



# Unlocking Ceph's Potential with NVMe-oF Integration

Orit Wasserman  
Aviv Caro

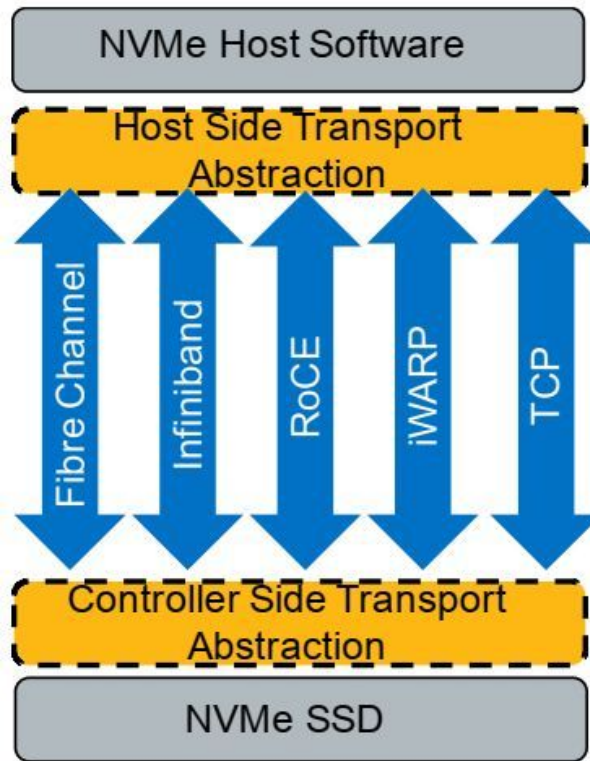


# NVMe Over Fabric

# NVMe Over Fabric



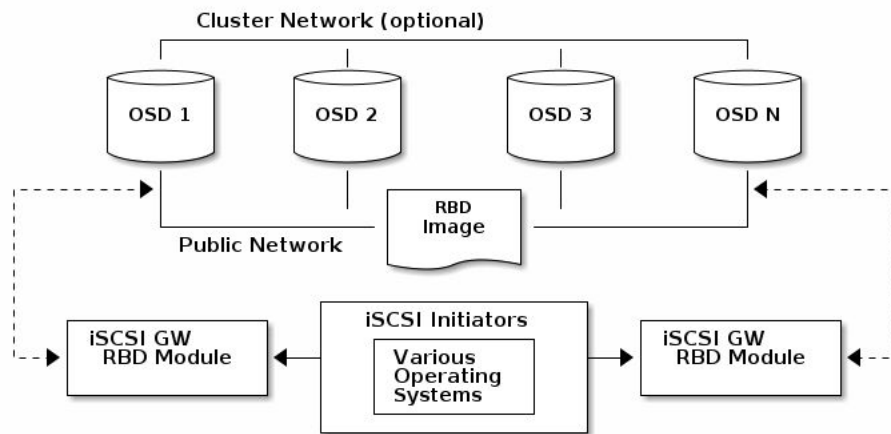
- Non Volatile Memory Express (NVMe)
  - PCIe attached storage
  - Local storage
- Expand NVMe efficiency and performance over network fabrics
- Eliminate unnecessary protocol translations
- Enable low-latency and high IOPS remote NVMe storage
- TCP:
  - Well-known and common transport
  - No networking infrastructure requirements and constraints
  - Ratified Nov, 2018



# Why not Ceph iSCSI?



- Legacy
- Performance
  - Higher throughput and IOPS
    - 30-70%
  - Reduced latency
    - 30-40%
- Reduced CPU usage
  - 30-40%
- Scalability



# Why NVMe-over-Fabrics?



## RADOS Block Device (RBD)

- RADOS protocol
- Distributed n-to-m protocol
- Reliable object access to sharded and replicated/erasure coded storage

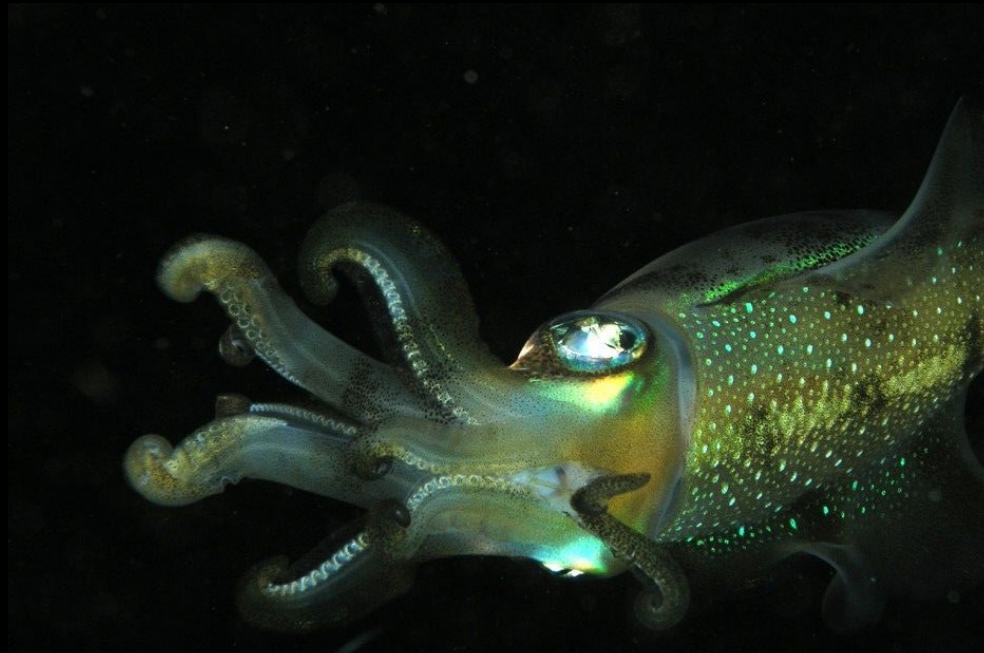
## Why do we need another protocol to access block storage in Ceph?

### NVMe-over-Fabrics (NVMe-oF)

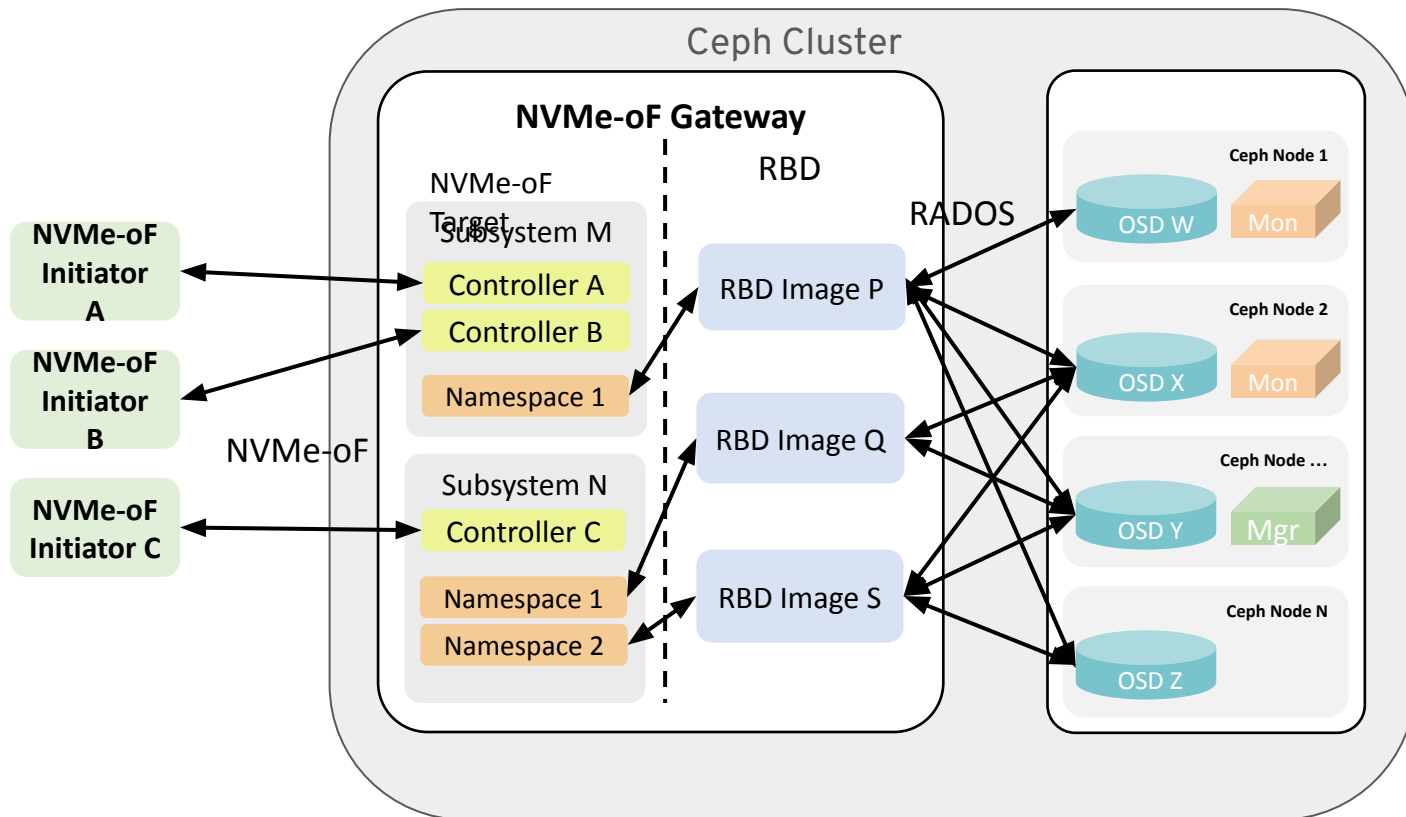
- Open, widely adopted ***industry standard***
- Enable use-cases where NVMe-oF is already part of ***ecosystem***
- Take advantage of NVMe-oF ***offloading*** in DPUs



# Ceph NVMe-of Gateway



# NVMe-oF and Ceph

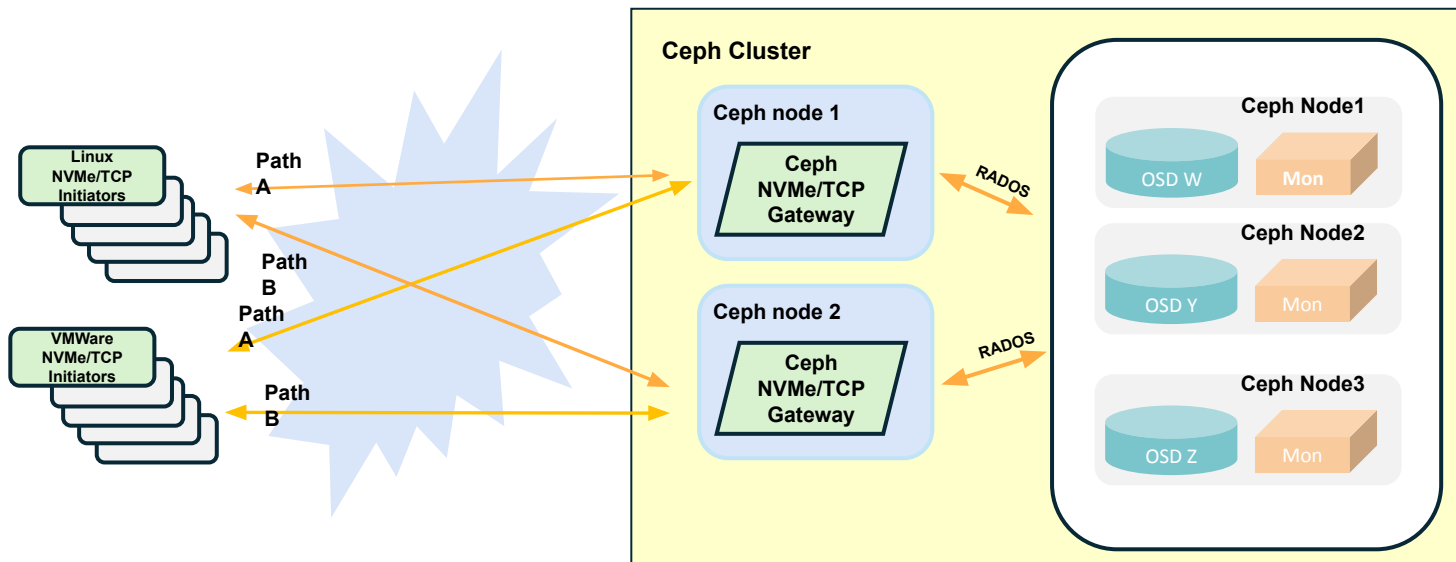


- Namespace mapped to an RBD image
- Subsystems logical grouping of Namespaces
- Each initiator get a Controller

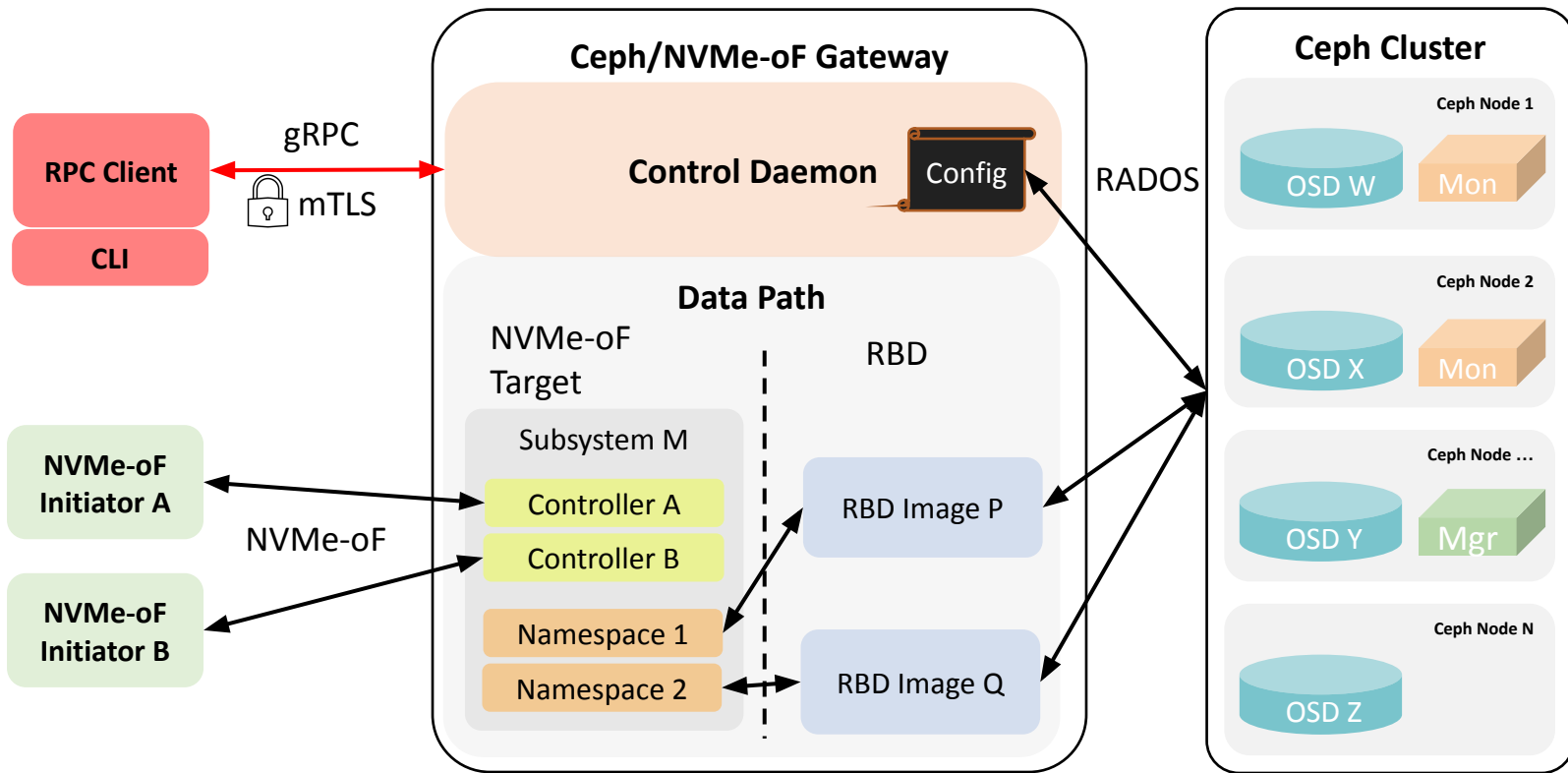
# Ceph NVMe/TCP Gateway



- Multiple GWs can be deployed on the same Ceph cluster to provide HA, and Load balancing
- Multiple NVMe Subsystems to allow Access control
- Both Linux and ESX initiators



# NVMe-oF Gateway Configuration

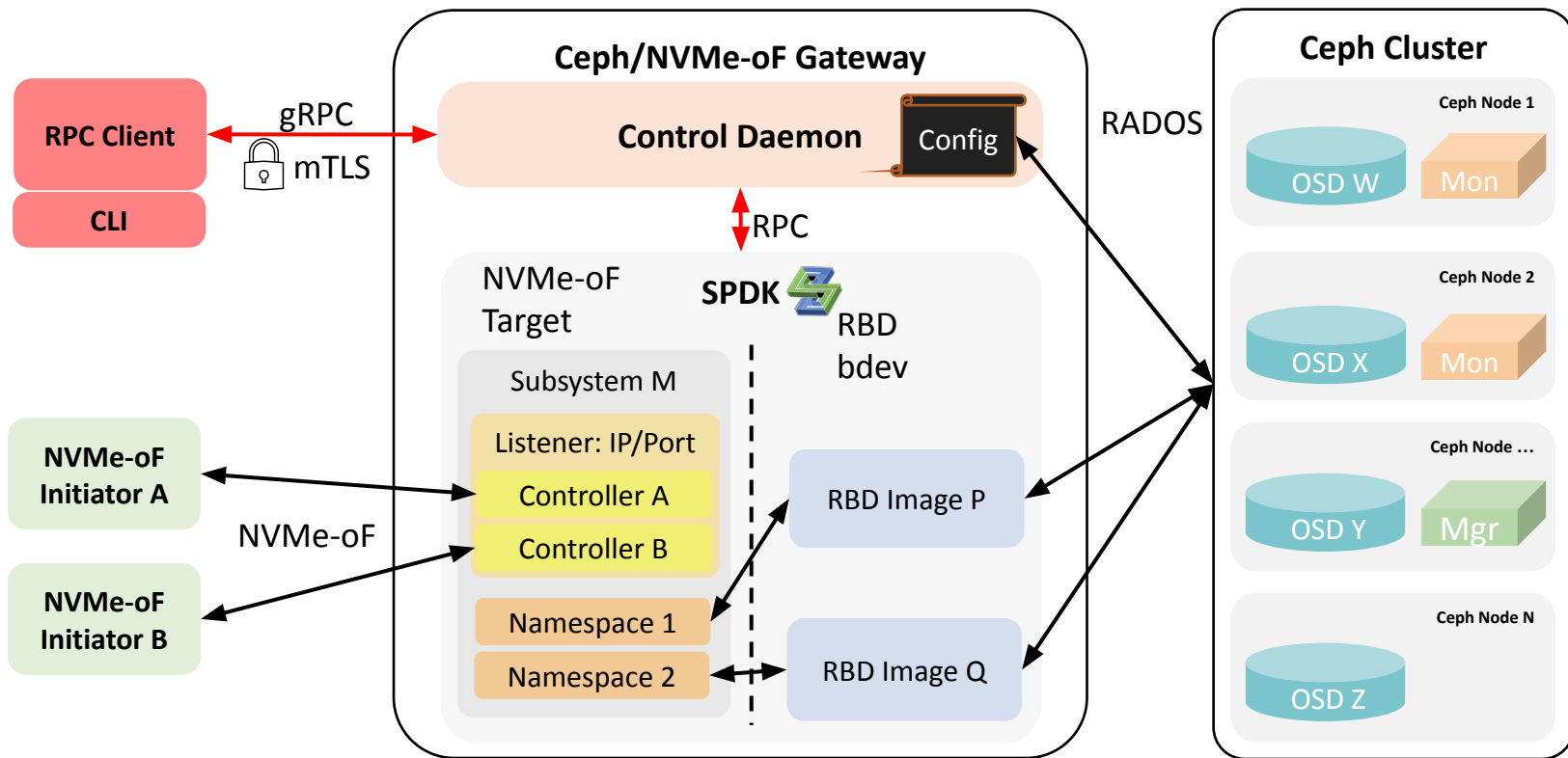




- Storage Performance Development Kit (SPDK)
- <https://spdk.io/>
- Provides a tools and libraries for writing high performance, scalable, user-mode storage applications
- Userspace NVMe Over Fabric target
- Support for Ceph RBD with bdev\_rbd
- Open source (BSD)



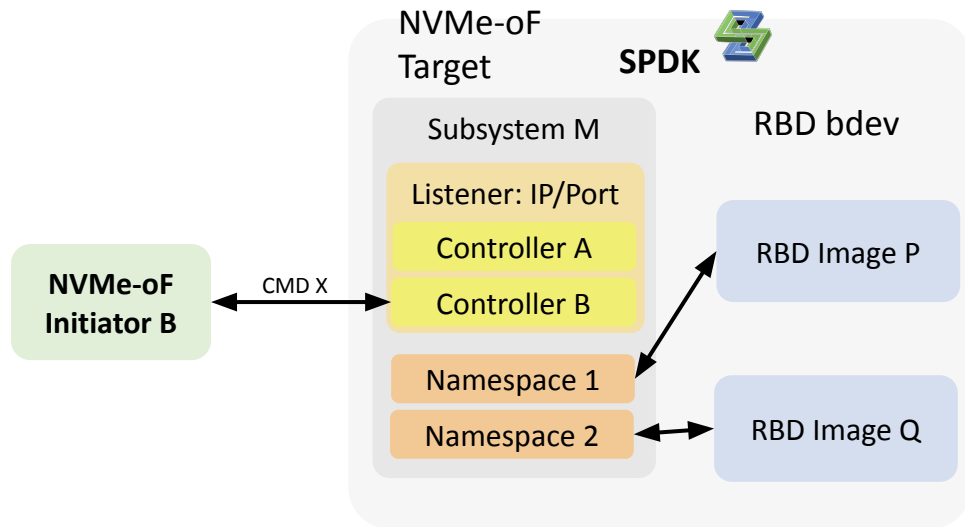
# Data Path: SPDK



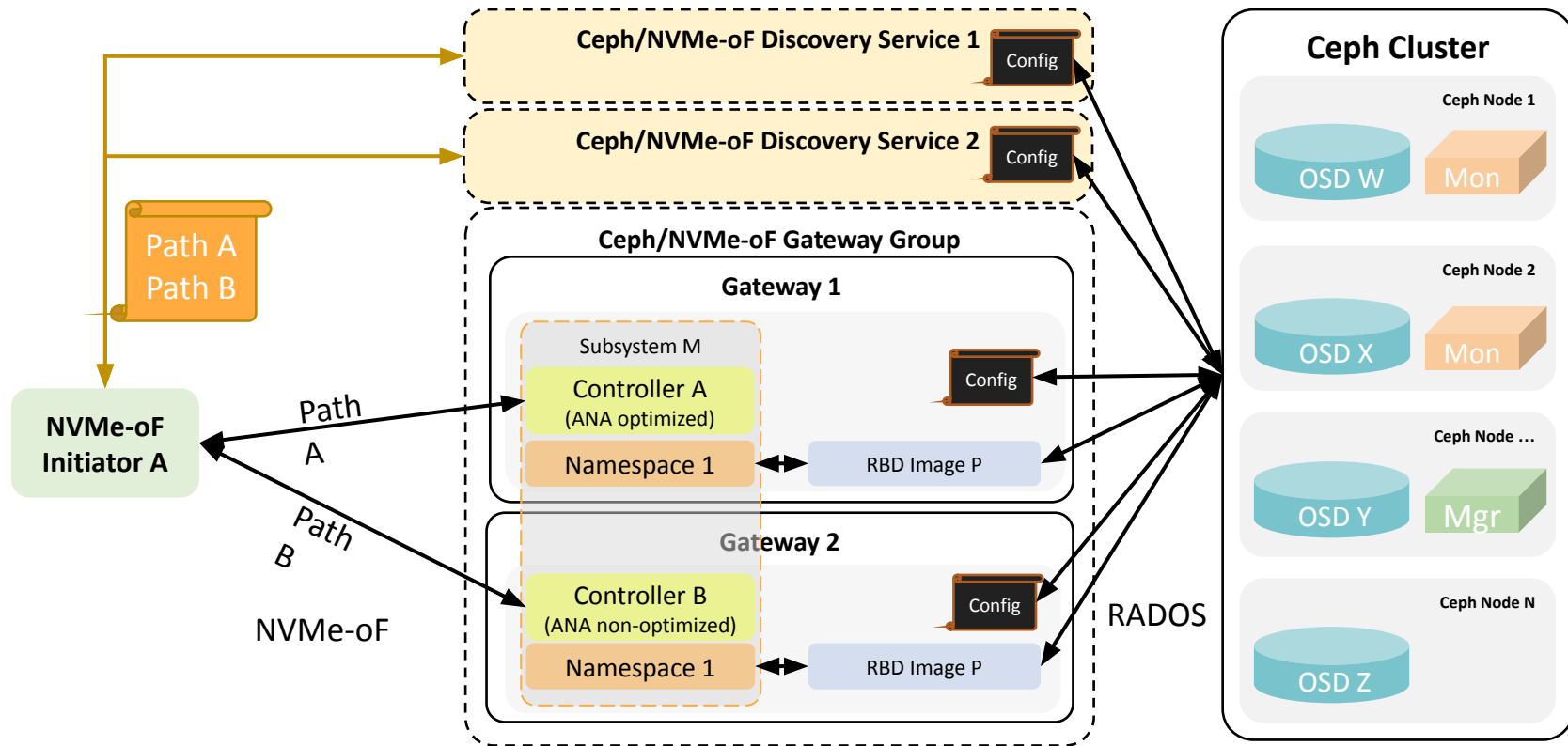
# NVMe to RBD Commands Mapping



- RBD backend in SPDK maps NVMe operations to RBD API
- Natively supported
  - Read
  - Write
  - Unmap
  - Flush
  - Write zeroes
  - **Compare and write**
- Emulated
  - Compare
  - Copy
  - **Abort\*\***



# Discovery





- NVMe Discovery service is available as a part of each gateway
  - It has access to the OMAP configuration
  - The Discovery service on all GWs will report the same data, allowing to use any Discovery service.
  - It is assigned to port 8009 (the default NVMe discovery port).
- Query will return a list of all available gateways IPs in the GW group
- Discover and Connect in one command:
  - Linux: `nvme connect-all`
  - ESX: `esxcli nvme fabrics discover -c`
- It is highly recommended to use these commands, and not manually connect the host to each GW separately.



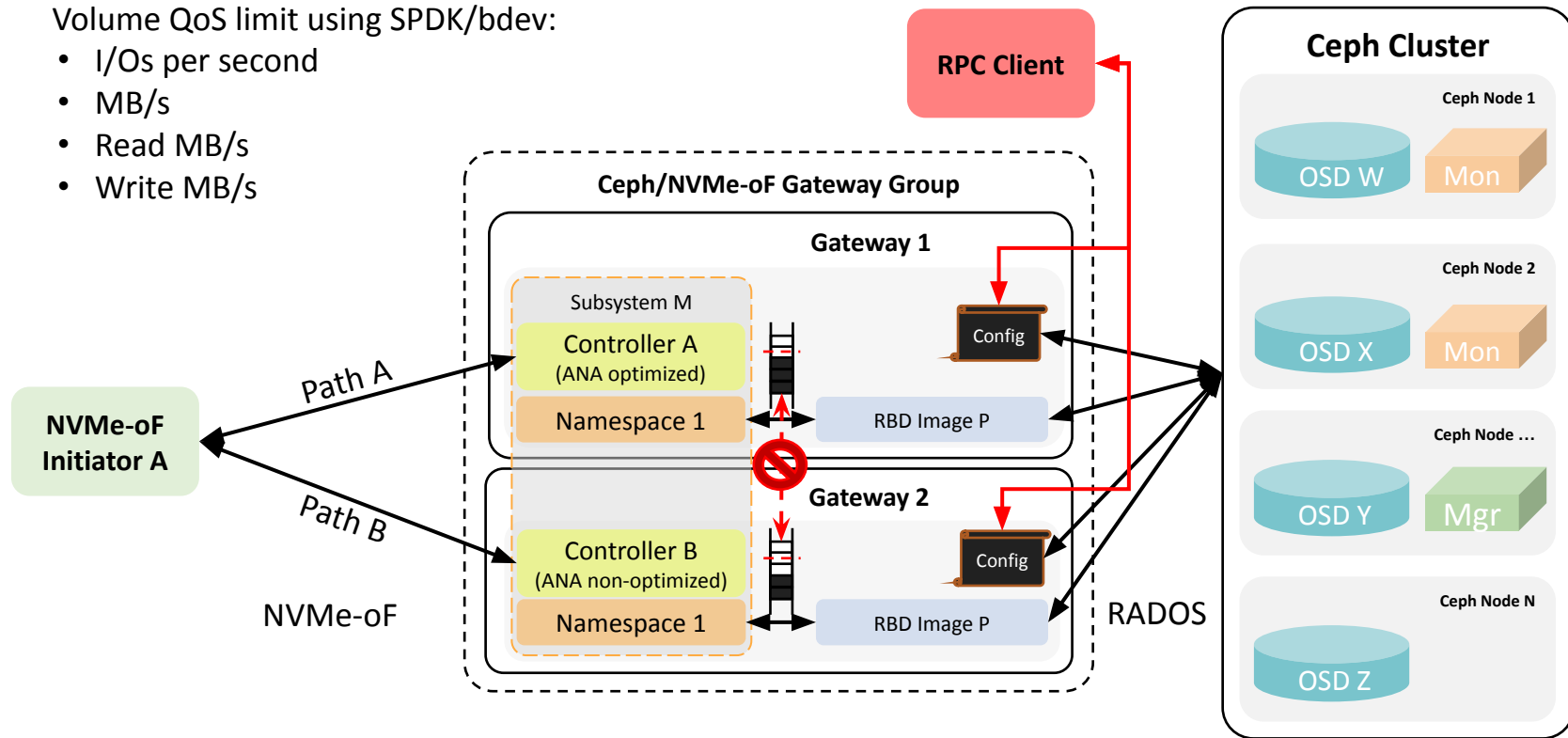
- Use shared volumes to create single storage pool
- VMware vSphere Storage APIs – Array Integration (VAAI):  
Set of storage primitives that enable storage offloading
  - **Atomic Test & Set (ATS)**
    - Support in using NVMe compare & write fused operation
    - Cmp & write limited to RBD object size resp. stripe size (alignment\*) => only 4K/1 block required
  - **XCOPY (extended copy)**
    - Copy NVMe command is supported in SPDK but QoS difficult
  - **Write same (zero)**
    - Write zeroes NVMe command => maps directly to RBD operation
  - **Unmap (delete)**
    - Supported as dataset management command => discard in RBD

# Quality of Service (QoS)



Volume QoS limit using SPDK/bdev:

- I/Os per second
- MB/s
- Read MB/s
- Write MB/s

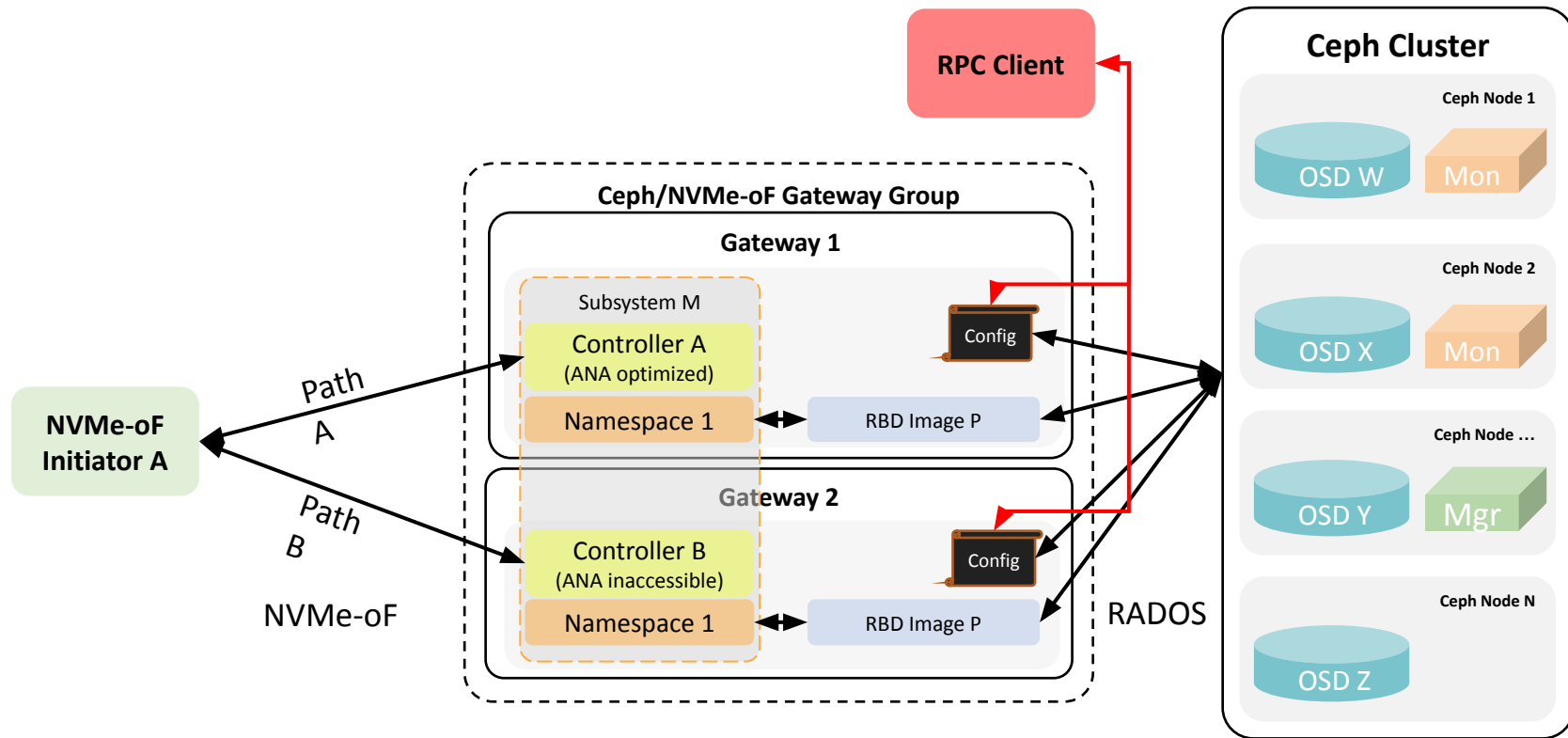


\*Global QoS across gateways in a group is not planned



**NVMe-oF High Availability**

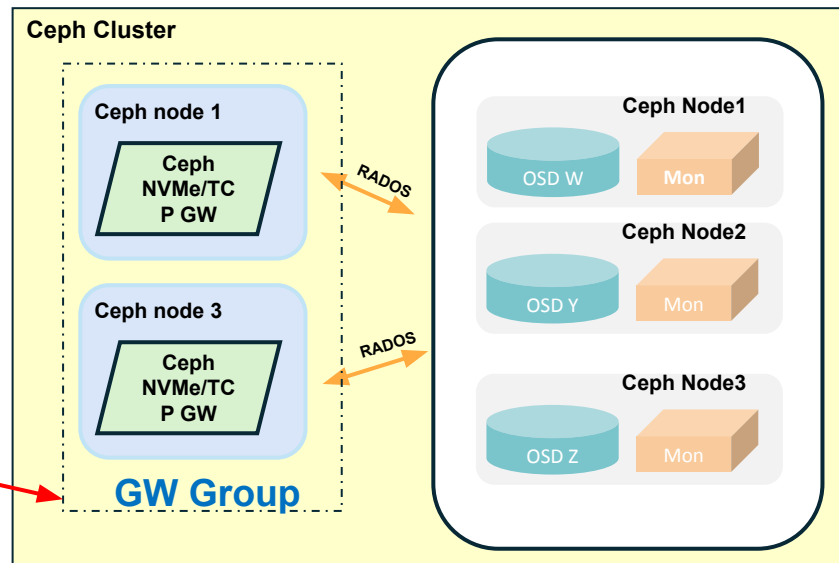
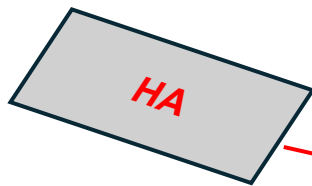
# Gateway Groups & Multi-pathing



# High Availability



- Requires at least 2 Gateways deployed in the Ceph cluster.
- HA is available within the NVMe-OF GW group domain.
- Currently only a single NVMe-oF GW group is supported
- In this release up to 4 GWs can be a part of a GW group (there is no hard restriction).



# High Availability Configuration



- All Gateways within the same GW group share the same configuration
  - All the NVMe Subsystems and namespaces are presented to the hosts (NVMe initiators) by all Gateways in the group
- Configuration can be managed:
  - Gateway CLI
  - Ceph REST API
  - gRPC
- Any Configuration change will be automatically reflected across all the Gateways in the group
  - Stored in an OMAP file in Ceph
  - Changes trigger updates to all Gateways in the group

# High Availability Connectivity



- Hosts must be connected to all of the Gateways in the GW group
- Each connected Gateway provides a different path to the Subsystem and Namespaces
- Only one active path to each namespace at a given time
- All other paths are in Stand-by and can become Active during a Failover

```
nvme-subsys5 - NQN=nqn.2016-06.io.spdk:cnode1
\
+- nvme5 tcp traddr=10.243.64.5,trsvcid=4420 live
+- nvme6 tcp traddr=10.243.64.10,trsvcid=4420 live
+- nvme7 tcp traddr=10.243.64.11,trsvcid=4420 live
+- nvme8 tcp traddr=10.243.64.12,trsvcid=4420 live
```

```
[root@init-nvme-vm5 ~]# nvme list-subsys /dev/nvme1n12
nvme-subsys1 - NQN=nqn.2016-06.io.spdk:cnode10
\
+- nvme1 tcp traddr=10.243.64.5,trsvcid=4420 live inaccessible
+- nvme2 tcp traddr=10.243.64.10,trsvcid=4420 live optimized
+- nvme3 tcp traddr=10.243.64.11,trsvcid=4420 live inaccessible
+- nvme4 tcp traddr=10.243.64.12,trsvcid=4420 live inaccessible
```

```
[root@ceph-nvme-vm4 ~]# ceph nvme-gw show mypool ''
{
  "epoch": 725,
  "pool": "mypool",
  "group": "",
  "num gws": 4,
  "Anagrp list": "[ 4 3 2 1 ]"
}
{
  "gw-id": "client.nvmeof.mypool.ceph-nvme-vm10.zaihd",
  "anagrp-id": 4,
  "performed-full-startup": 1,
  "Availability": "AVAILABLE",
  "ana states": " 4: ACTIVE , 3: STANDBY , 2: STANDBY , 1: STANDBY ,"
}
{
  "gw-id": "client.nvmeof.mypool.ceph-nvme-vm2.fwdklo",
  "anagrp-id": 3,
  "performed-full-startup": 1,
  "Availability": "AVAILABLE",
  "ana states": " 4: STANDBY , 3: ACTIVE , 2: STANDBY , 1: STANDBY ,"
}
{
  "gw-id": "client.nvmeof.mypool.ceph-nvme-vm3.unrawc",
  "anagrp-id": 2,
  "performed-full-startup": 1,
  "Availability": "AVAILABLE",
  "ana states": " 4: STANDBY , 3: STANDBY , 2: ACTIVE , 1: STANDBY ,"
}
{
  "gw-id": "client.nvmeof.mypool.ceph-nvme-vm4.wyests",
  "anagrp-id": 1,
  "performed-full-startup": 1,
  "Availability": "AVAILABLE",
  "ana states": " 4: STANDBY , 3: STANDBY , 2: STANDBY , 1: ACTIVE ,"
}
}
```

# High Availability Connectivity



- All Gateways in the GW group will be actively performing IOs in a Load-balanced way
- Currently the load balancing distributes Namespaces during creating between the Gateways using Round Robin
- In the event of a Failover, Namespaces are automatically re-assigned to one of the the remaining gateways

```
nvme-subsys5 - NQN=nqn.2016-06.io.spdk:cnode1
\
+- nvme5 tcp traddr=10.243.64.5,trsvcid=4420 live
+- nvme6 tcp traddr=10.243.64.10,trsvcid=4420 live
+- nvme7 tcp traddr=10.243.64.11,trsvcid=4420 live
+- nvme8 tcp traddr=10.243.64.12,trsvcid=4420 live
```

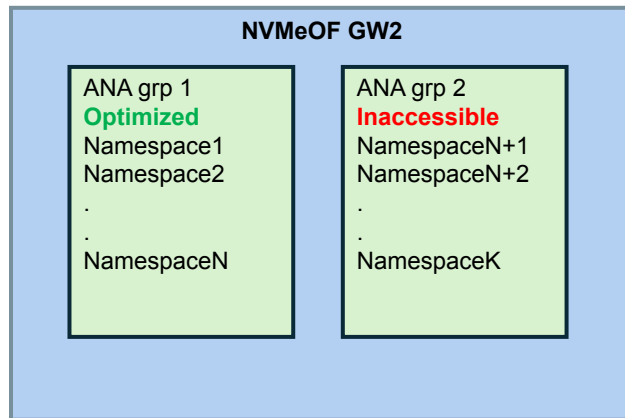
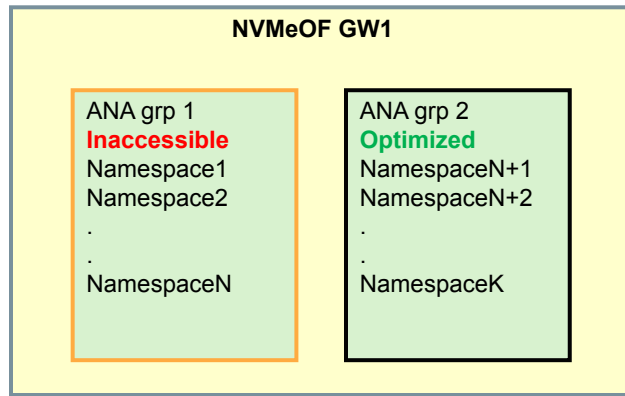
```
[root@init-nvme-vm5 ~]# nvme list-subsys /dev/nvme1n2
nvme-subsys1 - NQN=nqn.2016-06.io.spdk:cnode10
\
+- nvme1 tcp traddr=10.243.64.5,trsvcid=4420 live inaccessible
+- nvme2 tcp traddr=10.243.64.10,trsvcid=4420 live optimized
+- nvme3 tcp traddr=10.243.64.11,trsvcid=4420 live inaccessible
+- nvme4 tcp traddr=10.243.64.12,trsvcid=4420 live inaccessible
```

```
[root@ceph-nvme-vm4 ~]# ceph nvme-gw show mypool ''
{
  "epoch": 725,
  "pool": "mypool",
  "group": "",
  "num gws": 4,
  "Anagrp list": "[ 4 3 2 1 ]"
}
{
  "gw-id": "client.nvmeof.mypool.ceph-nvme-vm10.zaihd",
  "anagrp-id": 4,
  "performed-full-startup": 1,
  "Availability": "AVAILABLE",
  "ana states": " 4: ACTIVE , 3: STANDBY , 2: STANDBY , 1: STANDBY ,"
}
{
  "gw-id": "client.nvmeof.mypool.ceph-nvme-vm2.fwdklo",
  "anagrp-id": 3,
  "performed-full-startup": 1,
  "Availability": "AVAILABLE",
  "ana states": " 4: STANDBY , 3: ACTIVE , 2: STANDBY , 1: STANDBY ,"
}
{
  "gw-id": "client.nvmeof.mypool.ceph-nvme-vm3.unrawc",
  "anagrp-id": 2,
  "performed-full-startup": 1,
  "Availability": "AVAILABLE",
  "ana states": " 4: STANDBY , 3: STANDBY , 2: ACTIVE , 1: STANDBY ,"
}
{
  "gw-id": "client.nvmeof.mypool.ceph-nvme-vm4.wyests",
  "anagrp-id": 1,
  "performed-full-startup": 1,
  "Availability": "AVAILABLE",
  "ana states": " 4: STANDBY , 3: STANDBY , 2: STANDBY , 1: ACTIVE ,"
}
}
```

# ANA (Asymmetric Namespace Access)



- Defines Groups, and Groups Access
  - A group can contain 1..N namespaces
- An NVMe subsystem is exposing ANA groups to the host
- Each ANA group is defined with an Access property that is:
  - Optimized
  - Non-optimized
  - Inaccessible
- Ceph NVMe/TCP solution is using Optimized or Inaccessible on each ANA group



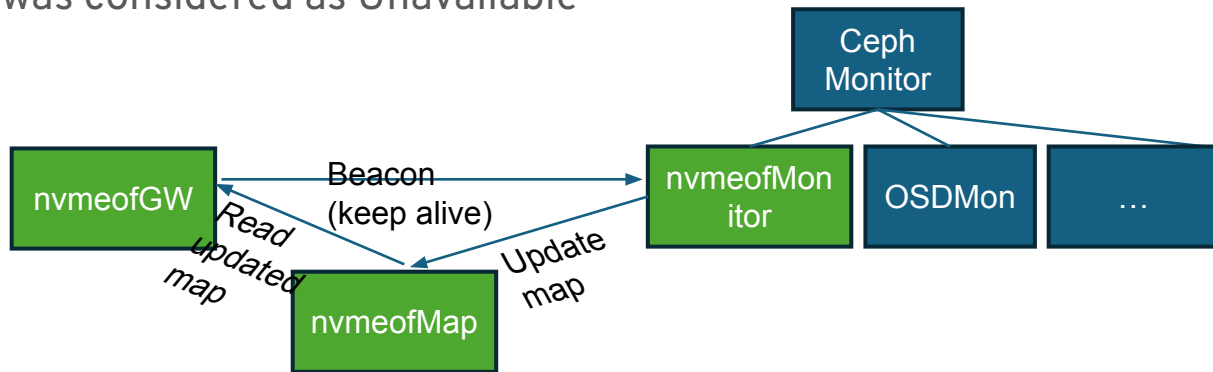


- Ceph NVMe-oF automatically defines and manages the ANA groups, and Accessibility. **The user is not allowed to change anything related to it**
- Each gateway is automatically assigned to manage a single ANA group
- Each gateway is Optimized on one of the ANA groups, and Inaccessible on the rest
- In the case of a Failover, one of the gateways will be automatically assigned to take over the ANA group that was assigned to the failing gateway
- A new Ceph monitor component running as part of the Mon for high availability management

# NVMe/TCP Monitor



- Assign an ANA group to each gateway, and to reassign as need in a Failover/Failback scenarios
- A new monitor client process in each gateway to send Beacons (i.e. Keepalive message) to the nvmeof monitor
- The monitor will decide to perform Failover in case it is not getting Beacons from the client for a while
- The monitor will decide to perform Failback in case it started getting Beacons from a GW that was considered as Unavailable



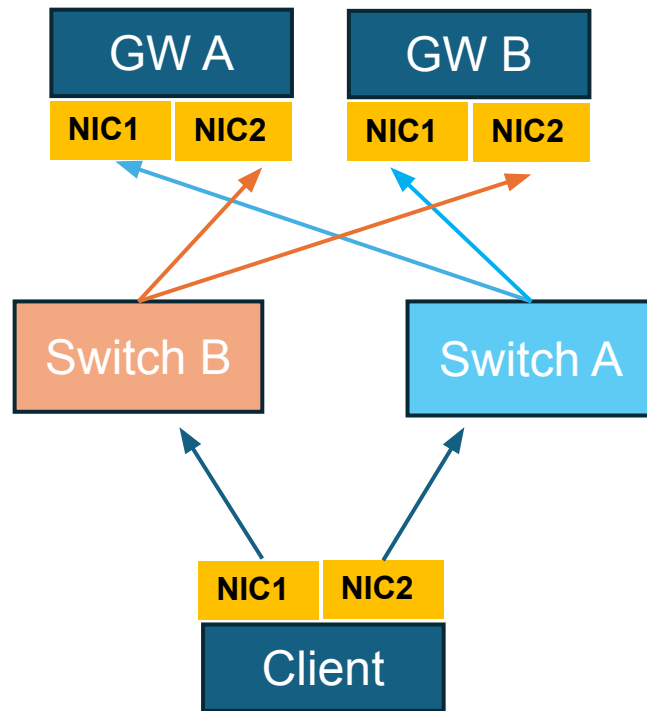


- Identifying that a gateway is down and performing a Failover will take ~15 seconds
- The Failover will be done by the nvmeof monitor as follows:
  - Set the Failing GW to be Inaccessible on the ANA group it 'owned'.
  - Block list (== Fence) the Cluster contexts that were used by the failed GW to access these volumes.
  - Assign the ANA group to the next less busy GW.
  - Set the ANA group to be Optimized on the taking over GW.

# Network Redundancy



- Failover will be performed only in the case that GW is down or seems to be down to the nvmeof monitor.
- From the host perspective, it is critical to have a fully redundant network between the host and the GWs, so that there will always be at least one network path between the host and the GWs.





## Planned/Future Features



# Planned/Future Features



- NVMe-oF CLI as a part of Ceph CLI
- Nvme in-band authentication
- Namespace masking
- Scale improvements:
  - More GWs in a GW group
  - Support for few GW groups in a cluster
  - Support for more Subsystems in a GW group
  - Support of more namespaces for a GW/GW group
- Performance improvements

# Thank You!



Jim Harris, Jonas Pfefferle, Danny Harnik, Scott Peterson, Yue Zhu, Ernesto Puerta, Bharti Wadhwa, Ilya Dryomov, Josh Durgin, Greg Farnum, Patrick Donnelly, Adam King, Sandy Kaur, Rebecca Cloe, Sanjeev Gupta, Brett Niver, Guifeng Tang, Mykola Golub, Congmin Yin, TJ Harris, Adam King, Redouane Kachach, Rahul Lepakshi, Aviv Caro, Alexander Indenbaum, Leonid Chernin, Gil Bergman, Barak Davidov, Roy Sahar, Vallari Agrawal, Nizamudeen A, Redouane Kachach Elhichou, ....



IBM, the IBM logo, and [other IBM trademark listed on the IBM Trademarks List] are trademarks or registered trademarks of IBM Corp., in the U.S. and/or other countries.

Intel, Intel logo, Intel Inside logo, Intel Centrino, Intel Centrino logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium, and Pentium are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

Red Hat®, JBoss®, OpenShift®, Fedora®, Hibernate®, Ansible®, CloudForms®, RHCA®, RHCE®, RHCSA®, Ceph®, and Gluster® are trademarks or registered trademarks of Red Hat, Inc. or its subsidiaries in the United States and other countries.

# Join the Community



<https://github.com/ceph/ceph-nvmeof>

[https://pad.ceph.com/p/rbd\\_nvmeof](https://pad.ceph.com/p/rbd_nvmeof)

Ceph Slack channel: **#nvmeof**

**Weekly meeting:** every Tuesday at 7am PT

<https://meet.jit.si/ceph-nvmeof>