимума возможного в производительности вашей базы в Sun Solaris.

Дескрипторы файлов в Solaris:

rlim_fd_cur. "Soft"-лимит числа файловых дескрипторов (и сокетов), которые может открыть один процесс.

rlim_fd_max. "Hard"-лимит числа файловых дескрипторов (и сокетов), которые может открыть один процесс.

Установка этих величин в 1024 НАСТОЯТЕЛЬНО рекомендуется, для того, чтобы избежать «падений» базы данных в результате нехватки ресурсов в Solaris.

Установки "maxusers" в Solaris kernel:

Параметр ядра Solaris `maxusers` управляет распределением некоторых важных ресурсов ядра, таких, как максимальный размер таблицы процессов (process table) и максимального числа процессов на пользователя (processes per user).

3.2.4. Сеть в Solaris — Full Duplex и Autonegotiation

Следующие настройки верны для случая непосредственного включения сервера Sun в систему хранения NetApp, без использования коммутатора.

Карты GbE под Solaris должны иметь выключенную установку autonegotiation, а параметр transmit flow control - включенным. Это так для карт Sun «ge», и скорее всего верно и для более новых карт Sun «ce».

NetApp рекомендует отключить autonegotiation, принудительно задать тип flow control, и принудительно задать режим full duplex.

3.2.5. Сеть в Solaris — адаптеры Gigabit Ethernet

Sun предлагает карты Gigabit Ethernet как для PCI, так и для SBUS. Карты на PCI имеют более высокую производительность чем версии на SBUS.

NetApp рекомендует использовать карты на PCI, когда это возможно.

Все базы данных, использующие систему хранения NetApp, должны использовать Gigabit Ethernet на обоих концах соединения, как на системе хранения, так и на сервере, для достижения оптимальной производительности.

SysKonnect это независимый производитель NIC, поставляющий карты Gigabit Ethernet. Версия карт на PCI обеспечивает наивысшую производительность.

Необходимо убедиться в том, что сервера Sun с картами Gigabit Ethernet работают в режиме full flow control (некоторые требуют независимой установки режимов для «send» и «receive»).

На сервере Sun установка flow control может быть произведена добавлением следующих строк в инициализационный скрипт (такой как, например, `/etc/rc2.d/S99*`) или изменением этих значений, если они уже существуют:

```
ndd -set /dev/ge instance       0
ndd -set /dev/ge ge_adv_pauseRX 1
ndd -set /dev/ge ge_adv_pauseTX 1
ndd -set /dev/ge ge_intr_mode   1
ndd -set /dev/ge ge_put_cfg     0
```

Внимание: **instance** может отличаться от 0, если на системе более одного интерфейса Gigabit Ethernet.

Продублируйте установки для каждого instance, который подключен к NetApp.

Для серверов, использующих `/etc/system`, добавьте следующие строки:

```
set ge:ge_adv_pauseRX=1
set ge:ge_adv_pauseTX=1
set ge:ge_intr_mode=1
set ge_ge_put_cfg=0
```

Отметьте, что помещение этих настроек в `/etc/system` влияет на все Gigabit-интерфейсы сервера. Коммутаторы, и другие подключаемые устройства, также должны быть соответствующим образом сконфигурированы.

3.2.6. Сеть в Solaris — Jumbo Frames в GbE

SysKonnect поставляет карты типа SK-98xx, поддерживающие jumbo frames. Для включения jumbo frames, выполните следующие шаги:

1. Отредактируйте `/kernel/drv/skge.conf` и раскомментируйте строку:
   `JumboFrames_Inst0="On";`
2. Отредактируйте `/etc/rcS.d/S50skge` и добавьте строку:
   `ifconfig skge0 mtu 9000`
3. Перезагрузите систему.

Если вы используете jumbo frames с картами SysKonnect NIC, используйте коммутаторы, которые поддерживают jumbo frames и включите поддержку jumbo frames на сетевом интерфейсе системы хранения NetApp.

3.2.7. Сеть в Solaris — Увеличение производительности сети

Настройка приведенных ниже параметров может дать выигрыш в производительности сети. Большинство из этих настроек отображается с помощью команды Solaris ndd, и устанавливается при использовании ndd или редактировании файла `/etc/system`.

`/dev/udp udp_recv_hiwat`. Определяет максимальную величину приемного буфера UDP. Это количество места в буферах, выделенное под принимаемые данные UDP. Значение по умолчанию 8192 (8kB). Должно быть установлено в 65535 (64kB).

`/dev/udp udp_xmit_hiwat`. Определяет максимальную величину буфера передачи UDP. Это количество места в буферах, выделенное под передаваемые данные UDP. Значение по умолчанию 8192 (8kB). Должно быть установлено в 65535 (64kB).

`/dev/tcp tcp_recv_hiwat`. Определяет максимальную величину приемного буфера TCP. Это количество места в буферах, выделенное под принимаемые данные TCP. Значение по умолчанию 8192 (8kB). Должно быть установлено в 65535 (64kB).

`/dev/tcp tcp_xmit_hiwat`. Определяет максимальную величину буфера передачи TCP. Это количество места в буферах, выделенное под передаваемые данные TCP. Значение по умолчанию 8192 (8kB). Должно быть установлено в 65535 (64kB).

`/dev/ge adv_pauseTX 1`. Задает принудительный контроль потока (flow control) на передачу для адаптера Gigabit Ethernet. Transmit flow control provides a means for the transmitter to govern the amount of data sent; Значение «0» это величина по умолчанию для Solaris, до тех пор, пока он не будет включен, в результате процесса автоопределения (autonegotiation) между NIC-ами. NetApp настоятельно рекомендует чтобы контроль потока на передачу был включен. Установка этой величины в «1» поможет избежать потерь пакетов или повторной и передачи, так как эта величина заставляет NIC включить контроль потока передачи. Если NIC переполняется данными, она сигналит передающему установить паузу. Иногда бывает нужно установить это параметр в «0», чтобы определить ситуацию, когда посылающий (NetApp system) переполняет потоком клиента. Рекомендованные настройки описаны в разделе 2.2.6 этого документа.

`/dev/ge adv_pauseRX 1`. Принудительно устанавливает контроль потока (flow control) приема для адаптера Gigabit Ethernet. Receive flow control provides a means for the receiver to govern the amount of data received. Установка «1» это значение по умолчанию для Solaris.

/dev/ge adv_1000fdx_cap 1. Задает принудительный full duplex для адаптера Gigabit Ethernet. Full duplex позволяет данным передаваться и приниматься одновременно. Это нужно установить одинаков как на стороне сервера Solaris, так и на системе хранения NetApp. Ошибки в определении режима duplex приводят к сетевым ошибкам и ошибкам базы данных.

sq_max_size. Устанавливает максимальное количество сообщений (messages), допустимых для каждой очереди IP (IP queue) (STREAMS synchronized queue). Увеличение этого параметра улучшает сетевую производительность. Безопасная величина для этого параметра 25 для каждых 64MB физической памяти на системе Solaris, вплоть до максимального значения в 100. Параметр следует оптимизировать начав с 25 и увеличивая на 10, пока сетевая производительность не достигнет пика.

Nstrpush. Определяет максимальное количество модулей, которые могут быть отправлены в поток, и должен быть установлен в 9.

Ncsize. Определяет размер DNLC (directory name lookup cache). DNLC хранит информацию о файлах на томе NFS. Непопадание в кэш может вызвать дисковую операцию ввода-вывода чтения директории.

Чтение этой информации из кэша может значительно улучшить производительность NFS; getattr, setattr, и lookup обычно составляют более 50% всех вызовов NFS. Если запрошенная информация не нашлась в кэше, то запускается дисковая операция, что в результате негативно сказывается на общей производительности. Единственное ограничение размера кэша DNLC это доступная ядру память. Каждый элемент DNLC занимает примерно 50 bytes дополнительной памяти ядра.

NetApp усановить величину `ncsize` на 8000.

nfs:nfs3_max_threads. Максимальное число потоков (threads), которые может использовать клиент NFS V3. Рекомендумая величина 24.

nfs:nfs3_nra. Размер упреждающего чтения (read-ahead count) для клиента NFS V3. Рекомендуемая величина 10.

nfs:nfs_max_threads. Максимальное число потоков (threads), которые может использовать клиент NFS V2. Рекомендумая величина 24.

nfs:nfs_nra. Размер упреждающего чтения (read-ahead count) для клиента NFS V2. Рекомендуемая величина 10.

3.2.8. Solaris IP Multipathing (IPMP)

Solaris имеет средства, обеспечивающие использование нескольких соединений IP в конфигурации, похожей на NetApp virtual interface (VIF). В некоторых случаях использование этой возможности может быть выгодным.

IPMP может быть сконфигурировано или в отказоустойчивой (failover configuration) или в совместно работающей (load-sharing) конфигурации.

Отказоустойчивая (failover) конфигурация достаточно очевидно устроена и несложно устанавливается. Два интерфейса используют один IP-адрес, один из интерфейсов находится в «standby» (в документации на Solaris это называется «deprecated»), а другой интерфейс - активный. Если соединение обрывается, то Solaris прозрачно перенаправляет трафик на второй интерфейс. Когда это проделано ядром Solaris, то приложение просто использует интерфейс, и не заботится о том, как производится переключение.

NetApp протестировал конфигурацию failover для Solaris IPMP и рекомендует ее использование в тех случаях, когда нужна отказоустойчивость, есть достаточное количество интерфейсов, а стандартный транкинг (например Cisco Etherchannel) не доступен.

Конфигурация load-sharing использует трюк, при котором исходящий трафик на отдельные IP-адреса разделяется по интерфейсам, но весь исходящий трафик содержит обратный адрес одного, «primary» интерфейса. При больших объемах записи на систему хранения эта конфигурация может дать определенную выгоду в производительности. Но, так как весь обратный трафик проходит только через один интерфейс, то, в случае большого объема чтения, преимуществ в ускорении нет.

Кроме этого, механизм, который Solaris использует для обнаружения ошибок и переключения failover на работающую NIC несовместим с кластером NetApp.

NetApp не рекомендует использовать IPMP в режиме load-sharing, по причине текущей несовместимости с кластерной конфигурацией NetApp, ограниченной возможностью улучшения производительности на чтении, сложностью и вызванными этим дополнительными рисками.

3.2.9. Протокол NFS в Solaris — опции монтирования

Установка правильных опций монтирования NFS может оказывать значительное влияние на производительность и надежность подсистемы ввода-вывода. Ниже приводятся некоторые рекомендации, которые помогут выбрать правильные опции.

Опции монтирования могут быть установлены вручную, когда файловая система монтируется в Solaris, или, обычнее, определяются в /etc/vfstab для тех монтирований, которые осуществляются в момент загрузки. Последний способ настоятельно рекомендуется, так как вы можете быть уверены, что система запустится после перезагрузки по любой причине, и войдет в рабочее состояние без необходимости ручного вмешательства оператора. Чтобы настроить опции монтирования:

1. Отредактируйте `/etc/vfstab`.
2. Для каждого NFS mount, участвующего в высокоскоростной инфраструктуре, убедитесь, что опции монтирования задают TCP V3 с размером передачи 32kB:
   `…hard,bg,intr,vers=3,proto=tcp, rsize=32768, wsize=32768,…`

Внимание: Эти величины являются значением по умолчанию в NFS для Solaris 8 и 9. Определение их не является безусловно необходимым, но рекомендуется для определенности.

Hard. Опция «soft» не должна никогда использоваться для баз данных. Это может вызвать неполную запись данных, и проблемы с файлами базы данных. Опция «hard» определяет, что запросы ввода-вывода будут посланы повторно, в случае, если они были неудачны при первой попытке. Это принуждает приложение производить операцию ввода-вывода через NFS, пока затребованный файл не окажется доступным. Это особенно важно в случае использования отказоустойчивых и избыточных сетей и серверов (например, в случае кластера NetApp).

Bg. Определяет, что операция монтирования должна выполняться в фоновом режиме, если система хранения NetApp недоступна, что позволяет загрузке Solaris продолжаться в этом случае. **Так как процесс загрузки системы может быть выполнен даже при недоступности всех необходимых файловых систем, позаботьтесь о том, чтобы нужные файловые системы были смонтированы и присутствовали до начала запуска Oracle Database.**

Intr. Эта опция позволяет операциям ожидать прерывания NFS. Если эта опция не используется, и соединение NFS смонтированное с опцией «hard» оборвано и не восстановлено, то единственный способ восстановить работу для Solaris в таком случае это перезагрузка сервера.

rsize/wsize. Определяет размер запроса NFS для чтения/записи. Величины этих параметров должны соответствовать значению `nfs.tcp.xfersize` на системе хранения NetApp. Величина 32768 (32kB) рекомендуется для максимальной производительности базы данных при использовании NetApp и Solaris. По меньшей мере размер NFS read/write size должен быть равен или больше, чем размер блока (block size) Oracle. Например, определив DB_FILE_MULTIBLOCK_READ_COUNT в 4 умножаем на размер database block size равный 8kB, получаем размер read buffer size (rsize) равный 32kB.

NetApp рекомендует установить DB_FILE_MULTIBLOCK_READ_COUNT в значение от 1 до 4 для баз типа OLTP, и от 16 до 32 для DSS.

Vers. Устанавливает используемую версию NFS. Version 3 обеспечивает оптимальную производительность баз данных под Solaris.

Proto. Говорит Solaris использовать TCP или UDP для соединения. В настоящий момент только TCP поддерживается для файлов Oracle по NFS. Ранее, UDP давал лучшую производительность, но был ограничен в применении только очень надежными соединениями. TCP имеет больший overhead, но обрабатывает ошибки и лучше управляет потоком. В действующих версиях Solaris (8, 9 и 10) разница в производительности незначительна.

Forcedirectio. Новая опция, появившаяся в Solaris 8. Она позволяет приложению обходить кэш ядра Solaris, что оптимально для Oracle. Эта опция должна использоваться для томов, содерержащих файлы данных. Она не должна использоваться для томов, содержащих исполняемые файлы. Использование этой опции с томом, содержащим исполняемые файлы Oracle, будет препятствовать запуску всех исполняемых файлов, хранящихся на томе. Если программа, которая обычно нормально запускается, не хочет стартовать, или немедленно падает в «core dump», проверьте, не находится ли она на томе, смонтированном с опцией «forcedirectio».

Рекомендованные NetApp опции монтирования для Oracle single-instance database на Solaris:

```
rw,bg,vers=3,proto=tcp,hard,intr,rsize=32768,wsize=32768,forcedirectio
```

Рекомендованные NetApp опции монтирования для Oracle9i RAC на Solaris:

```
rw,bg,vers=3,proto=tcp,hard,intr,rsize=32768,wsize=32768,forcedirectio,noac
```

Появление forced direct I/O в Solaris 8 было огромным шагом вперед. Direct I/O обходит кэш файловой системы Solaris. Когда блок данных считывается с диска, он читается непосредственно в буфера кэша Oracle, а не в кэш файловой системы. Без direct I/O, блок данных сперва заносится в кэш чтения файловой системы, откуда затем переносится в кэш-буфера Oracle, двойное буферирование непродуктивно тратит память и ресурсы CPU. Oracle не использует кэш файловой системы OS.

Используя средства мониторинга и статистики, NetApp обнаружил, что без включенного direct I/O на смонтированном томе NFS, большое количество страниц файловой системы попадают в своп. Это добавляет системе оверхеда на переключение контекстов (context switches), и использование CPU увеличивается. Со включенным direct I/O, эти параметры заметно снижаются. В зависимости от нагрузки, заметно значительное увеличение обшей производительности. В некоторых случаях она превышала 20%.

Direct I/O for NFS это новинка Solaris 8, однако он был представлен в UFS еще в Solaris 6. Direct I/O должен использоваться для томов, которые хранят файлы Oracle Database, но не для прочих файлов или программ Oracle, или когда выполняются обычные файловые операции ввода-вывода, такие как «dd». Обычные файловые операции используют преимущества от кэширования на уровне файловой системы.

Отдельный том может быть смонтирован больше одного раза, так что есть возможность использовать преимущества «forcedirectio», в то время, когда другие будут использовать данные тома без нее. Однако, это чревато ошибками, так что будьте внимательны.

NetApp рекомендует использовать «forcedirectio» на отдельных томах, на которых характер паттерна ввода-вывода не требует кэширования на клиенте NFS. In general these will be data files with access patterns that are mostly random as well as any online redo log files and archive log files. Опция «forcedirectio» не должна использоваться для томов, содержащих исполняемые файлы, такие, как директория ORACLE_HOME. Использование опции «forcedirectio» на томах, содержащих исполняемые файлы, вызовет их неверную работу.

Множественные точки монтирования (Multiple Mountpoints)

Чтобы достичь максимальной производительности транзакционная база типа OLTP может использовать возможности множественного монтирования базы данных и распределения нагрузки по этим точкам монтирования. Улучшение производительности в среднем оказывается в районе от 2% до 9%.

Чтобы это настроить, создайте дополнительную точку монтирования на ту же файловую систему на системе хранения NetApp. После чего переименуйте датафайлы базы данных (при помощи команды ALTER DATABASE RENAME FILE) или создайте симлинк со старой точки монтирования на новую.

3.2.10. iSCSI Initiators для Solaris

В настоящий момент, NetApp не поддерживает iSCSI initiators на Solaris[†]. Этот раздел будет обновлен в будущем, когда станут доступны iSCSI initiators на Solaris.

3.2.11. Fibre Channel SAN для Solaris

NetApp поставляет первую в отрасли систему унифицированного доступа, позволяющую обслуживать данные как NAS или SAN. NetApp предлагает решения Fibre Channel SAN для всех платформ, включая Solaris, Windows, Linux, HP/UX, и AIX. Решение NetApp Fibre Channel SAN предлагает тот же фреймворк управления, и богатую функциональность, что отличает наши NAS системы.

Пользователь может выбрать NAS или FC SAN для использования в Solaris, в зависимости от нагрузки и имеющегося оборудования. Для конфигурации FC SAN, настоятельно рекомендуется использовать *SAN host attach kit 1.2 for Solaris*. Комплект поставляется с Fibre Channel HBA, драйверами, firmware, утилитами и документацией. Для инсталляции и конфигурации консультируйтесь с документацией, поставляемой с комплектом.

[†] Устарело на момент публикации перевода

NetApp проверил и одобрил решение FC SAN с Solaris под Oracle. Консультируйтесь с руководством Oracle integration guide и NetApp FC SAN in a Solaris environment ([6]) для подробностей. Для выполнения резервного копирования и восстановления Oracle Database в SAN, см [7].

NetApp рекомендует использовать Fibre Channel SAN для Oracle Databases на Solaris там, где уже присутствует развитая инфраструктура Fibre Channel. NetApp также рекомендует обратить внимание на решение с использованием Fibre Channel SAN для Solaris там, где величины постоянного трафика ввода-вывода для сервера Oracle превышают 1Gb в секунду (~110MB в секунду).

3.3. Microsoft® Windows Operating Systems

3.3.1. OS Windows — Рекомендованные версии

Microsoft Windows NT® 4.0, Windows 2000 Server и Advanced Server, Windows 2003 Server

3.3.2. OS Windows — Сервиспаки

Microsoft Windows NT 4.0: Service Pack 5

Microsoft Windows 2000: SP2 или SP3

Microsoft Windows 2000 AS: SP2 или SP3

Microsoft Windows 2003: Standard или Enterprise

3.3.3. OS Windows — Настройки реестра

Следующие настройки в реестре улучшат производительность и надежность Windows. Эти установки потребуют перезагрузки сервера:

Опция /3GB не должна быть установлена в C:\boot.ini.

```
\\HKEY_LOCAL_MACHINE\SYSTEM\CurrentControlSet\Services\LanmanServer\Parameters\MaxMpxCt

Datatype: DWORD

Value: установить соответствующую значению  cifs.max_mpx


\\HKEY_LOCAL_MACHINE\SYSTEM\CurrentControlSet\Services\Tcpip\Parameters\TcpWindow

Datatype: DWORD

Value: 64240 (0xFAF0)
```

Таблица описывает некоторые из этих ключей и некоторые принципы их тонкой настройки:

Ключ	Описание
MaxMpxCt	Максимальное количество одновременных запросов, которые создает клиент Windows к системе хранения NetApp. Должен соответствовать cifs.max_mpx. Проследите по performance monitor величину параметра redirector/current. Если она постоянно находится в районе заданной величины

	MaxMpxCt, то увеличьте его значение.
`TcpWindow`	Максимальный размер окна передачи данных по TCP-сети. Значение должно быть установлено в 64240 (0xFAF0).

3.3.4. Сеть в Windows — Autonegotiation и Full Duplex

Перейдите в Control Panel -> Network -> Services tab -> Server и щелкните кнопку Properties. Установите опцию «maximize network applications» для максимальной сетевой производительности.

3.3.5. Сеть в Windows — Gigabit Ethernet Network Adapters

Все базы данных, использующие системы хранения NetApp должны использовать Gigabit Ethernet на обоих концах, как на стороне системы хранения, так и на стороне сервера баз данных, для достижения максимально возможной производительности.

NetApp тестировал Intel PRO/1000 F Server Adapter. Следующие настройки могут быть установлены для этого адаптера. Каждая настройка должна быть протестирована и настроена для достижения максимальной производительности.

Параметр	Описание
Coalesce buffers = 32	Число буферов, используемых для ускорения передачи.
Flow control = receive pause frame	Используемый метод контроля потока (flow control). Должен соответствовать установленному для адаптера Gigabit Ethernet в системе хранения NetApp.
Jumbo frames = disable	Разрешает передачу больших пакетов Ethernet. NetApp поддерживает их с версии Data ONTAP 6.1 и позднее.
Receive descriptors = 32	Число буферов передачи и дескрипторов, которые создает драйвер для приема пакетов.
Transmit descriptors = 32	Число буферов передачи и дескрипторов, которые создает драйвер для посылки пакетов.

3.3.6. Сеть в Windows — Jumbo Frames и GbE

Внимание: Будьте очень осторожны, когда используете jumbo frames с Microsoft Windows 2000. Если включены jumbo frames на системе хранения, и Oracle работает на сервере Windows, а аутентификация осуществляется в домене Windows 2000 domain, то процесс аутентификации может пойти через интерфейс со включенными jumbo frames к контроллеру домена, который, как обычно бывает, не сконфигурирован на использование jumbo frames. Это приведет к тому, что возникнут большие задержки или ошибки процесса аутентификации при использовании CIFS.

3.3.7. iSCSI Initiators для Windows

NetApp рекомендует использовать Microsoft iSCSI initiator или NetApp iSCSI host attach kit 2.0 for Windows по выделенной высокоскоростной сети Gigabit Ethernet на таких платформах, как Windows 2000, Windows 2000 AS, и Windows 2003 с Oracle Databases. Для платформ, например Windows NT, которые не имеют iSCSI, NetApp поддерживает CIFS для использования с Oracle Database и хранилища приложений. Однако рекомендуется обновиться до Windows 2000 или новее, и использовать iSCSI initiator (программный или аппаратный). NetApp в настоящий момент поддерживает Microsoft iSCSI initiator 1.02 и 1.03, доступный на www.microsoft.com.

3.3.8. FCP SAN Initiators для Windows

NetApp поддерживает использование Fibre Channel SAN в Windows для Oracle Databases. NetApp рекомендует использовать Fibre Channel SAN для Oracle Databases под Windows там, где уже присутствует развитая инфраструктура Fibre Channel. NetApp также рекомендует обратить внимание на решение с использованием Fibre Channel SAN для Windows там, где величины постоянного траффика ввода-вывода для сервера Oracle превышают 1GB в секунду (~110MB в секунду).

4. Настройки Oracle

Этот раздел описывает настройки, которые делаются в Oracle Database, обычно через установки в файле `init.ora`. Подразумевается, что у читателя есть соответствующие знания того, как правильно устанавливать эти опции, и о том, какой эффект они вызывают. Установки, описанные здесь, одни из наиболее часто используемых при настройке систем хранения NetApp с Oracle Databases.

4.1. DISK_ASYNCH_IO

Разрешение или запрещение режима асинхронного ввода-вывода (asynchronous I/O) в Oracle. Режим асинхронного ввода-вывода позволяет процессам выполнять следующую операцию, не ожидая завершения выполнения операции записи, что улучшает производительность системы и уменьшает время ожидания (idle time). Эта установка может улучшить производительность, в зависимости от характера базы данных. Если параметр DISK_ASYNCH_IO установлен в TRUE, тогда DB_WRITER_PROCESSES и DB_BLOCK_LRU_LATCHES (версии Oracle до 9i) или DBWR_IO_SLAVES должны использоваться так, как написано ниже. Правило вычисления значения таково:

```
DB_WRITER_PROCESSES = 2 * number of CPUs
```

Тесты для Solaris 8 с наложенным патчем 108813-11 и позднее, и для Solaris 9, показывают, что настройки:

```
DISK_ASYNCH_IO = TRUE

DB_WRITER_PROCESSES = 1
```

могут давать лучшие результаты, по сравнению с DISK_ASYNCH_IO установленным в FALSE.

NetApp рекомендует использовать ASYNC_IO для Solaris 2.8 и новее.

4.2. DB_FILE_MULTIBLOCK_READ_COUNT

Определяет максимальное количество блоков базы данных, читаемое за одну операцию ввода-вывода во время full table scan. Число читаемых байтов базы данных вычисляется умножением DB_BLOCK_SIZE * DB_FILE_MULTIBLOCK_READ_COUNT. Установка этого параметра уменьшает число вызовов операции ввода-вывода, требуемых для full table scan, что улучшает производительность. Увеличение этого значения может улучшить производительность базы данных, которая совершает много full table scans, но ухудшает производительность для баз данных OLTP, где full table scans довольно редок.

Установка этого числа кратным величине NFS READ/WRITE, определенным при монтировании, уменьшит величину фрагментации, которая возникает при вводе-выводе. Обратите внимание, что этот параметр определен в «блоках базы данных», а установки NFS - в «байтах», так что необходимо выравнивание значений с учетом этого. Например, установка значения DB_FILE_MULTIBLOCK_READ_COUNT в 4, умноженное на DB_BLOCK_SIZE равное 8kB, даст размер буфера чтения в 32kB.

NetApp рекомендует, чтобы значение DB_FILE_MULTIBLOCK_READ_COUNT было установлено от 1 до 4 для OLTP-баз, и от 16 до 32 для DSS.

4.3. DB_BLOCK_SIZE

Для наилучшей производительности базы данных, DB_BLOCK_SIZE должен быть кратен размеру блока в OS. Например, если размер блока в Solaris равен 4096:

```
DB_BLOCK_SIZE = 4096 * n
```

Опции NFS `rsize` и `wsize`, определяемые при монтировании файловой системы, также должны быть кратны этой величине. Ни в коем случае они не должны быть меньше. Например, если DB_BLOCK_SIZE в Oracle равен 16kB, то NFS `rsize` и `wsize` (размеры блоков чтения и записи) должны быть 16kB или 32kB, но не 8kB или 4kB.

4.4. DBWR_IO_SLAVES и DB_WRITER_PROCESSES

DB_WRITER_PROCESSES важен для систем, которые изменяют много данных. Он определяет исходное число процессов записи базы данных (DBWR – database writer processes), для соответствующего экземпляра (instance). Если используется DBWR_IO_SLAVES, то будет разрешен только один процесс записи (database writer process), независимо от величины DB_WRITER_PROCESSES. Множественные DBWR и DBWR IO slaves не могут существовать в системе одновременно. Рекомендуется использовать или ту или другую опцию для компенсации снижения производительности при выключении DISK_ASYNCH_IO. Статья на Metalink 97291.1 дает руководство по их использованию.

NetApp рекомендует использовать DBWR_IO_SLAVES для систем с одним CPU, и DB_WRITER_PROCESSES для многопроцессорных систем.

Главное правило настройки опций – всегда включать DISK_ASYNCH_IO, если это поддерживается операционной системой. Следующим шагом проверьте, поддерживается ли это в NFS, или только для блочного доступа (FC/iSCSI). Если поддерживается в NFS, тогда включите async I/O на уровне Oracle, и на уровне OS и замерьте изменение производительности. Если прирост заметен, то используйте async I/O для NFS. Если async I/O не поддерживается для NFS, или прирост производительности незначителен, то тогда попробуйте вулючить множественные DBWRs и DBWR IO slaves, как описано далее.

4.5. DB_BLOCK_LRU_LATCHES

Число DBWRs не может превышать величину в параметре DB_BLOCK_LRU_LATCHES:

```
DB_BLOCK_LRU_LATCHES = DB_WRITER_PROCESSES
```

Начиная с версии Oracle9i, параметр DB_BLOCK_LRU_LATCHES отсутствует, и не требует установки.

5. Резервное копирование, восстановление, катастрофоустойчивость

Для дополнительной информации о разработке стратегии резервного копирования, восстановления и катастрофоустойчивости, смотрите [8], [9], и [10].

Для дополнительной информации об организации быстрого резервного копирования и восстановления Oracle Database под UNIX®, смотрите [3] и [7].

5.1. Как создать резервную копию данных с системы хранения NetApp

Данные, хранящиеся на системе хранения NetApp, могут быть скопированы на другую систему хранения, на систему типа nearline, или магнитную ленту. Всегда следует учитывать протокол,

используемый для доступа к данным, когда создается резервная копия. Когда для доступа к данным используются NFS и CIFS, то использование Snapshot и SnapMirror® всегда даст в результате консистентную копию файловой системы. Но они должны координироваться с состоянием базы данных Oracle, чтобы получить консистентность также и для базы данных.

В случае использования протоколов Fibre Channel или iSCSI, снэпшот-копии и команды SnapMirror должны также быть скоординированы с сервером. Файловая система на сервере должна быть залокирована, и все данные в памяти сброшены на диски, до того, как мы отдадим команду на создание снэпшота.

Данные могут быть сохранены на ту же систему хранения, на другую такую же, на систему NearStore® или на устройство хранения на магнитной ленте. Устройство на ленте может быть подключено как непосредственно к системе хранения, так и находиться в сети, SAN (Fibre Channel) или LAN (Ethernet), и система хранения может проводить резервное копирование своих данных через сеть на такое устройство.

Возможные варианты для резервного копирования данных системы хранения NetApp таковы:

- Использовать SnapManager for Oracle для создания онлайн- или оффлайн-бэкапа
- Использовать авоматически создаваемые снэпшоты для создания онлайн-бэкапа
- Использовать скрипты на сервере, работающие на системе хранения по rsh, для того, чтобы создавать снэпшоты и производить онлайн-бэкап
- Использовать SnapMirror для репликации данных на другую систему хранения или систему типа NearStore
- Использовать SnapVault® для сохранения данных на другой системе хранения NetApp или системе NearStore
- Использовать команды серверной OS для копирования данных и создания резервной копии
- Использовать команды NDMP для сохранения данных на другой системе хранения NetApp или системе NearStore
- Использовать команды NDMP для резервного копирования на устройство хранения на магнитной ленте
- Использовать дополнительные «third-party» инструменты резервного копирования, для копирования данных с системы хранения или NearStore на ленты или другие устройства хранения

5.2. Создание онлайн-копий с помощью Snapshot

Технология NetApp Snapshot позволяет максимально эффективно использовать пространство системы хранения сохраняя только блоки изменений между созданными копиями Snapshot. Так как снэпшоты создаются практически мгновенно, резервное копирование данных является простым и быстрым делом. Снэпшоты могут создаваться по расписанию, или они могут быть созданы скриптом с сервера, а также с помощью SnapDrive™ или SnapManager®.

В Data ONTAP включен сервис расписаний, предназначенный для автоматизированного создания снэпшотов. Используйте автоматическое создание снэпшотов для резервных копий таких данных, как, например, данные «домашних директорий».

Базы данных и другие приложения, должны копироваться, будучи переведенными в режим резервной копии. Для Oracle Databases это означает перевод базы данных в hot backup mode перед

созданием снэпшота. У NetApp есть несколько документов, описывающих детально процесс резервного копирования Oracle Database.

Для подробностей относительно вопросов защиты данных см. [11].

NetApp рекомендует использовать снэпшоты для создания «холодных» и «горячих» копий баз данных Oracle. Использование снэпшотов не ухудшает производительности системы хранения. Рекомендуется выключить автоматическое расписание создания снэпшотов, и координировать процесс создания снэпшотов с состоянием базы данных Oracle.

Для дополнительной информации об интеграции технологий снэпшотов и резервного копирования баз данных Oracle, смотрите [3] и [7].

5.3. Восстановление отдельных файлов из Snapshot

Отдельные файлы и директории могут быть восстановлены из снэпшотов, используя обычные команды сервера, такие как команда cp в UNIX, или перетаскивание мышкой в Microsoft Windows. Данные также могут быть восстановлены с помощью команды single-file SnapRestore. Используйте тот метод, что быстрее работает.

5.4. Восстановление данных с помощью SnapRestore

SnapRestore позволяет быстро восстановить файловую систему целиком, в состояние, сохраненное в сделанном снэпшоте. SnapRestore можно использовать как для восстановления дискового тома целиком, так и отдельных файлов с этого тома.

При использовании SnapRestore для восстановления тома данных, данные на этом томе должны принадлежать одному приложению. В противном случае, эта операция может повлиять также и на данные другого приложения.

Опция SnapRestore восстановления одного файла, позволяет выбрать отдельный файл для восстановления, без восстановления всех файлов тома. Учтите, что файл, который восстанавливается с помощью SnapRestore, не может уже присутствовать где-либо в активной файловой системе. В этом случае система молча переключит single-file SnapRestore в операцию копирования. В результате операция single-file SnapRestore займет гораздо больше времени, чем ожидалось (в обычном виде команда выполняется за доли секунды) и также потребует, чтобы достаточное для копирования количество места присутствовало на активной файловой системе.

NetApp рекомендует использовать SnapRestore для мгновенного восстановления баз данных Oracle. SnapRestore может восстановить том целиком на конкретный момент времени взятия снэпшота в прошлом, или восстановить отдельный файл. Преимущества использования SnapRestore на уровне тома в том, что том может быть восстановлен в считанные минуты, что снижает время недоступности баз данных при выполнении восстановления Если вы планируете использовать SnapRestore на уровне томов, то рекомендуется организовать хранение лог-файлов, архивных логов, и копий контрол-файлов в отдельных томах, отдельно от собственно данных базы, и использовать SnapRestore только томе, содержащем эти данные.

Для подробностей о применении SnapRestore для восстановления Oracle Database, см. [3] и [7].

5.5. Консолидирование резервных копий с помощью SnapMirror

Зеркалирование данных с помощью SnapMirror возможно из одного тома или qtree, на одну или более удаленных систем хранения NetApp одновременно. Оно обновляет зеркалированные данные, сохраняя их доступными и актуальными.

SnapMirror особенно полезный инструмент для задач, связанных с сокращением окна резервного копирования на основной системе. SnapMirror можно использовать для непрерывного зеркалирования данных с основной системы хранения на выделенную систему типа nearline storage. Операции резервного копирования передаются тем самым на систему, на которой устройство на магнитной ленте может работать в течение всего дня, без необходимости прерывать работу основной системы хранения. Так как операции резервного копирования в таком случае не проводятся на основной, первичной системе, то «окно резервного копирования» при этом не играет важной роли.

5.6. Создание катастрофоустойчивой системы с помощью SnapMirror

SnapMirror непрерывно обновляет реплицированные данные, чтобы поддерживать их актуальными. SnapMirror это правильный инструмент для использования при построении катастрофоустойчивой (disaster recovery) системы. Тома могут быть реплицированы как асинхронно, так и синхронно, на систему в удаленном датацентре. Сервера приложений также должны быть реплицированы в этот датацентр.

В случае наступления события катастрофы, для DR-датацентра важно быть работоспособным, приложения должны корректно переключиться на сервера в DR-датацентре, и весь трафик приложений должен быть перенаправлен на эти сервера, до тех пор пока основной датацентр не восстановит работу. Когда основной датацентр вернется в работу, то SnapMirror можно использовать для эффективной передачи данных обратно, на основные системы хранения. После того, как основной датацентр вновь заработает, обычную работу по передаче изменений на DR-датацентр можно продолжить без необходимости проводить полную передачу всех данных «с нуля» еще раз.

Для подробного рассмотрения темы использования SnapMirror для DR-задач с Oracle, смотрите [12].

5.7. Создание оперативных резервных копий с помощью SnapVault

SnapVault предлагает решение централизованного дискового резервного копирования для гетерогенных сред хранения. Хранение резервных копий в виде множества снэпшотов во вспомогательной, «вторичной» системе хранения, позволяет предприятию хранить недели бэкапов в онлайне, доступных для быстрого восстановления данных из них. SnapVault также дает пользователям возможность выбирать, какие именно данные будут скопированы, частоту создания резервных копий, и то, как долго эти копии будут храниться.

SnapVault построен на базе принципах асинхронной блочной инкрементальной передачи, использованной в технологии SnapMirror, с добавлением технологий архивации. Это позволяет данным копироваться с помощью снэпшотов на основной системе хранения, и по расписанию передаваться на другую систему хранения или на устройство NearStore. Снэпшоты могут сохраняться на второй системе многие недели и даже месяцы, позволяя провести операцию восстановления данных на исходной системе практически мгновенно.

Для дополнительных сведений о защите данных с помощью SnapVault, смотрите [8], [9], и [11].

5.8. NDMP и резервное копирование-восстановление на лентах

Network Data Management Protocol, или NDMP, это открытый стандарт для централизованного управления процесса передачей данных. Архитектура NDMP позволяет производителям приложений резервного копирования управлять устройствами резервного копирования, подключенными к устройствам NetApp, или другим файл-серверам, обеспечивая единый интерфейс между приложением резервного копирования и файловыми серверами.

NDMP отделяет поток управления от данных резервного копирования или восстановления.

Это позволяет добиться большей гибкости в конфигурировании инфраструктуры, используемой для защиты данных на системах хранения NetApp.

Поскольку эти потоки разделены, то они могут поступать на различные устройства, также как и исходить с различных устройств, в результате позволяя предельно гибкую топологию, использующую NDMP. Доступные топологии NDMP обсуждаются подробнее в [13].

Если оператор не определил существующий снэпшот при запуске резервного копирования или копирования по NDMP, то Data ONTAP создаст его перед началом процесса. Этот снэпшот будет удален после окончания резервного копирования. Когда файловая система содержит данные FCP, снэпшот должен быть создан в момент времени, соответствующий консистентному состоянию данных. Как описывалось ранее, это наилучшим образом делается с помощью скрипта, который приостанавливает приложение, или переводит его в hot backup mode перед созданием снэпшота. После создания снэпшота, обычные операции могут быть продолжены, и резервная копия на ленту с этого снэпшота может быть перенесена за любое желаемое время.

При подключении системы хранения к Fibre Channel SAN для резервного копирования на ленту, сначала необходимо убедиться, что NetApp сертифицировал соответствующее ПО и оборудование. Полный список сертифицированных конфигураций доступен на портале защиты данных NetApp. Запасные соединения к коммутаторам Fibre Channel и ленточным библиотекам в настоящий момент не поддерживаются для Fibre Channel tape SAN. Кроме того для бэкапа на ленту необходимо использовать отдельный host bus adapter. Этот адаптер должен быть подключен в отдельный коммутатор Fibre Channel, в который включены только устройства NearStore и сертифицированные ленточные библиотеки и ленточные устройства.

Backup server должен также уметь соединяться с ленточной библиотекой по NDMP, или управлять роботом библиотеки, подключенной непосредственно к этому серверу.

5.9. Использование Tape Devices с системами NetApp

Системы хранения NetApp FAS и устройства NearStore поддерживают резервное копирование и восстановление для данных, на локально подключенные, а также использующие Fibre Channel, и Gigabit Ethernet устройства, включенные в SAN. Поддержка для наиболее распространенных устройств записи на магнитную ленту (tape drives), а также форматов магнитных лент, включена в поставку системы, а новые устройства добавляются при обновлениях системы. Кроме этого полностью поддерживается протокол RMT, позволяя проводить резервное копирование и восстановление для любой совместимой с ним системы. Образы резервных копий записываются в формат, производный от стандартного формата BSD dump, позволяя делать как полную копию файловой системы, так и девять уровней дифференциальной резервной копии.

5.10. Поддерживаемые инструменты резервного копирования других компаний

NetApp поддерживает партнерство со следущими производителями ПО рзервного копирования, использующими NDMP.

Atempo® Time Navigator www.atempo.com	Legato® NetWorker® www.legato.com
BakBone® NetVault® www.bakbone.com	SyncSort® Backup Express www.syncsort.com

CommVault® Galaxy www.commvault.com	VERITAS® NetBackup™ www.veritas.com
Computer Associates™ BrightStor™ www.ca.com	Workstation Solutions Quick Restore Enterprise www.worksta.com

5.11. Наилучшие методы резервного копирования и восстановления

Эта глава соединяет в себе технологии и продукты защиты данных NetApp, описанные выше, в набор наилучших методов и практик, для производства Oracle hot backups (онлайновых резервных копий) для резервного копирования, восстановления и архивации, с использованием как «первичных» систем хранения (системы хранения с высокопроизводительными дисками Fibre Channel) так и системы класса «nearline» (системы NearStore, с недорогими, емкими дисками ATA и SATA). Такая комбинация основной системы, для хранения рабочих баз данных, и вспомогательной, nearline, для бэкапов активного датасета, увеличивает производительность, и понижает стоимость операций. Периодическое перемещение данных с основного на вспомогательное хранилище увеличивает свободное место на основном и увеличивает производительность, одновременно заметно понижая стоимость хранения.

Отметьте: Если системы NetApp NearStore не являются частью вашей бэкап-стратегии, то ознакомьтесь с документом [6], для подробной информации о резервном копировании и восстановлении Oracle на системе хранения с использованием технологии Snapshot. Остальная часть этой главы описывает совместное использование основной системы хранения и системы NearStore.

5.11.1. SnapVault и резервное копирование базы

База данных Oracle может быть скопирована в то время, когда она работает и обрабатывает запросы в рабочем режиме, но сперва она должна быть переведена в так называемый hot backup mode.

Определенные действия должны быть произведены перед и после взятия снэпшот-копии с тома базы данных. Так как эти действия те же самые, что и при использовании иных методов резервного копирования базы, то многие администраторы базы данных, возможно, уже имеют соответствующие скрипты для выполнения этих операций.

Когда расписание SnapVault Snapshot может быть скоординировано с соответствующими действиями над базой данных, с помощью синхронизированных часов на системе хранения и сервере базы данных, можно упростить поиск и нахождение потенциальных проблем, если скрипт резервного копирования базы, создающий снэпшот-копию, использует команду SnapVault `snap create`.

В нашем примере консистентный образ базы данных создается каждый час, сохраняются снэпшот-копии за последних 5 часов (5 наиболее «свежих»). Создается и хранится также одна копия в день в течении недели, и одна «недельная», берущаяся в конце недели. В ПО SnapVault на вторичном хранилище сохраняется подобное же количество SnapVault Snapshot-копий.

Создание резервной копии базы Oracle в режиме hot backups с помощь SnapVault:

1. Установите связь системы NearStore с основной системой хранения.
2. Настройте расписание для задания количества сохраняемых снэпшот-копий на каждом из устройств хранения, как основной системе хранения так и NearStore.
3. Запустите процесс SnapVault между системой хранения и устройством NearStore.
4. Создайте скрипт, создающий Snapshot-копии с помощью SnapVault на системе хранения и NearStore, для выполнения Oracle hot backups.

5. Создайте на хост-системе скрипт расписания, выполняющийся с помощью cron, чтобы запускать скрипты hot backup снэпшот-копий SnapVault, как это описано выше.

Шаг 1: Установите соединение между системой NearStore и основной системой хранения.

Примеры в этом разделе подразумевают, что первичная система хранения, для размещения базы данных, носит имя `descent`, а устройство NearStore, предназначенное для архивирования базы данных, называется `rook`.

1. Введите лицензию и включите SnapVault на системе хранения («descent»):
```
descent> license add ABCDEFG
descent> options snapvault.enable on
descent> options snapvault.access host=rook
```

2. Введите лицензию и включите SnapVault на устройстве NearStore («rook»):
```
rook> license add ABCDEFG
rook> options snapvault.enable on
rook> options snapvault.access host=descent
```

3. Создайте том-получатель SnapVault на NearStore («rook»):
```
rook> vol create vault –r 10 10
rook> snap reserve vault 0
```

Шаг 2: Установка расписаний (выключение автоматических снэпшотов) на основной системе хранения и системе NearStore.

1. Отключите обычное расписание создания снэпшотов на системе хранения и системе NearStore, оно будет заменено расписанием SnapVault:
```
descent> snap sched oracle 0 0 0
rook> snap sched vault 0 0 0
```

2. Установите расписание снэпшотов SnapVault, создаваемое скриптом на системе хранения `descent`, для тома по имени `oracle`. Эта команда выключает расписание, а также определяет, сколько копий снэпшотов каждого вида будет храниться.
```
descent> snapvault snap sched oracle sv_hourly 5@-
```
Это расписание создает снэпшот по имени sv_hourly и оставляет пять наиболее свежих копий его, но не определяет время, которое будет задано скриптом cron ниже.
```
descent> snapvault snap sched oracle sv_daily 1@-
```
Также как выше, это расписание создает снэпшот по имени sv_daily и оставляет одну, наиболее свежую его копию, но также не определяет время его создания.
```
descent> snapvault snap sched oracle sv_weekly 1@-
```
Также как выше, это расписание создает снэпшот по имени sv_weekly и оставляет одну, наиболее свежую его копию. Также не задается время его создания.

3. Установите расписание снэпшотов SnapVault, создаваемое скриптом на системе хранения `rook`, для тома по имени `vault`. Эта команда выключает расписание, а также определяет, сколько копий снэпшотов каждого вида будет храниться.
```
rook> snapvault snap sched –x vault sv_hourly 5@-
```

Это расписание создает снэпшот по имени sv_hourly и оставляет пять наиболее свежих копий его, но не определяет время, котрое будет задано скриптом cron ниже.

```
rook> snapvault snap sched –x vault sv_daily 1@-
```

Также как выше, это расписание создает снэпшот по имени sv_daily и оставляет одну, наиболее свежую его копию, но также не определяет время его создания.

```
rook> snapvault snap sched –x vault sv_weekly 1@-
```

Также как выше, это расписание создает снэпшот по имени sv_weekly и оставляет одну, наиболее свежую его копию. Также не задается время его создания.

Шаг 3: Запуск процесса SnapVault между основной системой хранения и NearStore.

На этом этапе расписания сконфигурированы как на основной системе, так и на вторичной, NearStore, и SnapVault включен и работает. Однако, SnapVault не знает, какие тома или qtrees сохранены или где они хранятся на вторичной системе. Снэпшоты будут созданы на первичной системе, но данные пока не передаются на вторичную.

Чтобы дать SnapVault эту информацию, запустим следующую команду на вторичной системе:

```
rook> snapvault start -S descent:/vol/oracle/- /vol/vault/oracle
```

Шаг 4: Создание скрипта Oracle hot backup со SnapVault.

Пример скрипта `/home/oracle/snapvault/sv-dohot-daily.sh`:

```
#!/bin/csh -f
# Place all of the critical tablespaces in hot backup mode.
$ORACLE_HOME/bin/sqlplus system/oracle @begin.sql
# Create a new SnapVault Snapshot copy of the database volume on the primary filer
rsh -l root descent snapvault snap create oracle sv_daily
# Simultaneously 'push' the primary filer Snapshot copy to the secondary NearStore
system
rsh -l root rook snapvault snap create vault sv_daily
# Remove all affected tablespaces from hot backup mode.
$ORACLE_HOME/bin/sqlplus system/oracle @end.sql
```

Отметьте, что скрипты `@begin.sql` и `@end.sql`, содержат команды, переводящие таблицы баз данных в hot backup mode (begin.sql) и выводящие их из hot backup mode (end.sql).

Шаг 5: Использование cron для запуска скрипта Oracle hot backup, созданного для SnapVault на шаге 4.

Приложение выполнения по расписанию, такое как `cron` в UNIX-системах, или `task scheduler` в Windows, используется для создания снэпшота `sv_hourly` каждый час, каждого дня, кроме 2:00 в субботу; один снэпшот в день `sv_daily` каждый день в 2:00 кроме субботы, и один недельный `sv_weekly` в 2:00 субботы.

Sample cron script:

```
# sample cron script with multiple entries for Oracle hot backup
# using SnapVault, NetApp filer (descent), and NetApp NearStore (rook)
# Hourly Snapshot copy/SnapVault at the top of each hour
0 * * * *:
/home/oracle/snapvault/sv-dohot-hourly.sh
# Daily Snapshot copy/SnapVault at 2:00 a.m. every day except on Saturdays
0 2 * * 0-5:
/home/oracle/snapvault/sv-dohot-daily.sh
# Weekly Snapshot copy/SnapVault at 2:00 a.m. every Saturday
```

```
0 2 * * 6:
/home/oracle/snapvault/sv-dohot-weekly.sh;
```

На шаге 4 приведен пример скрипта для ежедневного бэкапа, `sv-dohot-daily.sh`. Скрипты hourly и weekly, для почасового и еженедельного бэкапа полностью идентичны, отличаясь только именем создаваемого снэпшота (sv_hourly и sv_weekly, соответственно).

5.12. SnapManager for Oracle – Практики резервного копирования и восстановления

SnapManager for Oracle это программный продукт, работающий на хост-системе и клиенте, выпущенный компанией NetApp, Inc. Который интегрирован с Oracle9i R2 и Oracle Database 10g R2. Он позволяет DBA или администраторам систем хранения управлять процессом создания резервных копий, восстановления из этих копий, и клонирования.

SnapManager for Oracle использует NetApp Snapshot, SnapRestore и FlexClone для интеграции с Oracle. SnapManager автоматизирует и упрощает сложные, и требующие ручных действий операции, обычно производимых Oracle DBA, что значительно улучшает условия service level agreements (SLA) для резервного копирования и восстановления.

SnapManager for Oracle протоколонезависим, и поэтому работает одинаково хорошо как с NFS так и с iSCSI. Он также интегрирован с собственными технологиями Oracle, такими как RAC, ASM и RMAN.

Для использования SnapManager for Oracle, вам нужно использовать следующие типы баз данных, приложений и лицензий:

- SnapManager for Oracle
- Red Hat Enterprise Linux 3.0 Update 4
- SnapDrive for UNIX V2.1 (Red Hat Enterprise Linux)
- NetApp Data ONTAP 7.0 или новее
- Oracle 9iR2 или Oracle 10gR2 использующих NFS или iSCSI LUN
- NetApp Host Agent 2.2.1
- Лицензия на FlexClone
- Лицензия на NFS или iSCSI
- Лицензия на SnapRestore

NetApp рекомендует использовать SnapManager for Oracle если вы используете Redhat Enterprise Linux 3.0 Update 4 (RHEL 3.0 Update 4) и выше, и вы хотите решить проблемы критичных приложений, относящихся к вопросам резервного копирования и восстановления Oracle, а также его клонирования.

Для подробностей об использовании SnapManager for Oracle в среде Oracle смотрите [14].

5.12.1 SnapManager for Oracle – копирование и восстановление с использованием ASM

SnapManager for Oracle обеспечивает средства для резервного копирования баз данных, расположенных на Oracle ASM на системах хранения NetApp. Это позволяет использующим базы Oracle 10g R2 на Automatic Storage Management (ASM) поверх iSCSI LUN настраивать использование снэпшотов NetApp и технологию SnapRestore с помощью SnapManager for Oracle. Это обеспечивает при обслуживании баз ту гибкость и простоту использования, ради которой создавался Oracle ASM.

NetApp требует, чтобы использовался драйвер ASMlib в Red Hat Linux Enterprise 3, когда вы пользуетесь ASM и SnapManager for Oracle. Драйвер ASMlib это необходимый компонент для работы SnapManager for Oracle и он не будет работать без него.

5.12.2 SnapManager for Oracle – копирование и восстановление с использованием RMAN

SnapManager for Oracle обеспечивает интеграцию с Oracle RMAN, позволяя заносить резервные копии, сделанные с помощью SnapManager в каталог RMAN. Это позволяет DBA использовать снэпшоты NetApp и технологию SnapRestore для резервного копирования и восстановления с помощью SnapManager, в то же время продолжая иметь доступ к средствам RMAN, которые он, возможно, лучше знает. Поэтому использование интеграции RMAN со SnapManager позволит вам проводить восстановление на блочном уровне не жертвуя RMAN.

NetApp требует, чтобы все датафайлы, логфайлы и файлы архивлогов забэкапленных баз данных, хранились на системе хранения NetApp на томе flexvol.

5.12.3 SnapManager for Oracle – клонирование

SnapManager for Oracle позволяет делать клонирование базы данных для версий Oracle 9i R2 и Oracle Database 10g R2.

Клонирование возможно с помощью технологии NetApp FlexClone, при использовании SnapManager. Для процесса клонирования базы данных предоставляется старый sid и новый sid, а также map-файл, которые позволяют DBA или администратору системы хранения определить для клона базы данных новое место размещения ее файлов, также как и новые их имена.

NetApp требует, чтобы клон базы данных был сделан с резервной копии, выполненной с находящейся в offline базы. Режим клонирования «hot database cloning» будет доступен в будущих релизах SnapManager for Oracle.

Ссылки

1. NetApp supported NFS mount options for Oracle database files.
   http://now.netapp.com/Knowledgebase/solutionarea.asp?id=kb7518
   http://www.netapp.com/library/tr/3373.pdf
2. Data ONTAP™ 7G—The Ideal Platform for Database Applications
   http://www.netapp.com/library/tr/3373.pdf
3. Database layout with FlexVol and FlexClones
   http://www.netapp.com/library/tr/3411.pdf
4. Oracle9i for UNIX: Backup and Recovery Using a NetApp Filer:
   http://www.netapp.com/library/tr/3130.pdf
5. Using the Linux NFS Client with NetApp: Getting the Best from Linux and NetApp
   http://www.netapp.com/library/tr/3183.pdf
6. Installation and Setup Guide 1.0 for Fibre Channel Protocol on Linux:
   http://now.netapp.com/NOW/knowledge/docs/hba/fcp_linux/fcp_linux10/pdfs/install.pdf
7. Oracle9i for UNIX: Integrating with a NetApp Filer in a SAN Environment:
   http://www.netapp.com/library/tr/3207.pdf
8. Oracle9i for UNIX: Backup and Recovery Using a NetApp Filer in a SAN Environment:
   http://www.netapp.com/library/tr/3210.pdf
9. Data Protection Strategies for Network Appliance Filers:
   http://www.netapp.com/library/tr/3066.pdf
10. Data Protection Solutions Overview:
    http://www.netapp.com/library/tr/3131.pdf
11. Simplify Application Availability and Disaster Recovery:
    http://www.netapp.com/partners/docs/oracleworld.pdf
12. SnapVault Deployment and Configuration:
    http://www.netapp.com/ibrary/tr/3240.pdf
13. Oracle8i™ for UNIX: Providing Disaster Recovery with NetApp SnapMirror Technology:
    http://www.netapp.com/library/tr/3057.pdf
14. NDMPCopy Reference:
    http://now.netapp.com/NOW/knowledge/docs/ontap/rel632/html/ontap/dpg/ndmp11.htm#1270498
15. SnapManager for Oracle:
    http://www.netapp.com/library/tr/3426.pdf

История изменений

Дата	Кто	Описание
06/03/08	Padmanabhan Sadagopan	Обновление
12/01/07	NetApp team	Создание