

NVMe(oF) prakticky

libnvme, nvme-cli, nvmetcli

Tomáš Bžatek

tbzatek@redhat.com

Mé nové NVMe SSD

- ▶ Namespace: /dev/nvme0n1
 - nsid = 123
 - eui64 = 0024cf014c00a2ef
- ▶ Controller: /dev/nvme0
 - cntlid = 1
 - transport = pcie
 - address = 0000:02:00.0
- ▶ Subsystem: nvme-subsys0
 - subsysnqn = nqn.2014.08.org.nvmexpress:1bb11bb17QF0122T
- ▶ číslování: async scan

nvme-cli

```
# nvme --help
```

```
list
```

List all NVMe devices and namespaces on machine

```
id-ctrl
```

Send NVMe Identify Controller

```
id-ns
```

Send NVMe Identify Namespace, display structure

```
fw-log
```

Retrieve FW Log, show it

```
smart-log
```

Retrieve SMART Log, show it

```
error-log
```

Retrieve Error Log, show it

```
device-self-test
```

Perform the necessary tests to observe the performance

```
self-test-log
```

Retrieve the SELF-TEST Log, show it

```
format
```

Format namespace with new block format

```
fw-commit
```

Verify and commit firmware to a specific slot

```
fw-download
```

Download new firmware

```
...
```

```
# nvme id-ns -H /dev/nvme0n1
```

```
NVME Identify Namespace 1:
```

```
nsze      : 0x7470c06    Total size in logical blocks
ncap      : 0x7470c06    Maximum size in logical blocks
nuse      : 0x7470c06    Current size in logical blocks
nsfeat    : 0
```

```
...
```

```
endgid    : 0
nguid     : 00000000000000000000000000000000
eui64     : 0024cf014c00a2ef
```

```
LBA Format 0 : Metadata Size: 0    bytes - Data Size: 512 bytes - Relative Performance: 0x2 Good
```

```
LBA Format 1 : Metadata Size: 0    bytes - Data Size: 4096 bytes - Relative Performance: 0x1 Better (in use)
```

```
# nvme format --verbose --lbaf=1 /dev/nvme0n1
```

```
Success formatting namespace:1
```

nvme smart-log -H /dev/nvme0

```
critical_warning          : 0
    Available Spare[0]    : 0
    Temp. Threshold[1]    : 0
    NVM subsystem Reliability[2] : 0
    Read-only[3]          : 0
    Volatile mem. backup failed[4] : 0
    Persistent Mem. R0[5] : 0
temperature              : 104 °F (313 K)
available_spare           : 100%
available_spare_threshold : 5%
percentage_used           : 18%
Data Units Read           : 8,457,129 (4.33 TB)
Data Units Written        : 54,379,806 (27.84 TB)
power_cycles              : 77
power_on_hours            : 13,098
media_errors              : 0
num_err_log_entries       : 273
```

nvme fw-log /dev/nvme0

Firmware Log for device:nvme0

afi : 0x2

frs1 : 0x34313043534e5453 (STNSC014)

frs2 : 0x36313043534e5453 (STNSC016)

man nvme-fw-download

man nvme-fw-commit

-a <action>, --action=<action>

Commit Action: This field specifies the action that is taken on the image downloaded with the Firmware Image Download command or on a previously downloaded and placed image.

- 0 | Downloaded image replaces the image indicated by the Firmware Slot field. This image is not activated.
- 1 | Downloaded image replaces the image indicated by the Firmware Slot field. This image is activated at the next reset.
- 2 | The image indicated by the Firmware Slot field is activated at the next reset.
- 3 | The image specified by the Firmware Slot field is requested to be activated immediately without reset.

...

-s <slot>, --slot=<slot>

libnvme

- ▶ C knihovna + Python bindings
- ▶ minimální závislosti, vhodná pro embedded
- ▶ nvme-cli = CLI wrapper nad libnvme
- ▶ kompletní a aktualizovaná implementace NVMe specifikací

<https://nvmexpress.org/specifications/>

217593	IO/Stora	NVMe	io_nvme	NEW	---	Need NVME_QUIRK_BOGUS for SAMSUNG MZ1WV480HCGL-000MV (Samsung SM-953 Datacenter SSD)
217649	IO/Stora	NVMe	io_nvme	NEW	---	Need NVME_QUIRK_BOGUS_NID for SAMSUNG MZAL41T0HBLB-00BL2 (PM9B1 1TB)
217829	IO/Stora	NVMe	io_nvme	NEW	---	"nvme nvme1: globally duplicate IDs for nsid 1" with AMD soft raid
217871	IO/Stora	NVMe	io_nvme	NEW	---	Got nvme timeout with KIOXIA 512GB + VMD mode
217954	IO/Stora	NVMe	io_nvme	NEW	---	Samsung 990 PRO NVMe through AXAGON PCEM2-ND PCIe extension card power down prematurely
217968	IO/Stora	NVMe	io_nvme	NEW	---	First resume after suspend works, second resume causes reboot (likely related to TEAMGROUP MP34 4TB SSD)
218186	IO/Stora	NVMe	io_nvme	NEW	---	divide error in blk_stack_limits() hit during fio workload against nvme target
218365	IO/Stora	NVMe	io_nvme	NEW	---	Encounter nvme timeout issue with WD SN740, may affect all WD nvme devices
218449	IO/Stora	NVMe	io_nvme	NEW	---	NVME Quirk needed
218455	IO/Stora	NVMe	io_nvme	NEW	---	nvme nvme0: Does your device have a faulty power saving mode enabled?
218777	IO/Stora	NVMe	io_nvme	NEW	---	deadlock in nvme_enable_ctrl
218930	IO/Stora	NVMe	io_nvme	NEW	---	block nvme0n1: No UUID available providing old NGUID
219220	IO/Stora	NVMe	io_nvme	NEW	---	nvme nvme0: I/O timeout, completion polled
219464	IO/Stora	NVMe	io_nvme	NEW	---	Critical medium error when accessing OPAL locked partition
219530	IO/Stora	NVMe	io_nvme	NEW	---	NVMe disks randomly disappear/timeout (Samsung related?)
219630	IO/Stora	NVMe	io_nvme	NEW	---	ana-log command returns incorrect values for namespace count greater than 1013"
219791	IO/Stora	NVMe	io_nvme	NEW	---	Namespaces from different controllers have the same SN.
219852	IO/Stora	NVMe	io_nvme	NEW	---	WD_BLACK SN850X dies after several days
219896	IO/Stora	NVMe	io_nvme	NEW	---	commit -a 3 + nvme reset may cause system hung
219916	IO/Stora	NVMe	io_nvme	NEW	---	Klevv C910 4TB RTS5765DL - Device not ready Issue
220049	IO/Stora	NVMe	io_nvme	NEW	---	SK Hynix PE8110 Series Duplicate cntlid 0
220073	IO/Stora	NVMe	io_nvme	NEW	---	failed to allocate host memory buffer with SSD Silicon Power XD80 512GB
220127	IO/Stora	NVMe	io_nvme	NEW	---	The host issued an abort to an IO which did not timeout
220329	IO/Stora	NVMe	io_nvme	NEW	---	Support for FPIN LI with nvme native multi-path
220585	IO/Stora	NVMe	io_nvme	NEW	---	Unwanted suspend wake from Mouse despite turned off in Bios: Lexar NM790 platform quirk: setting simple suspend
213133	IO/Stora	NVMe	io_nvme	REOP	---	Corsair MP600 CORE - timeout, aborting
216142	IO/Stora	NVMe	io_nvme	NEED	---	The whole system freezes while copying large files on NVMe
216948	IO/Stora	NVMe	io_nvme	NEED	---	boot problem: Samsung NVMe controller, can't found root

NVMe over Fabrics

NVMe over Fabrics

- ▶ součást NVM Express specifikace
- ▶ analogie např. k iSCSI
- ▶ fabric = transportní médium, různé topologie
- ▶ initiator = client
- ▶ target = server
- ▶ transport: `pcie`, `loop`, `fc`, `rdma`, `tcp`
- ▶ subsystem NQN (NVMe Qualified Name)
- ▶ host NQN, host ID = identifikace klienta

nvme-cli

```
# nvme --help
```

```
discover
```

Discover NVMeoF subsystems

```
connect-all
```

Discover and Connect to NVMeoF subsystems

```
connect
```

Connect to NVMeoF subsystem

```
disconnect
```

Disconnect from NVMeoF subsystem

```
disconnect-all
```

Disconnect from all connected NVMeoF subsystems

```
gen-hostnqn
```

Generate NVMeoF host NQN

```
show-hostnqn
```

Show NVMeoF host NQN

```
...
```

NVMe targets

- ▶ domácí NAS
- ▶ cloud storage, AI storage
- ▶ enterprise storage arrays (SAN)
- ▶ linux kernel `nvmet` target

NVMe initiator

```
# nvme connect \  
  --transport=tcp \  
  --traddr=192.168.123.1 \  
  --nqn=nqn.2014-08.org.nvmexpress.boot:test  
  
# nvme gen-hostnqn > /etc/nvme/hostnqn  
# cat /etc/nvme/hostnqn  
nqn.2014-08.org.nvmexpress:uuid:0eab259e-74eb-4a72-897a-d49b215807ea  
# cat /etc/nvme/hostid  
0eab259e-74eb-4a72-897a-d49b215807ea
```

NVMe initiator

```
# modprobe nvme_fabrics
```

```
# echo nqn=nqn.2014-08.org.nvmexpress.boot:test, \  
> transport=tcp, \  
> traddr=192.168.123.1, \  
> trsvcid=4420 > /dev/nvme-fabrics
```

```
# cat /dev/nvme-fabrics
```

```
instance=-1,cntlid=-1,transport=%s,traddr=%s,trsvcid=%s,nqn=%s,queue_size=%d,nr_io_queues=%d,reconnect_delay=%d,ctrl_loss_tmo=%d,keep_alive_tmo=%d,hostnqn=%s,host_traddr=%s,host_iface=%s,hostid=%s,duplicate_connect,disable_sqflow,hdr_digest,data_digest,nr_write_queues=%d,nr_poll_queues=%d,tos=%d,keyring=%d,tls_key=%d,fast_io_fail_tmo=%d,discovery,dhchap_secret=%s,dhchap_ctrl_secret=%s,tls,concat
```

nvmetcli

- ▶ frontend pro kernel configs
 - `/sys/kernel/config`
- ▶ podobnost s targetcli (LIO/iSCSI target)
- ▶ interaktivní shell
- ▶ backing store = file, block device
- ▶ konfigurace přes JSON
 - `# nvmetcli restore tcp-example.json`
- ▶ kernel moduly: `nvmet`, `nvmet_tcp`

```

{
  "ports": [
    {
      "addr": {
        "adrfam": "ipv4",
        "traddr": "192.168.122.1",
        "trsvcid": "4420",
        "trtype": "tcp"
      },
      "portid": 1,
      "subsystems": [
        "nqn.2014-08.org.nvmexpress.boot:test"
      ]
    }
  ],
  "subsystems": [
    {
      "attr": {
        "allow_any_host": "1"
      },
      "namespaces": [
        {
          "device": {
            "path": "/storage/disk1.bin"
          },
          "nsid": 123
        }
      ],
      "nqn": "nqn.2014-08.org.nvmexpress.boot:test"
    }
  ]
}

```



```
# tree -d /sys/kernel/config/nvmet/
/sys/kernel/config/nvmet/
├── hosts
├── ports
│   └── 1
│       ├── ana_groups
│       │   └── 1
│       ├── referrals
│       └── subsystems
│           └── nqn.2014-08.org.nvmeexpress.boot.poc:test-target ->
../../../../../../../../nvmet/subsystems/nqn.2014-08.org.nvmeexpress.boot.poc:test-target
└── subsystems
    └── nqn.2014-08.org.nvmeexpress.boot.poc:test-target
        ├── allowed_hosts
        ├── namespaces
        │   └── 123
        └── passthru
```

NVMe-oF Discovery

- ▶ well-known subsystem NQN:
`nqn.2014-08.org.nvmexpress.discovery`

```
# nvme discover \  
--transport=tcp \  
--traddr=172.18.60.20 \  
--trsvcid=4420
```

- ▶ `/etc/nvme/discovery.conf`
- ▶ `/etc/nvme/config.json`
- ▶ `nvmf-autoconnect.service`

Native NVMe multipath

- ▶ 2 a více controllerů v subsystému
 - různé transporty, IP adresy, porty, ...
- ▶ plně transparentní pro userspace
- ▶ automatická recovery dle connection parametrů:

`reconnect_delay=`

`ctrl_loss_tmo=`

`keep_alive_tmo=`

nvme-cli (TLS)

```
# nvme --help
```

gen-dhchap-key	Generate NVMeoF DH-HMAC-CHAP host key
check-dhchap-key	Validate NVMeoF DH-HMAC-CHAP host key
gen-tls-key	Generate NVMeoF TLS PSK
check-tls-key	Validate NVMeoF TLS PSK
tls-key	Manipulate NVMeoF TLS PSK

NVMe/TCP Boot

- ▶ iniciativa Timberland SIG
- ▶ ACPI NBFT table
- ▶ userspace boot: `dracut --> nvme-cli connect-all --nbft`

Attempt Configuration		
Enable DHCP	[]	↑ Must reboot system manually for changes to take place.
Host IP Address	10.0.2.15	
Host Subnet Mask	255.255.255.0	
Gateway	10.0.2.2	
Subsystem info from DHCP []		
NUM Subsystem NQN	nvme-test-target	
NUM Subsystem Address	192.168.0.94	
NUM Subsystem Port	[4420]	
NUM Subsystem NID	43b4d743-4876-4206-b76e-bc d39099037f	
Save Changes		
↓		
↑↓=Move Highlight	F9=Reset to Defaults <Enter>=Select Entry	F10=Save Esc=Exit

Thank you

tbzatek@redhat.com

<https://nvmexpress.org/>

<https://storaged.org/>