Общий процесс отказа

ZFS — не первый компонент в системе, который узнает о сбое диска. Когда диск выходит из строя, становится недоступным или имеет функциональную проблему, происходит следующий общий порядок событий:

1. Неисправный диск обнаруживается и регистрируется FMA.
2. Диск удаляется операционной системой.
3. ZFS видит измененное состояние и реагирует, выдавая ошибку устройства.

Состояние устройства ZFS (и виртуального устройства)

Общее состояние пула, как сообщает `zpool status`, определяется совокупным состоянием всех устройств в пуле. Вот несколько определений, которые помогут внести ясность в этот документ.

НЕ В СЕТИ

Только устройства нижнего уровня (диски) могут быть OFFLINE. Это ручное административное состояние, и исправные диски можно вернуть в оперативный режим и активировать в пуле.

НЕДОСТУПНО

Рассматриваемое устройство (или VDEV) не может быть открыто. Если VDEV имеет значение UNAVAIL, пул будет недоступен или не сможет быть импортирован. НЕДОСТУПНЫЕ устройства также могут сообщать о НЕИСПРАВНОСТИ в некоторых сценариях. С точки зрения эксплуатации UNAVAIL диски примерно эквивалентны НЕИСПРАВНЫМ дискам.

ДЕГРАДАЦИЯ

Произошла ошибка в устройстве, затронувшая все VDEV над ним. Пул по-прежнему работает, но в VDEV может быть потеряна избыточность.

УДАЛЕННЫЙ

Устройство было физически удалено во время работы системы. Обнаружение удаления устройства зависит от оборудования и может поддерживаться не на всех платформах.

НЕИСПРАВНОСТЬ

Все компоненты (верхние и резервные VDEV и диски) пула могут находиться в состоянии FAULTED. НЕИСПРАВНЫЙ компонент полностью недоступен. Серьезность ДЕГРАДИРОВАНИЯ устройства во многом зависит от того, какое это устройство.

В ИСПОЛЬЗОВАНИИ

Этот статус зарезервирован для запасных частей, которые использовались для замены неисправного привода.

Общий обзор замены дисков

На высоком уровне замена конкретного неисправного диска состоит из следующих шагов:

1. Определить FAULTED или UNAVAILABLE диск
2. zpool replace привод, о котором идет речь
3. Подождите, пока resilver закончится
4. zpool remove замененный диск
5. zpool offline удаленный диск
6. Выполните любую необходимую очистку

Эти шаги могут несколько различаться в зависимости от конкретного уровня резервирования и конфигурации оборудования.

Подробные шаги по замене диска

Давайте начнем с примера сценария, включающего несколько неисправных и деградировавших дисков:

```
[root@headnode (dc-example-1) ~]# zpool status
  pool: zones
 state: DEGRADED
status: One or more devices are faulted in response to persistent errors.
        Sufficient replicas exist for the pool to continue functioning in a
        degraded state.
action: Replace the faulted device, or use 'zpool clear' to mark the device
        repaired.
  scan: resilvered 7.64G in 0h6m with 0 errors on Fri May 26 10:45:56 2017
config:

        NAME              STATE     READ WRITE CKSUM
        zones             DEGRADED     0     0     0
          mirror-0        ONLINE       0     0     0
            c1t0d0        ONLINE       0     0     0
            c1t1d0        ONLINE       0     0     0
          mirror-1        DEGRADED     0     0     0
            c1t2d0        ONLINE       0     0     0
            c1t3d0        FAULTED      0     0     0  external device fault
          mirror-2        ONLINE       0     0     0
            c1t4d0        ONLINE       0     0     0
            c1t5d0        ONLINE       0     0     0
          mirror-3        DEGRADED     0     0     0
            1173487       UNAVAIL      0     0     0  was /dev/dsk/c1t16d0
            c1t6d0       ONLINE       0     0     0
          mirror-4        ONLINE       0     0     0
            c1t7d0        ONLINE       0     0     0
            c1t8d0        ONLINE       0     0     0
          mirror-5        DEGRADED     0     0     0
            spare-0       DEGRADED     0     0     0
              c1t10d0     REMOVED      0     0     0
              c1t11d0     ONLINE       0     0     0
            c1t9d0        FAULTED      0     0     0  external device fault
          mirror-6        ONLINE       0     0     0
            c1t12d0       ONLINE       0     0     0
            c1t13d0       ONLINE       0     0     0
        logs
          c1t14d0         ONLINE       0     0     0
        spares
          c1t15d0         INUSE     currently in use
      c1t16d0     ONLINE    0   0   0

errors: No known data errors
```

В приведенном выше примере есть два неисправных устройства и одно недоступное. С административной точки зрения эти два состояния функционально идентичны: вы хотите заменить их известными рабочими дисками.
ZFS будет знать, когда диск достигнет предела количества ошибок, и автоматически исключит его из пула. Это может произойти при любом типе отказа.

Определите физическое местонахождение НЕИСПРАВНОГО или НЕДОСТУПНОГО диска.

Используйте smartctl -d ata -a /dev/sdd для получения этой информации.

```
Extended self-test routine
recommended polling time:        ( 120) minutes.
Conveyance self-test routine
recommended polling time:        (   5) minutes.
SCT capabilities:              (0x303f) SCT Status supported.
                                        SCT Error Recovery Control supported.
                                        SCT Feature Control supported.
                                        SCT Data Table supported.

SMART Attributes Data Structure revision number: 16
Vendor Specific SMART Attributes with Thresholds:
ID# ATTRIBUTE_NAME          FLAG     VALUE WORST THRESH TYPE      UPDATED  WHEN_FAILED RAW_VALUE
  1 Raw_Read_Error_Rate     0x002f   200   200   051    Pre-fail  Always       -       369
  3 Spin_Up_Time            0x0027   141   138   021    Pre-fail  Always       -       3908
  4 Start_Stop_Count        0x0032   098   098   000    Old_age   Always       -       2530
  5 Reallocated_Sector_Ct   0x0033   200   200   140    Pre-fail  Always       -       0
  7 Seek_Error_Rate         0x002e   100   253   000    Old_age   Always       -       0
  9 Power_On_Hours          0x0032   081   081   000    Old_age   Always       -       13972
 10 Spin_Retry_Count        0x0032   100   100   000    Old_age   Always       -       0
 11 Calibration_Retry_Count 0x0032   100   100   000    Old_age   Always       -       0
 12 Power_Cycle_Count       0x0032   098   098   000    Old_age   Always       -       2136
192 Power-Off_Retract_Count 0x0032   200   200   000    Old_age   Always       -       285
193 Load_Cycle_Count        0x0032   086   086   000    Old_age   Always       -       343594
194 Temperature_Celsius     0x0022   101   086   000    Old_age   Always       -       42
196 Reallocated_Event_Count 0x0032   200   200   000    Old_age   Always       -       0
197 Current_Pending_Sector  0x0032   200   200   000    Old_age   Always       -       0
198 Offline_Uncorrectable   0x0030   200   200   000    Old_age   Offline      -       0
199 UDMA_CRC_Error_Count    0x0032   200   200   000    Old_age   Always       -       31447
200 Multi_Zone_Error_Rate   0x0008   200   200   000    Old_age   Offline      -       4

SMART Error Log Version: 1
No Errors Logged

SMART Self-test log structure revision number 1
Num  Test_Description    Status                  Remaining  LifeTime(hours)  LBA_of_first_error
# 1  Short offline       Completed without error       00%     13971         -

SMART Selective self-test log data structure revision number 1
 SPAN  MIN_LBA  MAX_LBA  CURRENT_TEST_STATUS
    1        0        0  Not_testing
    2        0        0  Not_testing
    3        0        0  Not_testing
    4        0        0  Not_testing
    5        0        0  Not_testing
Selective self-test flags (0x0):
  After scanning selected spans, do NOT read-scan remainder of disk.
If Selective self-test is pending on power-up, resume after 0 minute delay.
```

Используйте **zpool replace -f storage 4969025571654608094 /dev/disk/by-id/ata-WDC_WD5002ABYS-02B1B0_WD-WCASY7572574** для замена диска на запасной. Id диска можно посмотреть с помощью **zpool status**

```
root@[redacted]:/home/sa# zpool replace -f storage 4969025571654608094 /dev/disk/by-id/ata-WDC_WD5002ABYS-02B1B0_WD-WCASY7572574
root@[redacted]:/home/sa# zpool status
  pool: storage
 state: DEGRADED
status: One or more devices is currently being resilvered.  The pool will
        continue to function, possibly in a degraded state.
action: Wait for the resilver to complete.
  scan: resilver in progress since Sun Apr 23 17:04:41 2023
        39.1G scanned at 1.26G/s, 9.83M issued at 325K/s, 126G total
        0B resilvered, 0.01% done, no estimated completion time
config:

        NAME                                        STATE     READ WRITE CKSUM
        storage                                     DEGRADED     0     0     0
          mirror-0                                  DEGRADED     0     0     0
            ata-WDC_WD5000AADS-00S9B0_WD-WCAV93552277    ONLINE       0     0     0
            replacing-1                             DEGRADED     0     0     0
              4969025571654608094                   UNAVAIL      0     0     0  was /dev/disk/by-id/ata-WDC_WD5000AAKS-00YGA0_WD-WCAS80551974-part1
              ata-WDC_WD5002ABYS-02B1B0_WD-WCASY7572574  ONLINE       0     0     0

errors: No known data errors
root@[redacted]:/home/sa# zpool status
  pool: storage
 state: DEGRADED
status: One or more devices is currently being resilvered.  The pool will
        continue to function, possibly in a degraded state.
action: Wait for the resilver to complete.
  scan: resilver in progress since Sun Apr 23 17:04:41 2023
        126G scanned at 51.4M/s, 56.7G issued at 23.1M/s, 126G total
        57.6G resilvered, 44.86% done, 00:51:28 to go
config:

        NAME                                        STATE     READ WRITE CKSUM
        storage                                     DEGRADED     0     0     0
          mirror-0                                  DEGRADED     0     0     0
            ata-WDC_WD5000AADS-00S9B0_WD-WCAV93552277    ONLINE       0     0     0
            replacing-1                             DEGRADED     0     0     0
              4969025571654608094                   UNAVAIL      0     0     0  was /dev/disk/by-id/ata-WDC_WD5000AAKS-00YGA0_WD-WCAS80551974-part1
              ata-WDC_WD5002ABYS-02B1B0_WD-WCASY7572574  ONLINE       0     0     0  (resilvering)

errors: No known data errors
```

Проверка состояния жесткого диска в Linux

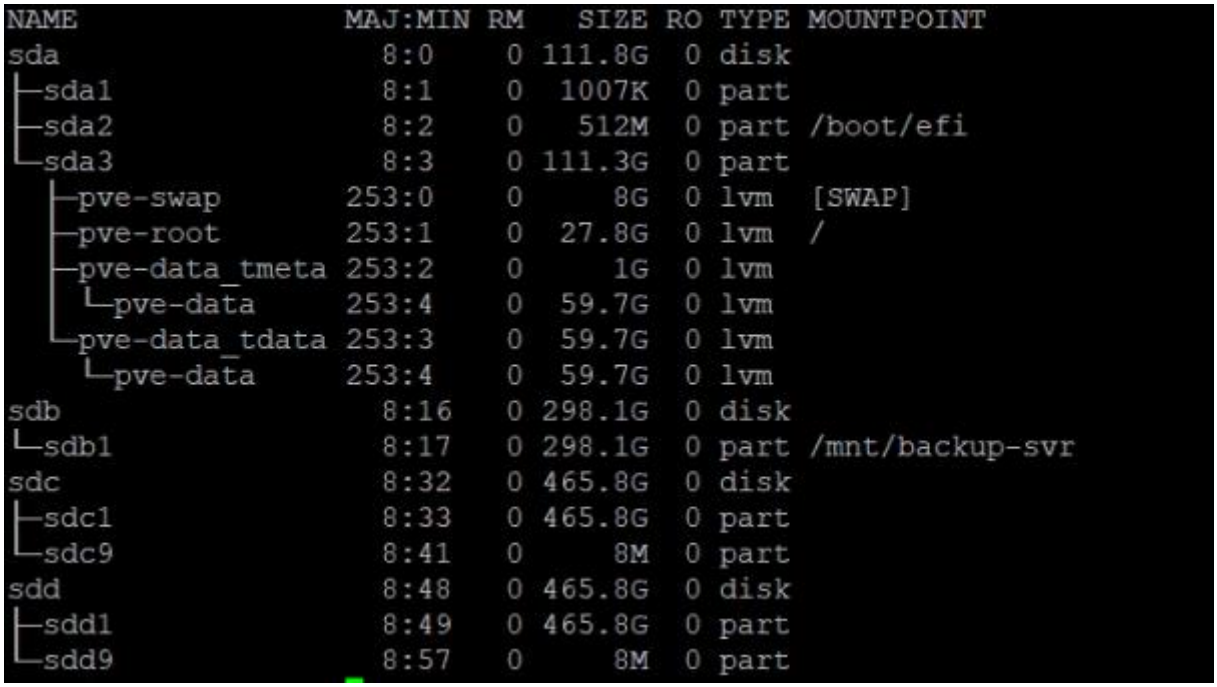
Инструмент, который мы собираемся использовать, называется smartmontools (который также доступен для Windows и OS X). Пакет smartmontools содержит две служебные программы (smartctl и smartd) для управления и мониторинга систем хранения с использованием технологии самоконтроля, анализа и отчетности (SMART), встроенной в большинство современных дисков ATA / SATA, SCSI / SAS и NVMe. Во многих случаях эти утилиты предоставляют предварительное предупреждение о деградации и сбое диска. Smartmontools был первоначально получен из пакета Linux smartsuite и фактически поддерживает диски ATA / ATAPI / SATA-3–8, а также диски SCSI и ленточные устройства.

Установка smartmontools

Для пользователей debian sudo apt install smartmontools Для пользователей Arch: sudo pacman -S smartmontools Вообще говоря, smartmontools доступен в большинстве дистрибутивов, просто установите с вашим менеджером пакетов, используя имя пакета «smartmontools».

Как сделать

После того, как он будет установлен, нам нужно выяснить, какой диск у нас сомнительный: sudo fdisk –l или lsblk



Как только мы узнаем диск, который хотим проверить, мы можем запустить три теста, в зависимости от того, насколько вы обеспокоены:

- Короткий тест, обычно достаточный для выявления проблем **(sudo smartctl -t short / dev / sdX)**
- Более длительный тест, если вас больше интересует, исследует всю поверхность диска **(sudo smartctl -t long / dev / sdX)**
- Испытание при транспортировке, которое используется для проверки наличия повреждений во время транспортировки устройства от производителя. **(sudo smartctl -t transport / dev / sdX)**

Следующий шаг — выяснить, какие типы тестов поддерживает наш диск, а также оценить, сколько времени потребуется на выполнение тестов.

sudo smartctl -c / dev / sdX (замените X соответствующей буквой)

Вам будет предоставлен большой объем вывода, как показано на этом снимке экрана. Тут видно, что это на этом диске короткий тест занимает 2 минуты, более длительный тест занимает 112 минут и испытание при транспортировке занимает 5 минут

```
=== START OF READ SMART DATA SECTION ===
General SMART Values:
Offline data collection status:  (0x82) Offline data collection activity
                                        was completed without error.
                                        Auto Offline Data Collection: Enabled.
Self-test execution status:      (   0) The previous self-test routine completed
                                        without error or no self-test has ever
                                        been run.
Total time to complete Offline
data collection:                ( 9480) seconds.
Offline data collection
capabilities:                    (0x7b) SMART execute Offline immediate.
                                        Auto Offline data collection on/off support.
                                        Suspend Offline collection upon new
                                        command.
                                        Offline surface scan supported.
                                        Self-test supported.
                                        Conveyance Self-test supported.
                                        Selective Self-test supported.
SMART capabilities:            (0x0003) Saves SMART data before entering
                                        power-saving mode.
                                        Supports SMART auto save timer.
Error logging capability:        (0x01) Error logging supported.
                                        General Purpose Logging supported.
Short self-test routine
recommended polling time:        (   2) minutes.
Extended self-test routine
recommended polling time:        ( 112) minutes.
Conveyance self-test routine
recommended polling time:        (   5) minutes.
SCT capabilities:              (0x303f) SCT Status supported.
                                        SCT Error Recovery Control supported.
                                        SCT Feature Control supported.
                                        SCT Data Table supported.
```

Заметка *:* Вы не получите никаких результатов прокрутки для вашего теста, кроме того, что вам будет указано, сколько времени займет тест. Если вы проводите длительный тест, вам, возможно, придется подождать час или два или дольше.

```
[redacted] # smartctl -t short /dev/sdc
smartctl 7.2 2020-12-30 r5155 [x86_64-linux-5.15.30-2-pve] (local build)
Copyright (C) 2002-20, Bruce Allen, Christian Franke, www.smartmontools.org

=== START OF OFFLINE IMMEDIATE AND SELF-TEST SECTION ===
Sending command: "Execute SMART Short self-test routine immediately in off-line mode".
Drive command "Execute SMART Short self-test routine immediately in off-line mode" successful.
Testing has begun.
Please wait 2 minutes for test to complete.
Test will complete after Wed May 10 14:46:23 2023 MSK
Use smartctl -X to abort test.
```

Как только тест закончен, самое время узнать результат! sudo smartctl -H / dev / sdX

```
[redacted]:~# smartctl -H /dev/sdc
smartctl 7.2 2020-12-30 r5155 [x86_64-linux-5.15.30-2-pve] (local build)
Copyright (C) 2002-20, Bruce Allen, Christian Franke, www.smartmontools.org

=== START OF READ SMART DATA SECTION ===
SMART overall-health self-assessment test result: PASSED
```