“Тихое” повреждение данных: обзор средств защиты

Обзор состояния проблемы неотслеживаемых повреждений данных (Silent Data Corruption). Приведены также предлагаемые для нее решения, выпущенные в рамках SNIA-стандарта – Data Integrity from Application to Storage, – в обеспечение целостности данных от порчи при передаче и хранении.

Сергей Платонов – менеджер по продуктам, компания AVRORAID.

Введение

Избыточность компонент, таких как жесткие диски, контроллеры, интерфейсы в дисковых массивах не гарантирует обеспечения корректности данных. Наряду со сбоями, которые отслеживаются микропрошивками аппаратной платформы, возникают так называемые неотслеживаемые повреждения данных (Silent Data Corruption – SDC). Подобные повреждения, возникнув однажды, могут оставаться незамеченными в течение многих месяцев, попадая в резервные копии и делая восстановление практически невозможным или крайне дорогим, при этом увеличивая время простоя до нескольких дней. Повреждения данных могут возникать как в момент передачи, так и во время нахождения на носителях. Частота возникновения подобных ошибок невелика (особенно при использовании жестких дисков с интерфейсом SAS) в сравнении с другими типами отказов, но стоимость восстановления данных при возникновении подобного типа отказа крайне высока.

Результаты исследований

Исследования компании NetApp совместно с университетами Винсконсина и Торонто показали следующее: из 1,5 млн проверенных жестких дисков – 8,5% устройств с интерфейсом SATA были подвержены ошибке. При этом 13% ошибок не были обнаружены при сканировании данных на соответствие контрольным суммам. Ошибок на дисках FC было обнаружено значительно меньше – 0,065% – от общего количества дисков. Подобные исследования были проведены CERN.

Исследования проблемы SDC были выполнены и в лаборатории компании AVRORAID. В исследованиях были использованы жесткие диски с интерфейсами SATA (в том числе модели, предназначенные для рабочих станций), SAS и nearline-SAS. Также проводились исследования с твердотельными накопителями с интерфейсом SATA, выполненных по технологии SLC.

В тестировании использовались следующие модели дисков:

- *SAS:* Seagate Cheetah 15k.5, Seagate Cheetah 15k.7, Hitachi Ultrastar;
- *SATA:* Seagate Constellation ES, Seagate Barracuda ES.2, Seagate Barracuda 7200.11, Western Digital RE3;
- *NL-SAS:* Toshiba MK1001TRKB, Seagate Constellation ES.2;
- *SSD с интерфейсом SATA:* Intel X25-E.

По завершении 12 недель тестирования было обнаружено, что SDC возникли на 0,5% дисков SATA и не были обнаружены на дисках SAS и nearline-SAS.

Наибольшее относительное количество SDC было обнаружено на твердотельных накопителях. Также большое количество SDC было обнаружено на жестких дисках класса desktop.

Решение проблемы. Вчера и сегодня

Развивая продукты, производители добавляли различные технологии, позволяющие гарантировать корректность данных на устройствах хранения. Производители СУБД ввели контрольные суммы, защиту RAM-памяти кодами коррекции ошибок (ECC), на шинах и в сетях был использован циклический избыточный код (CRC). Производители СХД также используют в своих решениях различные проприетарные технологии.

Ошибки, приводящие к порче данных, могут возникать на различных участках системы “приложение-СХД”:

- в самом приложении или ОС;
- в памяти хоста;
- в HBA хоста;
- в сети хранения данных;
- в памяти системы хранения данных;
- на участке backend СХД – жесткие диски;
- на жестких дисках.

Ошибки на жестких дисках и в памяти проявляются наиболее часто, и производители уделяют им максимальное внимание, но его оказалось недостаточно.

Выделяют следующие типы порчи данных:

- *непопадание* – данные были записаны или считаны с некорректного места на носители;
- *изменение данных* – данные были изменены на одном из участков;
- *потерянная операция* – операция записи не была выполнена, но было получено подтверждение. На практике автора публикации данная ошибка чаще всего происходила с SSD-носителями.

Данные могут быть повреждены во время чтения, записи или во время нахождения данных на конечном носителе.

Различают два способа борьбы с неотслеживаемым повреждением данных:

- определение с последующим исправлением или без него;
- предотвращение (также называемое “раннее определение”).

Производители программного и аппаратного обеспечения озабочены предоставлением сервиса, гарантирующего консистентность хранящихся данных.

Первой целевой программой, направленной на обеспечение целостности данных, стала инициатива HADR (Hardware Assisted Resilient Data Initiative), созданная компанией Oracle. Она была поддержана многими лидерами рынка систем хранения данных: EMC, NetApp, HP, NEC, Fujitsu, LSI и др. Участники программы могли использовать специализированные проверяющие данные алгоритмы, разработанные Oracle, перед выполнением записи на свои устройства данных полученных от СУБД. В случае обнаружения ошибки запрос на запись может быть отклонен или выполнен, но с записью информации об ошибке в журнале СХД.

Производители систем хранения данных могут выбирать все либо определенные проверки. Технология была впервые применена в Oracle 9i.

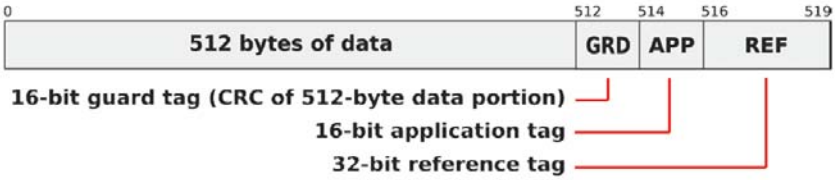

Рис. 1. Структура PI для стандартного 512-байтного блока.

В 2007 г. компании Emulex, LSI, Oracle и Seagate с целью создания решения, обеспечивающего контроль целостности данных "от двери до двери", основали Data Integrity Initiative (DII). Затем была организована рабочая группа SNIA Data Integrity Working Group (DITWG), куда вошли другие компании, заинтересованные в разработке комплексного решения. Вместе с тем, Oracle объявила о включении кода инфраструктуры целостности данных блочной подсистемы ввода-вывода в ядро Linux. Сейчас эту функциональность можно встретить в дистрибутиве Oracle Linux.

Сегодня существуют две ключевые технологии, относящиеся к инициативе DII:

– T10 Protection Information;

– Data Integrity Extentions.

Protection Information

Protection Information (PI) представляет собой набор дополнительных полей в блоке SCISI-команды. PI применяется при взаимодействии между HBA хоста и СХД, а также между бэкэнд-интерфейсом СХД и жесткими дисками. Модель PI, одобренная t10, определяет дополнительные 8 байт информации к стандартному 512b сектору или 164 байта к 4kb сектору. В этих полях содержится информация позволяющая выполнить валидацию данных, находящихся в стандартных 512 килобайтах. При записи данных СХД, поддерживающая данную модель, выполняет проверку данных перед подтверждением приема с использованием метаданных. В процессе чтения метаданные проверяются на стороне HBA-хоста.

При использовании проприетарных технологий корректность данных не может быть проверена до того, как они будут прочитаны. Пример структуры PI для стандартного 512-байтного блока приведен на рис. 1.

Блок *Guard Tag* защищает от порчи данных и представляет собой CRC-сумму для данных, находящихся в основных 512-килобайтах сектора. Размер этого поля составляет 2 байта (16 бит). При записи контроллер устройства проверяет, что CRC-сумма для данных корректна, и только затем выполняет конечную запись блока. В противном случае команда записи отклоняется. При чтении хост проверяет сумму и сверяет ее с данными.

Application Tag – двухбайтовое поле, которое используется приложением для хранения дополнительной информации. Данные в этом поле создаются и проверяются приложением, например, RAID-контроллеры связывают данные с конфигурацией массива. Проверка на уровне диска выполняется опционально.

Reference Tag – четырехбайтовое поле, содержащее информацию об адресе блока. В реальной жизни данные проходят через несколько промежуточных уровней, прежде чем будут записаны на носитель. Это может привести к тому, что данные будут получены в некорректной последовательности или с неверным адресом. Информация в данном поле позволит выполнить проверку корректно сети адреса и последовательности.

Сейчас очень небольшое число производителей систем хранения данных поддерживают стандарт T10 PI на уровне HBA-СХД. По-прежнему основное внимание уделяется борьбе с возникновением неотслеживаемых ошибок на жестких дисках.

Лидирующие производители жестких дисков, такие как Seagate, добавили в свои модели SAS-дисков поддержку PI для обеспечения целостности данных. При использовании подобных жестких дисков они должны быть отформатированы с поддержкой PI, с указанием одного из трех возможных режимов.

Data Integrity Extensions

Data Integrity Extensions (DIX) представляет собой набор требований по взаимодействию на уровне приложение-контроллер для обеспечения целостности данных. Данная технология создана как дополнительная к PI для обеспечения гарантии целостности данных "от двери до двери".

Для обеспечения целостности данных в DIX используется передача дополнительной информации для защиты данных в память и из памяти хоста. Буферы данных и метаданных разделены. Scatter-gather список для метаданных используется отдельный. Для контроллера используется набор дополнительных команд, которые указывают ему, как обрабатывать команды ввода-вывода.

Данная технологи никак не влияет на существующую архитектуру SCSI.

Сейчас основные работы, связанные с технологией DIX, ведутся в области разработки API для приложений.

На рис. 2 представлено сравнение технологий защиты данных от повреждений.

Интерфейс SATA и SDC

Указанные технологии можно применять только для устройств и дисков, использующих стандарт SCSI.

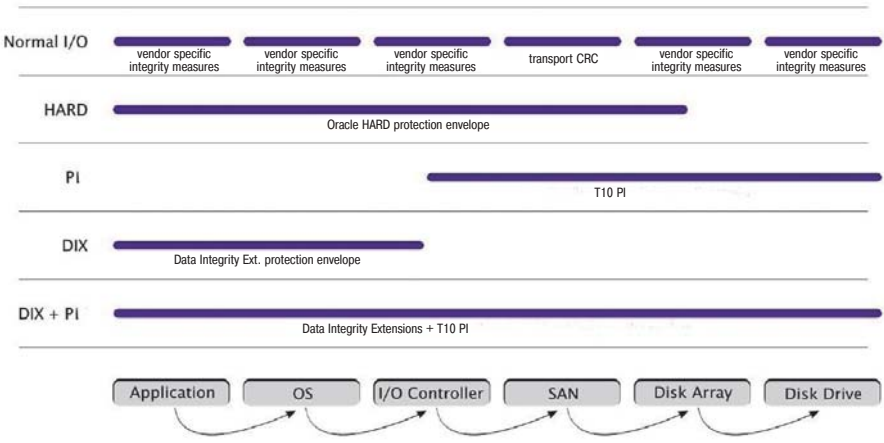

Рис. 2. Сравнение технологий защиты данных от повреждений.

Жесткие диски, оснащенные интерфейсом SATA, обладают большими возможностями по обеспечению защиты данных, чем параллельная шина ATA, обеспечивая проверку данных при передаче по шине. Однако этого не достаточно. SATA не предоставляет возможности применения дополнительных метаданных к каждому блоку. Вендоры используют дополнительные проприетарные технологии для обеспечения целостности данных на SATA-дисках, которые по-прежнему очень актуальны.

На текущий момент существует четыре основные технологии защиты от SDC в массивах, использующие жесткие диски с интерфейсом SATA:

1) запись дополнительных метаданных PI в соседний блок. Эта технология является одной из самой популярной. Но она несет следующие недостатки: значительная потеря полезного пространства и значительная потеря производительности из-за проблем с выравниванием блока;

2) группировка в кластеры по 64 (или 128) блока и использование 65 блока для PI. При использовании данной технологии наблюдается лучшая утилизация пространства, но минимальной единицей обращения с диском является кластер, что ограничивает возможности пользователя при выборе размера страйпа, а также накладывает дополнительные проблемы производительности при работе на случайных пантеонах доступа;

3) перепрочтение сектора после записи. Запись считается завершенной только после того, как данные будут снова прочитаны с сектора и сверены с оригиналом. При использовании данной технологии обслуживаются наиболее часто возникающие, но не все случаи SDC. Не страдает производительность чтения, ошибки обнаруживаются уже во время записи. Производительность записи может быть сильно снижена;

4) использование синдромов массива для проверки целостности данных. Данная технология имеет преимущества, заключающиеся в том, что не используются никакие дополнительные метаданные для проверки. Главными недостатками технологии являются ограничения по применяемым уровням массивов. Также SDC не может быть исправлена в случае, если один из дисков в массиве уровня 6 вышел из строя. Для определения SDC приходится выполнять перепрочтение всего страйпа и пересчет синдромов, что может привести к серьезным накладным расходам. При обнаружении SDC приходится выполнять восстановление страйпа, что у некоторых производителей требует выполнения восстановления всего жесткого диска, на котором была обнаружена SDC.

Для возможности определения и восстановления SDCна массиве RAID5, а также RAID6 с вышедшим из строя жестким диском, иногда используют дополнительные контрольные блоки, представляющие собой сумму не по страйпу, а по набору стрипов на каждом диске – так называемые “горизонтальные parity”.

Компанией AVRORAID была разработана технология, позволяющая с использованием 3-х синдромов массива избежать перечисленных недостатков. Эта технология будет включена в очередной релиз и будет доступна в системах хранения данных высокой производительности на платформе AVRORA 2.0.

Заключение

Технологии, обеспечивающие защиту данных "от двери до двери", по-прежнему актуальны и сегодня. Частичное применение решений по обеспечению защиты данных от возникновения неотслеживаемых повреждений позволило производителям значительно сократить возникновение подобных отказов и стандартизировать подходы к защите. Но для получения законченного решения рабочей группе предстоит выполнить еще много работы и, главное, привлечь на свою сторону производителей прикладного и системного программного обеспечения.

Мы рекомендуем внимательно отнестись к используемым для защиты данных технологиям. SATA имеет гораздо меньшие возможности по обеспечению корректности данных. При этом средняя стоимость дисков NL-SAS не превышает стоимость HDD с интерфейсом SATA более чем на 10%.

Сергей Платонов,
компания AVRORAID