

Замена сбойного диска в массивах ZFS

 interface31.ru/tech_it/2024/08/zamena-sboynogo-diska-v-massivah-zfs.html

ZFS все чаще применяется в системах хранения Linux благодаря своим широким возможностям и отличной надежности. Но очень часто пользователи не имеют практических навыков работы с этой файловой системой, отдавая работу с ней на откуп вышестоящим системам, например, системе виртуализации Proxmox. Первые сложности начинаются когда пользователь сталкивается с необходимостью обслуживания ZFS и не находит для этого графических инструментов. Одна из таких задач - это замена сбойного диска в массиве, задача серьезная и ответственная, но в тоже время простая. В этой статье мы расскажем как это сделать.



Онлайн-курс по устройству компьютерных сетей

На углубленном курсе "Архитектура современных компьютерных сетей" вы с нуля научитесь работать с Wireshark и «под микроскопом» изучите работу сетевых протоколов. На протяжении курса надо будет выполнить более пятидесяти лабораторных работ в Wireshark.

Как известно, нет ничего вечного и дисковые накопители не исключение. Они вырабатывают ресурс, выходят из строя, часто внезапно. Чтобы уберечь себя от подобных рисков давно были придуманы массивы с избыточностью, когда информация дублируется на несколько дисков и в случае отказа одного из них у нас останется рабочая копия, а мы сможем спокойно и без особых проблем заменить сбойный диск.

ZFS не исключение, сегодня она широко используется для хранилищ разного уровня и очень часто используется "из коробки", без полного понимания работы. Именно так, чаще всего, происходит в системе виртуализации Proxmox. Там можно легко создать ZFS в графическом интерфейсе, но практически невозможно им управлять и когда пользователь видит отказавший диск, то сразу возникает вопрос: что делать? Отказавший диск есть, а никаких инструментов работы с ним нет.

Status: stor1					
Reload					
Health DEGRADED					
Scan scrub in progress since Thu Aug 29 21:15:44 2024 133G / 1.24T scanned at 77.8M/s, 0B / 1.24T issued 0B repaired, 0.00% done, no estimated completion time					
Status One or more devices has experienced an error resulting in data corruption. Applications may be affected.					
Action Restore the file in question if possible. Otherwise restore the entire pool from backup.					
Errors 2 data errors, use '-v' for a list					
Devices					
Name	Health	READ	WRITE	CKSUM	Message
stor1	DEGRADED	0	0	0	
mirror-0	DEGRADED	0	0	0	
/dev/disk/by-id/wwn-0x5000035ac...-part1	ONLINE	0	0	117	
/dev/disk/by-id/wwn-0x5000035ac...-part1	DEGRADED	0	0	364	too many errors
special					
mirror-1	ONLINE	0	0	0	
/dev/disk/by-ata-Netac_SSD_120GB_AA...04-part1	ONLINE	0	0	1	
/dev/disk/by-ata-ADATA_SU650_4N...WK-part1	ONLINE	0	0	1	

Прежде всего не паниковать. Вся основная работа по администрированию Linux производится в консоли, веб-панели - это просто приятное дополнение, не более. Поэтому переходим в консоль с правами суперпользователя (root) и первым делом получаем список хранилищ (пулов) ZFS:

```
zpool list
```

После чего вы получите примерно такой вывод:

```
root@pve2:~# zpool list
```

NAME	SIZE	ALLOC	FREE	CKPOINT	EXPANDSZ	FRAG	CAP	DEDUP	HEALTH	ALTROOT
vm-store-nvme	238G	70.6G	167G	-	-	29%	29%	1.00x	ONLINE	-
vm-store-ssd	476G	19.4G	457G	-	-	1%	4%	1.00x	DEGRADED	-

Из чего мы делаем вывод, что у нас в данной системе два пула, один исправный - **ONLINE**, второй с отказавшей избыточностью - **DEGRADED**.

Теперь получим информацию о деградировавшем пуле:

```
zpool status vm-store-ssd
```

Где **vm-store-ssd** - имя интересующего нас пула.

```
root@pve2:~# zpool status vm-store-ssd
```

pool: vm-store-ssd					
state: DEGRADED					
status: One or more devices has been removed by the administrator. Sufficient replicas exist for the pool to continue functioning in a degraded state.					
action: Online the device using 'zpool online' or replace the device with 'zpool replace'.					
scan: scrub repaired 0B in 00:01:37 with 0 errors on Sun Jul 14 00:25:39 2024					
config:					
NAME	STATE	READ	WRITE	CKSUM	
vm-store-ssd	DEGRADED	0	0	0	
mirror-0	DEGRADED	0	0	0	
ata-ADATA_SU750_2M...QUA	REMOVED	0	0	0	
ata-ADATA_SU750_2M...ED9	ONLINE	0	0	0	

Как видим, отказавший массив содержал два диска, один из которых уже физически извлечен (REMOVED), ZFS оперирует именами дисков по id и идентификатор, как правило, уже содержит серийный номер, что позволяет быстро идентифицировать отказавший диск.

Если же диск окончательно вышел из строя и не определяется, либо определяется как-то не так, то рядом с диском будет указано как он именовался до того, как перестал работать.

```
root@pbs:~# zpool status stor1
pool: stor1
state: DEGRADED
status: One or more devices has experienced an error resulting in data
corruption. Applications may be affected.
action: Restore the file in question if possible. Otherwise restore the
entire pool from backup.
see: https://openzfs.github.io/openzfs-docs/msg/ZFS-8000-BA
scan: resilvered 30.1G in 00:05:15 with 0 errors on Mon Jul 1 10:35:19 2024
config:

    NAME                                STATE     READ WRITE CKSUM
    stor1                               DEGRADED     0     0     0
      mirror-0
        wwn-0x5000039...55              ONLINE       0     0     4
        wwn-0x5000039...60              DEGRADED     2     0    218  too many errors
      special
        mirror-1
          9107280285737058137            UNAVAIL     0     0     0  was /dev/disk/by-id/ata-ADATA_SU650_4N... T5-part1
          ata-ADATA_SU650_4N36235KYAWK    ONLINE       0     0     0
```

Такой диск помечается как **UNAVAIL** - недоступный, если же помеченный сбойным диск присутствует в системе и продолжает работать, то он помечается как **DEGRADED**.

В целом разобрались, id чаще всего содержит серийный номер диска, что позволяет быстро идентифицировать виновника на физическом уровне. Но обратите внимание на две записи на скриншоте выше.

```
wwn-0x5000039***55
wwn-0x5000039***60
```

Никакими серийниками тут и не пахнет, поэтому давайте узнаем на какие именно физические устройства указывают данные идентификаторы.

```
ls -al /dev/disk/by-id
```

```
lrwxrwxrwx 1 root root 9 Aug 29 04:34 ata-TOSHIBA_HDWD240_51I1 -> ../../sdb
lrwxrwxrwx 1 root root 10 Aug 29 04:34 ata-TOSHIBA_HDWD240_51I1-part1 -> ../../sdb1
lrwxrwxrwx 1 root root 10 Aug 29 04:34 ata-TOSHIBA_HDWD240_51I1-part9 -> ../../sdb9
lrwxrwxrwx 1 root root 9 Aug 29 04:34 ata-TOSHIBA_HDWD240_51Q1 -> ../../sda
lrwxrwxrwx 1 root root 10 Aug 29 04:34 ata-TOSHIBA_HDWD240_51Q1-part1 -> ../../sda1
lrwxrwxrwx 1 root root 10 Aug 29 04:34 ata-TOSHIBA_HDWD240_51Q1-part9 -> ../../sda9
lrwxrwxrwx 1 root root 9 Aug 29 04:34 wwn-0x5000039...60 -> ../../sdb
lrwxrwxrwx 1 root root 10 Aug 29 04:34 wwn-0x5000039...60-part1 -> ../../sdb1
lrwxrwxrwx 1 root root 10 Aug 29 04:34 wwn-0x5000039...60-part9 -> ../../sdb9
lrwxrwxrwx 1 root root 9 Aug 29 04:34 wwn-0x5000039...55 -> ../../sda
lrwxrwxrwx 1 root root 10 Aug 29 04:34 wwn-0x5000039...55-part1 -> ../../sda1
lrwxrwxrwx 1 root root 10 Aug 29 04:34 wwn-0x5000039...55-part9 -> ../../sda9
```

Вот теперь сразу становится понятно, что:

```
wwn-0x5000039***55 -> ata-TOSHIBA_HDWD240_51I1*** -> ../../sdb
wwn-0x5000039***60 -> ata-TOSHIBA_HDWD240_51Q1*** -> ../../sda
```

Если данные тома не содержат корневой файловой системы, а такой случай мы в данной статье не рассматриваем, то просто выключаем сервер, удаляем сбойный накопитель и ставим на его место новый. Загружаемся, узнаем id нового диска той же командой:

```
ls -al /dev/disk/by-id
```

Теперь можем выполнить замену, для чего вам потребуется всего одна команда:

```
zpool replace vm-store-ssd /dev/disk/by-id/ata-ADATA_SU750_2M***UA /dev/disk/by-id/ata-Netac_SSD_512GB_AA***68
```

В нашем случае мы заменили в пуле **vm-store-ssd** отказавший диск **ata-ADATA_SU750_2M***UA** на новый диск **ata-Netac_SSD_512GB_AA***68**.

Теперь вам остается только дождаться окончания синхронизации. ZFS - умная система и не синхронизирует нули, поэтому данный процесс будет зависеть только от объема реальных данных на накопителе. Посмотреть состояние процесса можно командой:

```
zpool status vm-store-ssd
```

```
root@pve2:~# zpool status vm-store-ssd
pool: vm-store-ssd
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 Thu Aug 29 19:46:50 2024
19.4G / 19.4G scanned, 14.1G / 19.4G issued at 190M/s
14.3G resilvered, 72.90% done, 00:00:28 to go
config:

    NAME                                STATE        READ  WRITE CKSUM
    vm-store-ssd                        DEGRADED     0     0     0
      mirror-0
        replacing-0
          ata-ADATA_SU750_2M[REDACTED]UA    REMOVED      0     0     0
          ata-Netac_SSD_51[REDACTED]68    ONLINE      0     0     0 (resilvering)
          ata-ADATA_SU750_2M[REDACTED]D9    ONLINE      0     0     0
errors: No known data errors
```

После синхронизации можно сбросить ошибки массива выполнив:

```
zpool clear vm-store-ssd
```

Как видим, заменить сбойный диск в массиве ZFS совсем не сложно, главное - быть внимательным и правильно определить нужное физическое устройство.

- Категории:
- Виртуализация,
- Системному администратору,
- Хранение и защита данных
- Теги:

- Proxmox,
 - RAID,
 - ZFS,
 - Виртуализация
-