
ceph days
LONDON 2026

NVMe/TCP CSI Driver (Dev Preview)

Introduction

- NVMeoF requires a Ceph CSI driver for Open Shift and general Kubernetes user consumption
  - Dev Preview already merged to ceph csi project!
  - There are some limitations to the Dev Preview (not prod ready)
  - Rook does not support deployment of NVMeoF GWs yet (WIP)

Current status (Upstream)

- Code is already available here - https://github.com/ceph/ceph-csi
- There is no official ceph-csi build that includes nvmeof-csi driver yet
- Work in progress to add testing, security features, QoS and more

SPDK TOP Tool

```
nvmeof-top  Address: 192.168.1.1      Group Name: demo      Version: 1.2.0     LB Group: 1     Subsystems: 2    Namespaces: 20          2025-04-30 06:06:23
Reactor Cores:  4   Total CPU:  249%   AVG CPU: 62.2%   Min CPU:  48%   Max CPU:  83%
Subsystem: nqn.2016-06.io.spdk:cnode1                Total IOPS:   25,867    Throughput:  194.69 MiB/s   Namespaces: 10/128     Clients: 1/2

NSID   RBD pool/image                          IOPS      r/s    rMB/s   r_await   rareq-sz      w/s    wMB/s   w_await   wareq-sz   LBGrp   QoS
   1   rbd/cnode1_image_1                      3697     3316    25.91      0.30      16.00      381     1.49      2.62      16.00       1     No
   2   rbd/cnode1_image_2                         0        0     0.00      0.00       0.00        0     0.00      0.00       0.00       2     No
   3   rbd/cnode1_image_3                      7721     5825    91.02      0.17      16.00     1896    14.81      0.53       8.00       1     No
   4   rbd/cnode1_image_4                         0        0     0.00      0.00       0.00        0     0.00      0.00       0.00       2     No
   5   rbd/cnode1_image_5                      2570     2128    16.62      0.47      16.00      442     3.45      2.26       8.00       1     No
   6   rbd/cnode1_image_6                         0        0     0.00      0.00       0.00        0     0.00      0.00       0.00       2     No
   7   rbd/cnode1_image_7                      2075     1603     6.26      0.62      16.00      472     7.38      2.12       4.00       1     No
   8   rbd/cnode1_image_8                         0        0     0.00      0.00       0.00        0     0.00      0.00       0.00       2     No
   9   rbd/cnode1_image_9                      9804     7024    54.88      0.14       8.00     2780    21.72      0.36       4.00       1     No
  10   rbd/cnode1_image_10                        0        0     0.00      0.00       0.00        0     0.00      0.00       0.00       2     No
```

Ceph NVMeoF CLI tool to assist with exposing Namespace Image performance and availability details

- Visualize performance and QoS stats via CLI for ease and speed of troubleshooting

Default NVMeoF Metadata Pool (WIP)

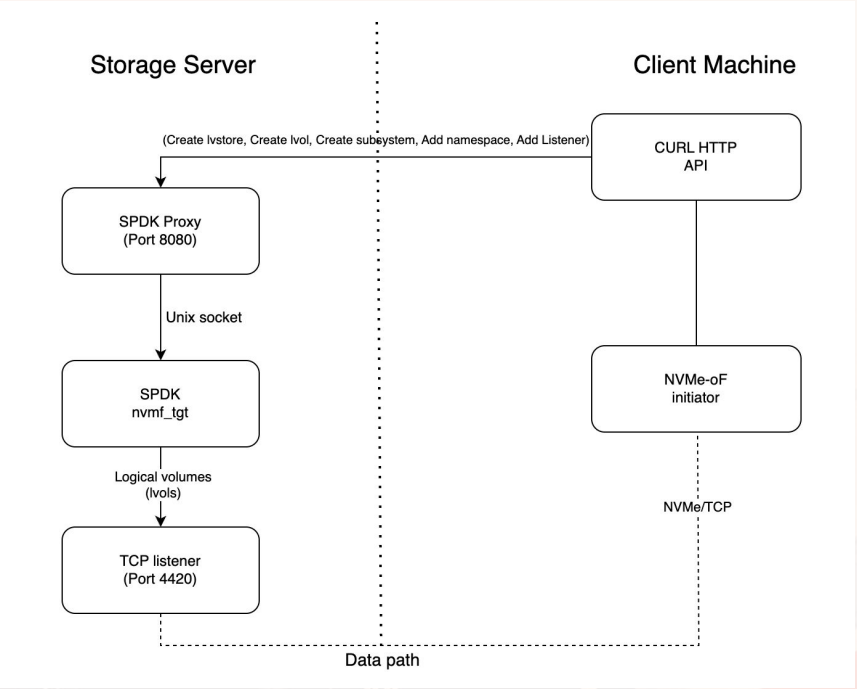
- Introduction
  - NVMeoF service requires a metadata pool to store service configuration details
    - Mandatory to enable NVMeoF service
    - Created only once per service, not a user data pool
    - Not intended to be user managed

- Current status (Upstream)
  - In Development

Listener Auto-creation (WIP)

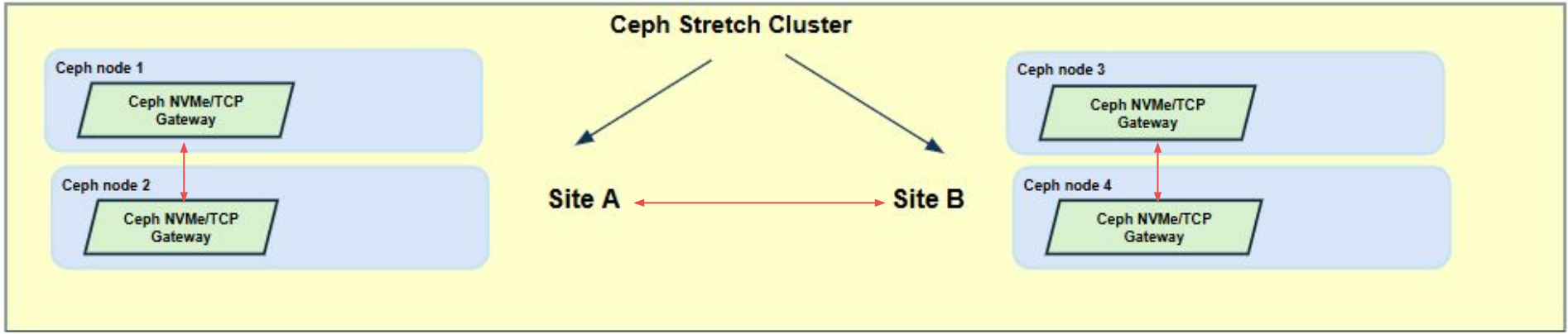
A “Listener” in the nvmeof GW is the IP address and port that will be used to receive the NVMe commands and IO.

- Every GW in the group must define at least one Listener per Subsystem, manually defined
- Users provide a CIDR (IP network mask + range)
  - E.g subsystem add -n nqn.2016-06.io.spdk:cnode21 --network-mask "10.242.64.0/24"
- Manual configuration still available
- The feature is backwards compatible with the “old” way of defining the listeners



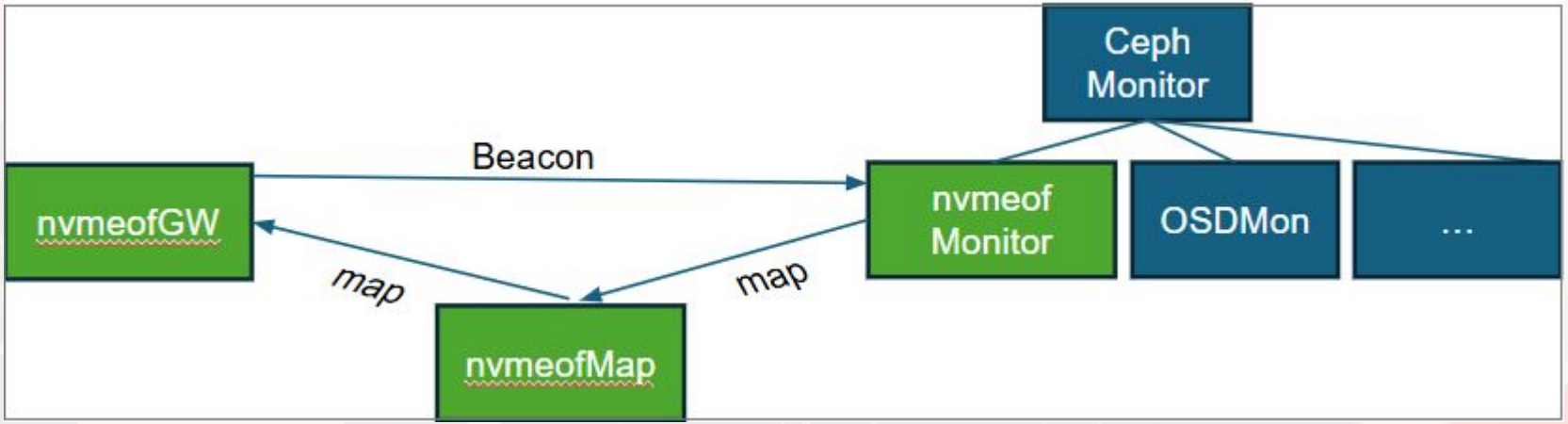
Stretch cluster optimization (WIP)

- Currently Gateways are unaware of cluster architecture and stretched clustering is not supported
- The following scenarios should be handled
  - Failover/Failback - select a GW that is in the same “location”
  - Namespace load balancing - optionally allow to set the preferred “location” per namespace, so that the auto lb will assign the right GW (i.e. ANA group) to that namespace
- Add an option to move GWs to standby (if user is interested to use the “other” location as a DR site only)
- Add an option to disable Auto failback and then manually trigger failback
- Add health alerts related to the new functionality



Shorten Failover Time

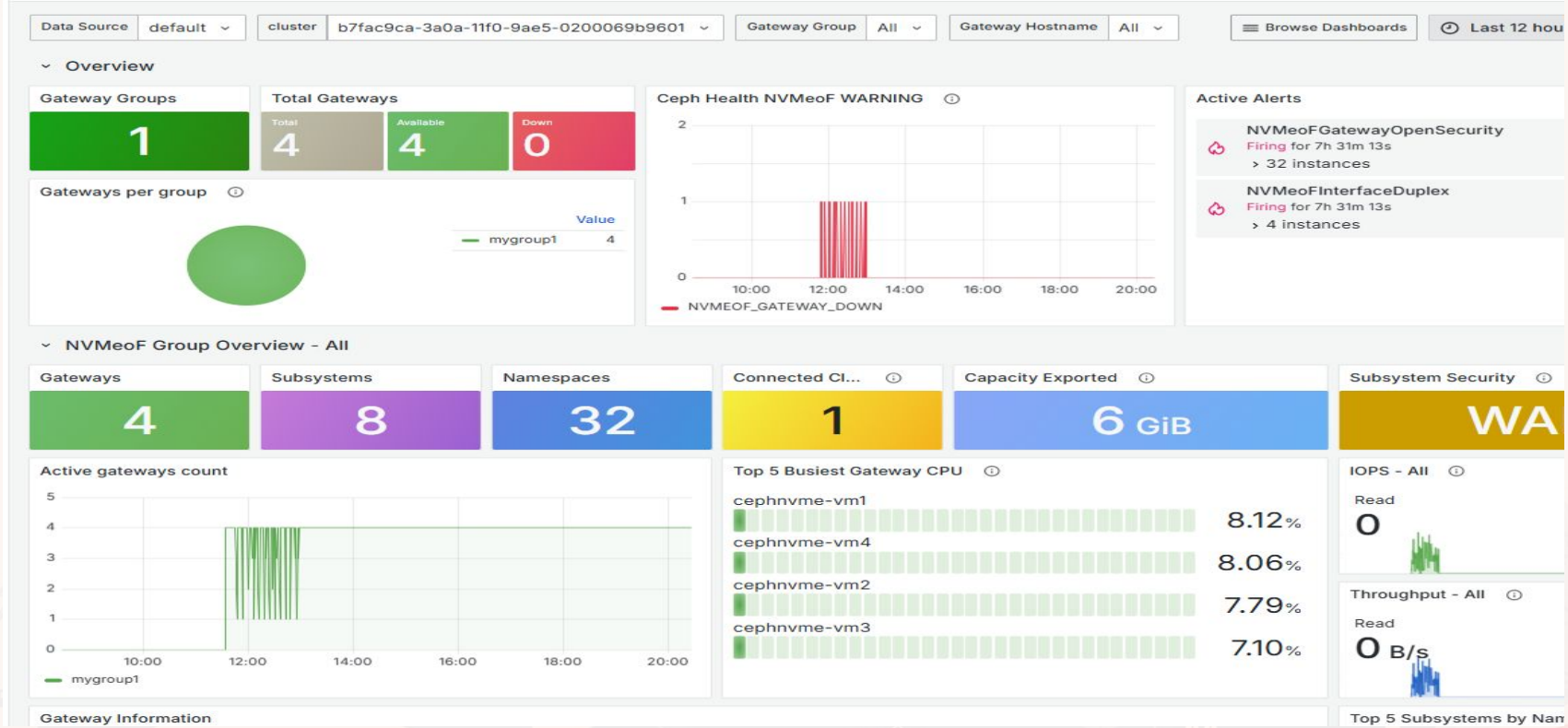
- Failover time between GWs is currently ~12 seconds.
- There is a requirement to shorten the time to ~4-6 seconds
- Currently nvmeof mon checks every 5 seconds, if last beacon received was <10sec
- There are few main changes that need to be made in the nvmeof mon and mon client to be able to have a better granularity of controlling the failover time, and to really be able to shorten the time



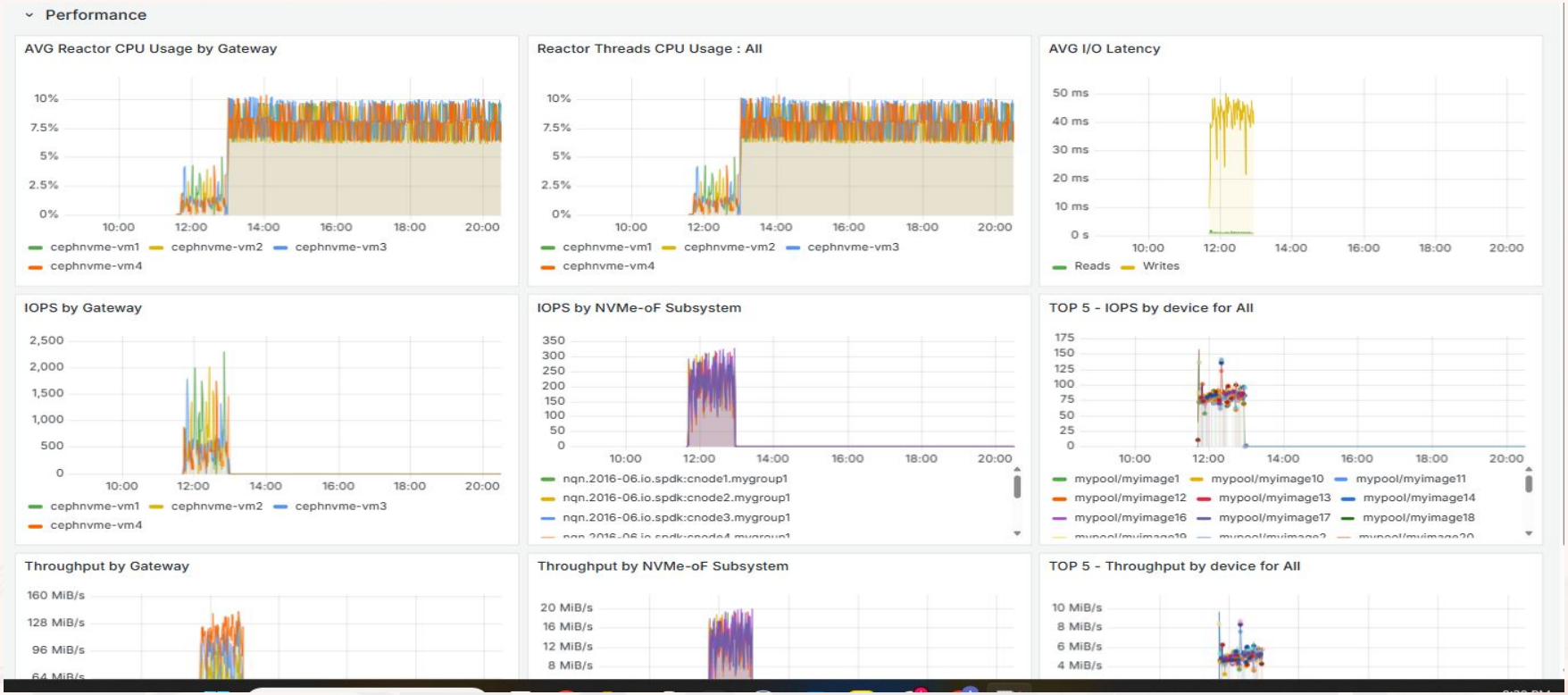
RBD Resuse of SPDK CRC (WIP/Reserching)

- This is a potential performance optimization.
  The idea is to reuse CRC calculated by SPDK, and to avoid recalculating the same CRC on data by RBD.
- This is not always helpful because:
  - The SPDK will only calculate CRC if the initiator connect with "--data-digest" option
  - The RBD cannot always reuse the CRC because in some cases the data is rearranged in a different way
- But this can be useful in many cases, specifically with small and aligned IO size
- WIP to check how much CPU we are saving
  - Initial results looks promising

Dashboard - NVMe-oF Gateways - Overview



Dashboard -NVMe-oF Gateways - Performance



ceph days
LONDON 2026

Questions?

ceph days LONDON 2026

Thank You!!!


ceph days
LONDON 2026