nvme-cli用户工具

由于数据中心需要很多监控SSD健康度和耐用度，以及更新firmware，安全擦除数据，读取设备日志等管理功能， NVMe组织 <https://nvmexpress.org/>`_ 开发了Linux用户空间命令行工具 `nvme-cli` ，可以在Linux系统中管理NVM-Express设备。

> **备注**
>
> 有关NVMe监控、管理和故障报告，请参考 SSD故障分析-NVMe SSD管理

安装

源代码编译安装

- clone源代码是需要 `libnvme` 子模块，所以使用以下命令clone:

```
git clone --recurse-submodules https://github.com/linux-nvme/nvme-cli
```

- 如果没有使用上述包含子模块方式clone但已经clone过源代码，则可以使用以下命令重新初始化并更新:

```
git submodule update --init
```

当然，上述命令也可以分为2个命令执行:

```
git submodule init
git submodule update
```

- 编译安装:

```
make
make install
```

发行版安装

主要的Linux发行版都包含了 `nvme-cli` ，可以使用对应包管理器安装:

- Debian/Ubuntu安装:

```
sudo apt install nvme-cli
```

使用

nvme-cli 命令案例和说明

命令	说明
nvme list	列出系统所有NVMe SSD:设备名,序列号,型号,namespace,使用量,LBA格式,firmware版本
nvme id-ctrl	NVMe控制器信息以及控制器支持功能
nvme id-ns	查看 NVMe namespaces, 优化, 功能, 和支持
nvme format	安全删除SSD上数据，格式化一个LBA大小或者为端到端数据保护信息
nvme sanitize	安全擦除SSD所有用户数据
nvme smart-log	输出NVMe SMART health status, temp, endurance, 以及更多的日志页面
nvme fw-log	输出firmware 日志页面
nvme error-log	输出 NVMe 错误日志页面
nvme reset	重置NVMe controller / NVMe SSD
nvme help	帮助信息输出，另外，还支持特定厂商的插件扩展信息帮助，例如 nvme help intel 会输出intel厂商的特定扩展(需要安装对应设备?)
nvme delete-ns	删除一个namespace
nvme create-ns	创建一个新的namespace, 例如创建一个较小的namespace 来过度配置 SSD 以提高耐用性、性能和延迟
nvme fw-download	将新固件下载到 NVMe 设备
nvme fw-commit	提交 (激活) firmware 立即运行或者在reset之后运行

- 列出系统所有安装的NVMe SSD:

```
sudo nvme list
```

在我的 HPE ProLiant DL360 Gen9服务器 上通过 PCIe bifurcation 安装了3根 三星PM9A1 NVMe存储 :

nvme list 输出

```
Node          SN             Model                      Namespace Usage                      Format
------------- -------------- -------------------------- --------- -------------------------- --------
/dev/nvme0n1  S676NF0R908202 SAMSUNG MZVL21T0HCLR-00B00 1         0.00   B /   1.02  TB  512   B
/dev/nvme1n1  S676NF0R908214 SAMSUNG MZVL21T0HCLR-00B00 1         0.00   B /   1.02  TB  512   B
/dev/nvme2n1  S676NF0R908144 SAMSUNG MZVL21T0HCLR-00B00 1         0.00   B /   1.02  TB  512   B
```

- 检查NVMe控制器以及支持的功能:

```
sudo nvme id-ctrl /dev/nvme0
```

输出类似:

nvme id-ctrl 输出

```
NVME Identify Controller:
vid       : 0x144d
ssvid     : 0x144d
sn        : S676NF0R908202
mn        : SAMSUNG MZVL21T0HCLR-00B00
fr        : GXA7401Q
rab       : 2
ieee      : 002538
cmic      : 0
mdts      : 7
cntlid    : 0x6
ver       : 0x10300
rtd3r     : 0x30d40
rtd3e     : 0x989680
oaes      : 0x200
ctratt    : 0x10
rrls      : 0
crdt1     : 0
crdt2     : 0
crdt3     : 0
oacs      : 0x17
acl       : 7
aerl      : 3
frmw      : 0x16
lpa       : 0xe
elpe      : 63
npss      : 4
avscc     : 0x1
apsta     : 0x1
wctemp    : 354
cctemp    : 358
mtfa      : 0
hmpre     : 0
hmmin     : 0
tnvmcap   : 1024209543168
unvmcap   : 0
rpmbs     : 0
edstt     : 35
dsto      : 0
fwug      : 0
kas       : 0
hctma     : 0x1
mntmt     : 318
mxtmt     : 356
sanicap   : 0x2
hmminds   : 0
hmmaxd    : 0
nsetidmax : 0
anatt     : 0
anacap    : 0
anagrpmax : 0
nanagrpid : 0
sqes      : 0x66
cqes      : 0x44
maxcmd    : 256
nn        : 1
oncs      : 0x57
fuses     : 0
fna       : 0
vwc       : 0x7
awun      : 1023
awupf     : 0
nvscc     : 1
nwpc      : 0
acwu      : 0
sgls      : 0
mnan      : 0
subnqn    : nqn.1994-11.com.samsung:nvme:PM9A1:M.2:S676NF0R908202
ioccsz    : 0
iorcsz    : 0
icdoff    : 0
ctrattr   : 0
msdbd     : 0
ps    0 : mp:8.37W operational enlat:0 exlat:0 rrt:0 rrl:0
          rwt:0 rwl:0 idle_power:- active_power:-
ps    1 : mp:8.37W operational enlat:0 exlat:200 rrt:1 rrl:1
          rwt:1 rwl:1 idle_power:- active_power:-
ps    2 : mp:8.37W operational enlat:0 exlat:200 rrt:2 rrl:2
          rwt:2 rwl:2 idle_power:- active_power:-
ps    3 : mp:0.0500W non-operational enlat:2000 exlat:1200 rrt:3 rrl:3
          rwt:3 rwl:3 idle_power:- active_power:-
ps    4 : mp:0.0050W non-operational enlat:500 exlat:9500 rrt:4 rrl:4
          rwt:4 rwl:4 idle_power:- active_power:-
```

- 重要: 检查NVMe SMART健康状态，温度等:

```
sudo nvme smart-log /dev/nvme0n1
```

输出:

nvme smart-log 输出

```
Smart Log for NVME device:nvme0n1 namespace-id:ffffffff
critical_warning                    : 0
temperature                         : 36 C
available_spare                     : 100%
available_spare_threshold           : 10%
percentage_used                     : 0%
data_units_read                     : 58
data_units_written                  : 0
host_read_commands                  : 1,299
host_write_commands                 : 0
controller_busy_time                : 0
power_cycles                        : 13
power_on_hours                      : 149
unsafe_shutdowns                    : 10
media_errors                        : 0
num_err_log_entries                 : 0
Warning Temperature Time            : 0
Critical Composite Temperature Time : 0
Temperature Sensor 1                : 36 C
Temperature Sensor 2                : 39 C
Thermal Management T1 Trans Count   : 0
Thermal Management T2 Trans Count   : 0
Thermal Management T1 Total Time    : 0
Thermal Management T2 Total Time    : 0
```

`smart-log-add` 输出信息中参数说明(不是所有设备都支持， `smart-log` 包含了部分参数):

nvme smart-log输出参数说明

参数	含义	说明
Available Spare	可用备用	包含可用剩余备用容量的标准化百分比（0 到 100%）
Available Spare Threshold	可用的备用阈值	当可用备用容量低于此字段中指示的阈值时，可能会发生异步事件完成。该值表示为标准化百分比（0 到 100%）
Percentage Used	使用百分比	根据实际使用情况和制造商对 NVM 寿命的预测，包含供应商对所用 NVM 子系统寿命百分比的具体估计。（注意：如果使用存储的时间超过其计划寿命，该数字可能会超过 100%。）
Data Units Read/Data Units Written	数据单元读取/数据单元写入	这是读/写的 512 字节数据单元的数量，但它以一种不寻常的方式测量。第一个值对应于 512 字节单元中的 1000 个。因此，可以将此值乘以 512000 以获得以字节为单位的值。它不包括元数据访问。
Host Read/Write Commands	主机读/写命令	发出的适当类型的命令数。使用此值以及以下值，可以计算"物理"读取和写入的平均 IO 大小。
Controller Busy Time	控制器忙碌时间	控制器忙于服务命令的时间（以分钟为单位）。这可用于衡量长期存储负载趋势。
Unsafe Shutdowns	不安全的停机	在未发送关机通知的情况下发生断电的次数。根据使用的 NVMe 设备，不安全的关机可能会损坏用户数据。
Warning Temperature Time/Critical Temperature Time	警告温度时间/临界温度时间	设备在警告或临界温度以上运行的时间（以分钟为单位）。它应该是零。
Wear_Leveling	磨损级别	这显示了使用了多少额定电池寿命，以及不同电池的最小/最大/平均写入计数。在这种情况下，看起来单元的额定写入次数为 1800 次，平均使用了大约 1100 次
Timed Workload Media Wear	定时工作负载介质磨损	媒体因当前的"工作量"而磨损。除了显示设备生命周期值外，该设备还允许从重置它们时测量一些统计数据（称为"工作负载"）。
Timed Workload Host Reads	定时工作负载主机读取	已读取的 IO 操作的百分比（因为工作负载计时器已重置）。
Thermal Throttle Status	热油门状态	这显示设备是否因过热而受到限制，以及过去何时发生限制事件。
Nand Bytes Written	写入 Nand 字节	写入 NAND 单元的字节。对于此设备，测量单位似乎是 32MB 值。其他设备可能会有所不同。
Host Bytes Written	主机字节写入	从系统写入 NVMe 存储的字节数。这个单位也是 32MB 的值。这些值的大小不是很重要，因为它们最有助于找到工作负载的写入放大。该比率以对 NAND 的写入和对 HOST 的写入来衡量。对于此示例，写入放大因子 (WAF) 为 16185227 / 6405605 = 2.53

- 安装 `smartmontools` 工具之后，使用 `smartctl` 可以更为方便检查SMART信息:

```
smartctl --all /dev/nvme0n1
```

输出:

smartctl -all /dev/nvme0n1 输出信息

```
smartctl 7.1 2019-12-30 r5022 [x86_64-linux-5.4.0-121-generic] (local build)
Copyright (C) 2002-19, Bruce Allen, Christian Franke, www.smartmontools.org

=== START OF INFORMATION SECTION ===
Model Number:                       SAMSUNG MZVL21T0HCLR-00B00
Serial Number:                      S676NF0R908202
Firmware Version:                   GXA7401Q
PCI Vendor/Subsystem ID:            0x144d
IEEE OUI Identifier:                0x002538
Total NVM Capacity:                 1,024,209,543,168 [1.02 TB]
Unallocated NVM Capacity:           0
Controller ID:                      6
Number of Namespaces:               1
Namespace 1 Size/Capacity:          1,024,209,543,168 [1.02 TB]
Namespace 1 Utilization:            530,768,371,712 [530 GB]
Namespace 1 Formatted LBA Size:     512
Namespace 1 IEEE EUI-64:            002538 b911b37f97
Local Time is:                      Wed Aug 31 12:01:28 2022 CST
Firmware Updates (0x16):            3 Slots, no Reset required
Optional Admin Commands (0x0017):   Security Format Frmw_DL Self_Test
Optional NVM Commands (0x0057):     Comp Wr_Unc DS_Mngmt Sav/Sel_Feat Timestmp
Maximum Data Transfer Size:         128 Pages
Warning  Comp. Temp. Threshold:     81 Celsius
Critical Comp. Temp. Threshold:     85 Celsius

Supported Power States
St Op     Max   Active     Idle   RL RT WL WT  Ent_Lat  Ex_Lat
 0 +     8.37W       -        -    0  0  0  0        0       0
 1 +     8.37W       -        -    1  1  1  1        0     200
 2 +     8.37W       -        -    2  2  2  2        0     200
 3 -   0.0500W       -        -    3  3  3  3     2000    1200
 4 -   0.0050W       -        -    4  4  4  4      500    9500

Supported LBA Sizes (NSID 0x1)
Id Fmt  Data  Metadt  Rel_Perf
 0 +     512       0         0

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

SMART/Health Information (NVMe Log 0x02)
Critical Warning:                   0x00
Temperature:                        43 Celsius
Available Spare:                    100%
Available Spare Threshold:          10%
Percentage Used:                    4%
Data Units Read:                    8,781,303 [4.49 TB]
Data Units Written:                 16,708,935 [8.55 TB]
Host Read Commands:                 210,063,865
Host Write Commands:                1,235,247,326
Controller Busy Time:               18,409
Power Cycles:                       18
Power On Hours:                     3,903
Unsafe Shutdowns:                   11
Media and Data Integrity Errors:    0
Error Information Log Entries:      0
Warning  Comp. Temperature Time:    0
Critical Comp. Temperature Time:    0
Temperature Sensor 1:               43 Celsius
Temperature Sensor 2:               56 Celsius

Error Information (NVMe Log 0x01, max 64 entries)
No Errors Logged
```

- 检查firmware 日志页面:

```
sudo nvme fw-log /dev/nvme0n1
```

输出:

nvme fw-log 输出

```
Firmware Log for device:nvme0n1
afi  : 0x1
frs1 : 0x5131303437415847 (GXA7401Q)
```

- 输出NVMe错误日志页面:

```
sudo nvme error-log /dev/nvme0n1
```

输出部分案例:

nvme error-log 输出

```
Error Log Entries for device:nvme0n1 entries:64
.................
 Entry[ 0]
.................
error_count  : 0
sqid         : 0
cmdid        : 0
status_field : 0(SUCCESS: The command completed successfully)
parm_err_loc : 0
lba          : 0
nsid         : 0
vs           : 0
cs           : 0
.................
 Entry[ 1]
.................
error_count  : 0
sqid         : 0
cmdid        : 0
status_field : 0(SUCCESS: The command completed successfully)
parm_err_loc : 0
lba          : 0
nsid         : 0
vs           : 0
cs           : 0
...
```

- (警告：我没有使用过)重置NVMe controller / NVMe SSD:

```
sudo nvme reset /dev/nvme0n1
```

NVMe Namespace解析

NVMe的namespace就是NVMe技术中用于存储用户数据的结构。一个NVMe可以具有多个namespace。不过大多数情况下，现在NVMe只使用一个namespace。但是，如果是多租户(multi-tenant)应用程序，虚拟化以及安全要求等业务场景，需要使用多名字空间(multiple namespaces)。

所谓namespace就是一组逻辑块，这些逻辑块地址范围从0到这个namespace的size; 名字空间ID (namespace ID, NSID)是控制器用来访问该名字空间的标识。

你会发现namespace的size和namespace的utilization(使用)对于生成LBA使用的比例非常有用。在标识namespace的命令输出的有用数据中能够用来优化主机软件的性能、数据一致性，TRIM(回收),LBA大小(例如512B,4kB)等等

- 检查NVMe的namesapce:

```
sudo nvme id-ns /dev/nvme0n1
```

输出:

nvme id-ns 输出

```
NVME Identify Namespace 1:
nsze    : 0x773bd2b0
ncap    : 0x773bd2b0
nuse    : 0
nsfeat  : 0
nlbaf   : 0
flbas   : 0
mc      : 0
dpc     : 0
dps     : 0
nmic    : 0
rescap  : 0
fpi     : 0x80
dlfeat  : 1
nawun   : 0
nawupf  : 0
nacwu   : 0
nabsn   : 0
nabo    : 0
nabspf  : 0
noiob   : 0
nvmcap  : 1024209543168
nsattr  : 0
nvmsetid: 0
anagrpid: 0
endgid  : 1
nguid   : 00000000000000000000000000000000
eui64   : 002538b911b37f97
lbaf  0 : ms:0   lbads:9  rp:0 (in use)
```

更新firmware

SSD厂商通常会在SSD的生产周期内多次发布firmware更新，不过一个SSD的5年生命期发布4~5次更新则很少见。firmware更新可以提供安全补丁，bug修复以及可靠性提高。OEM通常使用自己的管理工具来更新，比且会加密签名firmware以确保匹配其OEM产品，不过NVMe SSD可以从渠道分销获得通用firmware进行更新。请联系SSD供应商获取最新firmware。

请参考 NVMe 1.4 规范的 `Firmware Update Process` 部分，可以详细了解在哪里需要reset，firmware slot的概念(一些NVMe SSD有多个firmware副本存储在设备上，可以通过激活指定副本来运行，这样出现问题可以回退)

- 查看当前firmware版本:

```
sudo nvme id-ctrl /dev/nvme0 | grep "fr "
```

可以看到firmware版本:

```
fr        : GXA7401Q
```

- 下载firmware到目标设备:

```
nvme fw-download /dev/nvme0 -
nvme fw-commit /dev/nvme0 -a 0
```

这里 `-a` 参数:

  - `0` 表示将镜像替换掉 `Firmware slot` 字段指定的镜像，这个镜像没有激活
  - `1` 将镜像替换掉 `Firmware slot` 字段指定的镜像，这个镜像在下次reset时激活
  - `2` 通过 `Firmware slot` 字段制定的镜像在下次reset时激活
  - `3` 立即激活 `Firmware slot` 字段指定的镜像，无需reset

- 完成firmware下载之后，需要reset设备(如果这个设备不支持无需reset设备就激活镜像):

```
sudo nvme reset /dev/nvme0
```

> **警告**
>
> 实际上firmware升级需要非常谨慎，我还没有机会实践，以上仅是一些资料整理，后续有机会再实践

参考

- NVMe management command line interface
- Open Source NVMe™ Management Utility – NVMe Command Line Interface (NVMe-CLI)
- Using NVMe Command Line Tools to Check NVMe Flash Health

Previous Next