東京 2025

LINUX PLUMBERS CONFERENCE

TOKYO, JAPAN / DECEMBER 11-13, 2025

Towards Real-time NVMe monitoring (nvme-top) for Linux

Speaker: Nilay Shroff, Daniel Wagner
Event: Linux Plumbers Conference (LPC) 2025

東京 2025
LINUX
PLUMBERS CONFERENCE
TOKYO, JAPAN / DEC. 11-13, 2025

Motivation

- NVMe monitoring today:
  - `nvme-cli` provides **only static snapshots**
  - Operators must manually rerun commands
  - Debugging multipath or fabrics issues is **slow and reactive**
- In NVMe-oF deployments:
  - Path performance changes dynamically
  - ANA state can fluctuate
  - Congestion, latency spikes, or link failures are common
- **Operators need *continuous* visibility — not snapshots.**

nvme-top UI Design

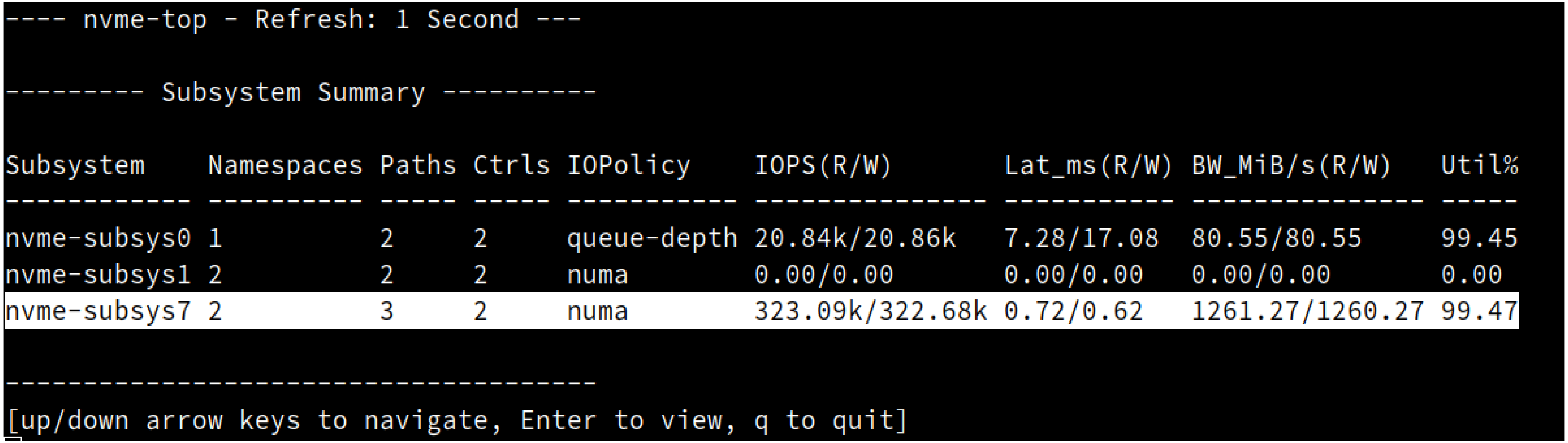
Two-level dashboard design:

- Level 1 — **Subsystem Summary**
  - Namespaces, controllers, paths
  - IOPolicy (numa, queue-depth, rr)
  - Aggregate IOPS, BW, latency
  - Utilization

- Level 2 — **Drill-Down Detail**
  - Namespace-level stats
  - Path performance
  - Path health (ANA, retries, failovers etc.)
  - Controller summary

All updated in real time.

Level 1: Subsystem Summary



```
---- nvme-top - Refresh: 1 Second ---

--------- Subsystem Summary ----------

Subsystem    Namespaces Paths Ctrls IOPolicy    IOPS(R/W)       Lat_ms(R/W) BW_MiB/s(R/W)   Util%
------------ ---------- ----- ----- ----------- --------------- ----------- --------------- -----
nvme-subsys0 1          2     2     queue-depth 20.84k/20.86k   7.28/17.08  80.55/80.55     99.45
nvme-subsys1 2          2     2     numa        0.00/0.00       0.00/0.00   0.00/0.00       0.00
nvme-subsys7 2          3     2     numa        323.09k/322.68k 0.72/0.62   1261.27/1260.27 99.47


--------------------------------------
[up/down arrow keys to navigate, Enter to view, q to quit]
```

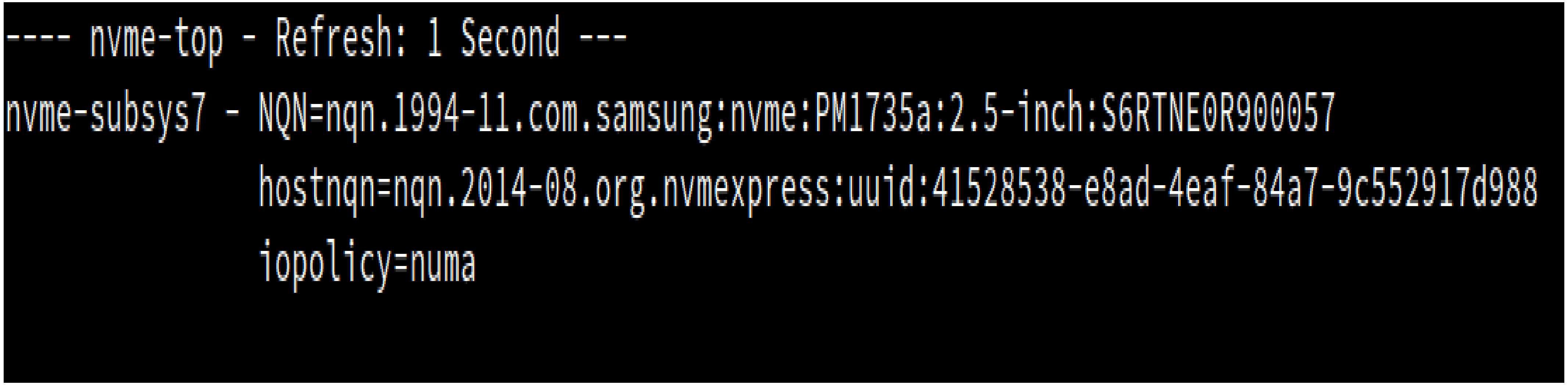
Displayed Metrics:

- Namespaces: Num of namespaces
- Paths : Num of paths
- Controllers : Num of controllers
- I/O Policy : numa, round-robin, queue-depth
- IOPS : Total IOPS aggregated across all ns per subsystem
- Latency : Max latency observed across all ns during the sample interval
- Bandwidth : Total bw aggregated across all ns per subsystem
- Utilization% : Max utilization among all ns during the sample interval.

User Interaction:

- Up/down arrow key to navigate
- Enter to expand
- q to quit

Level 2: Drill-Down (Header)

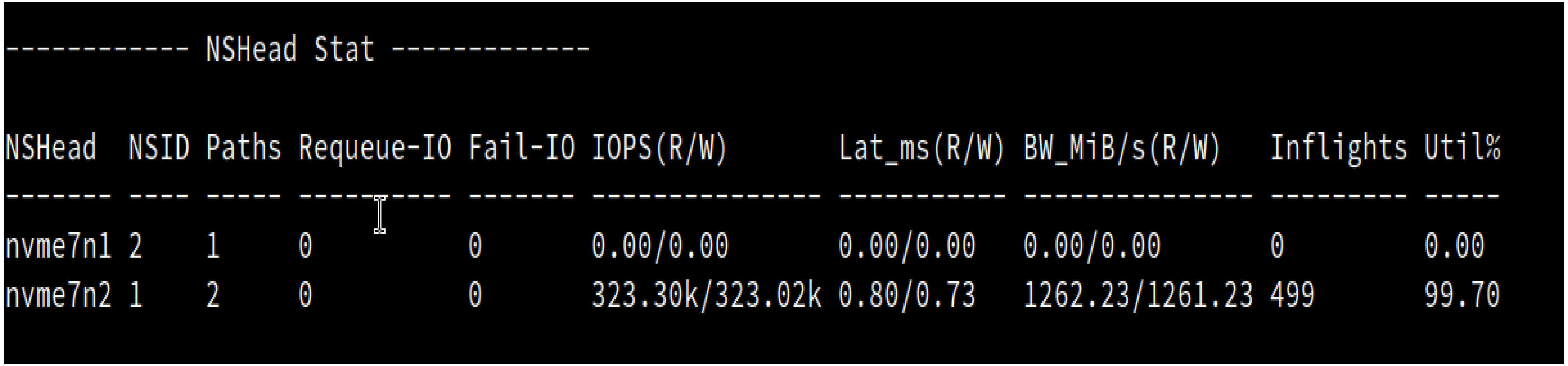


```
---- nvme-top - Refresh: 1 Second ---
nvme-subsys7 - NQN=nqn.1994-11.com.samsung:nvme:PM1735a:2.5-inch:S6RTNE0R900057
               hostnqn=nqn.2014-08.org.nvmexpress:uuid:41528538-e8ad-4eaf-84a7-9c552917d988
               iopolicy=numa
```

Displayed Metrics:

- Refresh : Interval in seconds, user wants to refresh the stat
- NQN : NVM Subsystem NVMe Qualified Name
- hostnqn : Host NVM Subsystem NVMe Qualified Name
- iopolicy : numa, round-robin, queue-depth

Level 2: Drill-Down: NSHead Statistics

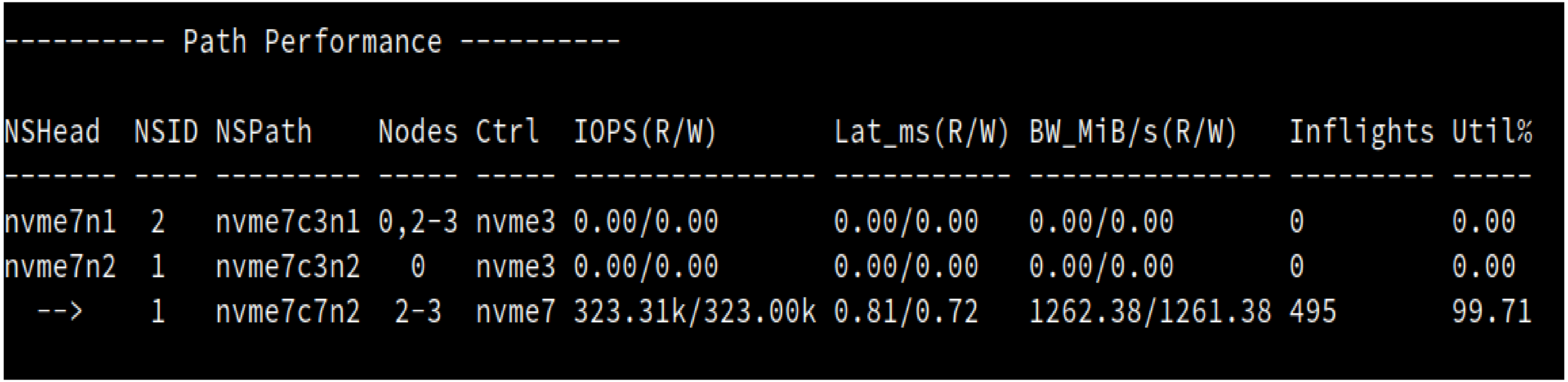


```
------------ NSHead Stat -------------

NSHead  NSID Paths Requeue-IO Fail-IO IOPS(R/W)       Lat_ms(R/W) BW_MiB/s(R/W)   Inflights Util%
------- ---- ----- ---------- ------- --------------- ----------- --------------- --------- -----
nvme7n1 2    1     0          0       0.00/0.00       0.00/0.00   0.00/0.00       0         0.00
nvme7n2 1    2     0          0       323.30k/323.02k 0.80/0.73   1262.23/1261.23 499       99.70
```

Displayed Metrics:

- NSHead : Namespace head
- NSID : Namespace ID
- Paths : Num of I/O paths associated with this namespace head
- Requeue-IO : Num of I/Os re-queued due to none of the available paths could process I/O currently (maybe due to transient error)
- Fail-IO : Num of I/Os forced to fail due to no available paths
- IOPS : Total read/write IOPS aggregated across all paths under this namespace head.
- Latency : Avg. read/write latency across all paths for this NSHead during the last sampling interval.
- Bandwidth : Total read/write bandwidth (in MiB/s) aggregated across all paths.
- Inflights : Total number of in-flight I/Os aggregated across all paths.
- Utilization% : Avg. device utilization across all paths associated with this NSHead.

Level 2: Drill-Down: Path Performance

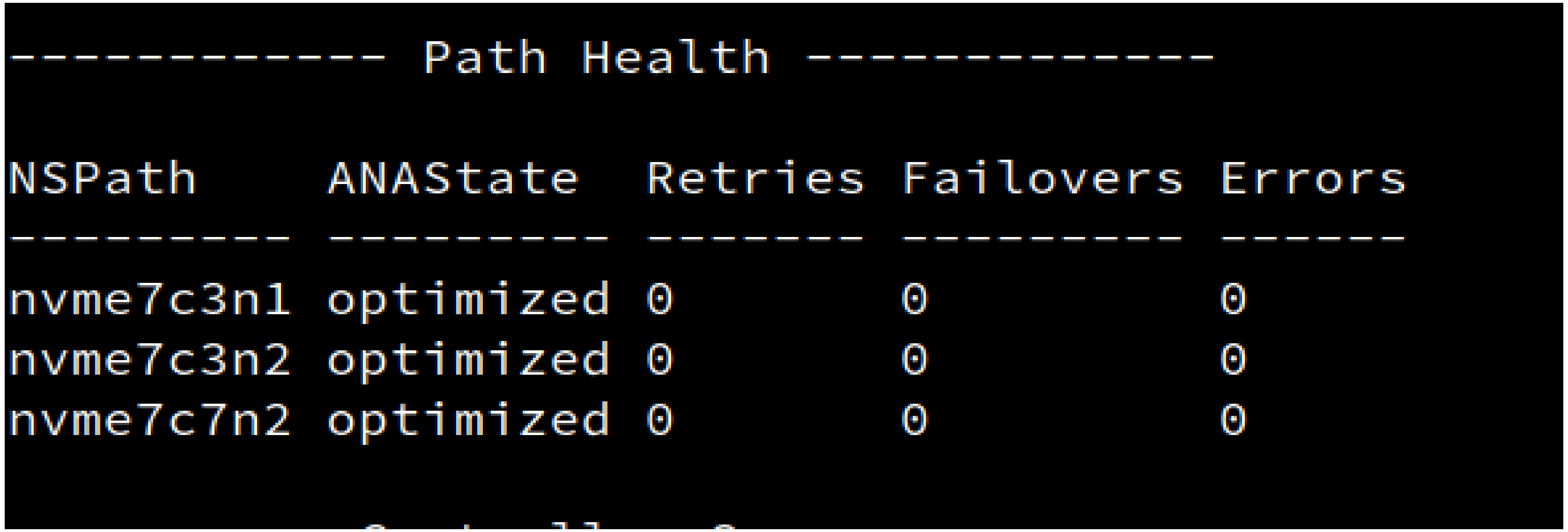


```
---------- Path Performance ----------

NSHead  NSID NSPath    Nodes Ctrl  IOPS(R/W)       Lat_ms(R/W) BW_MiB/s(R/W)   Inflights Util%
------- ---- --------- ----- ----- --------------- ----------- --------------- --------- -----
nvme7n1  2   nvme7c3n1 0,2-3 nvme3 0.00/0.00       0.00/0.00   0.00/0.00       0         0.00
nvme7n2  1   nvme7c3n2   0   nvme3 0.00/0.00       0.00/0.00   0.00/0.00       0         0.00
  -->    1   nvme7c7n2  2-3  nvme7 323.31k/323.00k 0.81/0.72   1262.38/1261.38 495       99.71
```

Displayed Metrics:

- NSHead : Name of the namespace head
- NSID : Namespace Identifier
- NSPaths : Path name. If multiple paths exist, the same NSHead appears multiple times (it is represented with --> symbol).
- Nodes : I/O originating from the list of NUMA nodes selects this path. (Displayed only when the I/O policy is numa)
- Ctrl : Controller name to which this path belongs
- IOPS : Read/write IOPS for the specific path.
- Latency : Read/write Avg. I/O latency (in milliseconds) for this path during the last sample.
- Bandwidth : Read/Write I/O bandwidth (in MiB/s) for this path.
- Inflights : Current number of in-flight I/Os on this path.
- Utilization% : Percent utilization of the specific path (fraction of time the controller was busy servicing I/O on this path).

Level 2: Drill-Down: Path Health

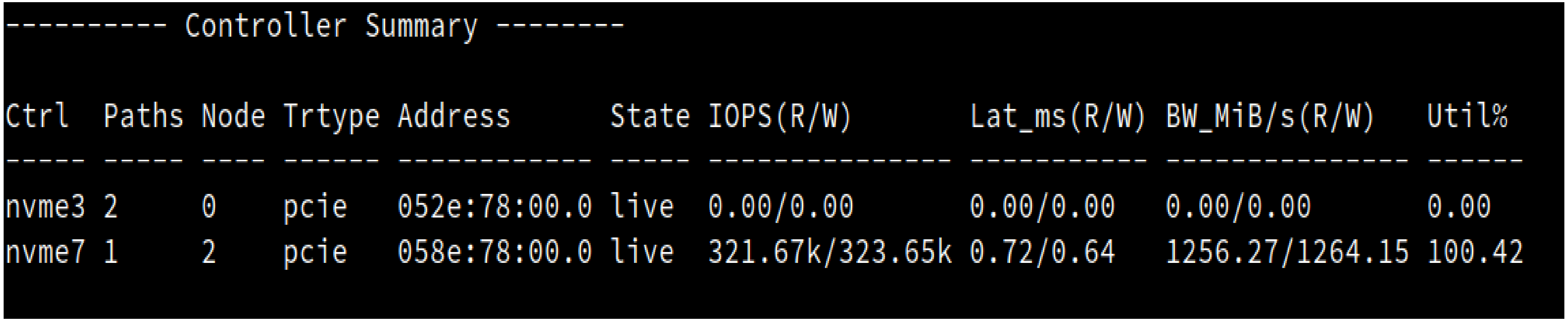


```
------------ Path Health --------------

NSPath     ANAState   Retries Failovers Errors
--------- --------- ------- --------- ------
nvme7c3n1 optimized 0       0         0
nvme7c3n2 optimized 0       0         0
nvme7c7n2 optimized 0       0         0
```

Displayed Metrics:

- NSPath : Path name
- ANAState : ANA state value of the path
- Retries : Number of I/O retries observed on this path
- Errors : Number of I/O errors reported on this path
- Failovers : Number of times I/O switched from this path to another path (meaningful only when a ns is reachable from multiple paths)

Level 2: Drill-Down: Controller Summary

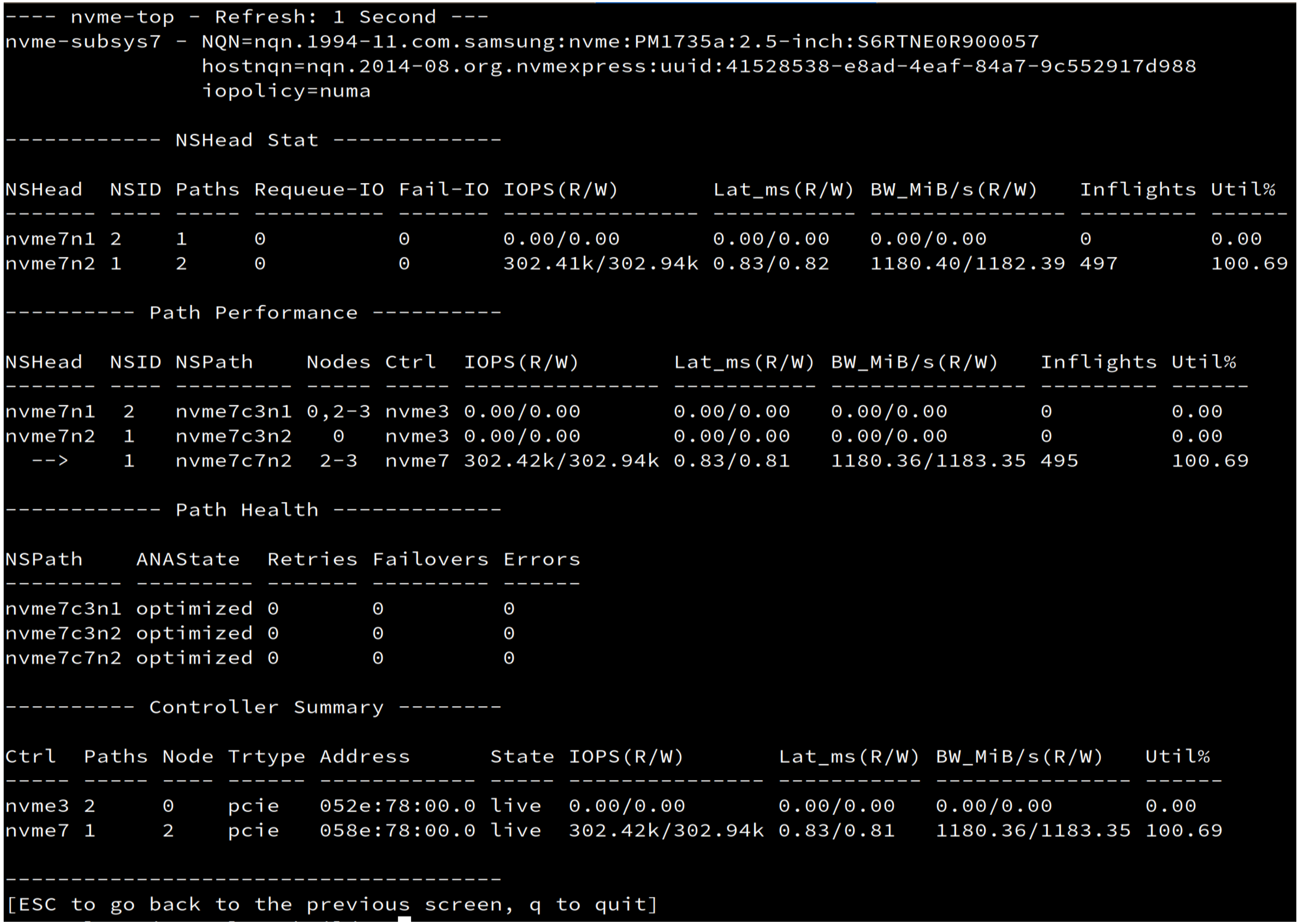


```
---------- Controller Summary --------

Ctrl  Paths Node Trtype Address      State IOPS(R/W)       Lat_ms(R/W) BW_MiB/s(R/W)   Util%
----- ----- ---- ------ ------------ ----- --------------- ----------- --------------- ------
nvme3 2     0    pcie   052e:78:00.0 live  0.00/0.00       0.00/0.00   0.00/0.00       0.00
nvme7 1     2    pcie   058e:78:00.0 live  321.67k/323.65k 0.72/0.64   1256.27/1264.15 100.42
```

Displayed Metrics:

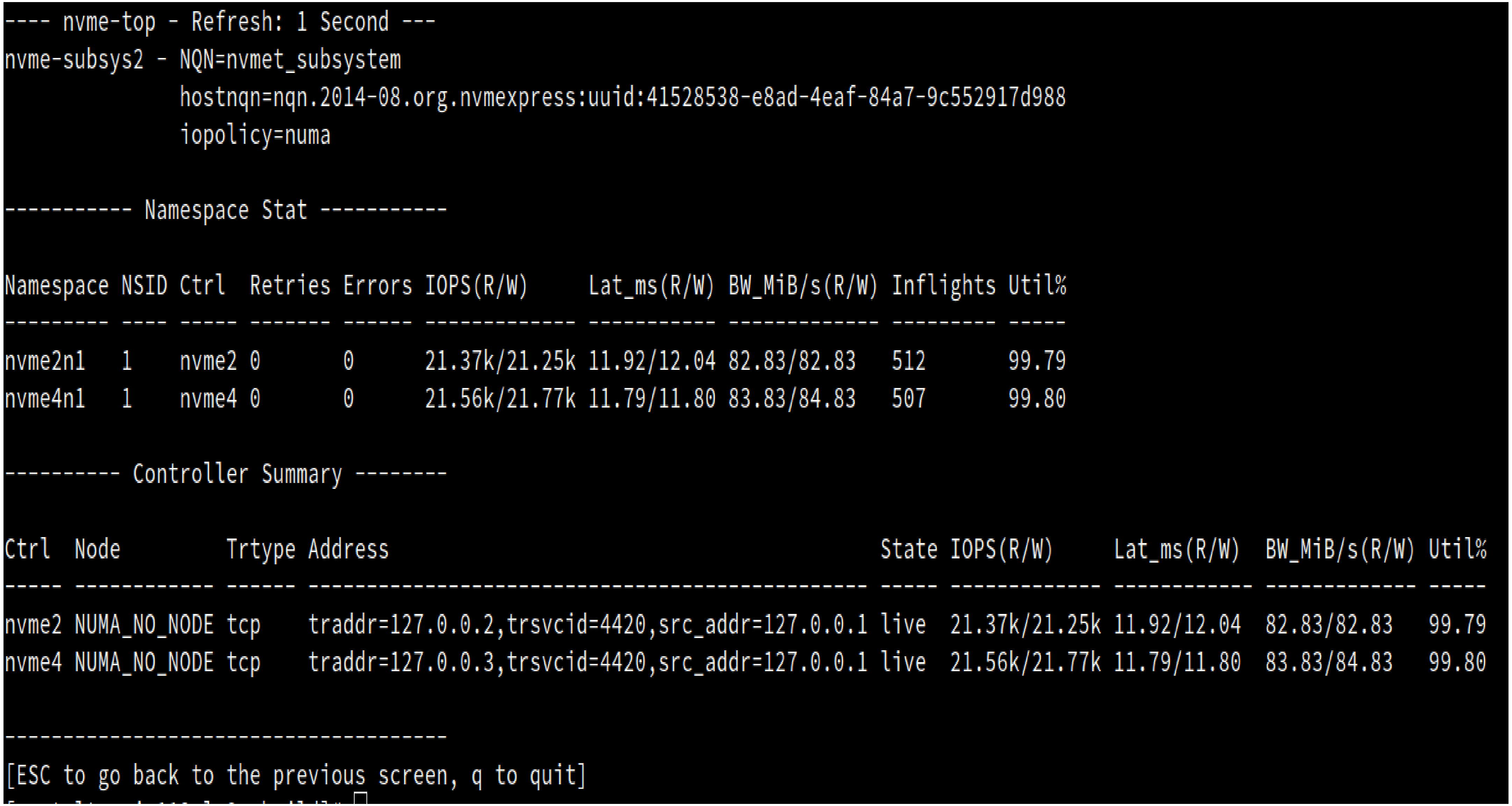
- Ctrl : Controller name
- Paths : Number of I/O paths associated with the controller
- Node : NUMA node local to the controller
- Trtype : Transport type (e.g. pcie, tcp, rdma etc.)
- Address : Transport address
- State : Controller state (live, reconnecting etc.)
- IOPS : Total read/write IOPS aggregated across all paths associated with the controller.
- Latency : Max read/write I/O latency (in milliseconds) observed across all paths for the controller during the last sample interval
- Bandwidth : Total read/write bandwidth aggregated across controller paths
- Utilization% : Maximum disk utilization in percentage among the paths under this controller

Level 2: Dashboard (multipath)



```
---- nvme-top - Refresh: 1 Second ---
nvme-subsys7 - NQN=nqn.1994-11.com.samsung:nvme:PM1735a:2.5-inch:S6RTNE0R900057
               hostnqn=nqn.2014-08.org.nvmexpress:uuid:41528538-e8ad-4eaf-84a7-9c552917d988
               iopolicy=numa

------------ NSHead Stat -------------

NSHead  NSID Paths Requeue-IO Fail-IO IOPS(R/W)       Lat_ms(R/W) BW_MiB/s(R/W)   Inflights Util%
------- ---- ----- ---------- ------- --------------- ----------- --------------- --------- ------
nvme7n1 2    1     0          0       0.00/0.00       0.00/0.00   0.00/0.00       0         0.00
nvme7n2 1    2     0          0       302.41k/302.94k 0.83/0.82   1180.40/1182.39 497       100.69

---------- Path Performance ----------

NSHead  NSID NSPath    Nodes Ctrl  IOPS(R/W)       Lat_ms(R/W) BW_MiB/s(R/W)   Inflights Util%
------- ---- --------- ----- ----- --------------- ----------- --------------- --------- ------
nvme7n1  2   nvme7c3n1 0,2-3 nvme3 0.00/0.00       0.00/0.00   0.00/0.00       0         0.00
nvme7n2  1   nvme7c3n2   0   nvme3 0.00/0.00       0.00/0.00   0.00/0.00       0         0.00
  -->    1   nvme7c7n2  2-3  nvme7 302.42k/302.94k 0.83/0.81   1180.36/1183.35 495       100.69

------------ Path Health -------------

NSPath    ANAState  Retries Failovers Errors
--------- --------- ------- --------- ------
nvme7c3n1 optimized 0       0         0
nvme7c3n2 optimized 0       0         0
nvme7c7n2 optimized 0       0         0

---------- Controller Summary --------

Ctrl  Paths Node Trtype Address      State IOPS(R/W)       Lat_ms(R/W) BW_MiB/s(R/W)   Util%
----- ----- ---- ------ ------------ ----- --------------- ----------- --------------- ------
nvme3 2     0    pcie   052e:78:00.0 live  0.00/0.00       0.00/0.00   0.00/0.00       0.00
nvme7 1     2    pcie   058e:78:00.0 live  302.42k/302.94k 0.83/0.81   1180.36/1183.35 100.69

--------------------------------------
[ESC to go back to the previous screen, q to quit]
```

Level 2: Dashboard (non-multipath)



```
---- nvme-top - Refresh: 1 Second ---
nvme-subsys2 - NQN=nvmet_subsystem
               hostnqn=nqn.2014-08.org.nvmexpress:uuid:41528538-e8ad-4eaf-84a7-9c552917d988
               iopolicy=numa

----------- Namespace Stat -----------

Namespace NSID Ctrl  Retries Errors IOPS(R/W)     Lat_ms(R/W) BW_MiB/s(R/W) Inflights Util%
--------- ---- ----- ------- ------ ------------- ----------- ------------- --------- -----
nvme2n1   1    nvme2 0       0      21.37k/21.25k 11.92/12.04 82.83/82.83   512       99.79
nvme4n1   1    nvme4 0       0      21.56k/21.77k 11.79/11.80 83.83/84.83   507       99.80

---------- Controller Summary --------

Ctrl  Node        Trtype Address                                    State IOPS(R/W)     Lat_ms(R/W)  BW_MiB/s(R/W) Util%
----- ----------- ------ ------------------------------------------ ----- ------------- ------------ ------------- -----
nvme2 NUMA_NO_NODE tcp    traddr=127.0.0.2,trsvcid=4420,src_addr=127.0.0.1 live  21.37k/21.25k 11.92/12.04  82.83/82.83   99.79
nvme4 NUMA_NO_NODE tcp    traddr=127.0.0.3,trsvcid=4420,src_addr=127.0.0.1 live  21.56k/21.77k 11.79/11.80  83.83/84.83   99.80


--------------------------------------
[ESC to go back to the previous screen, q to quit]
```

Implementation Details

- Pure termios-based UI
    - No ncurses dependency
    - Low overhead
- Why not ncurses?
    - Simpler deployment
    - No external libraries
    - Faster redraw for small dashboards
    - Easier integration into nvme-cli
- Uses ANSI escape codes for:
    - Cursor movement
    - Screen clearing
    - Row highlighting
- Efficient delta (time interval) based redraw
- Stats collection via libnvme / sysfs / ioctls

Call for Feedback

- What metrics are missing/redundant ?
- Suggestions for interactive features?
- How should we handle large fabrics (>50 controllers)?
- Update existing libnvme APIs ?
  - fetch the latest attribute value (instead of re-using cached value)
  - introduce new APIs for real-time update?
- Polling vs. Notification-based Models
  - Acceptable overhead?
  - Should nvme-top wait for kernel events (inotify/fanotify or uevents)?
- Keep simple ASCII TUI?
  - Support curses / btop-style interface?
  - Exporter mode for Prometheus/Grafana?
- Qdepth is per-controller attribute
  - Fix kernel to export qdept through controller sysfs attribute instead of per-path sysfs attribute?
- Modify kernel to export following error counters?
  - Failovers
  - Errors
  - Requeue-IOs
  - Fail-IOs

Summary

- **nvme-top aims to:**
  - Provide real-time NVMe visibility
  - Multipath-aware dynamic drill-down
  - Improve debugging of NVMe-oF multipath
  - Highlight path/controller imbalances
  - Enable fast triage during fabric issues
  - Lightweight terminal UI
  - A practical debugging tool for NVMe-MP deployments

- **The goal of this BoF is to shape its future direction together with the community.**

THANK YOU!

Feedback and collaboration are welcome!

BACKUP

Existing Gaps

Tool	Limitation
nvme-cli	Static data, not real-time
iostat	Not NVMe multipath/topology aware
iotop	Not NVMe multipath/topology aware
Perf/eBPF	Low level not user friendly

Missing: A tool combining *top-style interactivity* with *NVMe-specific awareness*.

Why nvme-cli Alone Is Not Enough

Today:

- *nvme list*
- *nvme list-subsys*
- *nvme get-log*
- *nvme ana-log*

…but:

- Each output is **static**
- Structure is rebuilt from sysfs **once per invocation**
- No persistent in-memory state
- No incremental updates
- No live reactions to changing multipath conditions

To monitor fabric NVMe reliably:

We need *continuous updates*.

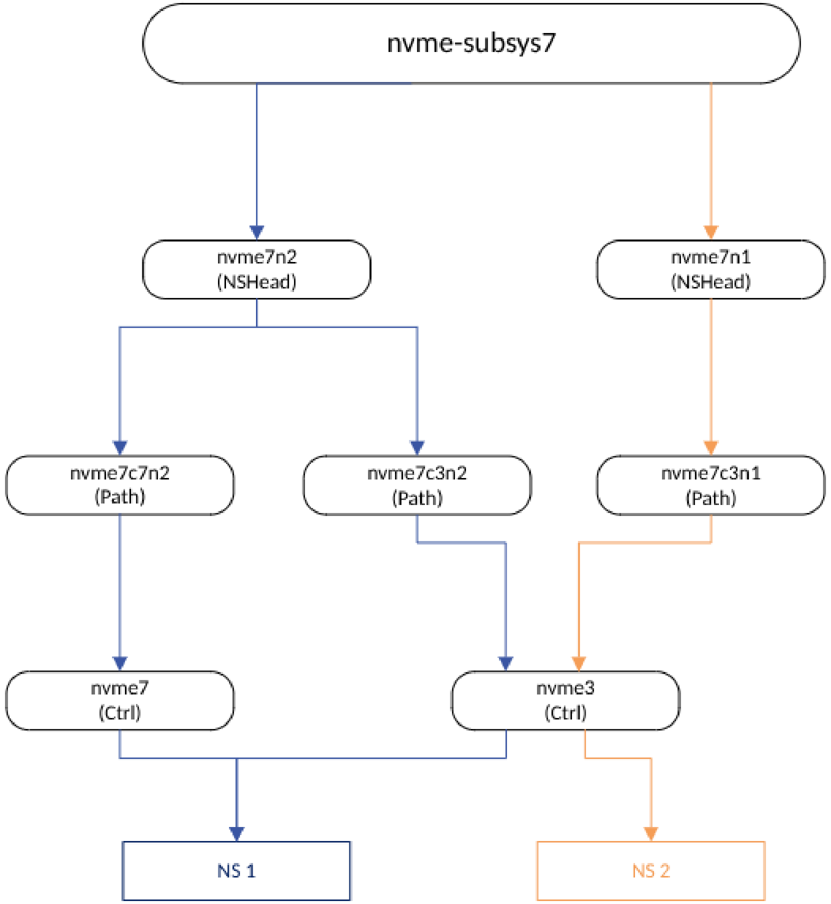
What is nvme-top?

- A proposed tool that provides:
  - Real-time NVMe monitoring (e.g., refresh every second or at configured interval)
  - Continuously updating dashboard, similar to `top` / `iotop`
  - NVMe-aware views: subsystem, namespace, path, controller

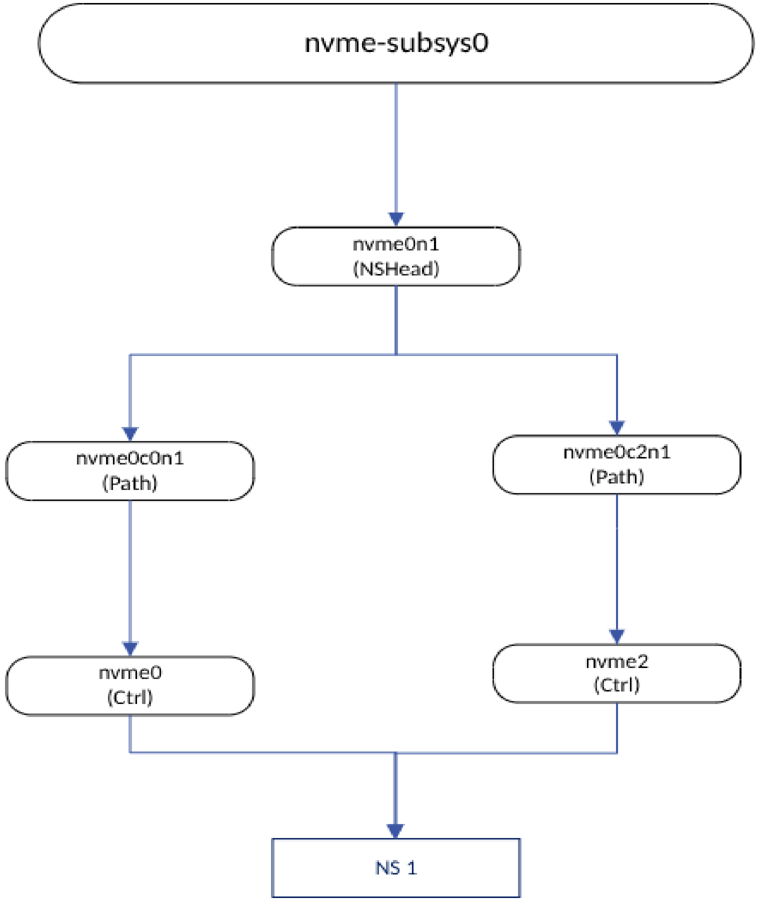
- Multipath intelligence:
  - ANA state
  - NUMA node affinity
  - Queue depth (QDepth)
  - Per-path latency + bandwidth

- Goal: Improve operational visibility and reduce debugging time.

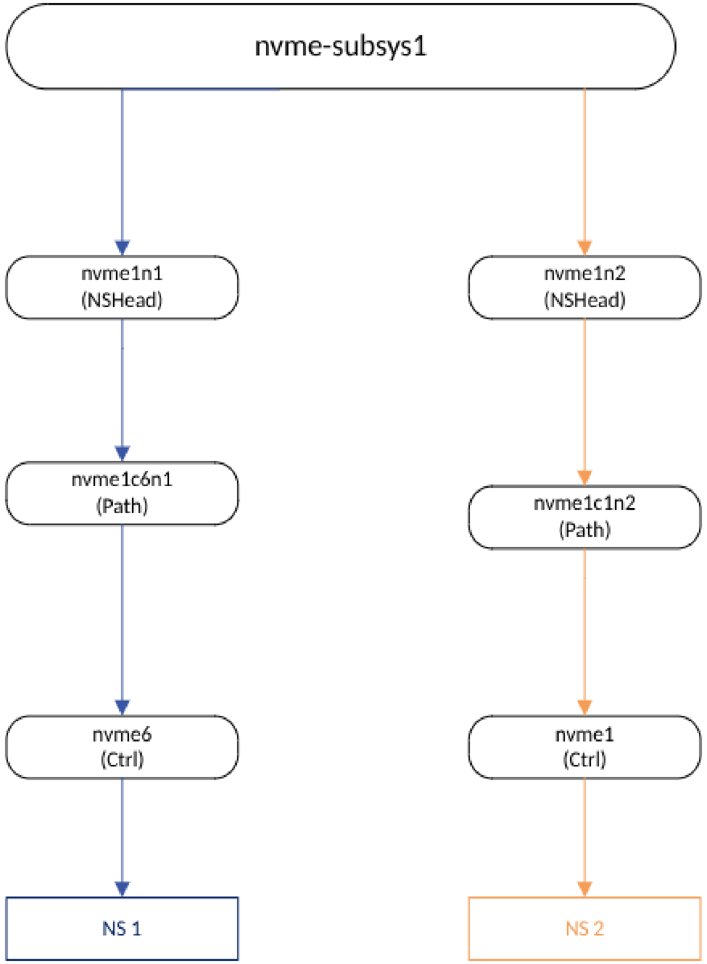
Topology multipath (nvme-subsys7)



Topology multipath (nvme-subsys0)



Topology multipath (nvme-subsys1)



nvme-subsys1

nvme1n1
(NSHead)

nvme1c6n1
(Path)

nvme6
(Ctrl)

NS 1

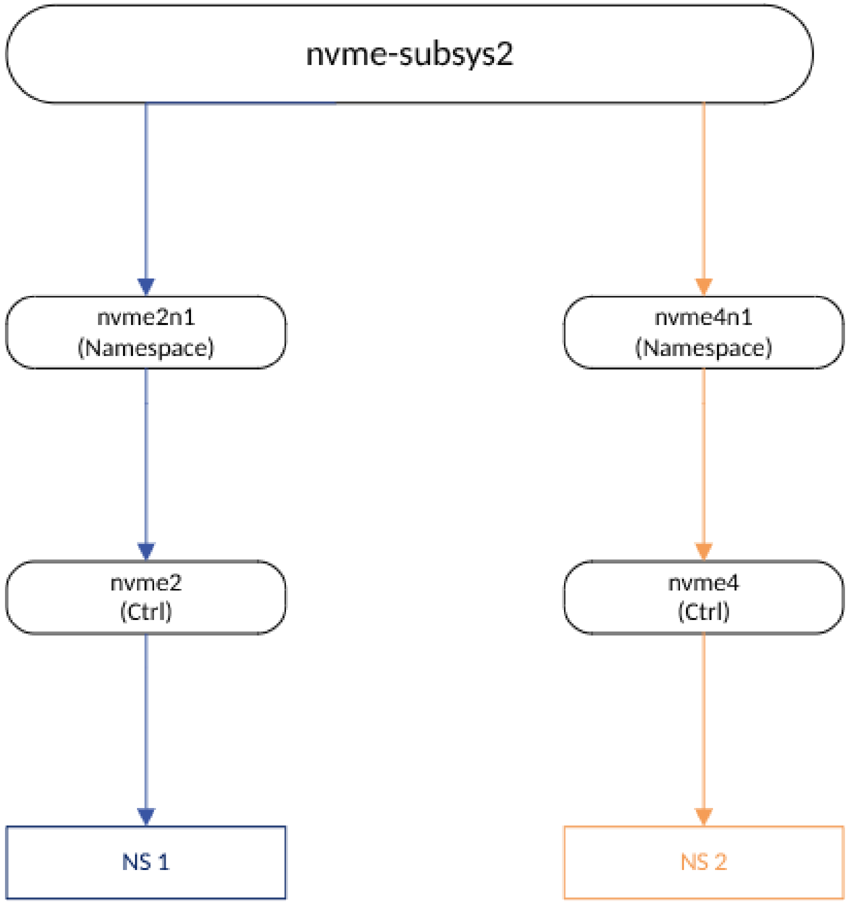
nvme1n2
(NSHead)

nvme1c1n2
(Path)

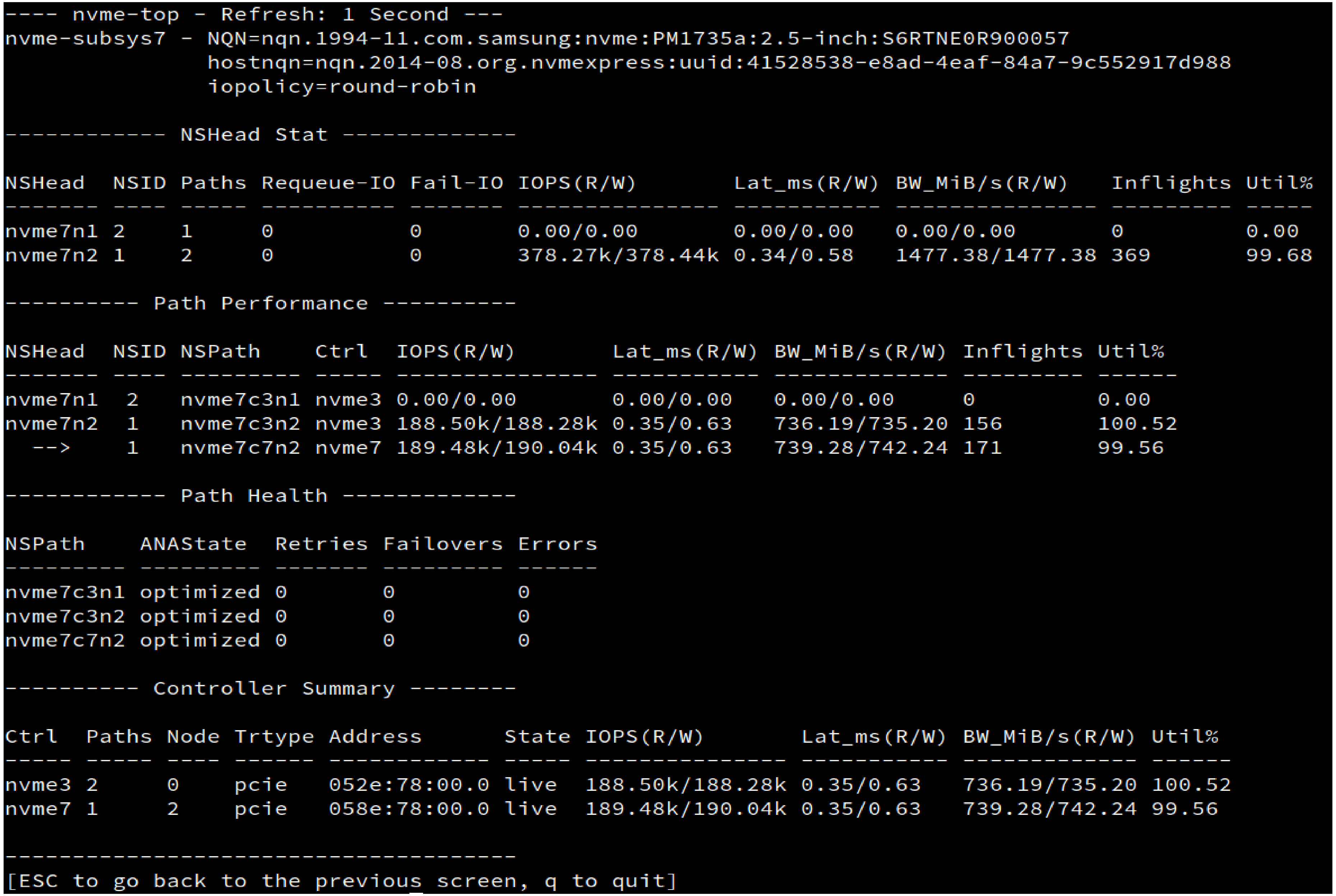
nvme1
(Ctrl)

NS 2

Topology non-multipath (nvme-subsys2)

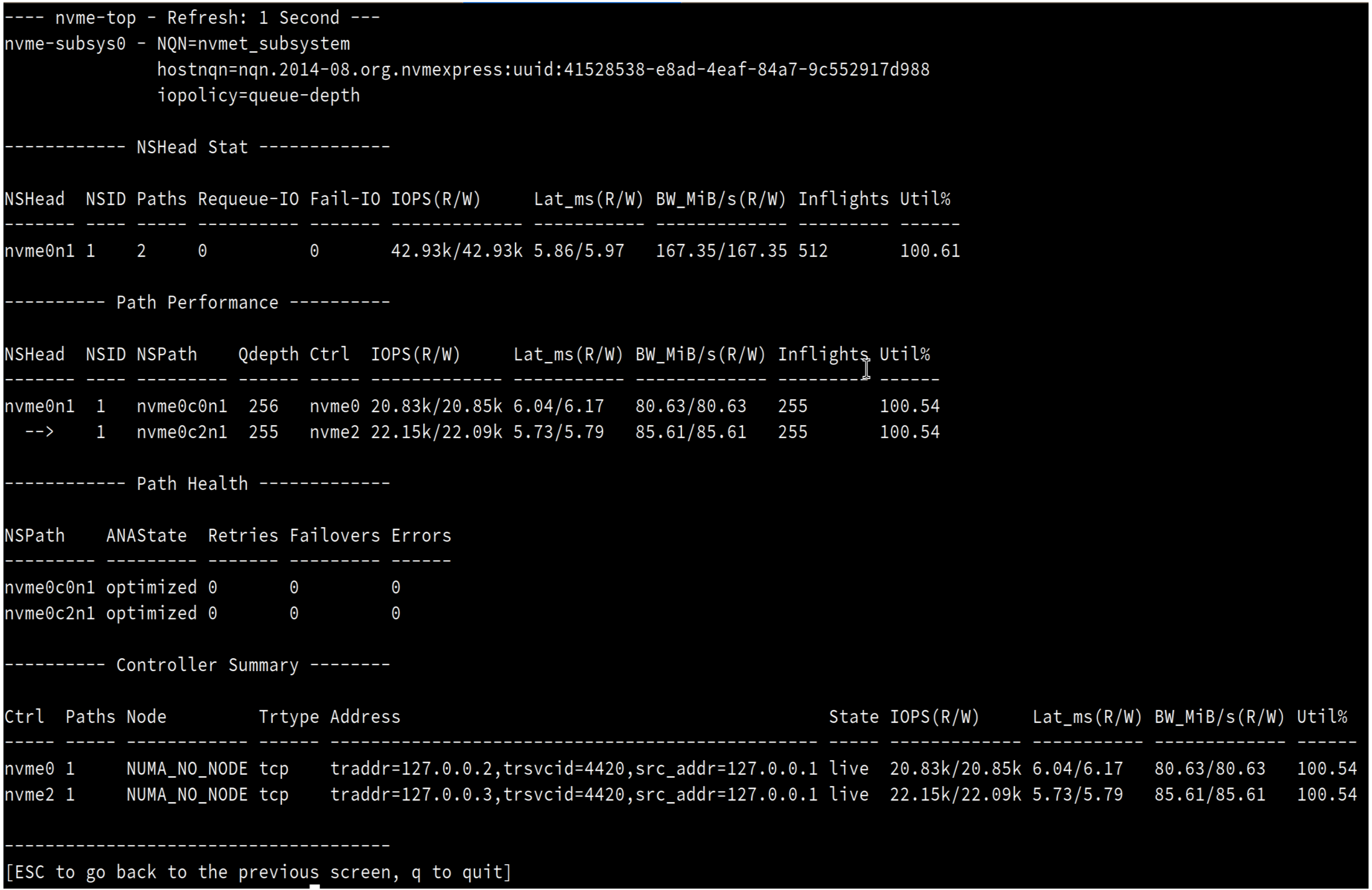


Level 2: Drill-Down (iopolicy=round-robin)



```
---- nvme-top - Refresh: 1 Second ---
nvme-subsys7 - NQN=nqn.1994-11.com.samsung:nvme:PM1735a:2.5-inch:S6RTNE0R900057
               hostnqn=nqn.2014-08.org.nvmexpress:uuid:41528538-e8ad-4eaf-84a7-9c552917d988
               iopolicy=round-robin

------------ NSHead Stat -------------

NSHead  NSID Paths Requeue-IO Fail-IO IOPS(R/W)        Lat_ms(R/W) BW_MiB/s(R/W)    Inflights Util%
------- ---- ----- ---------- ------- ---------------- ----------- ---------------- --------- -----
nvme7n1 2    1     0          0       0.00/0.00        0.00/0.00   0.00/0.00        0         0.00
nvme7n2 1    2     0          0       378.27k/378.44k  0.34/0.58   1477.38/1477.38  369       99.68

---------- Path Performance ----------

NSHead  NSID NSPath    Ctrl  IOPS(R/W)        Lat_ms(R/W) BW_MiB/s(R/W) Inflights Util%
------- ---- --------- ----- ---------------- ----------- ------------- --------- ------
nvme7n1  2   nvme7c3n1 nvme3 0.00/0.00        0.00/0.00   0.00/0.00     0         0.00
nvme7n2  1   nvme7c3n2 nvme3 188.50k/188.28k  0.35/0.63   736.19/735.20 156       100.52
  -->    1   nvme7c7n2 nvme7 189.48k/190.04k  0.35/0.63   739.28/742.24 171       99.56

------------ Path Health -------------

NSPath    ANAState  Retries Failovers Errors
--------- --------- ------- --------- ------
nvme7c3n1 optimized 0       0         0
nvme7c3n2 optimized 0       0         0
nvme7c7n2 optimized 0       0         0

---------- Controller Summary --------

Ctrl  Paths Node Trtype Address      State IOPS(R/W)        Lat_ms(R/W) BW_MiB/s(R/W) Util%
----- ----- ---- ------ ------------ ----- ---------------- ----------- ------------- ------
nvme3 2     0    pcie   052e:78:00.0 live  188.50k/188.28k  0.35/0.63   736.19/735.20 100.52
nvme7 1     2    pcie   058e:78:00.0 live  189.48k/190.04k  0.35/0.63   739.28/742.24 99.56

--------------------------------------
[ESC to go back to the previous screen, q to quit]
```

Level 2: Drill-Down (iopolicy=qdepth)



```
---- nvme-top - Refresh: 1 Second ---
nvme-subsys0 - NQN=nvmet_subsystem
               hostnqn=nqn.2014-08.org.nvmexpress:uuid:41528538-e8ad-4eaf-84a7-9c552917d988
               iopolicy=queue-depth

------------ NSHead Stat -------------

NSHead  NSID Paths Requeue-IO Fail-IO IOPS(R/W)     Lat_ms(R/W) BW_MiB/s(R/W) Inflights Util%
------- ---- ----- ---------- ------- ------------- ----------- ------------- --------- ------
nvme0n1 1    2     0          0       42.93k/42.93k 5.86/5.97   167.35/167.35 512       100.61

---------- Path Performance ----------

NSHead  NSID NSPath    Qdepth Ctrl  IOPS(R/W)     Lat_ms(R/W) BW_MiB/s(R/W) Inflights Util%
------- ---- --------- ------ ----- ------------- ----------- ------------- --------- ------
nvme0n1  1   nvme0c0n1  256   nvme0 20.83k/20.85k 6.04/6.17   80.63/80.63   255       100.54
  -->    1   nvme0c2n1  255   nvme2 22.15k/22.09k 5.73/5.79   85.61/85.61   255       100.54

------------ Path Health -------------

NSPath    ANAState  Retries Failovers Errors
--------- --------- ------- --------- ------
nvme0c0n1 optimized 0       0         0
nvme0c2n1 optimized 0       0         0

---------- Controller Summary --------

Ctrl  Paths Node         Trtype Address                                          State IOPS(R/W)     Lat_ms(R/W) BW_MiB/s(R/W) Util%
----- ----- ------------ ------ ------------------------------------------------ ----- ------------- ----------- ------------- ------
nvme0 1     NUMA_NO_NODE tcp    traddr=127.0.0.2,trsvcid=4420,src_addr=127.0.0.1 live  20.83k/20.85k 6.04/6.17   80.63/80.63   100.54
nvme2 1     NUMA_NO_NODE tcp    traddr=127.0.0.3,trsvcid=4420,src_addr=127.0.0.1 live  22.15k/22.09k 5.73/5.79   85.61/85.61   100.54

--------------------------------------
[ESC to go back to the previous screen, q to quit]
```