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 созданию отказоустойчивого хранилища данных

Jay White, NetApp
Сентябрь 2011 | TR-3437

Коротко о главном:

Этот документ содержит технические рекомендации и наилучшие практики по созданию надежного, отказоустойчивого и высокодоступного хранилища данных на системе хранения NetApp®. Рассматриваемые в этом документе вопросы и темы важны при планировании и проектировании системы хранения данных с использованием устройств NetApp FAS, с целью получить надежное хранилище, удовлетворяющее ожиданиям и потребностям пользователей.

Оглавление

1 Введение

Готовя этот документ, мы подошли к вопросу надежности хранения с акцентом на ключевые функции и возможности, которые позволяют настроить систему хранения на максимальную отказоустойчивость в рамках соответствующих классов доступности данных. Это явилось основой для данного документа. Однако не всегда возможно, а в ряде случаев даже не всегда необходимо конфигурировать систему на достижение максимальной возможной отказоустойчивости, это зависит от задачи и требований к данной конфигурации хранилища. Кроме того, конечной целью любой системы хранения не обязательно является обеспечение отказоустойчивости как таковой, любой ценой, гораздо важнее обеспечение доступности данных. Насколько «отказоустойчивой» с точки зрения бизнеса может считаться система хранения, если в результате сбоя она, хоть и сохранила данные, но вследствие падения уровня производительности, нормальная работа с этими данными для приложений стала невозможна?

В ситуациях, подобной вышеописанной, простой отказоустойчивости, как таковой, недостаточно. Отказоустойчивость следует рассматривать с позиции доступности данных и того, как это влияет на систему в целом.

1.1 Доступность данных

Ключевое свойство системы хранения NetApp это доступность данных. С точки зрения этого документа можно оценить доступность данных исходя из трех факторов:

- **Производительность:** Понятие производительности может быть разбито на два ключевых аспекта с позиции доступности данных. Во-первых, пользователи могут иметь определенные требования по производительности системы хранения, которые обязательно должны выполняться, для удовлетворения требований приложения к данным. Сбой в доступности данных с этой точки зрения означают, что хотя система хранения продолжает отвечать на запросы ввода-вывода, но реакция ее и уровень производительности становятся неудовлетворительными и ниже допустимого для данной задачи и ее функций лимита. Во-вторых, если производительность системы страдает настолько, что система перестает совсем отвечать на запросы ввода-вывода, тогда сложившаяся ситуация безусловно является проблемой доступности данных.
- **Отказоустойчивость**: Отказоустойчивость, с точки зрения доступности данных, это способность системы переживать отдельные или множественные сбои, продолжая обслуживать операции ввода-вывода в *degraded* состоянии. Существует множество различных опций и специальных возможностей, отвечающих за способность системы переживать отказы; они будут обсуждаться на протяжении всего документа.
- **Recoverability:** *Recoverability* понимается как способность системы автоматически восстанавливаться после сбоя и продолжать обслуживание рабочей нагрузки даже во время восстановительных процедур.

Вышеперечисленные три фактора применимы к трем уровням доступности данных:

- **Подсистема хранения:** Уровень «Подсистемы хранения» относится ко всем аппаратным и программным компонентам, которые составляют внутреннее устройство системы хранения. В первую очередь это все, что находится от HBA «вниз», до физических компонент подключенных к ним массивов дисков, и программный код RAID, как часть Data ONTAP®: другими словами, все то, что позволяет внутри системы взаимодействовать между контроллером и подключенным к ним массивом дисков.

- **IT-Система:** Уровень «IT-системы» относится к возможности системы хранения в целом переживать сбои и отказы. В первую очередь это относится к сбоям уровня контроллера, влияющих на способность системы продолжать работу с внешними потребителями данных. Это относится к отдельным контроллерам и High Available-парам их, а также к компонентам, которые отвечают за внешние коммуникации данных контроллера, например сетевые и FC-интерфейсы.
- **Сайт:** Уровень «сайта» относится к возможности группы совместно размещенных систем хранения переживать сбои и отказы. Это, в первую очередь, относится к тем аспектам архитектуры распределенной системы хранения, которые позволяют пережить отказ всей системы, например в результате аварии и отказа датацентра целиком, в результате стихийного бедствия или террористического акта.

Для дальнейшего количественного определения наилучших практик и рекомендаций эта информация должна применяться к определенным уровням доступности данных. Это необходимо, так как, например, вы не можете обеспечить максимальной отказоустойчивости и *recoverability* одновременно, не жертвуя при этом чем-то в производительности.

В зависимости от требований каждого из уровней доступности данных, достигается баланс между тремя перечисленными выше факторами, что ведет к различным рекомендациям и наилучшим практикам в этих уровнях доступности. Уровни доступности данных, а также рекомендации для каждого из них приведены в разделе **6 Уровни доступности данных**, в этом документе.

1.2 Цели документа

Главная цель данного документа рассмотреть уровень подсистемы хранения с точки зрения обеспечения высокой доступности данных, учтя при этом три фактора – производительность, отказоустойчивость и способность к восстановлению (*recoverability*), определение которым дано в главе **1.1 Доступность данных**.

2 Надежность

Наиболее широко применяемый параметр измерения надежности в сегодняшней индустрии, это так называемый *mean time between failure* (MTBF) или «ожидаемое время до сбоя». Проблема заключается в том, что MTBF не обладает практической точностью в предсказании сбоя, как, например, его производные *average return rate* (ARR) или *average failure rate* (AFR), оба эти параметра отслеживаются компаниями, но, в большинстве случаев, эти данные не публикуются.

NetApp отслеживает параметры ARR и AFR для критичных компонентов системы хранения. Хотя ARR и AFR это параметры сформулированные яснее, чем MTBF, они также не идеальны. В случае, когда методы математической статистики используются для вычисления надежности, следует понимать, что они имеют практический смысл только в применении к весьма большой популяции устройств.

2.1 Измерение надежности

Для измерения надежности электронных компонентов существует три общепринятых сегодня метрики. Это так называемое «ожидаемое время наработки на отказ» (*mean time between failures*, MTBF), «усредненная частота возврата (диска по гарантии)» (*average return rate*, ARR), и «усредненная величина отказов» (*average failure rate*, AFR). Эти метрики детально рассмотрены в главе ниже, но основные их особенности вынесем отдельно:

- Запланированный срок жизни жестких дисков составляет пять лет. NetApp настоятельно рекомендует заменять диски по прошествии пяти лет срока их жизни. Эта рекомендация также следует установленному пятилетнему сроку гарантии на диски.
- Чем больше дисков вы используете в вашей конфигурации, тем выше вероятность того, что вы встретитесь с их отказами в течение срока жизни используемых дисков.
- MTBF это наименее точный параметр измерения надежности.
- AFR это наилучший параметр для измерения надежности, но он требует времени для определения точного действующего объема данных.

Эта глава в первую очередь рассматривает диски, но те же методы применимы и к другим компонентам системы хранения и иных устройств.

Mean Time Between Failures (MTBF)

MTBF – это наименее точная метрика измерения надежности. MTBF обычно ошибочно интерпретируется как срок практической жизни устройства. Поскольку производители оборудования не могут надежно протестировать их на весь их срок жизни, они тестируют множество однотипных устройств с тем, чтобы оценить коэффициент отказов в течение предполагаемого срока службы. Традиционно используемая формула для этого такова:

Длительность теста * число дисков / число отказавших в ходе тестирования дисков = MTBF

Наиболее часто величина MTBF для подсистемы хранения публикуется для дисков. Диски SSD, SATA, SAS, и FC имеют различные величины MTBF:

- SSD (SLC) – 2 миллиона часов
- SAS и FC – 1,6 миллиона часов
- SATA – 1,2 миллиона часов

Гарантийный срок жизни дисков составляет 5 лет (43 800 часов), что гораздо короче, чем 1,6 миллиона, или даже 1,2 миллиона часов.

Повторю это снова: MTBF это не параметр, указывающий на срок жизни диска, но параметр, говорящий о вероятности отказов, на протяжении его заданного производителем срока жизни.

Основываясь на MTBF, мы можем вычислить, что для дисков SATA (1,2 миллиона часов MTBF) следует ожидать, что примерно 0,73% ваших используемых дисков выйдут из строя каждый год. Для дисков FC и SAS (1,6 миллиона часов MTBF) следует ожидать примерно 0,55% отказов каждый год. Для дисков SSD (2,0 миллиона часов MTBF) - 0,44% отказов среди всех используемых дисков этого типа в год.

SATA: 1 200 000 часов MTBF / 8760 часов в год = 136,9863 года
1 отказ / 136,9863 года = 0,0073 * 100 = 0,73% отказов в год

FC и SAS: 1 600 000 часов MTBF / 8760 часов в год = 182,6484 года
1 отказ / 182,6484 года = 0,00537 * 100 = 0,55% отказов в год

SSD: 2 000 000 часов MTBF / 8760 часов в год = 228,3105 лет
1 отказ / 228,3105 лет = 0,00438 * 100 = 0,44% отказов в год

Рис. 1) Вычисление уровня отказов дисков, основываясь на MTBF.

Используя эти данные, давайте посчитаем три различных примерных конфигурации:

- 30 дисков SAS с ожидаемым временем эксплуатации 5 лет
- 300 дисков SAS с ожидаемым временем эксплуатации 5 лет
- 3 000 дисков SAS с ожидаемым временем эксплуатации 5 лет

Вычислим, сколько отказавших дисков мы ожидаем на протяжении срока службы этих конфигураций:

30 дисков SAS * 0,55% = 0,165 отказов/год * 5 лет = 0,825 отказов за 5 лет
300 дисков SAS * 0,55% = 1,65 отказов/год * 5 лет = 8,25 отказов за 5 лет
3000 дисков SAS * 0,55% = 16,5 отказов/год * 5 лет = 82,5 отказов за 5 лет

Рис. 2) Ожидаемое число отказов за 5 лет эксплуатации, основываясь на опубликованных данных MTBF и числе эксплуатируемых дисков.

В первую очередь вы должны видеть, что чем больше дисков вы эксплуатируете, тем выше вероятность того, что один из ваших дисков откажет во время его эксплуатации. При запланированном сроке пятилетней гарантии (три года стандартной, и два года расширенной) обычно применяемой для дисков *enterprise-class* сегодня, будет разумно сказать, что ожидаемый срок надежной жизни для диска составляет пять лет, после чего следует ожидать значительного роста отказов, тем большего, чем дольше диск находится в работе.

Средняя величина возвратов (Average Return Rate, ARR)

Показатель ARR для устройства это значительно лучший показатель надежности, чем MTBF так как он основывается на реальной величине отказов и возвратов дисков в сервисную службу. К сожалению и он также не является лучшим параметром измерения надежности, так как не различает устройства, которые были возвращены не по причине ошибок, а по каким-то другим причинам, не связанными с отказами. Некоторые примеры возвратов не связанных с отказами это ложные отказы (*false positives*), превентивные замены, сделанные до фактического отказа, или же возвраты, связанные с ошибками отгрузки. Хотя ARR и не является наилучшим показателем для определения надежности, для компании полезно отслеживать его, чтобы понимать наличие проблем с операционной эффективностью и иными бизнес-ориентированными причинами.

Средняя величина отказов (Average Failure Rate, AFR)

Это наиболее точный параметр измерения надежности устройства, так как он основывается на данных устройства, которое было возвращено производителю, и действительно проверено на наличие ошибки. К сожалению, для определения AFR необходимо время, так как его результат должен быть усреднен (*average* в названии) на протяжении определенного периода времени. В результате такого усреднения AFR становится тем более точным, чем на более длительном периоде времени он усредняется. Устройство может выйти из строя по множеству причин, некоторые из которых рассматриваются далее в этом документе.

Цель данного документа не разбор, чему равны параметры ARR или AFR для различных устройств, поставляемых NetApp (это непубличная информация), но, скорее, пояснение и помещение в контекст тех метрик, которые публично доступны или потенциально доступны для пользователя NetApp.

2.2 Надежность системы

Часто спрашивают, каков MTBF для контроллера или дисковой полки. Есть несколько причин, почему MTBF не называется для агрегата из нескольких связанных устройств:

- В основе принципа вычисления MTBF лежит использование отдельного устройства, или группы интегрированных между собой устройств. Контроллеры и дисковые полки состоят из множества компонентов, часть из которых опциональна (карты расширения, модули ввода-вывода в дисковой полке, и так далее), а также из компонентов, которые состоят из множества более мелких устройств. Таким образом, конфигурация этих устройств весьма вариативна, с точки зрения компонентов, вовлеченных в работу устройства и системы в целом.
- Величина MTBF считается, принимая во внимание работу всех используемых в устройстве компонентов, но не все компоненты контроллера или дисковой полки критичны для выполнения их основных функций. Например, отказ индикации на дисковой полке будет отказом, но не влияющим на основную задачу дисковой полки – отдавать и принимать данные дисков.
- Как уже указано выше в главе **2.1 Измерение надежности**, MTBF это наименее точный из параметров измерения надежности.

Добавление дополнительных устройств в ходе жизненного цикла устройства также уменьшает точность и ценность полученного результата. Более важно для понимания надежности вести учет параметров ARR и AFR, которые устраняют необходимость в вычислении MTBF.

Среди дисковых полок, модулей ввода-вывода в полках, и дисков, наименее надежными компонентами должны считаться, безусловно, диски. Но это не значит, что дисковые полки и модули ввода-вывода дисковых полок более надежны, чем диски. Логика этого заключения следующая:

- В дисковой полке находится гораздо больше дисков, чем любых иных устройств. Например, дисковая полка DS4243 имеет 2 или 4 блока питания (PSU), 2 модуля ввода-вывода IOM3, 1 собственно корпус дисковой полки, и 24 диска.
- Диски содержат много разной электроники и гораздо сложнее, чем любые другие компоненты, кроме того они содержат механические движущиеся части (исключая SSD).

В результате этих рассуждений, мы придем к выводу, что когда мы говорим о надежности, в первую очередь мы говорим о надежности дисков.

2.3 Надежность и наилучшие практики ее обеспечения

Некоторые наилучшие практики, следование которым повысит надежность вашей системы хранения:

- Удаляйте из системы сбойный диск как можно скорее и заменяйте его исправным, это поможет устранить вероятность возможного распространения проблем на исправные устройства и компоненты системы.
- Заменяйте или выводите из активной эксплуатации оборудование, которое отработало свой гарантийный срок.
- Следуйте рекомендованным практикам при работах по замене оборудования, чтобы избежать физических и электрических (электростатических) его повреждений.

- Помните и понимайте, что отказы оборудования неизбежны, и убедитесь, что запасные компоненты (spares) имеются и легко доступны. Следуйте опубликованным наилучшим практикам в отношении *hot/cold spares* и хорошо понимайте процедуры оборота (возврата и приема) запчастей для оборудования вашего сайта.
- Наличие *cold spares* не является поводом для отказа от *hot spares* в системе. Чем дольше аппаратный компонент системы лежит на полке склада (*cold spare*), тем больше шансов, что он пострадает от случайного физического повреждения или просто откажется работать, будучи вставленным на замену. Диск, установленный и работающий в составе системы хранения (*hot spare*) находится в рабочем состоянии высокой готовности, и готов немедленно принять на себя роль замены для вышедшего из строя диска.

3 Ошибки и сбои

Этот раздел рассматривает некоторые дополнительные детали о ряде основных ошибок и отказов, случающихся с системой хранения. Он не включает в себя описания всех возможных ошибок, которые только могут случиться; вместо этого основное внимание будет уделено тому, что затрагивает операции обеспечения отказоустойчивость системы, например процесс *RAID reconstruction*. Понятие *Single points of failure* (SPOF) также будет рассмотрено, так как оно относится к вопросам отказоустойчивости системы.

3.1 Единая точка отказа (Single Point of Failure, SPOF)

Некоторые потенциальные «точки отказа» устраняются путем правильной конфигурации самой системы. Например, каждая дисковая полка NetApp использует более одного модуля ввода-вывода, блока питания и дисков. Другие возможные SPOF могут присутствовать в зависимости от выбранной конфигурации системы:

- **Контроллер:** NetApp поддерживает работу в так называемой *single-controller configuration*, в которой в системе хранения используется один контроллер, и этот контроллер сам по себе будет являться SPOF. Использование пары контроллеров так называемом *HA cluster*, в котором используется два контроллера, устраняет проблему в контроллере как SPOF.
- **Host bus adapter (HBA):** Включают в себя интегрированные порты, и карты расширения, которые называются *port group*. *Port group* это любой набор взаимосвязанных портов. Например, интегрированные в контроллер порты A и B могут использовать другую микросхему ASIC, чем порты C и D, но и те и другие зависят от материнской платы, которая обеспечивает их функционирование. Одна четырехпортовая HBA обычно имеет два ASIC, но сама карта HBA, как отдельное устройство, является SPOF. По этой причине NetApp обычно рекомендует подключать петли к дискам (FC-AL) и стеки (SAS) к более одной *port group*. Например это могут быть два HBA или комбинация из интегрированных портов и одного и более HBA на карте расширения. NetApp всегда рекомендует как минимум делать подключения в раздельные ASIC.
- **Кабели:** Существует множество типов кабелей, используемых для подключения в системе хранения. Некоторые кабели более надежны и устойчивы к физическим повреждениям, чем другие; например, оптические кабели гораздо более уязвимы к механическим повреждениям, чем медные кабели Ethernet. Для предотвращения появления SPOF в ваших кабельных соединениях, NetApp рекомендует (и во многих случаях требует) использовать подключение типа *multipath high-availability* (MPHA). MPHA обеспечивает вторичный путь для данных для всех дисковых полок, подключенных в системе хранения.

- **Модули ввода-вывода полок:** Хотя полные отказы модулей ввода-вывода дисковых полок и являются весьма редкими событиями, они все же возможны. Практическим подходом для защиты от такой ситуации будет такая схема размещения дисков в группе, когда не более двух дисков в каждой их RAID-групп располагаются на одной полке (подразумеваем использование RAID-DP®). Это не является решением отказоустойчивости уровня полки и обеспечивает только сохранность данных. Получающаяся в результате деградация производительности системы в результате отказа полки (без использования зеркалирования) может парализовать работу системы и возможность продолжения операций. Рекомендуемый метод защиты против отказов уровня полки это использование *local SyncMirror®* ли иного метода зеркалирования данных, которые позволят быстро получит доступ к данным в случае отказа. Зеркалирование данных также поможет и в случае множественных отказов. Обратите внимание, что *Data ONTAP 8.1 Cluster-Mode* в настоящий момент не поддерживает SyncMirror.

3.2 Диски

Ошибки и отказы, связанные с дисками, обычно весьма сложны. В результате широко распространены заблуждения, относительно типов отказов, того, как они возникают, и как должны быть решены.

В некоторых случаях может создаться впечатление, что система хранения NetApp весьма агрессивно выводит из строя диски, по причинам не всегда выглядящим критичными. Например, после обнаружения одного плохо считываемого блока NetApp может целиком забраковать и вывести из работы диск, что может выглядеть несколько чрезмерным. Понятие «плохого блока» (*bad block*) в целом весьма простое. Диск возвращает код ошибки, связанный с неуспешной операцией диска, и эта ошибка показывает наличие серьезной проблемы. В зависимости от значительности ошибки, возвращенной диском, это также может показывать, что другие блоки на диске также должны считаться ненадежными. В этой ситуации безопаснее счесть диск вышедшим из строя целиком, и удалить его из активной файловой системы, чтобы не создавать проблем с надежностью для остальных данных.

Следующие пять состояний приводят к тому, что диск считается вышедшим из строя и запускаются процедуры восстановления RAID:

- Сам диск сообщил о фатальной ошибке.
- Уровень хранилища в Data ONTAP сообщил, что диск недоступен.
- Работающий диск уведомил Data ONTAP что теперь должен считаться сбойным.
- Слой хранилища и RAID в Data ONTAP рекомендовал считать, что диск должен считаться сбойным, что вызвано превышением заданного порога ошибок на нем.
- Возникло событие *Lost write protection* (LWP).

4 Действия по исправлению и предупреждению сбоев

Когда система регистрирует проблему, Data ONTAP проверяет текущее состояние RAID и состояние ошибки. В результате предпринимается одно из трех возможных действий:

- Запускается процесс *RAID reconstruction*.
- Запускается процесс *Rapid RAID Recovery* (может быть также результатом использования Maintenance Center).
- Ошибка игнорируется.

Процессы *RAID reconstruction* и *Rapid RAID Recovery* обсуждались детально выше. Ошибки потенциально могут игнорироваться для RAID-групп только в случае, когда она уже находится в состоянии *degraded*. Это так потому, что Data ONTAP уже понимает, что проблема присутствует, и, скорее всего, идет процесс восстановления из состояния *degraded state*, который решит в том числе и эту проблему.

Ошибки не связанные с отказами дисков обычно детектируются в ходе превентивной процедуры, например, так называемого *RAID scrubs*, и вызывают одно из следующих действий:

- Перезаписать содержимое подозрительного блока в новый блок (*data block repair*).
- Перезаписать данные *parity* для блока (*parity repair*).

Понимание вами того, что Data ONTAP управляет процессом восстановления блоков данных и блоков parity достаточно для целей данного документа, так как эти операции неспецифичны для конкретно дисковых отказов, а, скорее, являются проблемами с индивидуальными блоками данных в файловой системе.

4.1 RAID Reconstruction

Когда происходит сбой диска и запускается процесс *RAID reconstruction*, несколько различных факторов определяют то, как долго будет идти процесс восстановления, и как он повлияет на производительность системы в итоге. Некоторые из факторов, влияющих на системную производительность при работе в *degraded mode* это:

- Профиль рабочей нагрузки системы (произвольная / последовательная и соотношение чтений / записей)
- Текущий уровень загрузки CPU и полосы ввода-вывода
- Размер RAID-группы
- Используемые типы дисковой полки и ее модулей ввода-вывода
- Тип дисков (SSD, SATA, FC, или SAS)
- Настройки опций RAID
- Назначение путей доступа к дискам
- Распределение дисков по стекам и петлям
- Одиночный или двойной сбой диска и избранный системой вариант *RAID reconstruction*

По причине множества действующих факторов бывает очень трудно точно предсказать влияние этого процесса на работу системы хранения.

Когда диск выходит из строя, все операции ввода-вывода, которые при нормальной работе шли на этот диск, перенаправляются на заменяющий его диск. Трафик процесса *RAID Reconstruction* распределяется между всеми дисками RAID-группы в состоянии *degraded*, так как чтение данных при этом процессе производится со всех дисков данных этой RAID-группы. Дополнительный ресурс полосы пропускания в стеке/петле может потребоваться в том случае, если *degraded* RAID-группа и диск, заменяющий сбойный, находятся в одном стеке/петле. Ввод-вывод процесса *RAID reconstruction* будет идти параллельно с основным рабочим вводом-выводом системы, в соответствии с заданными опциями приоритетов в настройки RAID. Более глубоко этот вопрос обсуждается в главе ***5.2 Опции RAID***, в этом документе.

Процедура Single-Drive Reconstruction

Отказ одного диски и *RAID reconstruction* для группы RAID-DP происходит тем же образом, что и для RAID-группы с типом RAID с одним parity-диском (RAID-4), информация *double-parity* не используется. Процесс ребилда RAID состоит в считывании данных со всех оставшихся дисков и данных с диска *parity* (за исключением случая восстановления самого диска *parity*) и восстановления на основе этой информации утраченных блоков данных на вышедшем из строя диске.

Процедура *Single-Drive Reconstruction* удваивает объемы ввода-вывода в стеке/петле дисковой полки, содержащей RAID-группу, так как на каждую пользовательскую, рабочую операцию ввода-вывода, направленную на восстанавливающуюся RAID-группу, необходимо восстановить запрошенные данные. Этот трафик добавляется к трафику ребилда, созданному вычислениями из parity и записью на замещающий диск.

Процедура Double-Drive Reconstruction

Процедура *Double-Drive Reconstruction* применяется, когда в группе RAID-DP необходимо восстановить данные с использованием как *single-parity* так и *double-parity*. Такой тип ребилда считывает блоки со всех оставшихся дисков данных, а также с дисков *single-* и *double-parity*. Объем ввода-вывода по стеку/петле дисковой полки, содержащей восстанавливаемый RAID, утраивается. Data ONTAP действует достаточно интеллектуально, и не требует нескольких чтений при вычислениях восстанавливаемых данных и из *parity* и из *double-parity*; для обоих вычислений достаточно всего одной операции чтения.

Процедура *Double-Disk Reconstruction* утраивает объемы ввода-вывода в стеке/петле дисковой полки, содержащей RAID-группу, так как на каждую пользовательскую, рабочую операцию ввода-вывода, направленную на восстанавливающуюся RAID-группу, нужно восстановить данные для двух различных вышедших из строя дисков. Этот трафик добавляется к трафику ребилда, созданному вычислениями из parity и записью на замещающий диск.

4.2 Rapid RAID Recovery

Процесс под названием *Rapid RAID Recovery* похож на обычный *RAID reconstruction*, но без необходимости восстанавливать данные из *parity*, так как диск остается доступным и данные с него считываются. Некоторым блокам на диске может понадобиться восстановление из данных *parity*, но большая часть диска может быть скопирована поблочно на замещающий его диск. Так как это копирование на блочном уровне, все блоки копируются вне зависимости от того, находятся в них какие-то данные, или они пусты.

Процесс *Rapid RAID Recovery* конечно увеличивает общий объем ввода-вывода на стеке/петле, так как со сбойного диска считываются оставшиеся на нем данные, и переносятся на заменяющий его. Однако, влияние на оставшиеся диски RAID-группы гораздо меньше, чем при традиционном *RAID Reconstruction*, так как восстановление данных, путем вычисления parity, не требуется для подавляющего большинства данных сбойного диска. Процесс *Rapid RAID Recovery* также выполняется значительно быстрее, чем полный *RAID reconstruction*.

4.3 Maintenance Center

Когда он включен на вашей системе, *Maintenance Center* работает вместе с *Rapid RAID Recovery* для того, чтобы оценить состояние сбойного диска, перед тем как он будет возвращен в NetApp. Когда диск поступает в распоряжение процесса *Maintenance Center*, инициируется *Rapid RAID Recovery*, выводя диск из RAID-группы. Сбойный диск оценивается и тестируется Data ONTAP с

помощью ее внутренних процедур. Если диск оказывается рабочим, он возвращается в системный пул дисков *spare*. Если диск будет сочтен не рабочим, он остается в состоянии сбойного, и требует замены.

Maintenance Center для своей работы требует минимум двух дисков *hot spare*, доступных системе, и включенного *Rapid RAID Recovery* (опция `raid.disk.copy.auto.enable`). NetApp рекомендует установить опцию `raid.min_spare_count` в значение 2, чтобы позволить системе уведомить администратора, когда требования *Maintenance Center* по количеству *spare disks* перестанут выполняться.

4.4 Защита от «потерянных записей» (Lost Write Protection)

Lost write protection, или «защита от потери записи» это возможность Data ONTAP, которая реализуется процессом при каждом чтении с WAFL®. Данные проверяются с помощью дополнительной информации контрольной суммы блока (в контексте WAFL) и данных RAID parity. Если регистрируется ошибка, то возможно два пути реакции:

- Диск, содержащий данные сбойный.
- Aggregate, содержащий данные, отмечается как неконсистентный.

Если *aggregate* помечен как *неконсистентный* (inconsistent), потребуется использование процедуры *WAFL iron* для возвращения *aggregate* в рабочее состояние. Если диск вышел из строя, то он подлежит обычной процедуре, как в случае любого вышедшего из строя диска в системе.

Ошибка, обнаруженная в результате *lost write protection* это весьма редкое событие в жизни системы хранения. Основная цель этого механизма заключается в обнаружении ситуаций, которые могут вызвать в дальнейшем сложные и трудно исправимые проблемы с данными, и вовремя принять меры по исправлению и поддержанию целостности данных.

4.5 Фоновое сканирование дисков (Background Media Scans)

Фоновое сканирование дисков это диагностический процесс, который непрерывно работает на всех дисках в RAID-группе.

Этот тип сканирования (*media scrub*) используется для обнаружения ошибок чтения непосредственно с магнитных дисков. Цель этого процесса не столько в том, чтобы убедиться в целостности содержимого блока данных, сколько в том, что блок в принципе читается с диска.

Производительность системы снижается в результате работы этого процесса менее чем на 4% в среднем. Это так потому, что этот процесс выполняется, в основном, за счет внутренних возможностей жестких дисков, и не требует заметной нагрузки на CPU или ввод-вывод системы.

4.6 Сканирование RAID Parity (RAID Parity scrubs)

Сканирование *RAID parity scrubs* используется для проверки целостности данных в целом. Для очень активных данных на дисках, выгода от использования *RAID parity scrubs* не слишком высока, так как данные сами по себе читаются достаточно часто, и Data ONTAP может проконтролировать целостность данных и другими способами. Чаще всего *RAID parity scrubs* пригождается для архивных и редко считываемых в обычном рабочем порядке данных, где у Data ONTAP нет возможности вовремя обнаружить проблемы с чтением у блоков содержащих такие данные.

Этот процесс проходит, прочитывая данные целиком. В результате такого чтения данные проверяются с помощью parity на считываемость и целостность. Если обнаруживается

некорректный блок, он помечается как плохой, и данные воссоздаются из информации *parity*, и записываются в новый блок. *RAID scrubs* минимально влияет на обычный рабочий ввод-вывод системы хранения, и можно оценить его влияние в менее 10% в среднем. Для больших объемов данных архивного типа доступа NetApp рекомендует увеличивать частоту и/или длительность работы процесса *RAID parity scrubs*.

Процесс сканирования RAID parity включен по умолчанию, и сканирование запускается на 360 минут (6 часов). Параметр влияния на производительность системы по умолчанию установлено в низкое значение, так что этот процесс использует только свободные ресурсы процессора в состоянии *idle*.

5 Дополнительные аспекты

Кроме уже названных аспектов отказоустойчивости и особенностей проведения восстановительных действий в системе, существуют еще ряд особенностей, которые важно понимать, при конфигурировании вашей системы хранения. Это раздел рассматривает ряд специальных аспектов и факторов, таких, как размер RAID-группы, опции RAID, и наилучшие практики для смешанных конфигураций.

5.1 Отказоустойчивость на уровне дисковой полки

Многие администраторы в прошлом планировали добиться отказоусточивости уровня дисковой полки путем организации структуры RAID-групп таким образом, чтобы на дисках каждой полки находилось не более двух дисков любой RAID-группы. Таким образом, в случае использования RAID-DP, если полка выйдет из строя, ни одна из RAID-групп системы не потеряет более 2 дисков разом, и сохранит данные за счет отказоустойчивости уровня RAID. Это непрактичный и не слишком реалистичный подход к отказоустойчивости. Оцените сами:

- С дисковой полкой DS14 (14 дисков) вы получите семь RAID-групп в состоянии *degraded* в результате отказа полки. Для полок DS4243 или DS2246 вы получите 12 degraded RAID-групп в результате отказа полки.
- По умолчанию Data ONTAP выполняет *RAID reconstruction* для только двух RAID-групп в каждый взятый момент времени. Это приводит к тому, что 5 (DS14) или 10 (DS4243/DS2246) групп в состоянии *degraded* находятся в зоне риска потери данных, пока ожидают ребилда RAID для этих двух групп.
- Большой объем трафика ребилда, вместе со значительным количеством RAID-групп в состоянии *degraded* , значительно ухудшает производительность системы для обработки основного рабочего трафика ввода-вывода.

Специальное решение NetApp для обеспечения отказоустойчивости уровня дисковой полки, называется SyncMirror. SyncMirror защищает систему хранения NetApp от отказов уровня полки, которые, хотя и редки, но все же возможны. Кроме отказоустойчивости на уровне полки SyncMirror позволяет также увеличить производительность на чтение для некоторых конфигураций, где такая производительность оказалась ограничена дисками, на величину до 80%. Отметьте, что в настоящее время *Data ONTAP 8.1 Cluster-Mode* не поддерживает SyncMirror.

5.2 RAID-группы

Конфигурация RAID-группы оказывает значительное влияние на отказоустойчивость системы хранения. NetApp настоятельно рекомендует использовать RAID-DP для всех конфигураций, так

как он обеспечивает наивысшую отказоустойчивость, и позволяет реализовать непрерывающее нормальную работу обновление *firmware* для дисков. Все, что мы будем рассматривать ниже в этой главе предполагает, что вы используете RAID-DP.

Существует соблазн всегда создавать RAID-группы максимально возможной длины, для минимизации расхода на диски parity и максимизации производительности, но в результате мы получаем:

- **Большой «домен отказа»:** Чем больше дисков вы включили в RAID-группу, тем больше вероятность того, что один или более этих дисков выйдут из строя по обычной причине конечности ресурса надежности, исчерпания срока службы и так далее. Надежность дисков есть решающий фактор в понимании риска одновременного выхода из строя группы дисков в пределах одной RAID-группы. В конечном счете, все вычисления есть только приблизительная оценка, так как никто не может гарантировать (или не гарантировать) то, что диски откажут одновременно, это произойдет в одной RAID-группе, и что они вообще откажут (в течение стандартного пятилетнего гарантийного срока).
- **Увеличенное время *RAID reconstruction*:** Чем больше дисков с данными составляет RAID-группу, тем больше вычислений необходимо при восстановлении информации в RAID из *parity* после сбоя диска. Каждый диск данных в группе содержит порцию данных, которую нужно включить в расчет *parity*. Чем больше таких блоков, тем больше операций расчета *parity*, тем больше время на восстановление RAID. По нашим оценкам, в случае RAID-группы размером от 12 до 20, каждый диск увеличивает это время примерно на 6%.

SATA или FC/SAS?

Часто диски SATA (1,2 миллиона часов MTBF) считают менее надежными, чем диски FC и SAS (1,6 миллиона часов MTBF). Данные NetApp, основывающиеся на его данных AFR и ARR говорят, что диски SATA enterprise-класса в реальной жизни не отличаются по уровню надежности от дисков FC и SAS, что может вызвать вопрос: «Почему же тогда максимальный размер RAID-группы для дисков SATA задан меньший, чем для SSD, FC, или SAS?»

Несмотря на то, что уровень надежности для дисков SATA и можно считать аналогичным FC и SAS, нельзя не принимать во внимание их емкость и быстродействие. Диски SATA имеют большую емкость, и они более медленные, чем диски FC/SAS, что означает значительно большее время *RAID reconstruction*, чем у дисков FC/SAS. В версии Data ONTAP 8.0.1 мы увеличили максимальный размер RAID-группы для дисков SATA с 16 до 20. Это решение было принято после анализа работы «боевых» систем и основано на данных о надежности дисков SATA, собранных NetApp (среди прочих факторов). Однако увеличение RAID-группы свыше 20 дисков (от 21 до 28), ведет к рискованной ситуации множественных отказов дисков в одной группе, и потенциально проблеме «непрерывного перестроения RAID» в ходе новых сбоев.

«Непрерывное перестроение RAID» это ситуация, когда процесс RAID reconstruction затягивается настолько, что в течение этого периода возможен выход из строя следующего диска, прежде чем завершится восстановление после предшествующего сбоя. Риск такой ситуации возрастает с увеличением размеров дисков, выходящих на рынок (3TB, 4TB, и больше).

Solid-State Drives (SSD)

Учитывая сравнительно небольшую емкость SSD и значительно лучшую производительность на уровне самого диска, использование больших RAID-групп из дисков SSD влечем за собой значительно меньшие риски. Наши данные показывают, что процесс *RAID reconstruction* для 100GB

SSD занимает менее 20 минут даже на нагруженной системе. Учитывая такое быстрое восстановление, будет вполне разумно ожидать достаточную отказоустойчивость системы в случае использования даже максимально большой RAID-группы, из 28 дисков SSD (в RAID-DP).

5.3 Опции RAID

Рассматривая ситуацию с точки зрения доступности данных, важно понимать то, как вы можете настроить конфигурацию хранилища для того, чтобы соответствовать вашим требованиям доступности данных. Например, для конфигурации архивного хранилища, будет лучше настроить систему таким образом, чтобы позволить операциям *RAID Reconstruction* иметь приоритет перед операциями рабочего ввода-вывода системы, и, тем самым, обеспечить максимально быстрое восстановление, пусть даже за счет оперативности рабочего доступа. Напротив, для конфигурации хранения данных Exchange может быть необходимо, чтобы рабочий ввод-вывод системы имел приоритет на использование системных ресурсов, перед операциями *RAID Reconstruction*, и не замедлял критически важный рабочий ввод-вывод системы даже во время операций восстановления.

Опции RAID это основной доступный для конфигурирования метод указать Data ONTAP как и в каких пропорциях основные, рабочие операции ввода-вывода и операции ввода-вывода, вызванные процессами ребилда (*RAID reconstruction* и *Rapid RAID Recovery*) должны конкурировать за системные ресурсы.

Опция `raid.reconstruct.perf_impact` может быть установлена в `low`, `medium`, или `high`. По умолчанию значение установлено в `medium`. Изменение этого значения приводит к следующему поведению:

- **Low:** Настройка позволяет операциям ребилда и основным операциям ввода-вывода системы совместно использовать и конкурировать за 0% системных ресурсов, при пиковой загрузке. Эта настройка гарантирует, что основной, рабочий ввод-вывод системы может, при необходимости занять все 100% системных ресурсов. Операции ввода-вывода при ребилде RAID и прочих аналогичных действиях используют только ресурсы с уровнем *idle*.
- **Medium:** Настройка позволяет операциям ребилда и основным операциям ввода-вывода системы совместно использовать и конкурировать за 40% системных ресурсов, при пиковой загрузке. Эта настройка гарантирует, что основной, рабочий ввод-вывод системы может, при необходимости, занять 60% системных ресурсов без конкурирования за ресурсы с операциями ребилда RAID.
- **High:** Настройка позволяет операциям ребилда и основным операциям ввода-вывода системы совместно использовать и конкурировать за 90% системных ресурсов, при пиковой загрузке. Эта настройка гарантирует, что основной, рабочий ввод-вывод системы может, при необходимости занять 10% системных ресурсов без конкурирования за ресурсы с операциями ребилда RAID.

Для операций RAID Reconstruction и подобных ей, термин «системные ресурсы» включает в себя:

- Уровень использования CPU
- Полосу пропускания ввода-вывода
- Уровень использование дисков

Нет ограничения на то, сколько CPU и канала ввода вывода находящихся в idle могут задействовать процедуры восстановления; поэтому, 0% в значении опции Low означает, что при

нагруженной рабочими задачами системы выполняться будут только фоновые процессы. Это также значит, что эта опция имеет мало смысла на системе без рабочей нагрузки, так как на ней отсутствует рабочая нагрузка, с которой бы могла конкурировать за ресурсы нагрузка ребилда.

Указанные выше в описании опций процентные соотношения не означают и не гарантируют, что операции ребилда займут именно столько, это означает, что рабочая нагрузка и нагрузка, вызванная операциями ребилда будут совместно использовать и конкурировать за ресурсы в указанных процентных соотношениях. Установка этого значения в *high* при этом не будет означать, что в такой системе вы потеряете 90% обычного объема рабочего ввода-вывода, так как на самом деле фоновые процессы, такие как *RAID reconstruction* и рабочие для системы хранения нагрузки, будут вместе сосуществовать в пределах этой квоты. То есть задачи ребилда *могут* использовать этот процент ввода вывода, но это не значит, что *только они будут* его использовать, напротив, этот процент будет использоваться обеими этими задачами ввода-вывода совместно, а операции ребилда будут использовать этот процент объема ввода-вывода только в том случае, если он им потребуется.

5.4 Политика в отношении Spare Drives

Рекомендации по дискам *spare* варьируются в зависимости от конфигурации и ситуации. В прошлом NetApp основывал свои рекомендации по количеству *spares* исключительно на числе подключенных к системе дисков. Это, безусловно, важный фактор, но не единственный, который стоит принять во внимание. Системы хранения NetApp используются в широком спектре конфигураций.

Это требует определить более гибкий подход к определению числа *spares* для обслуживания вашей конфигурации системы хранения.

В зависимости от требований вашей конфигурации системы хранения, вы можете откорректировать политику выбора количества *spares* в соответствии со следующими принципами:

- Минимальное число *spares*: В конфигурациях, где уровень использования дисковой емкости является ключевым аспектом, вам, вероятнее всего, захочется использовать минимально возможное число *spare-дисков*. Такой вариант позволит вам пережить наиболее распространенные отказы. В случае множественных отказов вам, возможно, понадобится ручное вмешательство.
- Сбалансированное число *spares*: Эта конфигурация является «золотой серединой» между «минимальным» и «максимальным» вариантами. Он рассчитывает, что вы не столкнетесь со сценарием наихудшей ситуации, и обеспечивает достаточное количество дисков для большинства сценариев отказов.
- Максимальное число *spares*: Этот вариант обеспечивает достаточное число *spares* даже для наихудшей ситуации с отказами, которые могут потребовать разом максимально возможного числа резервных дисков. Термин «максимальное» не означает, что система не сможет работать с еще большим, чем указано как рекомендованное, числом *spares*. Вы всегда можете добавить столько дисков *hot spares* в систему, сколько сочтете нужным.

Выбор одного из приведенных подходов, в соответствии с требованиями вашей системы, к числу spares есть рекомендация NetApp и наилучшее решение.

Большинство системных архитекторов, проектирующих хранилище данных, предпочтут *сбалансированный* вариант, однако пользователи, особенно те, для кого потеря их данных является весьма чувствительным ударом, могут склониться и к *максимальному* варианту. Учитывая, что небольшие по объему системы хранения используют небольшое количество дисков, для такой конфигурации может подойти и *минимальный* вариант.

Для конфигураций, использующих RAID-DP, руководствуйтесь Таблицей 1 для определения рекомендованного числа *spare disk* системы.

Таблица 1) Определение рекомендованного числа *spares*.

Рекомендованное число spare drives		
Минимальный	**Сбалансированный**	**Максимальный**
Два на контроллер	Четыре на контроллер	Шесть на контроллер
Особые случаи		
FAS2000 platform	Системы *entry-level*-класса, использующие только установленные в их собственный корпус диски, могут использовать только один hot spare на контроллер (два на HA-кластер, из двух контроллеров).	
RAID-группы	Системы, использующие только одну RAID-группу, не нуждаются в более чем двух *hot spares* на систему (контроллер этой группы).	
Maintenance Center	*Maintenance Center* требует минимум двух дисков *hot spares* в системе (на контроллер).	
Свыше 48 часов на замену	Для систем, расположенных удаленно и с возможными задержками в сервисной доставке, где повышена вероятность множественных отказов, рекомендуется использовать двойное количество *spares* для каждого из вариантов.	
Более 1200 дисков	Для систем, использующих более 1200 дисков, дополнительные два диска *hot spares* следует добавить в каждый из трех названных вариантов.	
Менее 300 дисков	Для систем, использующих менее 300 дисков, вы можете уменьшить рекомендуемое число *spares* для *сбалансированного* и *максимального* вариантов до двух.	

Дополнительные факторы, влияющие на выбор *hot spares*:

- Рекомендации по числу *spares* действуют для каждого типа дисков в системе. Смотрите главу **5.5 Смешанные конфигурации**, для подробностей.
- Диски большей емкости могут работать как *spare* и заменять диски меньшей емкости (их емкость будет обрезана до необходимой величины).
- Более медленные диски, которые замещают более быстрые того же типа будут негативно влиять на производительность RAID-группы и aggregate в целом. Например, если 10K RPM SAS диск (из полки DS2246) заменяет 15K RPM SAS диск (из полки DS4243), это может дать вам неоптимальную конфигурацию.
- Хотя диски FC и SAS эквивалентны с точки зрения производительности, некоторые новые возможности повышения отказоустойчивости в дисковых полках SAS есть только для дисков SAS. По умолчанию Data ONTAP позволяет дискам FC и SAS взаимозаменять друг друга. Это поведение можно изменить, установив опцию RAID `raid.disk.type.enable` в значение on. Смотрите главу **5.5 Смешанные конфигурации**, для подробностей.

Диски «горячей» и «холодной» замены

NetApp не препятствует администраторам использовать так называемые *cold spare* диски. NetApp рекомендует удалять вышедшие из строя диски из системы как можно скорее, и наличие *cold spares* в этом случае может ускорить процедуру замены. Однако диски *cold spares*, то есть не установленные в систему, в дисковую полку, диски, а находящиеся «на складе», не заменяют и не отменяют необходимость наличия в системе дисков *hot spares*, установленных в системе в дисковые полки.

Hot spares, также как и *cold spares*, используются для замены вышедших из строя дисков, но иным образом. Диски cold spares могут использоваться для замены вышедших из строя дисков (тем самым ускоряя процессы замены неисправного диска и возврата его в NetApp), но *hot spares* применяются иначе: для немедленной реакции на отказ диска, путем предоставления места и возможности запустить процедуры *RAID reconstruction* или *Rapid RAID Recovery*. Трудно вообразить себе ситуацию, когда администратор бежит в датацентр, чтобы вставить диск из *cold spare*, когда вышел из строя один из дисков системы. *Cold spares* также имеют риск встретиться с проблемой «умер в момент замены», так как не установленный в дисковую полку диск имеет больше шансов подвергнуться нежелательному физическому воздействию во время хранения или установки. Как пример можно рассмотреть вариант воздействия электростатики, которое может губительным образом сказаться на работе диска при его установке в систему.

Вследствие такой разницы в применении и использовании *cold* и *hot spares*, ни в коем случае не следует рассматривать *cold spares* как замену, позволяющую обойтись без наличия в системе дисков *hot spares*.

Требования по необходимому минимуму spare disks в системе

Опция RAID `raid.min_spare_count` используется для задания минимального количества дисков spares, доступных контроллеру системы хранения. Если вы используете возможности Maintenance Center, следует установить значение этой опции в 2, чтобы система могла уведомить администратора, когда нехватка *hot spares* будет блокировать работу Maintenance Center.

NetApp рекомендует вам установить значение этой опции равное суммарному числу *spares*, которое определено для вашей системы в соответствии с заданной политикой количества *spares* на систему с тем, чтобы система могла уведомить вас в случае, если вы в процессе эксплуатации и создании новых томов и *aggregates* опуститесь ниже ранее определенного числа *spares*.

5.5 Смешанные конфигурации

Возможность делать смешанные (SAS + FC) конфигурации на системах хранения NetApp это значительное преимущество для многих наших пользователей. Целью этой главы является не удержать от использования смешанных конфигураций, а показать, что с изменением или появлением новых технологий появляется необходимость оценивать и пересматривать такие смешанные конфигурации с тем, чтобы обеспечить отсутствие недостатков или просчетов в оценке отказоустойчивости системы. Это не означает, что создавая смешанную конфигурацию хранилища вы априори ставите под удар надежность и отказоустойчивость, так как некоторые из таких конфигураций поддерживаются и обеспечивают тот же уровень отказоустойчивости, что и разделенные, «гомогенные» (состоящие только из дисков одной технологии) конфигурации.

Технологии дисковых полок

Так как NetApp проводит постепенный переход с полок семейства DS14 на полки с подключением по SAS (DS4243 и DS2246), часто можно встретить системы, использующие оба типа полок, и

технологии подключения как FC, так и SAS. Дисковые полки SAS имеют новые возможности в области отказоустойчивости, недоступные для дисковых полок семейства DS14.

Например, так называемый *Alternate Control Path* (ACP) имеется только для дисковых полок SAS. NetApp рекомендует разделять логические дисковые структуры, размещенные на дисковых полках SAS, от таковых на полках FC.

Эквивалентность для системы FC и SAS дисков

Опция `raid.disktype.enable` по умолчанию установлена в `off`. Это означает, что при создании aggregate и выборе spare, Data ONTAP рассматривает диски FC и SAS как одинаковые. Например, диск SAS может заменить отказавший диск FC и наоборот. Хотя с точки зрения производительности диски FC и SAS эквивалентны, есть определенная разница с точки зрения отказоустойчивости, так как дисковые полки, в которых размещаются диски с разным типом интерфейса, отличаются друг от друга. Диски FC устанавливаются в полки типа DS14mk2 и DS14mk4, использующие интерфейсные модули ESH2 и ESH4 соответственно. Диски SAS устанавливаются в полки типа DS4243, использующие интерфейсный модуль IOM3, и полки типа DS2246, использующие модуль IOM6. Полки SAS имеют ряд решений, направленных на повышение надежности, по сравнению с семейством полок DS14. Поэтому, если диск FC из полки DS14 заменяет диск, являвшийся частью RAID-группы, находящейся целиком на полке SAS, общий уровень отказоустойчивости этой RAID-группы снизится до уровня отказоустойчивости DS14.

С точки зрения отказоустойчивости, NetApp рекомендует установить опцию `raid.disktype.enable` в состояние on для того, чтобы указать системе различать диски FC и SAS.

Скорости вращения дисков

С появлением дисковой полки DS2246, у NetApp в дисковой номенклатуре появились диски SAS 10K RPM 2.5”, в то время, как в полке DS4243 используются диски SAS 15K RPM 3.5”. Некоторые диски FC 10K RPM также все еще встречаются в установленных у пользователей системах, хотя они уже и не продаются самим NetApp. Смешивание дисков 10K RPM с 15K RPM в одном общем aggregate ограничивает все диски скоростью работы, соответствующей дискам 10K RPM. Как результат это приводит к увеличению времени *RAID reconstructions*.

NetApp рекомендует не смешивать диски 10K RPM и 15K RPM в одном aggregate.

5.6 Риски системы с точки зрения My Autosupport

Для пользователей, использующих службу My AutoSupport™, имеется новая развернутая секция *Health Summary* , включающая в себя специальную возможность под названием *System Risk Details* (SRD). Секция данных SRD проактивно идентифицирует риски в текущей конфигурации системы хранения NetApp, которые могут негативно влиять на системную производительность, доступность данных и отказоустойчивость.

Каждый элемент этой секции содержит информацию об определенных рисках системы, потенциальных негативных эффектах, и ссылки на описания по их устранению. Упреждающе реагируя на обнаруженные риски, вы можете значительно снизить возможные непредвиденные остановки работы вашей системы хранения NetApp.

NetApp рекомендует использовать SRD для увеличения уровня отказоустойчивости, оценки рисков системы до того, как они приведут к незапланированному прерыванию ее работы. Больше подробностей вы можете найти на сайте NetApp Support (ранее NOW™) в разделе My AutoSupport, расположенной по адресу: https://now.netapp.com/NOW/asuphome.

6 Классы доступности данных

Этот раздел рассматривает требования к отказоустойчивости и соответствующие рекомендации для определенных классов данных и систем хранения этих данных, принятых и сформулированных для данного документа. Большая часть рекомендаций данного документа может быть применена ко всем перечисленным классам доступности данных.

Значения по умолчанию

Важно отметить, что значения по умолчанию не всегда соответствуют наилучшим практикам. В жизни значения по умолчанию назначаются исходя из следующих определяющих их моментов:

- Они соответствуют некоторому среднему значению соответствующей установки, ни «оптимизированному», ни «неоптимизированному».
- Их значения могут быть определены достаточно давно, и с тех пор не пересматриваться для соответствия их текущим практикам использования, решениям и новым возможностям сегодняшнего применения.
- Они могут быть определены для того, чтобы быть примерно подходящими для большинства известных производителю конфигураций.

Учитывая широту и глубину решений хранения данных предлагаемых NetApp, почти невозможно обеспечить, чтобы значения по умолчанию были приведены в соответствие с наилучшими практиками для данного случая применения. Во многих случаях для их установки просто не существует единственного ответа, это означает, что необходимо приложить должную осмотрительность и знания, чтобы надлежащим образом оптимизировать конкретную конфигурацию конкретной системы хранения для данных ваших клиентов.

6.1 Класс 1: *MISSION CRITICAL*

Этот тип соответствует сервисам, которые выполняют задачи, имеющие повышенные требования к надежности и доступности, и прекращение работы которых влечет за собой значительные потери для пользователя. Базы данных с обработкой онлайн-транзакций (online transaction processing, OLTP), некоторые системы виртуализации и облачной инфраструктуры – это только некоторые из возможных систем, которые можно рассматривать как системы задач класса *mission critical*. Этот класс доступности данных настраивается на приоритетность рабочих операций ввода-вывода, чтобы обеспечить работу зависящих от системы хранения приложений. Приоритезация основной рабочей нагрузки ввода-вывода над задачами восстановления RAID, в случае его сбоя и состояния *degraded*, увеличивает время завершения восстановления. Это повышает риск возникновения дополнительного сбоя в системе, до того, как ее работа будет полностью восстановлена после предшествующего сбоя; например, может произойти отказ еще одного диска прежде, чем будет восстановлена работа RAID после отказа ранее вышедшего из строя диска, и операция *RAID reconstruction* еще не завершена.

Таблица 2) Рекомендации и наилучшие практики для класса доступности *mission-critical*.

Mission-Critical	
Flash Cache	Используйте Flash Cache для улучшения производительности системы и минимизации влияния процессов в *degraded mode* на параметры и производительность рабочего ввода-вывода

SyncMirror	Используйте *local SyncMirror* чтобы обеспечить защиту от отказов уровня дисковой полки и улучшить производительность в *degraded mode*. Отметьте, что Data ONTAP 8.1 Cluster-Mode в настоящее время не поддерживает SyncMirror.
Диски Spares	Используйте вариант выбора количества дисков *hot spares* типа *maximum*, чтобы быть уверенным в том, что в системе будет доступно достаточно запасных дисков для любых возможных процедур исправления сбоев. Установите опцию RAID `raid.min_spares_count` на рекомендованное для этого типа значение, чтобы быть уверенным в том, что администратор будет своевременно уведомлен, когда количество дисков станет менее запланированного.
Типы дисков	Используйте диски высокой производительности (SAS, FC, или SSD), вместо дисков высокой емкости (SATA). Диски небольшой емкости со скоростью вращения 15K RPM или же диски SSD обеспечат минимальное время восстановления RAID после сбоя. Это особенно важно, когда операциям ввода-вывода рабочей нагрузки системы установлен приоритет перед операциями восстановления, что может значительно увеличить время восстановления. Производительные диски помогут сократить период восстановления при таких настройках.
Заполненность Aggregates	Наблюдайте за состоянием заполненности aggregate данными томов, так как излишнее заполнение может привести к заметному падению производительности (так как головками считывания дисков на таком aggregate потребуется больший «пробег» для операций ввода-вывода). Отказ диска также может заметнее снизить производительность ввода-вывода если этот диск был близок к заполнению.
Мониторинг загрузки	Наблюдайте за загрузкой CPU, дисков и полосы петель/стеков к полкам. Если загрузка выше 50%, то растет риск получить более заметное падение производительности ввода-вывода при операциях в *degraded mode*. Это также может увеличить сроки выполнения ребилда RAID.
Приоритезация ввода-вывода	Приоритезируйте рабочие операции ввода-вывода системы над операциями *RAID reconstruction* с помощью опции RAID `raid.reconstruct.perf_impact` установленной в *Low*.
Scrubs	Используйте значения по умолчанию для *RAID scrubs* и *media scrubs*. Система, скорее всего, будет работать под высокой нагрузкой, поэтому увеличение длительности операций *scrubs*, скорее всего, будет иметь невысокую полезность для целостности данных, однако займет при этом какое-то количество ценных системных ресурсов.
Maintenance Center	Использование Maintenance Center рекомендуется для того, чтобы обеспечить интеллектуальный контроль и тестирование подозрительных дисков непосредственно в системе хранения. Он также облегчает процесс RMA (возврата неисправных дисков производителю), и обеспечивает возврат системы в рабочее состояние за минимально возможное время.

6.2 Класс 2: *BUSINESS CRITICAL*

Этот класс доступности данных в первую очередь распространяется на данные, хранимые в рамках регулятивных актов и иных подобных требований к бизнесу. Хотя обслуживание доступа пользователей к таким данным на системе хранения является также важным, однако сохранность является приоритетной и потеря данных этого класса доступа будет чревата тяжелыми и пагубными последствиями для бизнеса пользователя.

Никому из пользователей не нравится терять данные. Но в ряде случаев (например для данных, хранение которых определено регуляторными государственными требованиями) потери могут быть сопряжены еще и со значительными и тяжелыми санкциями и штрафами, налагаемыми на бизнес пользователя системы хранения. Это может быть также хранилище для данных, представляющих собой ценную интеллектуальную собственность компании. Медицинские записи и истории болезни пациентов, исходные коды разрабатываемых компанией программ, деловая переписка в системах электронной почты – это примеры информации данного класса доступности. Этот класс настраивается на приоритет корректирующих действий и связанного с ним трафика ввода-вывода и баланс приоритетов с рабочим трафиком ввода-вывода. Повысив приоритет для операций и ввода-вывода *RAID reconstruction* над рабочим трафиком ввода-вывода системы, вы увеличите влияние операций восстановления данных на рабочую производительность системы.

Таблица 3) Рекомендации и наилучшие практики для класса доступности *business-critical*.

Business-Critical	
Flash Cache	Используйте Flash Cache для улучшения производительности системы и минимизации влияния процессов в *degraded mode* на параметры и производительность рабочего ввода-вывода
SyncMirror	Используйте *local SyncMirror* чтобы обеспечить защиту от отказов уровня дисковой полки и улучшить производительность в *degraded mode*. Отметьте, что Data ONTAP 8.1 Cluster-Mode в настоящее время не поддерживает SyncMirror.
Диски Spares	Используйте вариант выбора количества дисков *hot spares* типа *maximum*, чтобы быть уверенным в том, что в системе будет доступно достаточно запасных дисков для любых возможных процедур исправления сбоев. Установите опцию RAID `raid.min_spares_count` на рекомендованное для этого типа значение, чтобы быть уверенным в том, что администратор будет своевременно уведомлен, когда количество дисков станет менее запланированного.
Типы дисков	Используйте диски высокой производительности (SAS, FC, или SSD), вместо дисков высокой емкости (SATA). Диски небольшой емкости со скоростью вращения 15K RPM или же диски SSD обеспечат минимальное время восстановления RAID после сбоя. Это особенно важно, когда операциям ввода-вывода рабочей нагрузки системы установлен приоритет перед операциями восстановления, что может значительно увеличить время восстановления. Производительные диски помогут сократить период восстановления при таких настройках.
Заполненность Aggregates	Наблюдайте за состоянием заполненности aggregate данными томов, так как излишнее заполнение может привести к заметному падению производительности (так как головками считывания дисков на таком aggregate потребуется больший «пробег» для операций ввода-вывода). Отказ диска также может заметнее снизить производительность ввода-вывода если этот диск был близок к заполнению.
Мониторинг загрузки	Наблюдайте за загрузкой CPU, дисков и полосы петель/стеков к полкам. Если загрузка выше 50%, то растет риск получить более заметное падение производительности ввода-вывода при операциях в *degraded mode*. Это также может увеличить сроки выполнения ребилда RAID.
Приоритезация ввода-вывода	Установите опцию RAID `raid.reconstruct.perf_impact` в *Medium* для баланса между рабочими операциями ввода-вывода системы и операциями *RAID reconstruction*.
Scrubs	Рекомендуется увеличить частоту выполнения *RAID scrubs* для повышения уровня контроля целостности данных.

Maintenance Center	Использование Maintenance Center рекомендуется для того, чтобы обеспечить интеллектуальный контроль и тестирование подозрительных дисков непосредственно в системе хранения. Он также облегчает процесс RMA (возврата неисправных дисков производителю), и обеспечивает возврат системы в рабочее состояние за минимально возможное время.

6.3 Класс 3: *REPOSITORY*

Класс хранилища типа *Repository* используется для хранения совместно используемых данных, или пользовательских данных, которые некритичны для выполнения бизнес-операций. Это могут быть научные или инженерные вычисления, документы рабочей группы, пользовательская «домашняя директория», это все примеры данных, попадающих в этот класс хранения и уровня доступности. Этот уровень требует компромисса между рабочими операциями системы хранения и операциями восстановления при сбое (если такие требуются). Установки по умолчанию в системе хранения хорошо соответствуют этому классу хранилища.

Таблица 4) Рекомендации и наилучшие практики для класса доступности *repository*.

Repository	
Flash Cache	Используйте Flash Cache для улучшения производительности системы и минимизации влияния процессов в *degraded mode* на параметры и производительность рабочего ввода-вывода
SyncMirror	Используйте *local SyncMirror* чтобы обеспечить защиту от отказов уровня дисковой полки и улучшить производительность в *degraded mode*. Отметьте, что Data ONTAP 8.1 Cluster-Mode в настоящее время не поддерживает SyncMirror.
Диски Spares	Используйте вариант выбора количества дисков *hot spares* типа *balanced*, чтобы меть возможность использовать больше дисков системы под полезную нагрузку. Установите опцию RAID `raid.min_spares_count` на рекомендованное для этого типа значение, чтобы быть уверенным в том, что администратор будет своевременно уведомлен, когда количество дисков станет менее запланированного.
Типы дисков	Используйте диски SATA (возможно, расположенные «за» Flash Cache) для этого типа конфигурации.
Заполненность Aggregates	Наблюдайте за состоянием заполненности aggregate данными томов, так как излишнее заполнение может привести к заметному падению производительности (так как головками считывания дисков на таком aggregate потребуется больший «пробег» для операций ввода-вывода). Отказ диска также может заметнее снизить производительность ввода-вывода если этот диск был близок к заполнению.
Мониторинг загрузки	Наблюдайте за загрузкой CPU, дисков и полосы петель/стеков к полкам. Если загрузка выше 50%, то растет риск получить более заметное падение производительности ввода-вывода при операциях в *degraded mode*. Это также может увеличить сроки выполнения ребилда RAID.
Приоритезация ввода-вывода	Установите опцию RAID `raid.reconstruct.perf_impact` в *Medium* для баланса между рабочими операциями ввода-вывода системы и операциями *RAID reconstruction*..
Scrubs	Рекомендуется увеличить частоту выполнения *RAID scrubs* для повышения уровня контроля целостности данных..

Maintenance Center	Использование Maintenance Center рекомендуется для того, чтобы обеспечить интеллектуальный контроль и тестирование подозрительных дисков непосредственно в системе хранения. Он также облегчает процесс RMA (возврата неисправных дисков производителю), и обеспечивает возврат системы в рабочее состояние за минимально возможное время.

6.4 Класс 4: *ARCHIVAL*

Этот тип системы характеризуется большими объемами записи информации, в дальнейшем редко считываемой.

Уровень загрузки системы в среднем не ожидается значительным. Так как записанные данные такого типа обычно сравнительно редко считываются, важно настроить систему на контроль и поддержание максимальной целостности записанных данных. Так как приоритетной задачей для системы хранения этого класса данных является сохранение целостности данных, ее конфигурация настраивается на приоритет операций контроля и восстановления, и минимизацию времени ребилда RAID. Примеры данных этого класса это данные резервной копии, архивы, *near-line*, а также эталонные и справочные данные.

Таблица 5) Рекомендации и наилучшие практики для класса доступности *archival*.

Archival	
Диски Spares	Используйте вариант выбора количества дисков *hot spares* типа *maximum*, чтобы быть уверенным в том, что в системе будет доступно достаточно запасных дисков для любых возможных процедур исправления сбоев. Установите опцию RAID `raid.min_spares_count` на рекомендованное для этого типа значение, чтобы быть уверенным в том, что администратор будет своевременно уведомлен, когда количество дисков станет менее запланированного.
Типы дисков	Используйте диски SATA (возможно, расположенные «за» Flash Cache) для этого типа конфигурации.
Заполненность Aggregates	Наблюдайте за состоянием заполненности aggregate данными томов, что может привести к заметному падению производительности (так как головками считывания дисков на таком aggregate потребуется больший «пробег» для операций ввода-вывода). Отказ диска также может заметнее снизить производительность ввода-вывода если этот диск был близок к заполнению..
Мониторинг загрузки	Наблюдайте за загрузкой CPU, дисков и полосы петель/стеков к полкам. Если загрузка выше 50%, то растет риск получить более заметное падение производительности ввода-вывода при операциях в *degraded mode*. Это также может увеличить сроки выполнения ребилда RAID.
Приоритезация ввода-вывода	Используйте значение по умолчанию, равное *Medium* для опции RAID `raid.reconstruct.perf_impact` для обеспечения баланса между основным рабочим вводом-выводом и операциями восстановления RAID.
Scrubs	Рекомендуется увеличить длительность выполнения *RAID scrub* (задается опцией `raid.scrub.duration`) для того, чтобы обеспечить целостность данных. Рекомендуется увеличить величину *media scrub rate* (задается опцией `raid.media_scrub.rate`) для увеличения целостности блоков уровня диска.

Maintenance Center	Использование Maintenance Center рекомендуется для того, чтобы обеспечить интеллектуальный контроль и тестирование подозрительных дисков непосредственно в системе хранения. Он также облегчает процесс RMA (возврата неисправных дисков производителю), и обеспечивает возврат системы в рабочее состояние за минимально возможное время.

6.5 Класс 5: *MULTIPURPOSE*

Одна из множества сильных сторон систем хранения NetApp заключается в возможности хранить на одной системе (HA-кластере) различные классы данных, с различными требованиями по доступности. Однако соседство на одной системе таких различных уровней может вести к конфликту рекомендаций по конфигурированию. В *Data ONTAP 7G*, использующейся под многоцелевое (*multipurpose*) хранилище, aggregates и тома (volumes) обычно одновременно используют системные ресурсы. Для глобальных опций настройки NetApp рекомендует устанавливать их в соответствии с рекомендациями, соответствующими наиболее критичным для вашего бизнеса типам и классам данных, из размещенных на системе хранения. [С появлением *Data ONTAP 8.0 Cluster-Mode*, возможность разделять то, как логические конфигурации используют системные ресурсы, была значительно усовершенствована.]

Например, опция RAID `raid.reconstruct.perf_impact` может быть установлена в значение *Low* или *High*, если ваша система хранения хранит как данные класса *mission-critical*, так и класса *archival*. Но так как данные *mission-critical* более чувствительны и более важные, чем данные архивов и резервных копий класса *archival*, то рекомендуется установить значение вышеприведенной опции в *Low*, в соответствии с рекомендацией для класса 1 (*Mission-Critical*).

Таблица 6) Рекомендации и наилучшие практики для класса доступности *multipurpose*.

Multipurpose	
Рекомендации по приоритезации	В случае конфликта рекомендаций между классами выбирайте рекомендацию для более высокого класса доступности.
FlexShare®	Рекомендуется использовать *FlexShare* для приоритезации уровня использования системных ресурсов между томами данных системы хранения.
Физическое разделение	Разделяйте полки с дисками разных технологий и структуры дисков под разные классы хранений. Например, если у вас есть и диски SAS и SATA (дисковые полки DS4243), подключенные к одной и той же системе, вы можете использовать диски SAS для размещения данных класса *mission-critical*, а диски SATA для данных *archival*. Хотя вы и можете смешивать дисковые полки DS4243 с дисками SAS с дисковыми полками DS4243 с дисками SATA в одном стеке полок, NetApp рекомендует разделять полки с разными типам дисков в разные стеки, например, чтобы отказ на физическом уровне, случившийся на одном из классов хранения и доступности, не влиял в результате напрямую на оба класса (в данном примере).