IBM

ADVANCED TECHNOLOGY GROUP (ATG)

Accelerate with ATG Webinar: IBM Storage Ceph Deep Dive for NVMe over Fabrics

Speaker:

John Shubeck, Senior Storage Technical Specialist, ATG

Date: July 17, 2025

IBM

Accelerate with ATG Technical Webinar Series

Advanced Technology Group experts cover a variety of technical topics.

Audience: Clients who have or are considering acquiring IBM Storage solutions. Business Partners and IBMers are also welcome.

To automatically receive announcements of upcoming Accelerate with ATG webinars - Clients, Business Partners and IBMers are welcome to send an email request to accelerate-join@hursley.ibm.com.

2025 Upcoming Webinars – Register Here!

Content Aware Storage (CAS) with IBM Fusion – July 22nd, 2025

Forging Ahead - IBM Storage Virtualize 9.1.0 Technical Update - August 12th, 2025

Important Links to Bookmark:

Accelerate with ATG - Click here to access the Accelerate with ATG webinar schedule for 2025, view presentation materials, and watch past replays dating back two years. **https://ibm.biz/BdSUFN**

ATG MediaCenter Channel - This channel offers a wealth of additional videos covering a wide range of storage topics, including IBM Flash, DS8, Tape, Ceph, Fusion, Cyber Resiliency, Cloud Object Storage, and more. **https://ibm.biz/BdfEgQ**

Offerings

Client Technical Workshops

- **IBM Fusion & Ceph: August 6-7 (Coppell, TX)**
- **IBM Storage Scale & Storage Scale Functions: August 20-21 (San Jose, CA)**
- **IBM DS8000 G10 Advanced Functions: August 26-27 (Chicago, IL)**
- **IBM FlashSystem Deep Dive & Advanced Functions: September 10-11 (Durham, NC)**
- **Cyber Resilience with IBM Storage Defender**

TechZone Test Drive / Demo's

- IBM Cloud Object Storage Test Drive - (VMware based)
- IBM DS8900F Safeguarded Copy (SGC) Test Drive
- IBM DS8900F Storage Management Test Drive
- IBM Storage Scale and Storage Scale System GUI
- IBM Storage Virtualize Test Drive
- IBM Storage Ceph Test Drive
- IBM Storage Ceph Test Drive - (VMware based)
- IBM Storage Protect Live Test Drive
- Managing Copy Services on the DS8000 Using IBM Copy Services Manager Test Drive

Please reach out to your IBM Representative or Business Partner for more information.

***IMPORTANT* The ATG team serves clients and Business Partners in the Americas, concentrating on North America.**

Announcing the 2025 IBM TechXchange Conference

Our theme this year is simple but powerful: **GO / Explore. Build. Launch.**

For more information, please visit - https://www.ibm.com/community/ibm-techxchange-conference/

Accelerate with ATG Survey

Please take a moment to share your feedback with our team!

You can access this 6-question survey via Menti.com with code 5151 0447 or

Direct link https://www.menti.com/alhsf3bgvxu6
Or

QR Code

IBM

ADVANCED TECHNOLOGY GROUP (ATG)

Accelerate with ATG Webinar: IBM Storage Ceph Deep Dive for NVMe over Fabrics

Speaker:

John Shubeck, Senior Storage Technical Specialist, ATG

Date: July 17, 2025

IBM

About the Presenter

John Shubeck is an information technology professional with over 42 years of industry experience spanning both the customer and technology provider experience. John is currently serving as a Senior Storage Technical Specialist on IBM Object Storage platforms across all market segments in the Americas.

Introducing our panelists

Shu Mookerjee is a Level 2 Certified Technical Specialist with over twenty years at IBM, working in a variety of roles including sales, management and technology. For the last decade, he has focused exclusively on storage and has been the co-author of four (4) Redbooks. Currently, Shu is part of the Advanced Technology Group where he provides education, technical guidance, Proofs of Concept and Proofs of Technology to IBMers, business partners and clients.

Introducing our panelists

Jerrod Carr is an IBM Principal Storage Technical Specialist in IBM Storage Solutions. Jerrod Carr has been in the Storage industry for over 21 years selling hardware and software for various large technology companies. With beginnings in the Cleversafe IBM team for 8 years providing expertise in Cloud Object Storage, the last 3 years working on the Americas SWAT team as a Senior Storage Specialist providing unstructured data experience to the various markets.

Summary of topics

- Overview of the Ceph RADOS Block Device (RBD)
- Configuring Ceph RBD in the Dashboard UI
- Configuring Ceph RBD in the CLI
- The Ceph RBD client experience and live demonstration
- Overview of Ceph NVMe over Fabrics (i.e. NVMe/TCP)
- INTERMISSION
- Overview of Ceph NVMe over Fabrics (NVMe-oF)
- Configuring Ceph NVMe/TCP in the Dashboard UI
- The Ceph RBD client experience and live demonstration
- Day 2 and day 3 considerations

RADOS Block Device (RBD) overview

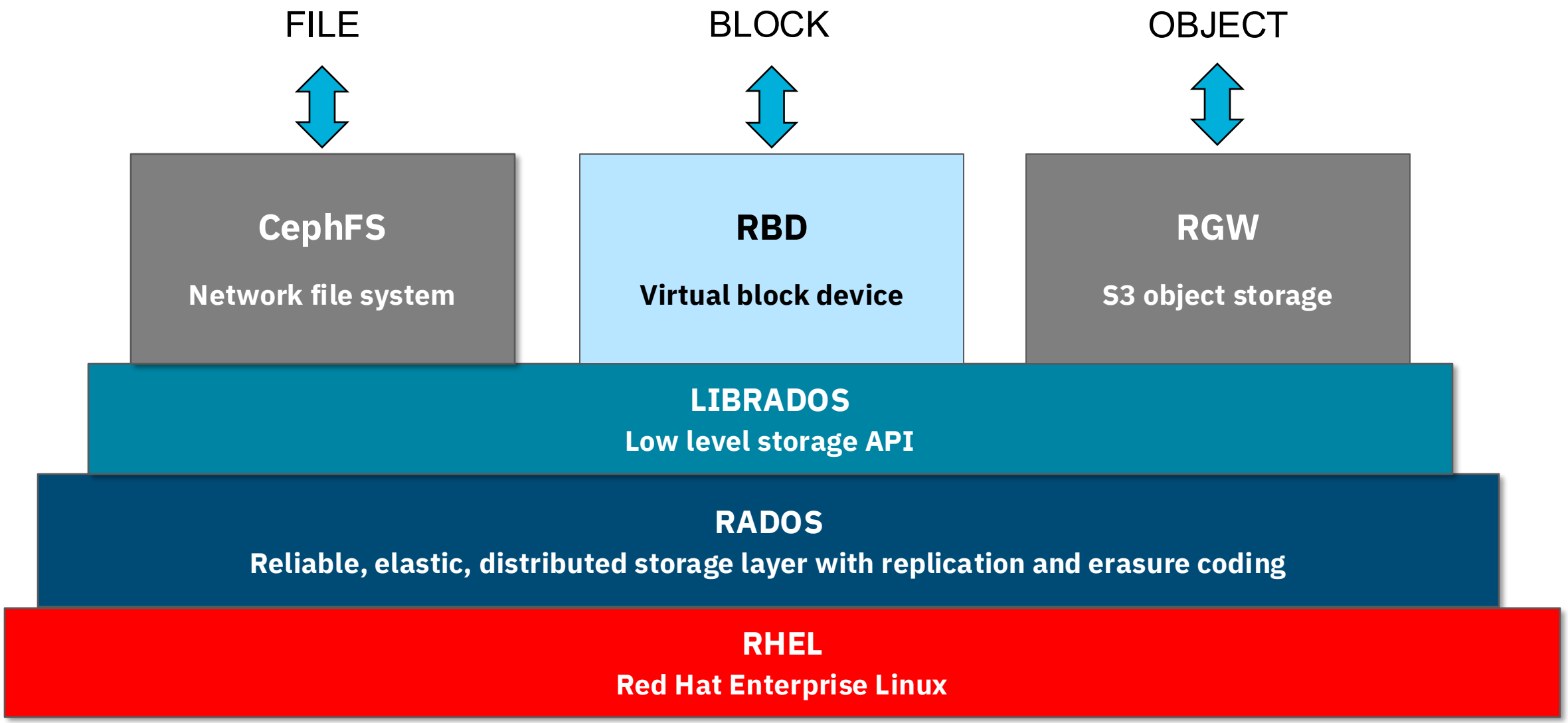
Ceph block storage overview (RBD)

- The *RADOS Block Device (RBD)* feature provides block storage from the Red Hat Ceph Storage cluster. RADOS provides virtual block devices stored as *RBD images* in pools in the Red Hat Ceph Storage cluster.

- *Block devices* are the most common long-term storage devices for servers, laptops, and other computing systems. They store data in fixed-size blocks. Block devices include both hard drives, based on spinning magnetic platters, and solid-state drives, based on nonvolatile memory. To use the storage, format a block device with a file system and mount it on the Linux file system hierarchy.

- The *RADOS Block Device (RBD)* feature does not run as a daemon or service, rather, it is an intrinsic access method of the Ceph cluster itself

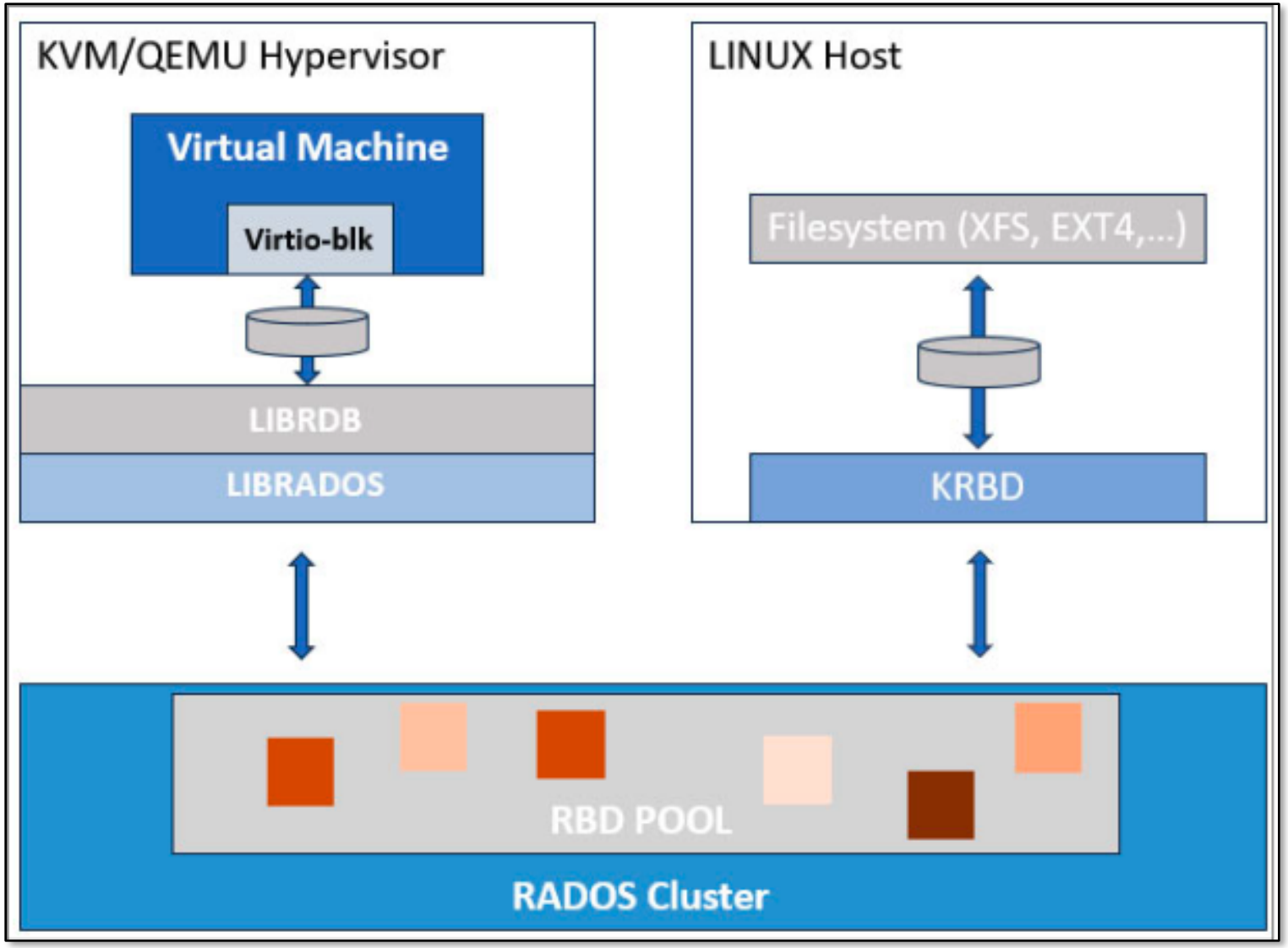
IBM Storage Ceph block storage components (RBD)



RADOS Block Device (RBD) ecosystem components

Containerized workloads

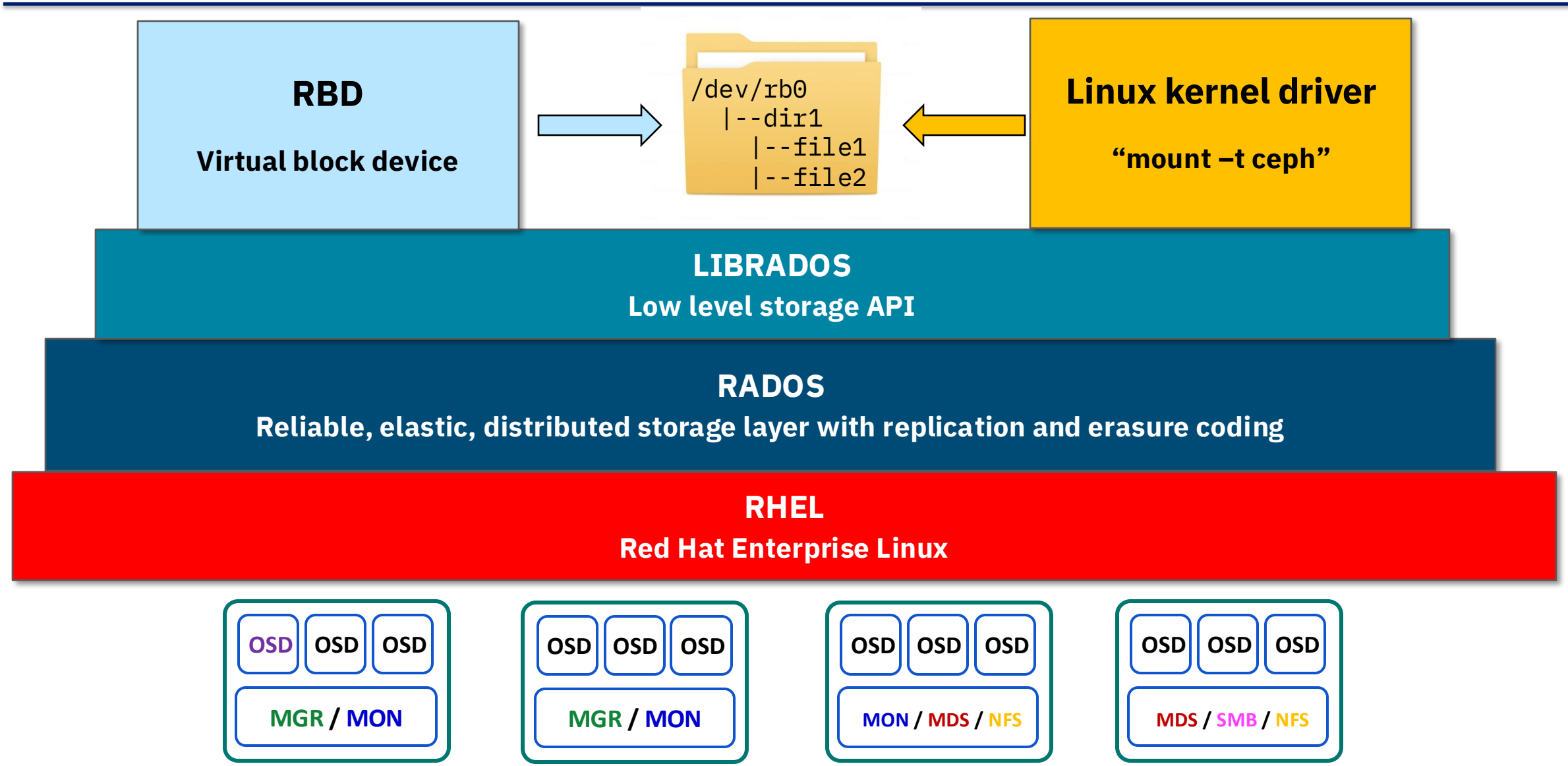
"librbd" API
(e.g. Persistent Volume Claims)



Linux kernel module

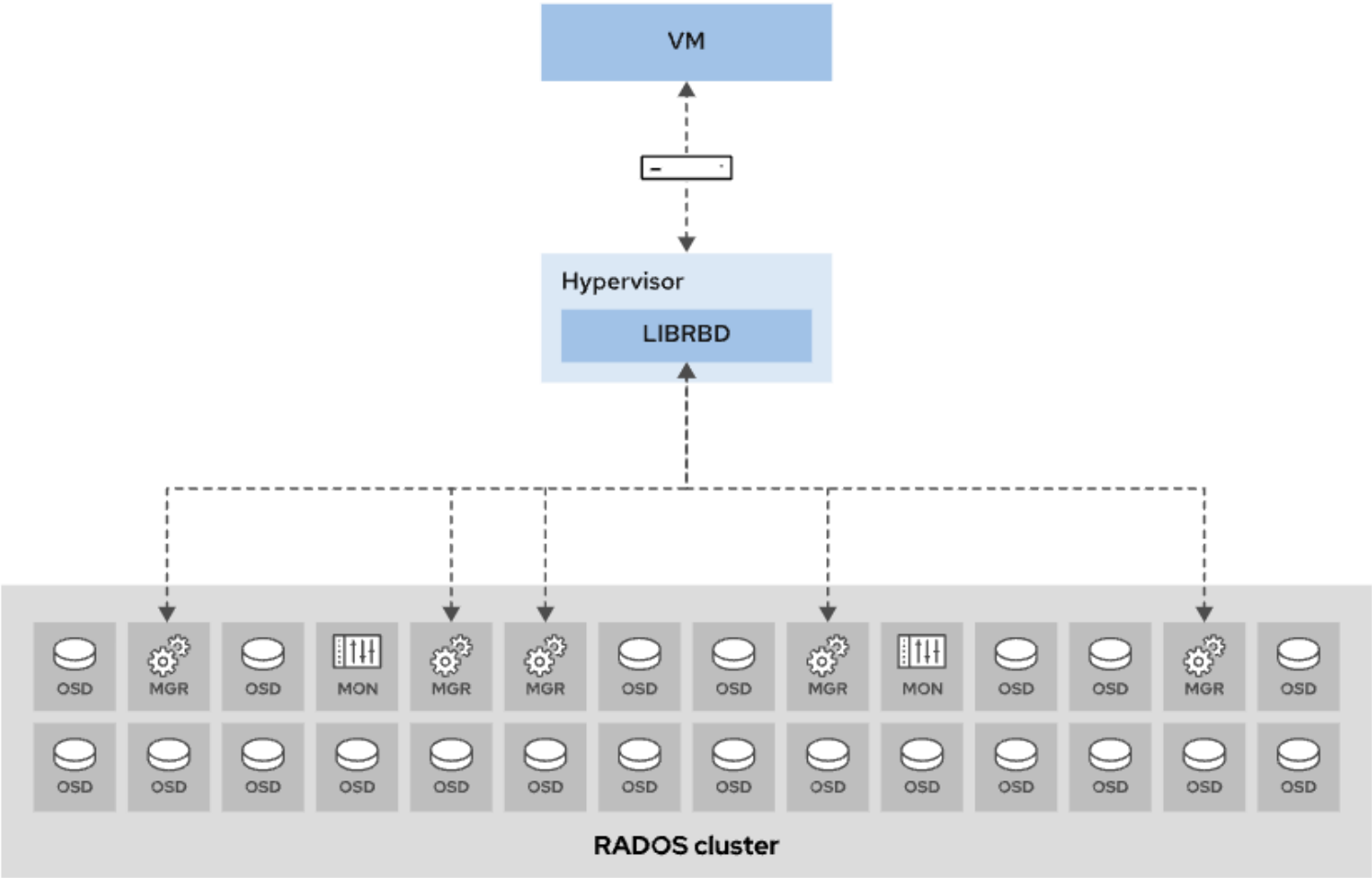
- RHEL
- Rocky
- Ubuntu
- Suse
- Debian
- CentOS

IBM Storage Ceph block components (RBD)



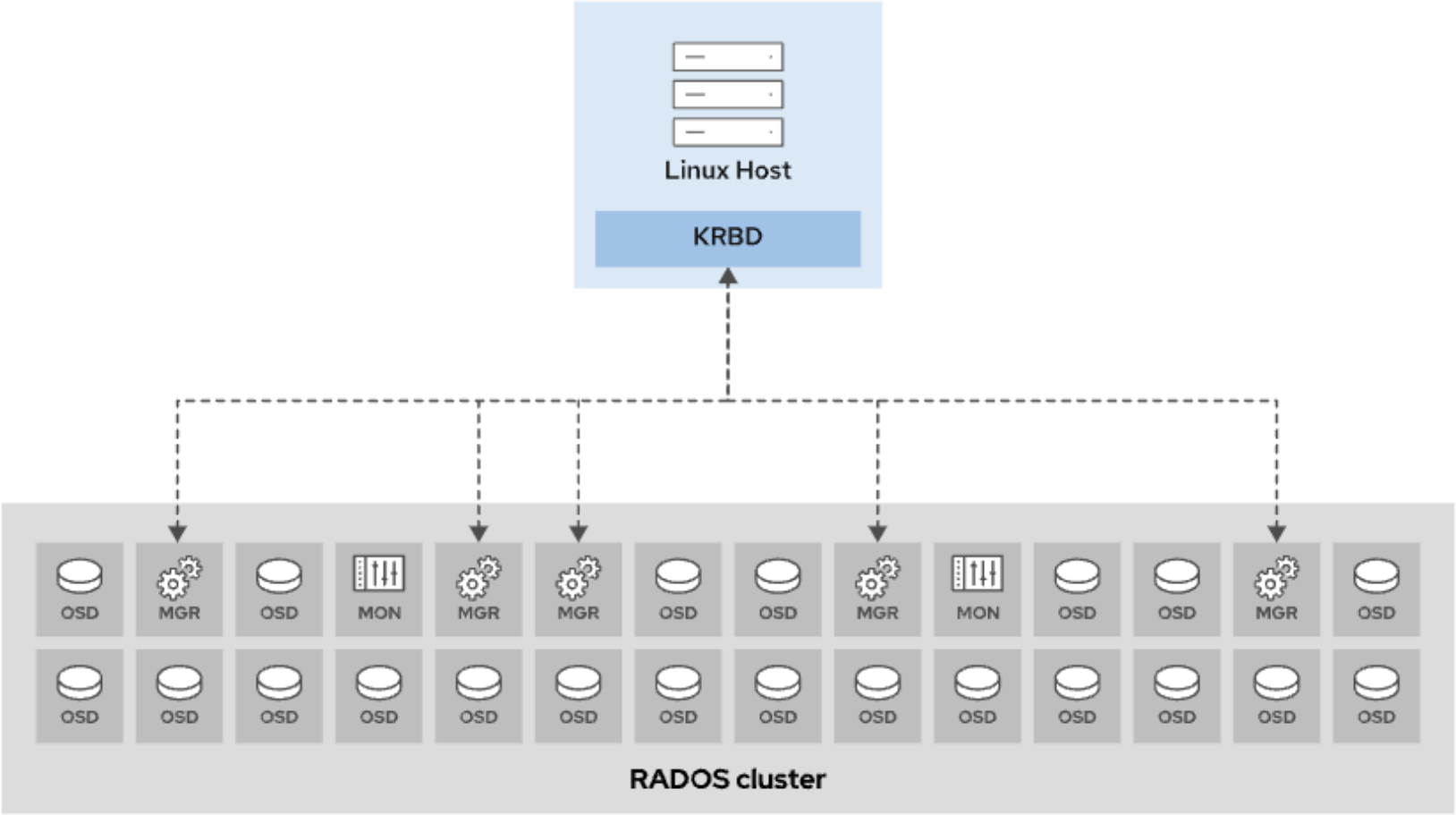
Accessing Ceph Storage with librbd-based Clients

The librbd library provides direct access to RBD images for user space applications. Cloud and virtualization solutions, such as OpenStack and libvirt, use librbd to provide RBD images as block devices to cloud instances and the virtual machines that they manage



Linux kernel environment access

Ceph clients typically mount an RBD image using the native Linux kernel module, krbd. This module maps RBD images to Linux block devices with names such as /dev/rbd0.



Kernel RBD access method

- On a Linux client machine, the
  - The kernel module maps the RBD block device to a kernel block device
  - The kernel block device appears as a regular block device in Linux (e.g. /dev/rbd0)
  - The Linux client formats the device and mounts a file system (e.g. /mnt/rbd0)

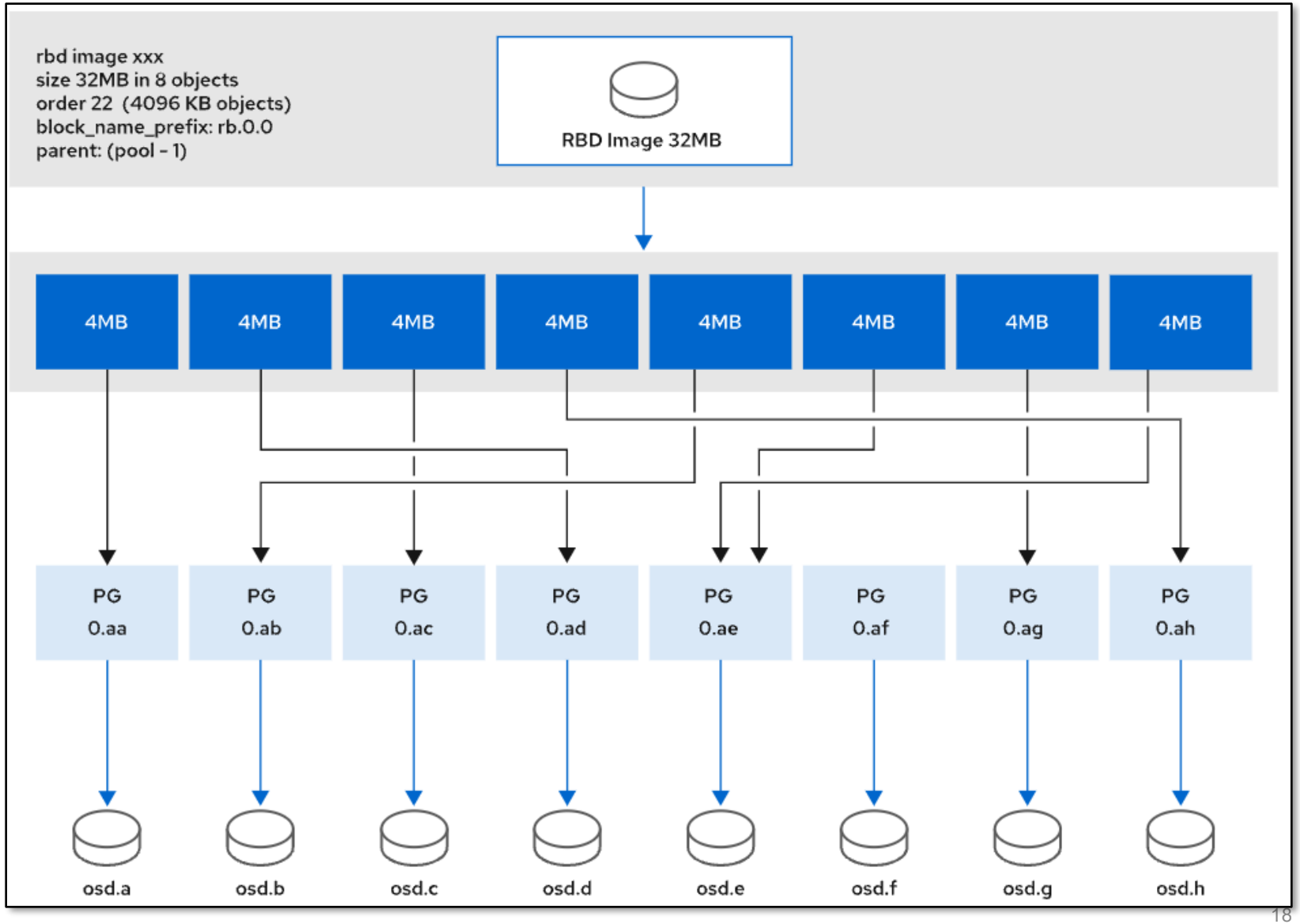
- On the Ceph cluster:
  - RBD images are striped over objects in a RADOS object store (i.e. OSDs)
  - The Ceph Pool is the point of control for Compression, Quotas, and QoS
  - Low level RBD storage attributes follow the Ceph Pool attributes

RBD striping

Ceph block devices allow storing data striped over multiple Object Storage Devices (OSD) in a Red Hat Ceph Storage cluster.

RBD stripe size and stripe unit can be tuned for specific application requirements (i.e. stripe size * stripe unit = object size)

RBD object size can be 4KiB < (4MiB) < 32MiB



IBM Storage Ceph virtual storage constructs

POOL

A Ceph Pool is a logical partition within a Ceph storage cluster where data is stored

- Assigned a data protection type (e.g. “replicated” or “erasure”)
- Mapped to one and only one Ceph “Application” (CephFS, RBD, RGW)

IMAGE

The Ceph block storage (i.e. RBD) construct and capacity allocation

- Mapped by a Linux kernel client to a local device (e.g. “*/dev/rb0*”)

VOLUME

The Ceph file system (i.e. CephFS) construct with sharing options

- Mapped by a Linux kernel client to a mounted device (e.g. “*/mnt/fsdemo*”)

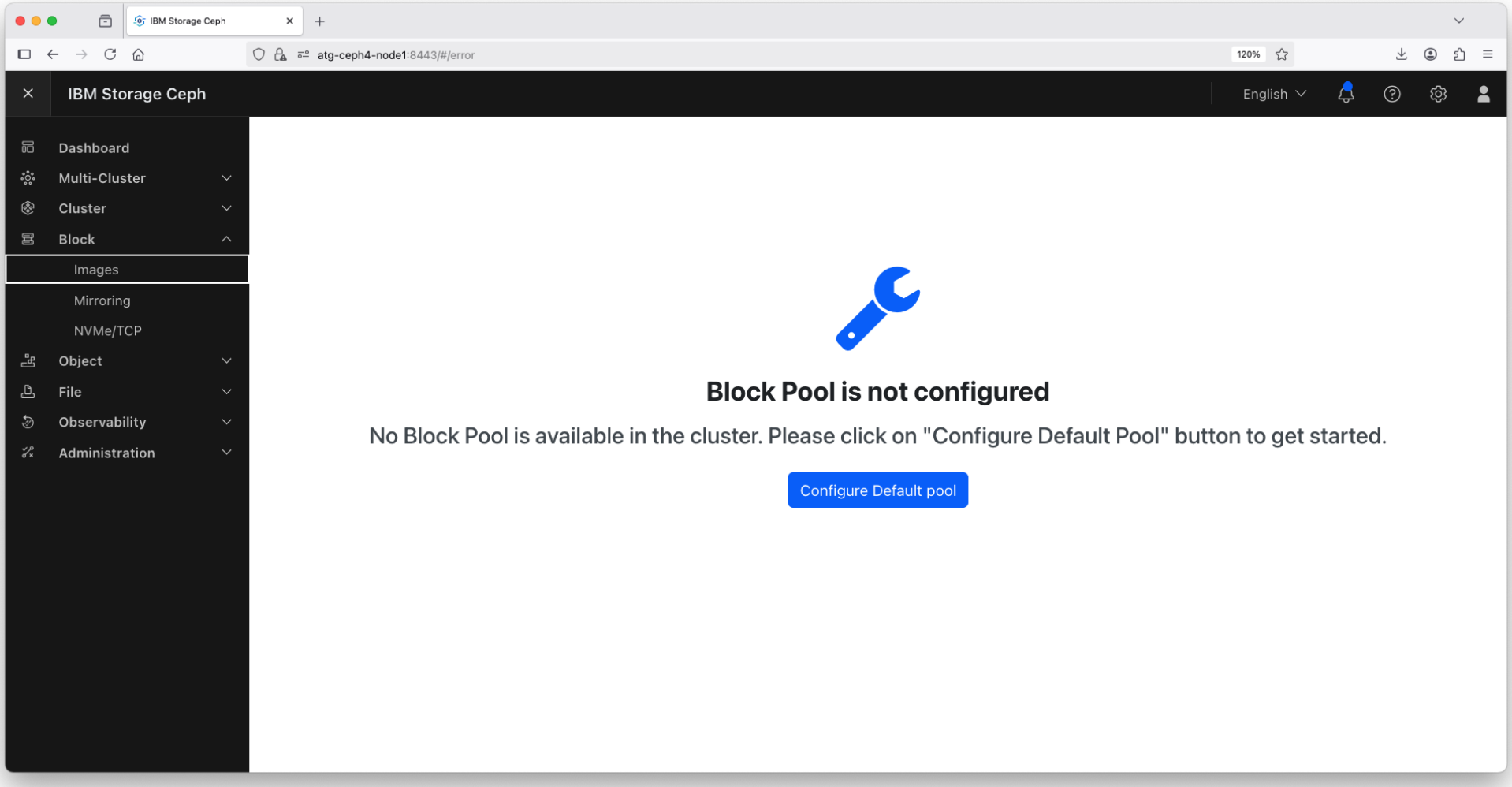
BUCKET

The Ceph S3 object storage (i.e. RGW) construct plus advanced features

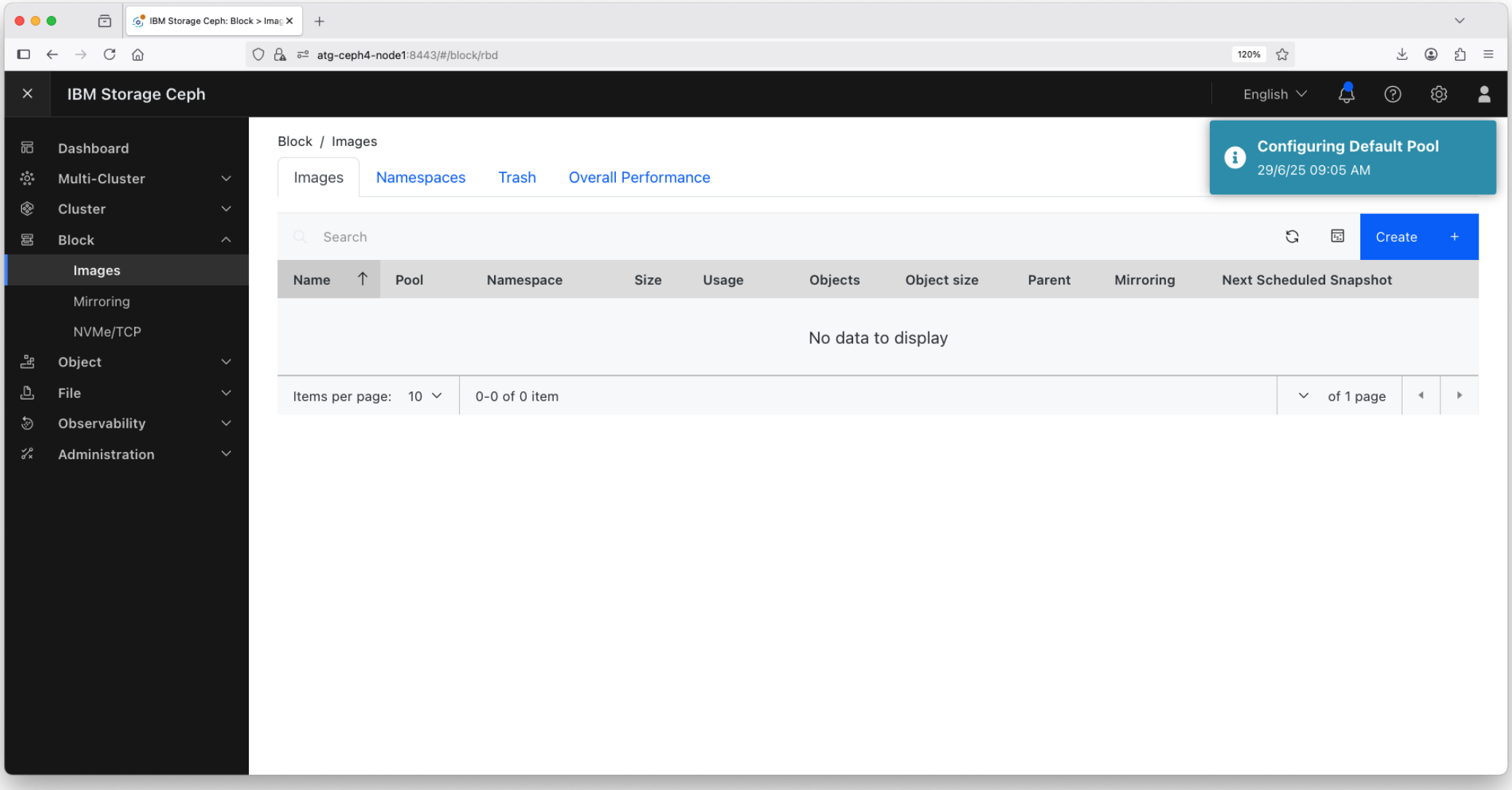
- Addressed by an S3 API client as a “*container*” for objects (e.g. “*/mnt/fsdemo*”)

RADOS Block Device (RBD) configuration

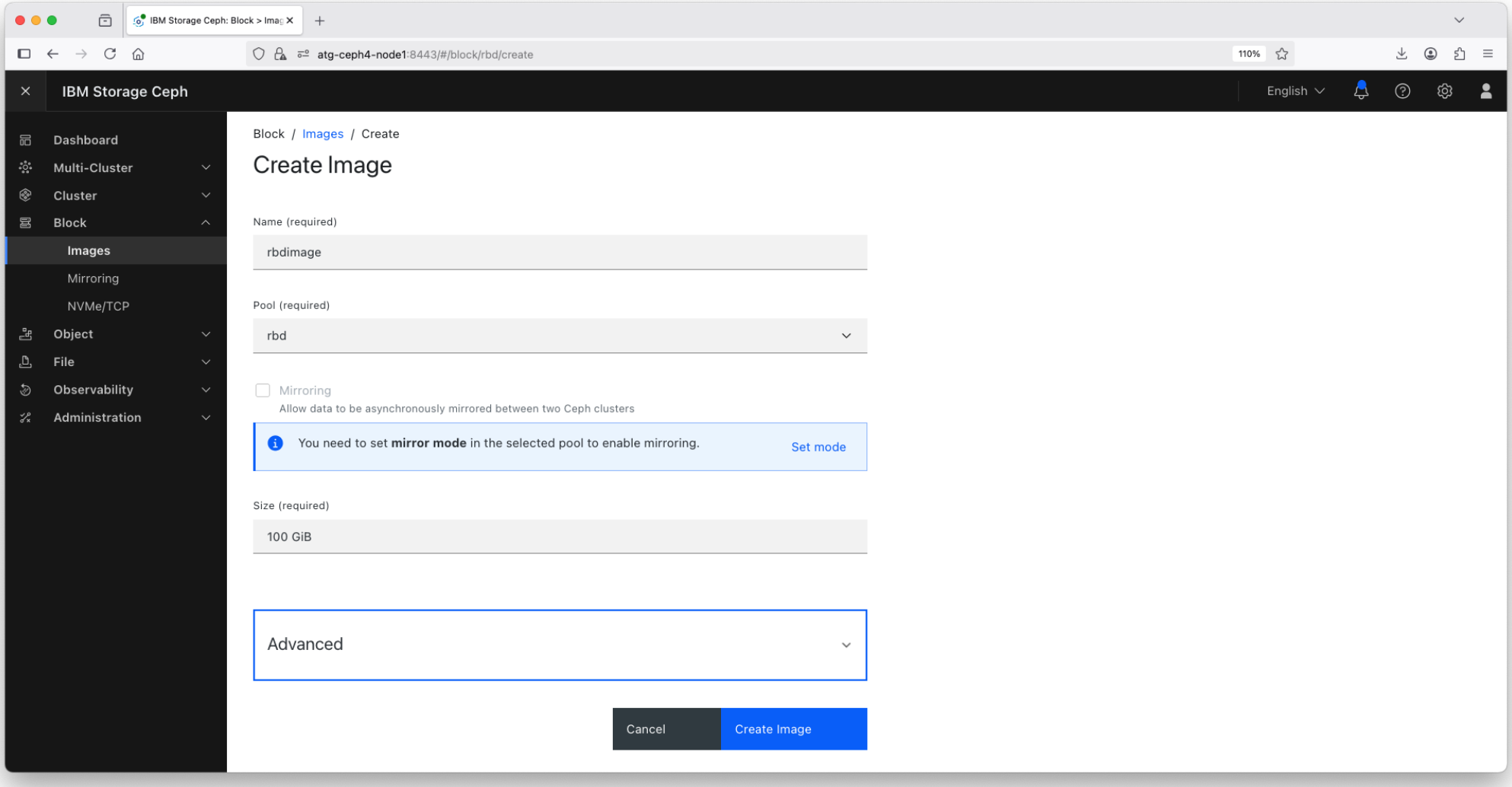
First touch of the Ceph Dashboard *Navigation Pane -> Block group -> Images*



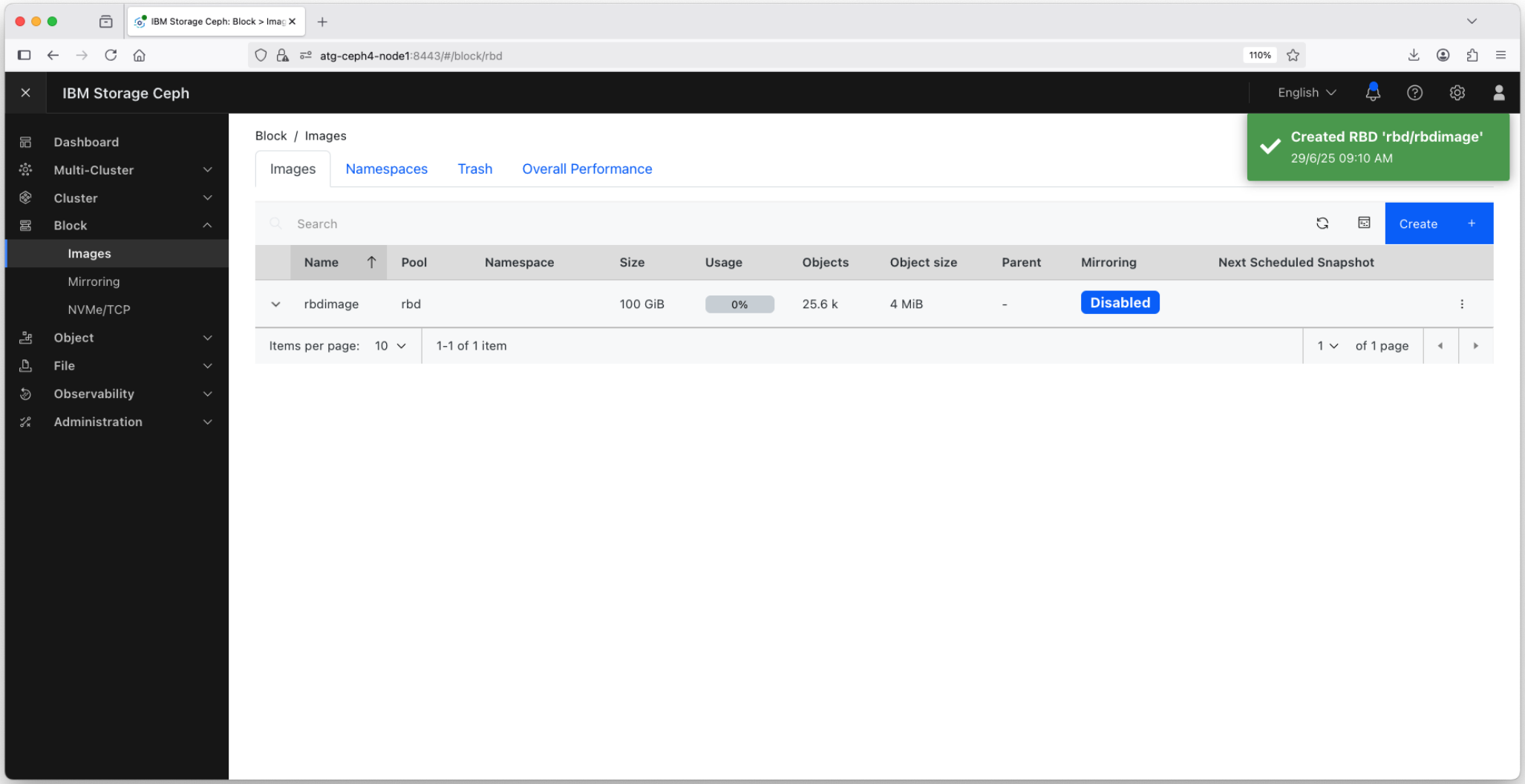
First touch of the Ceph Dashboard – Pool *"RBD"* created



First touch of the Ceph Dashboard – Create Image "*rbdimage*"



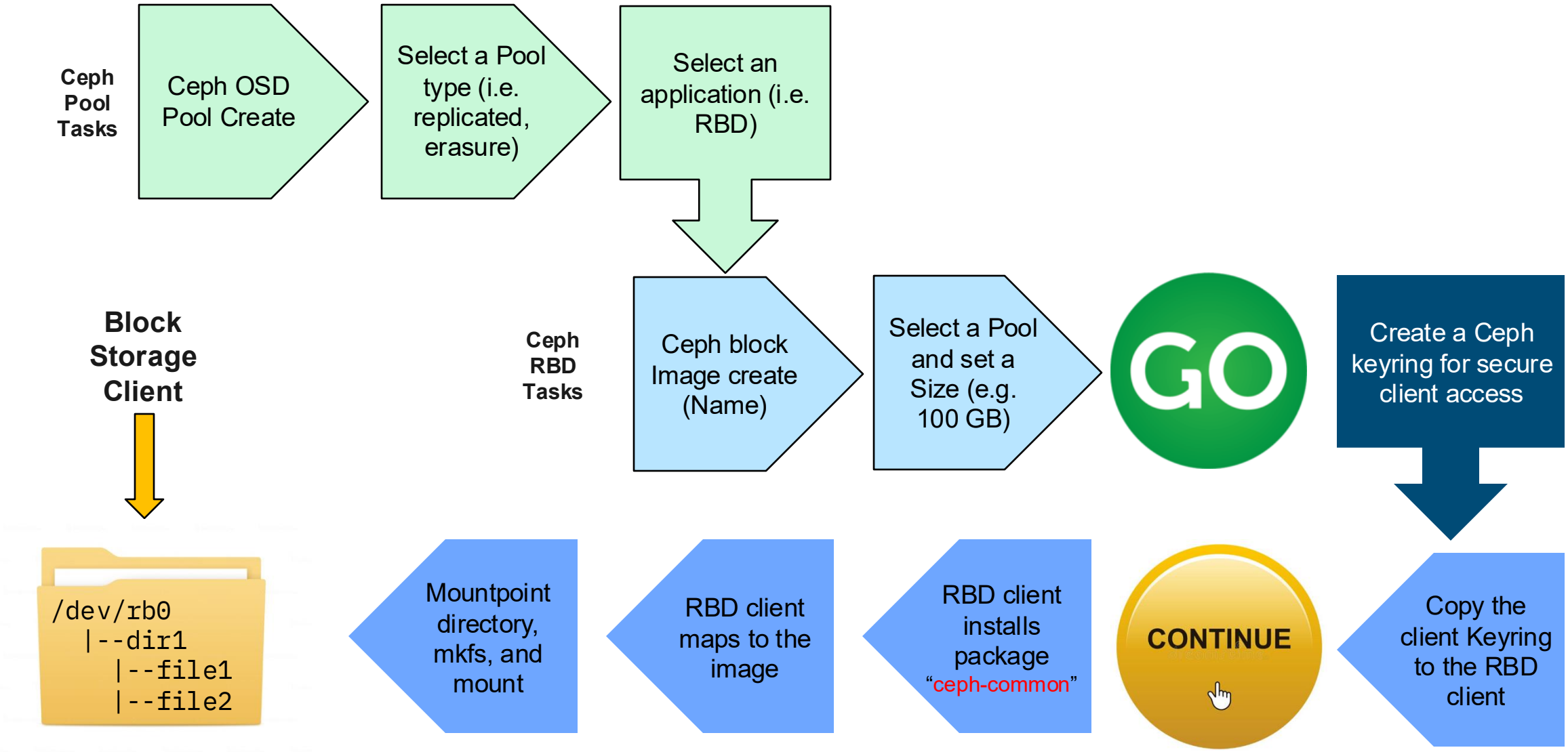
Ceph Dashboard – RBD image view



Ceph command line (*cephadm*) block storage management

IBM Storage Ceph block device (RBD) configuration tasks

LIVE



Ceph Pool and RBD Image creation steps

```
[ceph: root@node1 /]# ceph osd pool create rbdpool
pool ‘rbdpool' created

[ceph: root@node1 /]# ceph osd pool application enable rbdpool rbd
enabled application 'rbd' on pool 'rbdpool’

[ceph: root@node1 /]# rbd create rbdpool/rbdimage --size=100G
[ceph: root@node1 /]# rbd ls rbdpool
rbdimage

[ceph: root@node1 /]# ceph df
--- POOLS ---
POOL                    ID  PGS   STORED  OBJECTS     USED  %USED  MAX AVAIL
. . . Output omitted . . .
.mgr      1    1  449 KiB        2  1.3 MiB      0     81 GiB
rbd       5   32      0 B        0      0 B      0     81 GiB
rbdpool   6   32  6.3 KiB        5   48 KiB      0     81 GiB
testp     5   32      0 B        0      0 B      0     81 GiB
```

Create a block device user

```
[root@node1 ~]# ceph auth get-or-create client.atg \
mon "profile rbd" osd "profile rbd" \
-o ceph.client.atg.keyring

[root@node1 ~]# cat ceph.client.atg.keyring
[client.atg]
  key = AQA+IG1om3hYLxAAXI6mj2SiMBD3c+DLOzkblw==
  caps mon = "profile rbd"
  caps osd = "profile rbd”

[root@node1 ~]# scp ceph.client.atg.keyring root@atg-ceph4-node1:/etc/ceph
```

https://www.ibm.com/docs/en/storage-ceph/8.1?topic=file-systems

Ceph RBD client command example

```
[root@node1 ~]# dnf install -y ceph-common

[root@node1 ~]# export CEPH_ARGS="--id=admin"

[root@node1 ~]# ceph osd lspools
1 .mgr
5 rbd
6 rbdpool

[root@node1 ~]# rbd ls
myrbd

[root@node1 ~]# ceph df

. . . Output omitted . . .
--- POOLS ---
.mgr      1    1  449 KiB        2  1.3 MiB      0     81 GiB
rbd        5   32       0 B        0       0 B      0     81 GiB
rbdpool   6   32  6.3 KiB        5   48 KiB      0     81 GiB
testp     5   32      0 B        0      0 B      0     81 GiB
```

Ceph RBD client command example

```
[root@node1 ~]# ceph osd lspools
1 .mgr
5 rbd
6 rbdpool

[root@node1 ~]# rbd ls rbdpool --id=atg
rbdimage

[root@node1 ~]# ceph df

--- RAW STORAGE ---
CLASS     SIZE    AVAIL     USED  RAW USED  %RAW USED
hdd    256 GiB  255 GiB  818 MiB   818 MiB       0.31
TOTAL  256 GiB  255 GiB  818 MiB   818 MiB       0.31

--- POOLS ---
POOL     ID  PGS   STORED  OBJECTS     USED  %USED  MAX AVAIL
.mgr      1    1  449 KiB        2  1.3 MiB      0     81 GiB
rbd       5   32      0 B        0      0 B      0     81 GiB
rbdpool   6   32  6.3 KiB        5   48 KiB      0     81 GiB
```

Ceph RBD client command example

```
[root@client ~]# dnf install -y ceph-common

[root@client ~]# rbd ls rbdpool --id=admin    # Another way to reference the cephx keyring
rbdimage

[root@client ~]# rbd info rbdpool/rbdimage
 size 100 GiB in 25600 objects
  order 22 (4 MiB objects)
  snapshot_count: 0
  id: 36d4732280f41
  block_name_prefix: rbd_data.36d4732280f41
  format: 2
  features: layering, exclusive-lock, object-map, fast-diff, deep-flatten
  op_features:
  flags:
  create_timestamp: Wed Jul  9 12:31:45 2025
  access_timestamp: Wed Jul  9 12:31:45 2025
  modify_timestamp: Wed Jul  9 12:31:45 2025
```

Ceph RBD client commands (Linux kernel driver)

```
[root@client ~]# rbd list rbdpool
rbdimage


[root@client ~]# rbd info rbdpool/rbdimage
rbd image 'rbdimage':
    size 100 GiB in 25600 objects
    . . . output truncated . . .


[root@client ~]# rbd map rbdimage
/dev/rbd0


[root@client ~]# rbd showmapped
id  pool   namespace  image     snap  device
0   rbdpool          rbdimage  -     /dev/rbd0
```

Ceph RBD client demonstration - Let’s go live!

Ceph RBD client commands (continued)

```
[root@client ~]# mkfs.xfs /dev/rbd0
meta-data=/dev/rbd0              isize=512    agcount=16, agsize=1638400 blks
         =                       sectsz=512   attr=2, projid32bit=1
         =                       crc=1        finobt=1, sparse=1, rmapbt=0
         =                       reflink=1    bigtime=1 inobtcount=1 nrext64=0
data     =                       bsize=4096   blocks=26214400, imaxpct=25
         =                       sunit=16     swidth=16 blks
naming   =version 2              bsize=4096   ascii-ci=0, ftype=1
log      =internal log           bsize=4096   blocks=16384, version=2
         =                       sectsz=512   sunit=16 blks, lazy-count=1
realtime =none                   extsz=4096   blocks=0, rtextents=0
Discarding blocks...Done.
```

Ceph RBD client commands (continued)

```
[root@client ~]# mkdir /mnt/rbdimage

[root@client ~]# mount /dev/rbd0 /mnt/rbdimage

[root@client ~]# mkdir /mnt/rbdimage/dir1
[root@client ~]# mkdir /mnt/rbdimage/dir2

[root@client ~]# touch /mnt/rbdimage/dir1/atestfile

[root@client ~]# dd if=/dev/random of=/mnt/rbdimage/dir1/10MB.dat bs=1M count=10
10+0 records in
10+0 records out
10485760 bytes (10 MB, 10 MiB) copied, 0.0813964 s, 129 MB/s

[root@client ~]# echo 98333, Fox Island, WA >> /mnt/rbdimage/dir2/zip-codes.csv
[root@client ~]# echo 98335, Gig Harbor, WA >> /mnt/rbdimage/dir2/zip-codes.csv
```

Ceph RBD client commands (finished)

```
[root@client ~]# ls -al /mnt/rbdimage/dir1
drwxr-xr-x. 2 root root       39 Jun 29 11:38 .
drwxr-xr-x. 4 root root       30 Jun 29 11:38 ..
-rw-r--r--. 1 root root 10485760 Jun 29 11:38 10MB.dat
-rw-r--r--. 1 root root        0 Jun 29 11:38 atestfile

[root@client ~]# cat /mnt/rbdimage/dir2/zip-codes.csv
98333, Fox Island, WA
98335, Gig Harbor, WA

[root@client ~]# tree /mnt
/mnt
└── rbdimage
    ├── dir1
    │   ├── 10MB.dat
    │   └── atestfile
    └── dir2
        └── zip-codes.csv

3 directories, 3 files
```

Ceph RBD verification

```
[root@client ~]# df /mnt/rbdimage
Filesystem     1K-blocks   Used Available Use% Mounted on
/dev/rbd0      104792064 774716 104017348   1% /mnt/rbdimage

[root@client ~]# ceph df
--- RAW STORAGE ---
CLASS     SIZE    AVAIL     USED  RAW USED  %RAW USED
hdd    256 GiB  255 GiB  712 MiB   712 MiB       0.27
TOTAL  256 GiB  255 GiB  712 MiB   712 MiB       0.27

--- POOLS ---
POOL  ID  PGS   STORED  OBJECTS     USED  %USED  MAX AVAIL
.mgr   1    1  449 KiB        2  1.3 MiB      0     81 GiB
rbdpool 2   32   13 MiB       24   38 MiB   0.02     81 GiB
```

Client RBD cleanup

```
[root@client ~]# umount /mnt/rbdimage

[root@client ~]# rbd unmap rbdpool/rbdimage

[root@client ~]# rbd showmapped

[root@client ~]# rbd list rbdpool
rbdimage

#
# The rbdimage is still there but we are no longer mapped to it
#
```

A message from IBM Technology Zone

. . . 60 SECOND COMMERCIAL MESSAGE . . .

IBM TechZone for IBM Storage Ceph Test Drive (Ceph POC in the cloud)

North America ATG Storage - IB

techzone.ibm.com/collection/north-america-atg-storage-ibm-storage-ceph-test-drive

IBM Technology Zone | My TechZone | Support

- Overview
- Resources
- Environments
- Metadata
- Comments

North America ATG Storage - IBM Storage Ceph Test Drive

Overview

Gold

(0) Rate this resource

Edit

<< GO TO the Resources tab to see the run books and supporting material.

<< GO TO the Environments tab to RESERVE this collection.

The IBM Storage Ceph test drive (Software Defined offers IBMers, Business Partners, and Clients, hands-on access into an operational IBM Storage Ceph.

This lab starts with a single Ceph node bootstrapped to the first login. From there, the user will expand the cluster to four (4) nodes, add the OSDs for storage, and start various Ceph services. After the cluster has been expanded and storage has been provisioned, the lab guide can be followed to create Ceph Pools and configure block, file, and object storage services as desired, Finally, the lab document guides the user through the client experience for block, file, and object data operations.

Click HERE for a quick link to the Appliance based access instructions

This demo reservation is fully automated and usually takes less than one (1) hour to provision. The duration can be extended by the user after booking.

Author
John Shubeck (jshubeck@us.ibm.com)

Collaborators
BethAnn Stugis (bethann.stugis@us.ibm.com), Kavita Shah (kashah@us.ibm.com)

Cookie Preferences

NVMe over Fabrics (i.e. NVMe/TCP)

What is NVMe over Fabrics (NVMe-oF)?

- *NVMe over Fabrics* (NVMe-oF) is a protocol that extends the capabilities of NVMe storage to networked environments, allowing multiple hosts to access shared NVMe storage resources.

- Allows NVMe commands to be sent and received over a network fabric, rather than just through a direct PCIe connection within a single server.

Why?

- To enable greater scalability, flexibility, and performance for networked storage solutions. NVMe-oF allows NVMe commands to be transported over a network fabric, such as FibreChannel, Infiniband . . . or . . . Ethernet!

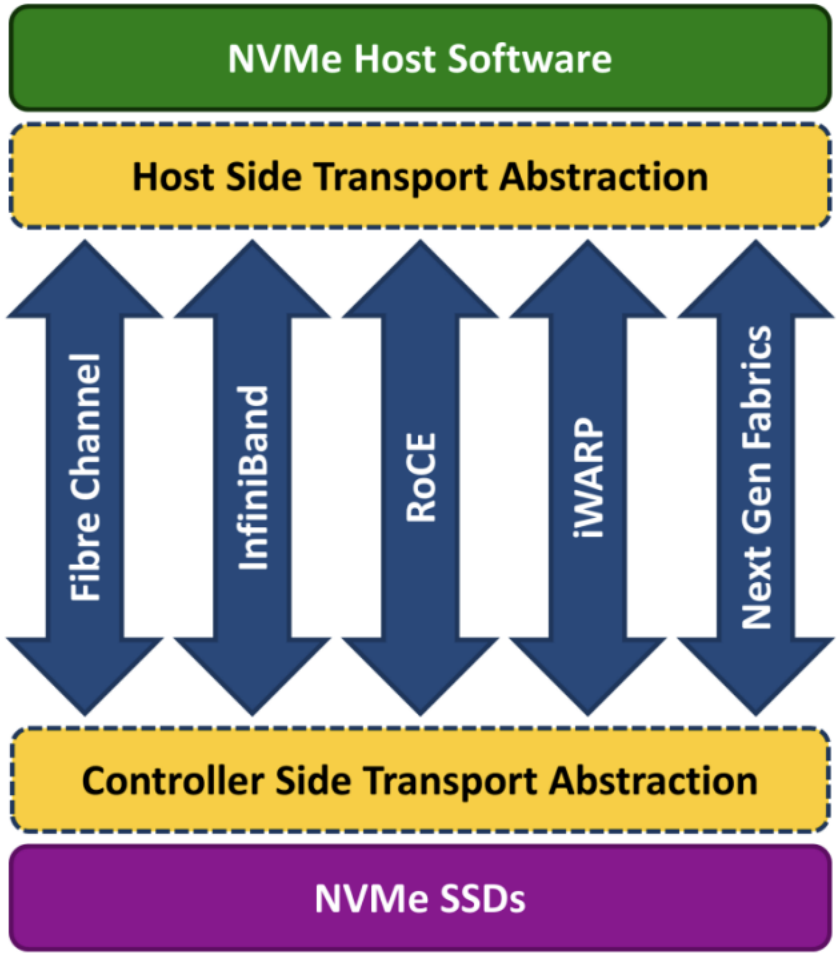
Why not iSCSI?

- NVMe-oF Extends the NVMe command set end to end from the physical NVMe (SSD) storage device to the client initiator
- NVMe-oF delivers lower latency and higher throughput than iSCSI
- NVMe-oF was designed for high speed scale-out storage clusters and networks

Ceph Community statement about iSCSI

- The iSCSI gateway is in maintenance as of November 2022. This means that it is no longer in active development and will not be updated to add new features. (https://docs.ceph.com/en/latest/rbd/iscsi-overview/)

Components of NVMe over Fabrics



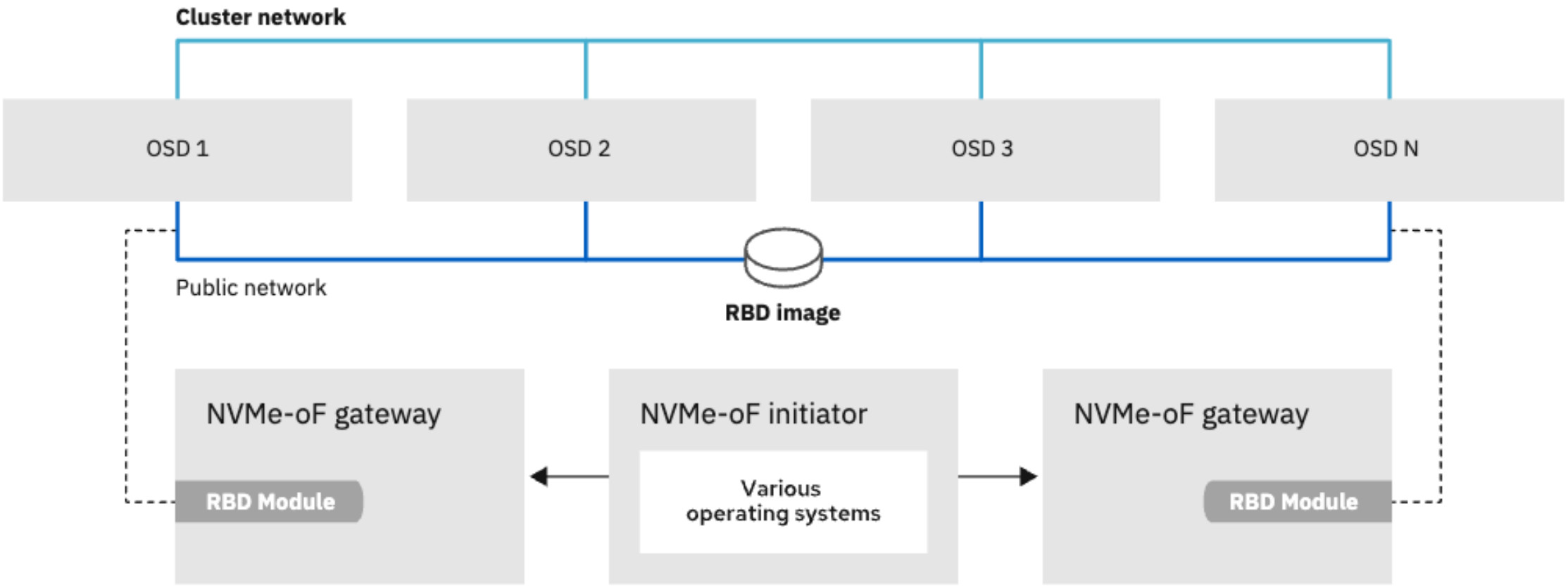
History of NVMe over Fabrics

- 2014 – Work begins
- 2016 – First NVMe-oF specification released
- 2018 – Revision 1.0 ratified
- 2021 – Revision 1.1 ratified

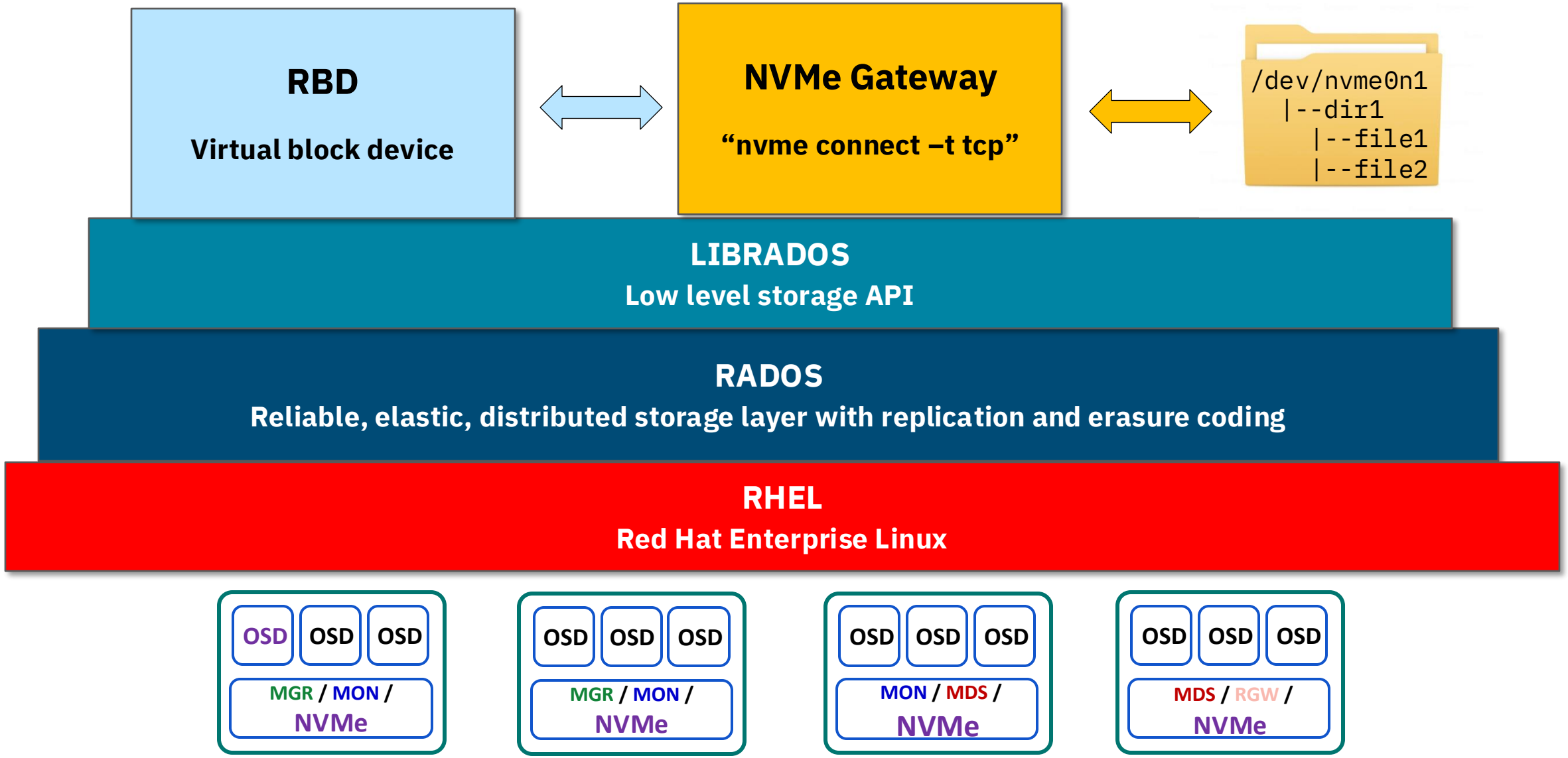
History of Ceph NVMe/TCP

- 2022 – Experimental feature work
- 2024 – Introduced in Reef a new feature NVMe/TCP
- 2025 – Ceph Dashboard support added (Squid)

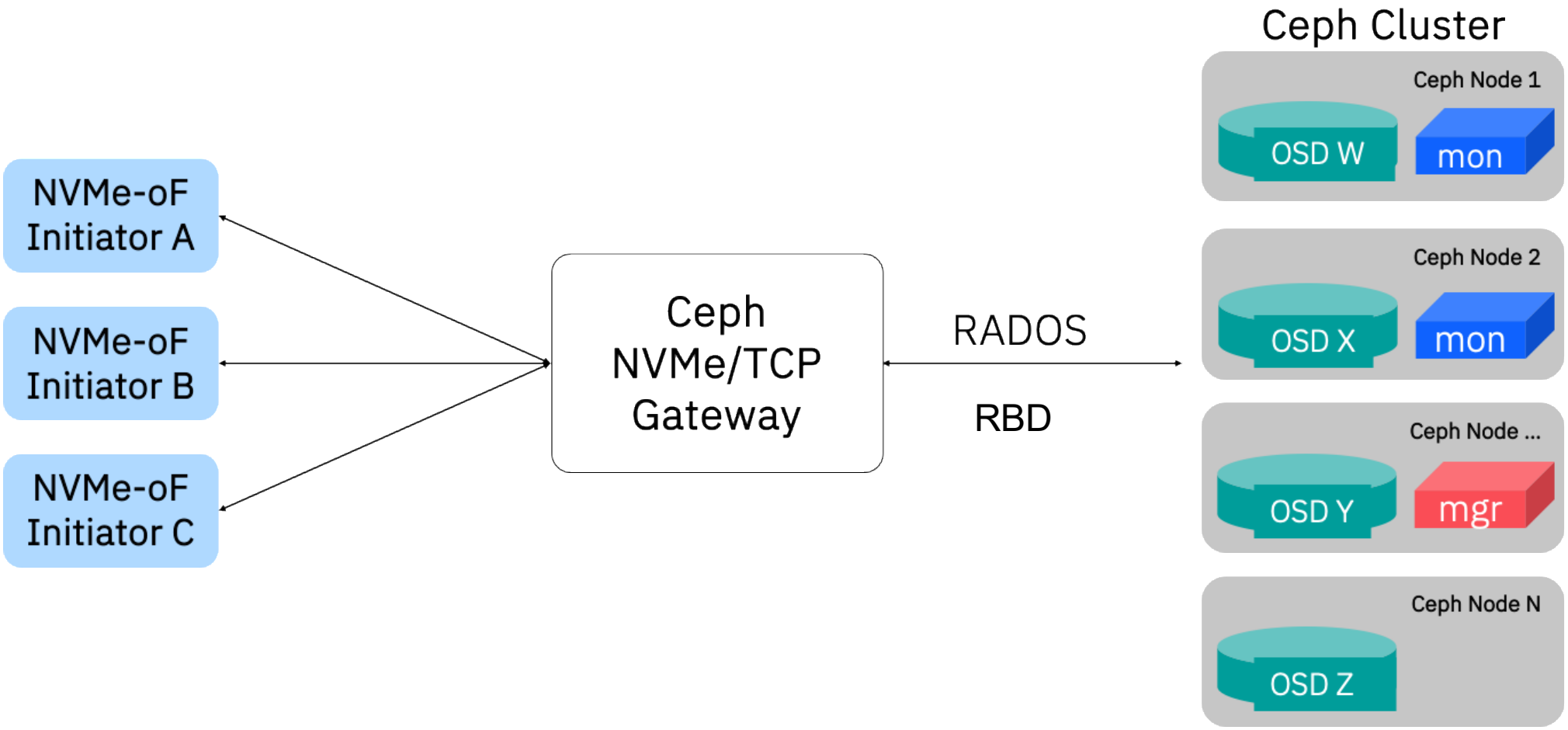
Components of Ceph NVMe over Fabrics

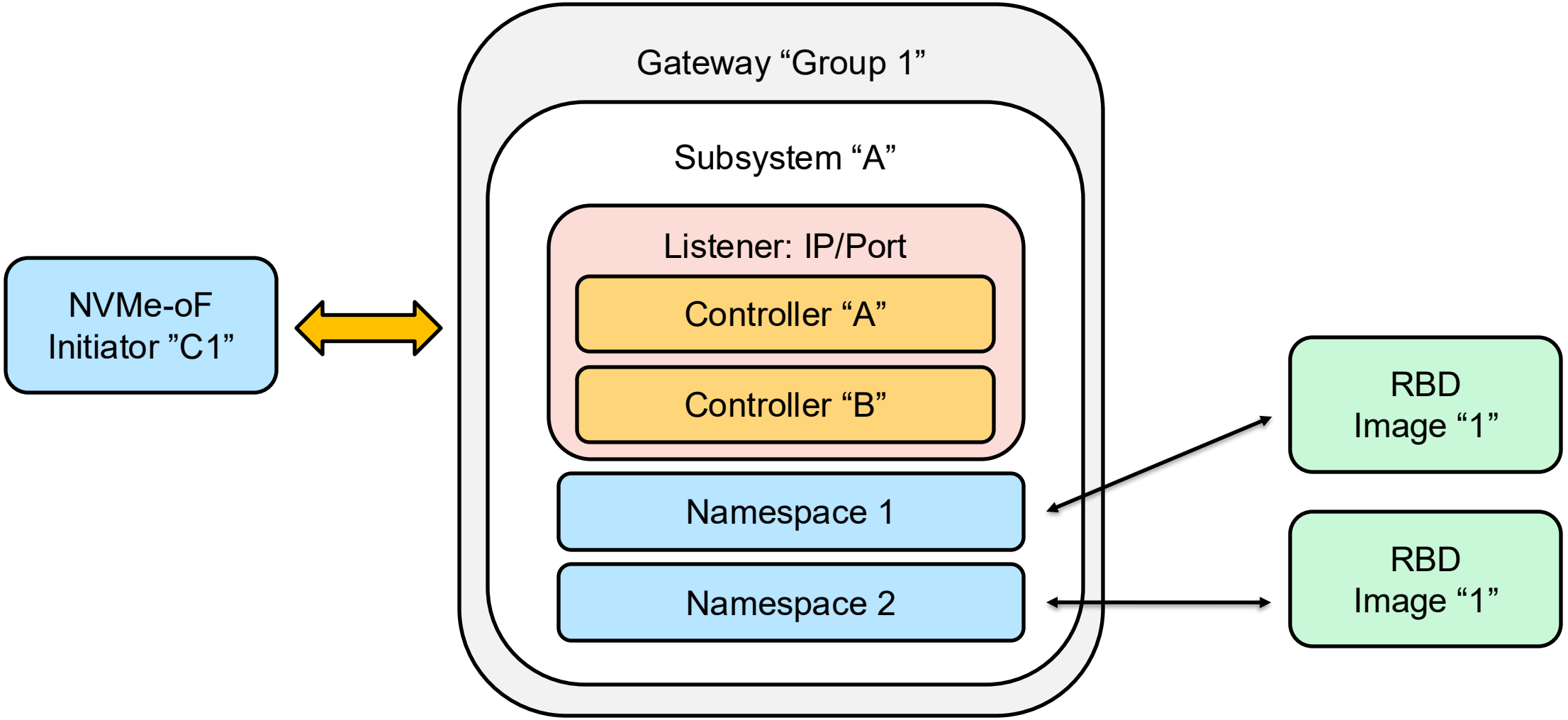


IBM Storage Ceph block components (RBD and NVMe)



Ceph NVMe/TCP fundamentals





Create a new Pool for NVMe images (*Navigation Pane -> Pools -> Create*)

IBM Storage Ceph: Cluster > Po

atg-ceph4-node1:8443/#/pool/create

IBM Storage Ceph

English

- Dashboard
- Multi-Cluster
- Cluster
  - Pools
  - Hosts
  - OSDs
  - Physical Disks
  - CRUSH Map
  - Monitors
- Block
  - Images
  - Mirroring
  - NVMe/TCP
- Object
- File
- Observability
- Administration

Cluster / Pools / Create

Create Pool

Name *	nvmepool
Pool type *	replicated
PG Autoscale	on
Replicated size *	3
Applications	rbd

☐ Mirroring

Check this option to enable Pool based mirroring on a Block(RBD) pool.

CRUSH

Crush ruleset	replicated_rule

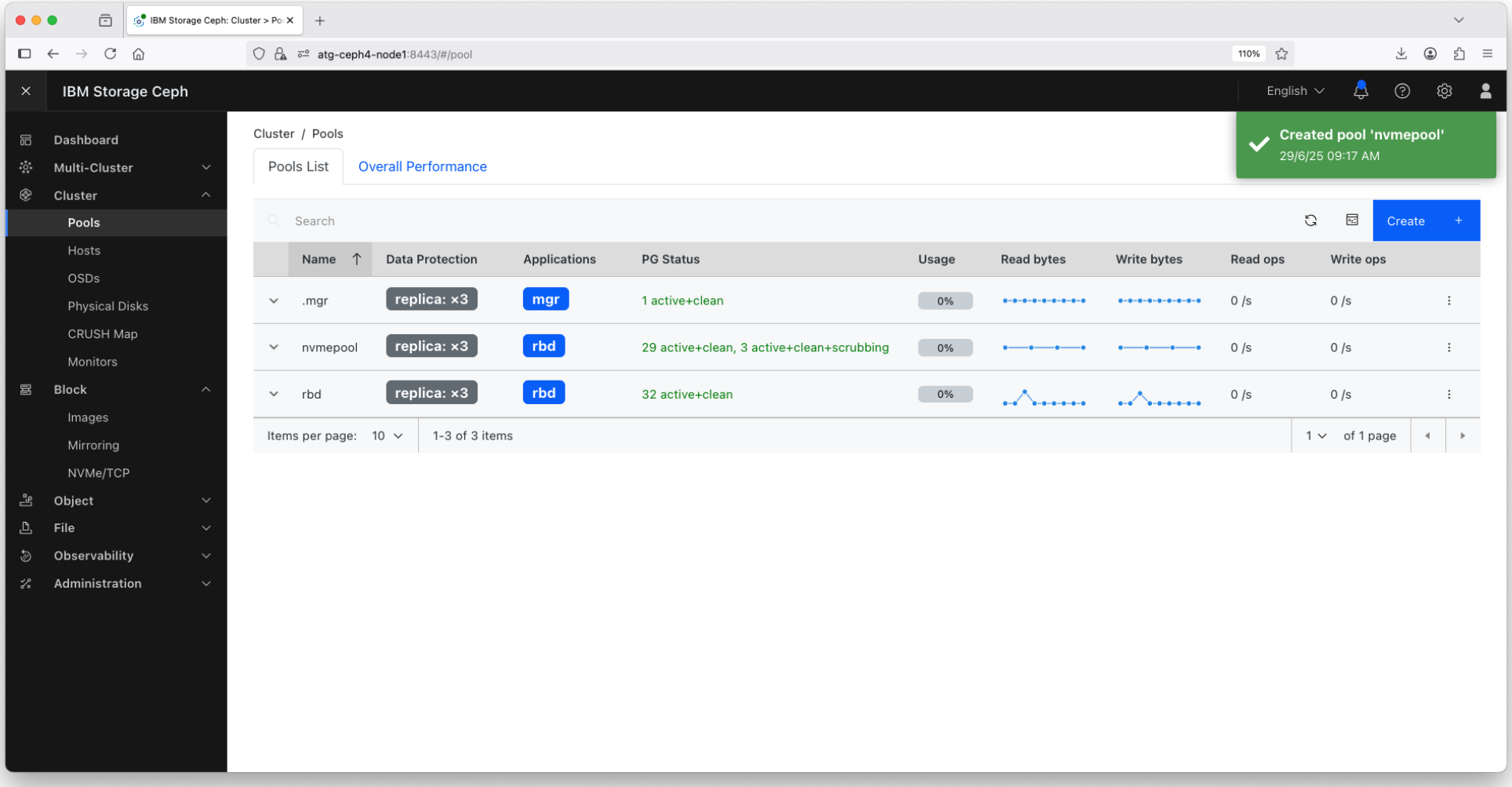
Compression

Mode	none

Quotas

Max bytes	e.g., 10GiB

Ceph Pool listing (*Navigation Pane -> Pools*)



Create the NVMe/TCP Service instances (*Administration-> Services -> Create*)

IBM Storage Ceph: Administrati

atg-ceph4-node1:8443/#/services/(modal:create)

IBM Storage Ceph

English

Dashboard
Multi-Cluster
Cluster
Block
Object
File
Observability
Administration
Services
Upgrade
Ceph Users
Manager Modules
Configuration

Administration / Services

Search

Create

Service	Placeme	Last Refreshed	Ports
alertmanager	count:1	8 minutes ago	9093,9094
ceph-exporter	*	9 minutes ago	-
crash	*	9 minutes ago	-
grafana	count:1	8 minutes ago	3000
mgr	atg-ceph	9 minutes ago	-
mon	atg-ceph	9 minutes ago	-
node-exporter	*	9 minutes ago	9100
osd.cost_capacity	*	9 minutes ago	-
prometheus	count:1	8 minutes ago	9095
rgw.s3service	atg-ceph	9 minutes ago	80

Items per page: 10 | 1-10 of 10 items | 1 of 1 page

Create Service

Type * nvmeof

Block Pool * nvmepool

An RBD application-enabled pool in which the gateway configuration can be managed.

Group Name * default

The name of the gateway group.

Service Name * nvmeof. nvmepool.default

☐ Unmanaged

If Unmanaged is selected, the orchestrator will not stop or stop any daemons associated with this service. Placement and all other properties will be ignored.

Placement Hosts

Hosts atg-ceph4-node3 ✖ atg-ceph4-node4 ✖

☐ Encryption

Enables mutual TLS (mTLS) between the client and the

Cancel | Create Service

Verify startup of the NVMe/TCP Service (*Administration-> Services -> Daemons*)

IBM Storage Ceph: Administrati

atg-ceph4-node1:8443/#/services

IBM Storage Ceph

English

- Dashboard
- Multi-Cluster
- Cluster
- Block
- Object
- File
- Observability
- Administration
  - Services
  - Upgrade
  - Ceph Users
  - Manager Modules
  - Configuration

ceph-exporter	*	4 / 4	9 minutes ago	-	
crash	*	4 / 4	9 minutes ago	-	
grafana	count:1	1 / 1	9 minutes ago	3000	
mgr	atg-ceph4-node1;atg-ceph4-node2	2 / 2	9 minutes ago	-	
mon	atg-ceph4-node1;atg-ceph4-node2;atg-ceph4-node3	3 / 3	9 minutes ago	-	
node-exporter	*	4 / 4	9 minutes ago	9100	
nvmeof.nvmepool.default	atg-ceph4-node3;atg-ceph4-node4	2 / 2	A minute ago	5500,4420,8009,10008	

Daemons | Service Events

Search | Hostname | Any

Hostname	Daemon name	Version	Status	Last Refreshed	CPU Usage	Memory Usage	Daemon Events
atg-ceph4-node3	nvmeof.nvmepool.default.atg-ceph4-node3.sysalm	1.4.14	running	A minute ago	13%	53 MiB	
atg-ceph4-node4	nvmeof.nvmepool.default.atg-ceph4-node4.zyosmc	1.4.14	running	A minute ago	6%	48.5 MiB	

Items per page: 10 | 1-2 of 2 items | 1 of 1 page

osd.cost_capacity	*	16 / 16	9 minutes ago	-	
prometheus	count:1	1 / 1	9 minutes ago	9095	

Items per page: 10 | 1-10 of 11 items | 1 of 2 pages

Show NVMe Gateways (*Navigation Pane -> Block -> NVMe/TCP -> Gateways*)

IBM Storage Ceph: Block > NVM

atg-ceph4-node1:8443/#/block/nvmeof/gateways

IBM Storage Ceph

English

- Dashboard
- Multi-Cluster
- Cluster
- Block
  - Images
  - Mirroring
  - NVMe/TCP
- Object
- File
- Observability
- Administration

Block / NVMe/TCP / Gateways

Subsystems | Gateways

Selected Gateway Group

default

Gateways | Overview | Performance

Gateways

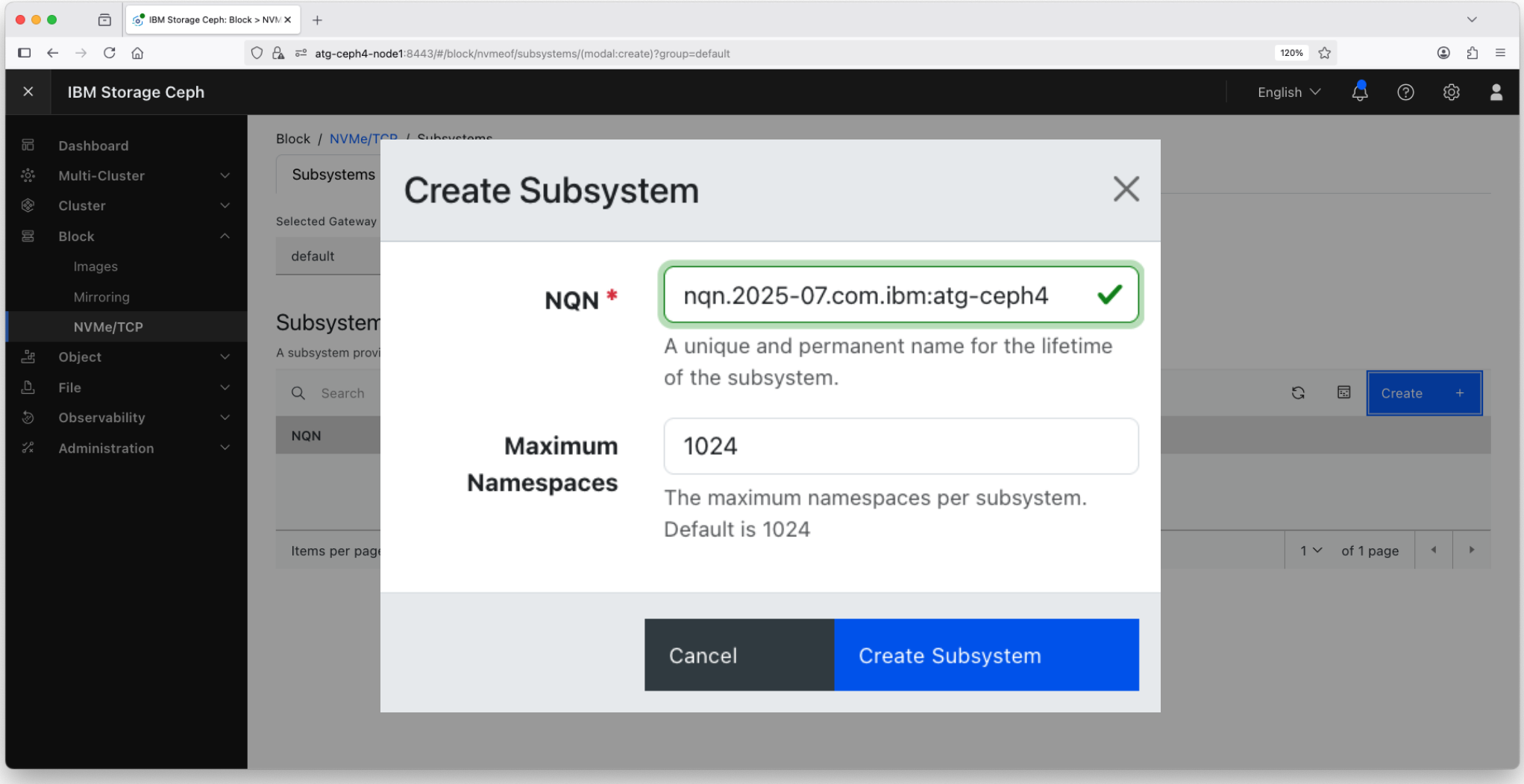
Ceph NVMe-oF gateways provide Ceph Block Device storage through NVMe/TCP. For VMware clients the NVMe/TCP volumes display as VMFS Datastores. For Linux clients the NVMe/TCP volumes display as as block devices.

Search

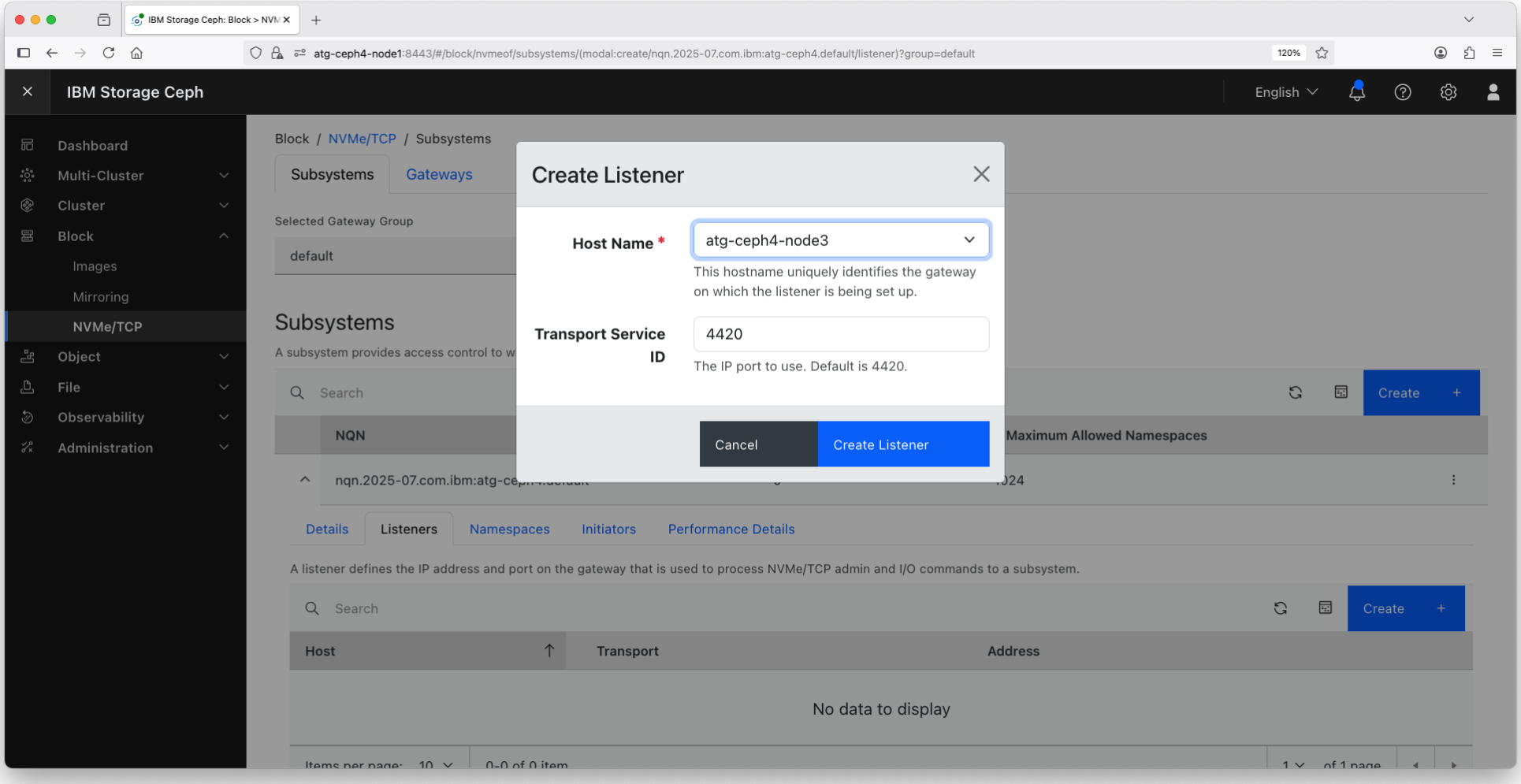
Gateway ID	Hostname	Status
client.nvmeof.nvmepool.default.atg-ceph4-node3.udxrbz	atg-ceph4-node3	running
client.nvmeof.nvmepool.default.atg-ceph4-node4.zcaesr	atg-ceph4-node4	running

Items per page: 10 | 1-2 of 2 items | 1 of 1 page

Create an NVMe subsystem (*NVMe/TCP -> Subsystems -> Create*)

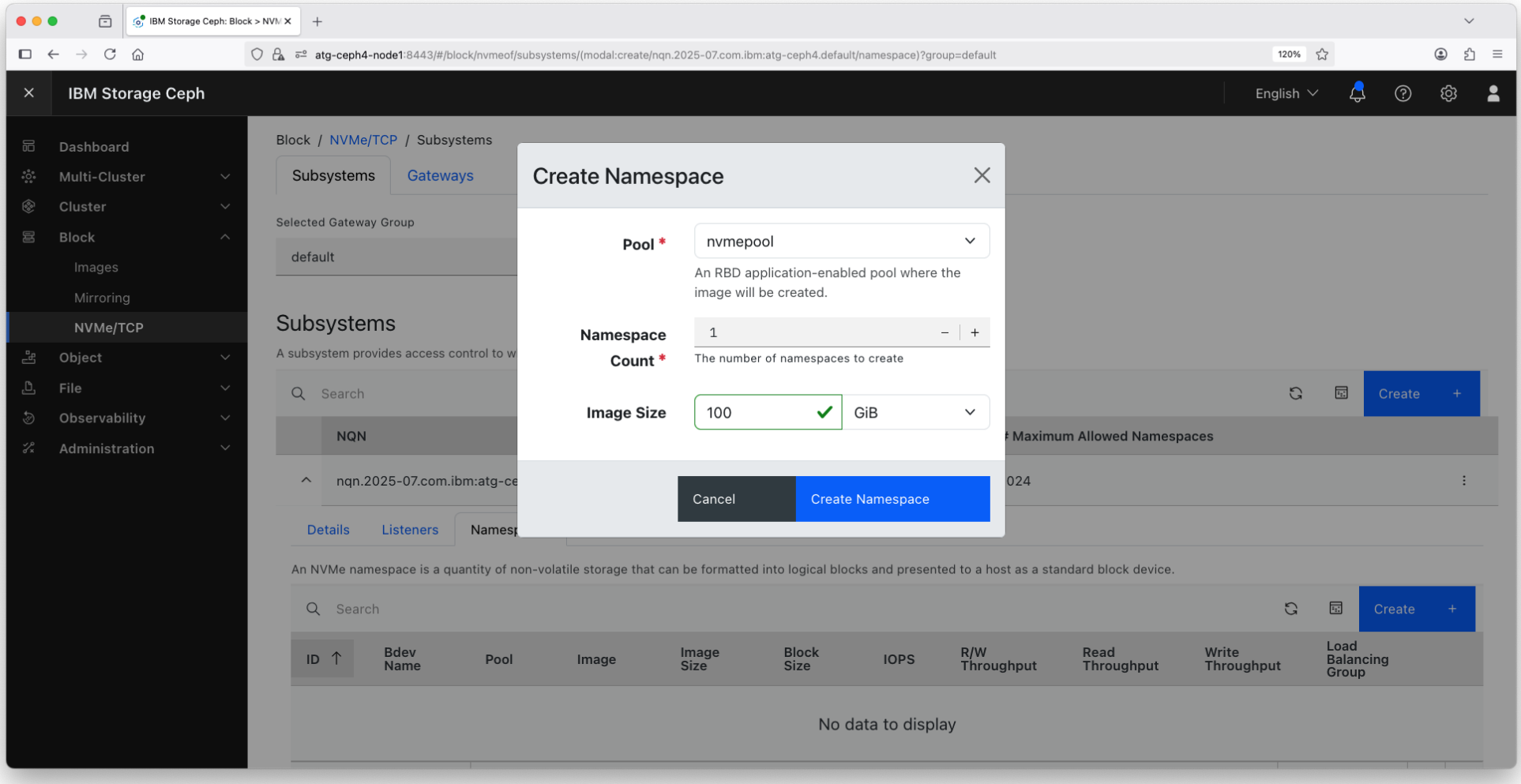


Create NVMe Subsystem Listeners (*NVMe/TCP -> Subsystems -> Listeners*)



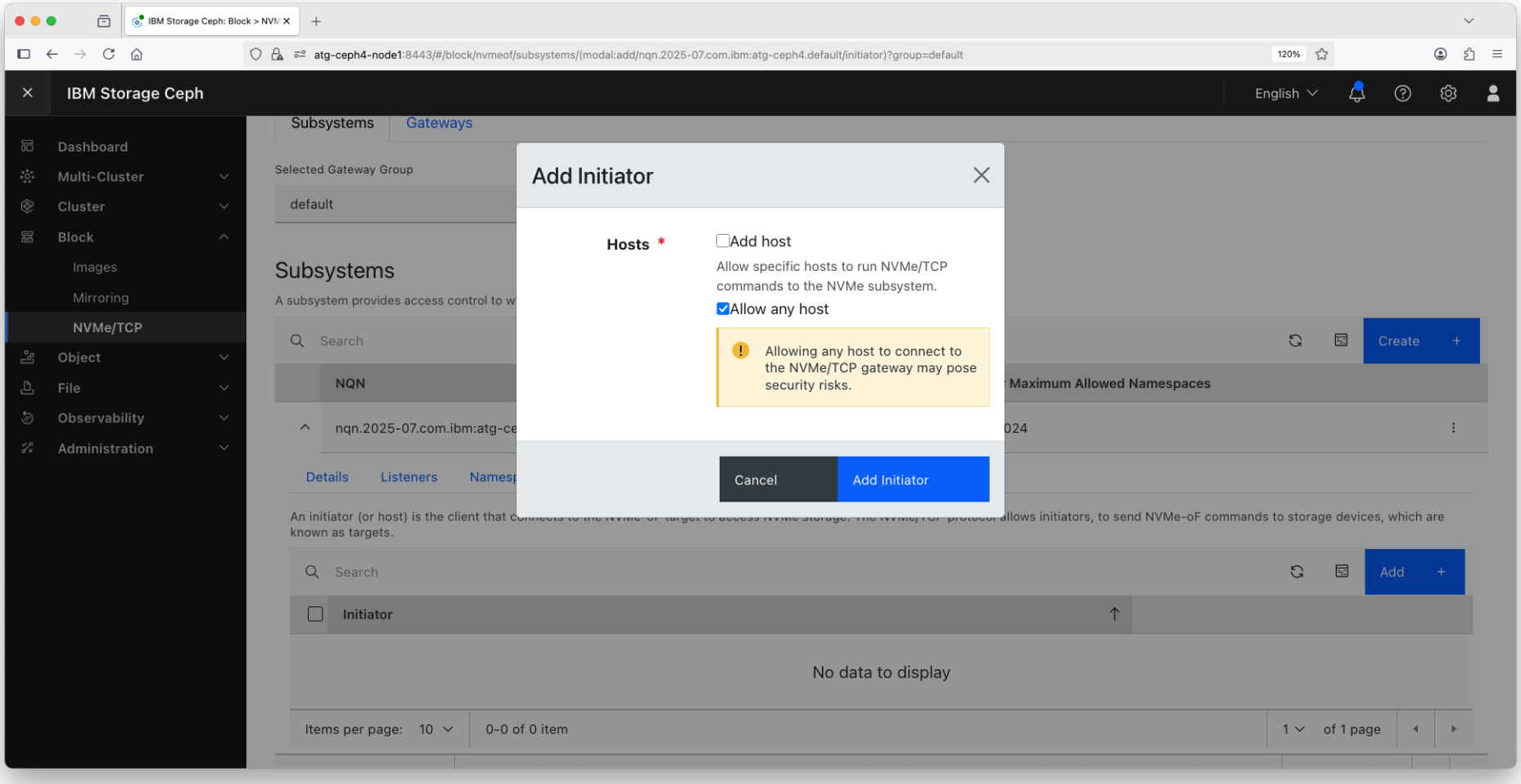
Best practice: Create two or more listeners for multi path high availability

Create NVMe Subsystem Namespaces (*Subsystems -> Namespace*)



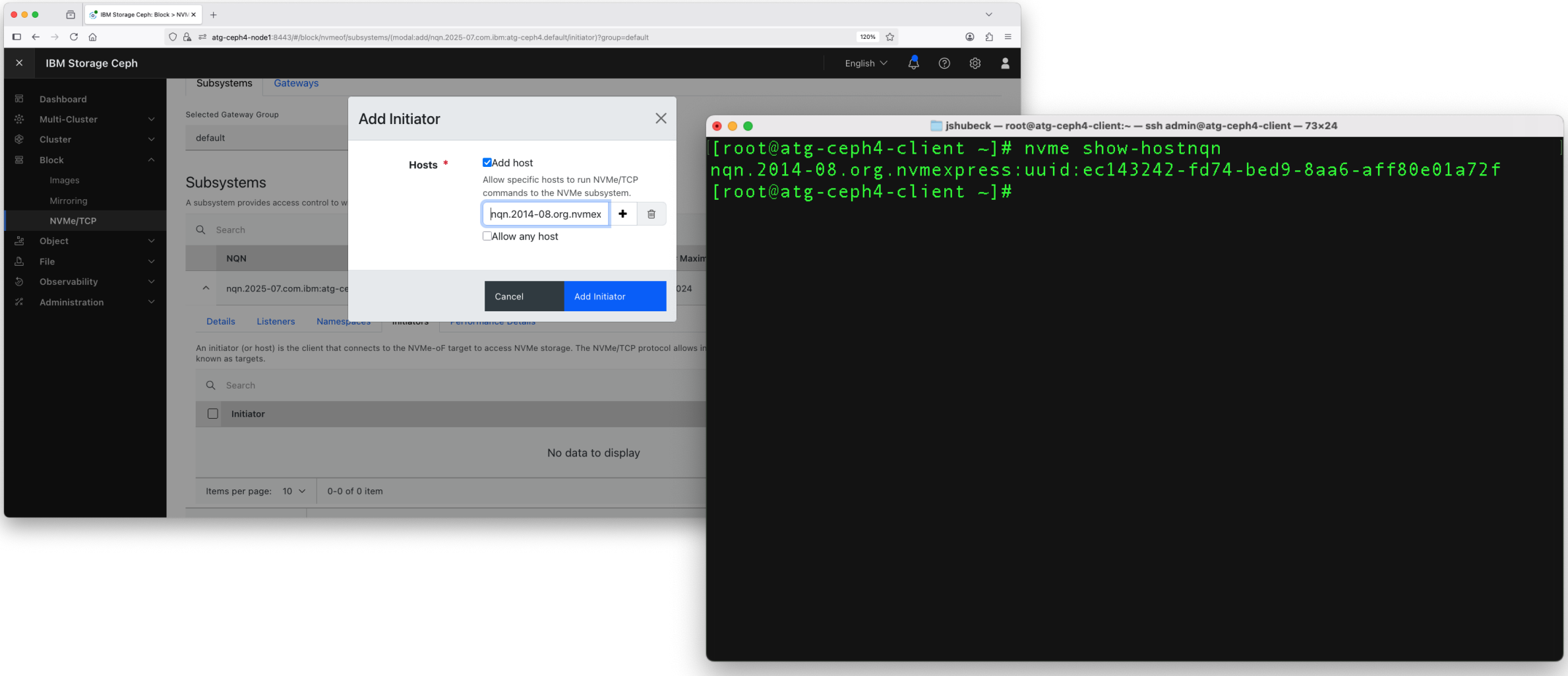
NOTE: The Ceph Dashboard will automatically create an RBD Image for each Namespace

Allow any NVMe Initiator (*Subsystems -> Initiators -> Add*)



ALERT: Allowing any host might be OK for a POC or an isolated network, but otherwise not recommended.

Add a named NVMe Initiator (*Subsystems -> Initiators -> Add*)



TIP: Use "*nvme show-hostnqn*" on the Linux client to get the NQN of the initiator.

Ceph Dashboard (*Navigation Pane -> Block -> Images*)

IBM Storage Ceph: Block > Imag

atg-ceph4-node1:8443/#/block/rbd

IBM Storage Ceph

English

- Dashboard
- Multi-Cluster
- Cluster
- Block
  - Images
  - Mirroring
  - NVMe/TCP
- Object
- File
- Observability
- Administration

Block / Images

Images | Namespaces | Trash | Overall Performance

Search

Create +

Name	Pool	Namespace	Size	Usage	Objects	Object size	Parent	Mirroring	Next Scheduled Snapshot
nvme_nvmepool_default_1s9viloo6t5	nvmepool		100 GiB	0%	25.6 k	4 MiB	-	Disabled	
rbdimage	rbdpool		100 GiB	0%	25.6 k	4 MiB	-	Disabled	

Items per page: 10 | 1-2 of 2 items | 1 of 1 page

NVMe/TCP client experience

IBM Storage Ceph NVMe client commands

```
# Install the NVMe CLI client
# dnf install -y nvme-cli
. . . Output omitted . . .
Last metadata expiration check: 2:46:07 ago on Mon 07 Jul 2025 11:20:01 AM EDT.
Package nvme-cli-2.11-5.el9.x86_64 is already installed.
Dependencies resolved.
Nothing to do.
Complete!
# Discover NVMe namespaces
# modprobe nvme-fabrics
# nvme discover -t tcp -a atg-ceph4-node3 -s 8009
=====Discovery Log Entry 0======
trtype:  tcp
adrfam:  ipv4
subtype: nvme subsystem
. . . output omitted . . .
subnqn:  nqn.2025-07.com.ibm:atg-ceph4.default
traddr:  192.168.65.113
. . . output truncated . . .
```

IBM Storage Ceph NVMe client commands (continued)

```
# Initiate a connection to the NVMe target
# nvme connect -t tcp -a atg-ceph4-node3 -n nqn.2025-07.com.ibm:atg-ceph4.default
connecting to device: nvme0

# nvme list
/dev/nvme0n1   /dev/ng0n1    Ceph36920816290045   Ceph bdev Controller

# Traditional local storage formatting and mount
# mkfs.xfs /dev/nvme0n1
meta-data=/dev/nvme0n1           isize=512    agcount=4, agsize=6553600 blks
. . . output omitted . . .
Discarding blocks...Done.

# mkdir /mnt/nvme0n1
# mount /dev/nvme0n1 /mnt/nvme0n1
```

IBM Storage Ceph NVMe client commands (continued)

```
# Create directories and write data
# mkdir /mnt/nvme0n1/dir1
# mkdir /mnt/nvme0n1/dir2

# touch /mnt/nvme0n1/dir1/atestfile
# dd if=/dev/random of=/mnt/nvme0n1/dir1/10MB.dat bs=1M count=10
10+0 records in
10+0 records out
10485760 bytes (10 MB, 10 MiB) copied, 0.0870174 s, 121 MB/s
# cat <<EOF > /mnt/nvme0n1/dir2/zip-codes.csv
98333, Fox Island, WA
98335, Gig Harbor, WA
EOF

# cat /mnt/nvme0n1/dir2/zip-codes.csv
98333, Fox Island, WA
98335, Gig Harbor, WA
```

IBM Storage Ceph NVMe client commands (finished)

```
# Verifcation commands
# tree /mnt
/mnt
└──── nvme0n1
    ├──── dir1
    │     ├──── 10MB.dat
    │     └──── atestfile
    └──── dir2
          └──── zip-codes.csv

# df /mnt/nvme0n1
Filesystem     1K-blocks   Used Available Use% Mounted on
/dev/nvme0n1   104792064 773932 104018132   1% /mnt/nvme0n1
```

Cephadm commands for NVMe

```
# Ceph verification command
[ceph: root@atg-ceph4-node1 /]# ceph df
--- RAW STORAGE ---
CLASS     SIZE    AVAIL     USED  RAW USED  %RAW USED
hdd    256 GiB  247 GiB  8.5 GiB   8.5 GiB       3.32
TOTAL  256 GiB  247 GiB  8.5 GiB   8.5 GiB       3.32

--- POOLS ---
POOL                    ID  PGS   STORED  OBJECTS     USED  %USED  MAX AVAIL
.mgr                     1    1  449 KiB        2  1.3 MiB      0     77 GiB
.rgw.root                2   32  1.9 KiB        6   72 KiB      0     77 GiB
cephfs.vol01.meta        3   32  8.4 KiB      242  792 KiB      0     77 GiB
cephfs.vol01.data        4   32      0 B        8      0 B      0     77 GiB
. . . output omitted . . .
default.rgw.buckets.data 7  512  2.3 MiB        7  7.0 MiB      0     77 GiB
nvmepool                 8   32   12 MiB       14   37 MiB   0.02     77 GiB
rbdpool                  9   32      0 B        0      0 B      0     77 GiB
```

Client NVMe-oF cleanup

```
[root@client ~]# umount /mnt/nvme0n1

[root@client ~]# nvme disconnect --nqn nqn.2001-07.com.ibm:atg-ceph4.group1

[root@client ~]# nvme list

Node             Generic       SN                   Model
-------------    -------------- -------------------- ----------------------
/dev/nvme0n1     /dev/ng0n1     Ceph36920816290045   Ceph bdev Controller


#
# The NVMe namespace is still there but we are no longer mapped to it
#
```

Ceph RBD and NVMe/TCP principles of operation – day 2

Monitoring dashboard for NVMe/TCP *(NVMe/TCP -> Overview -> Performance)*

IBM Storage Ceph: Block > NVM

atg-ceph4-node1:8443/#/block/nvmeof/gateways

IBM Storage Ceph

English

Dashboard
Multi-Cluster
Cluster
Block
Images
Mirroring
NVMe/TCP
Object
File
Observability
Administration

Block / NVMe/TCP / Gateways

Subsystems | Gateways

Selected Gateway Group

default

Gateways | Overview | Performance

Grafana Time Picker Last 1 hour (Default)

Overview

Gateway Groups	Total Gateways
1	2

Gateways per group

Value	
default	2

Ceph Health NVMeoF WARNING

0

Active Alerts

No alerts matching filters

NVMeoF Group Overview - default

Gateways	Subsystems	Namespaces	Connected Client:	Capacity Exported	Subsystem Security
2	1	1	1	100 GiB	OK

Active gateways count

Top 5 Busiest Gateway CPU

IOPS - All

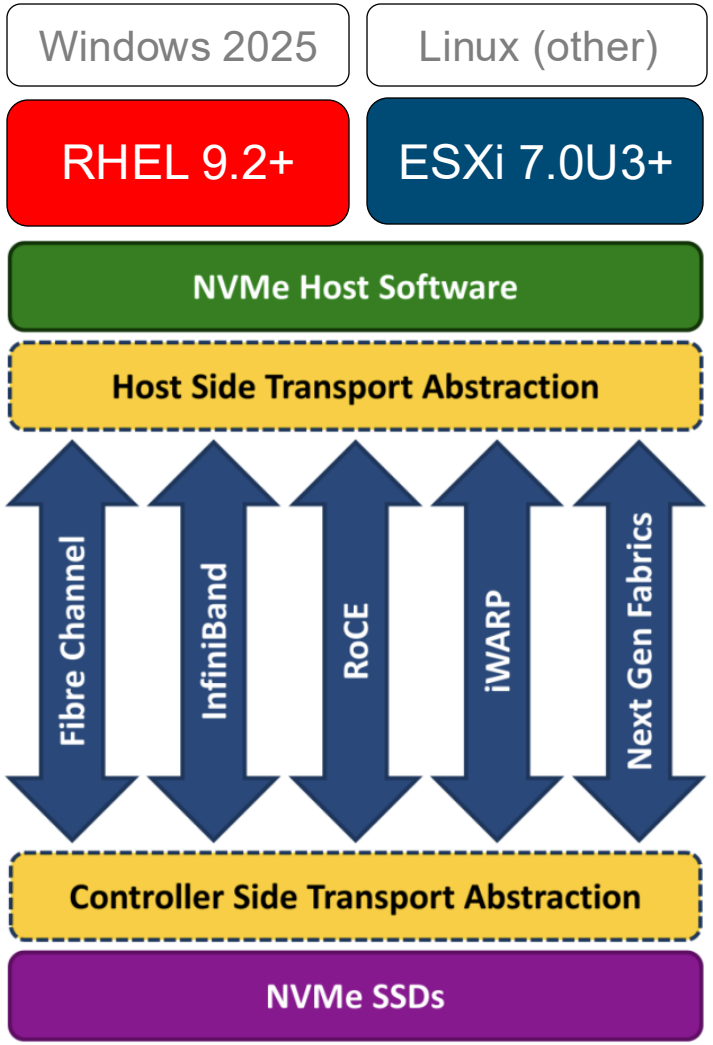
What about scalability?

Scaling-out with NVMe-oF gateway

- Four (4) NVMe gateway groups
- Eight (8) NVMe gateways in a group
- 64 100GbE interfaces (4 groups * 8 gateways per group * 2)
- 128 NVMe subsystems per gateway group
- 32 hosts per NVMe subsystem
- 1024 namespaces per NVMe per gateway group

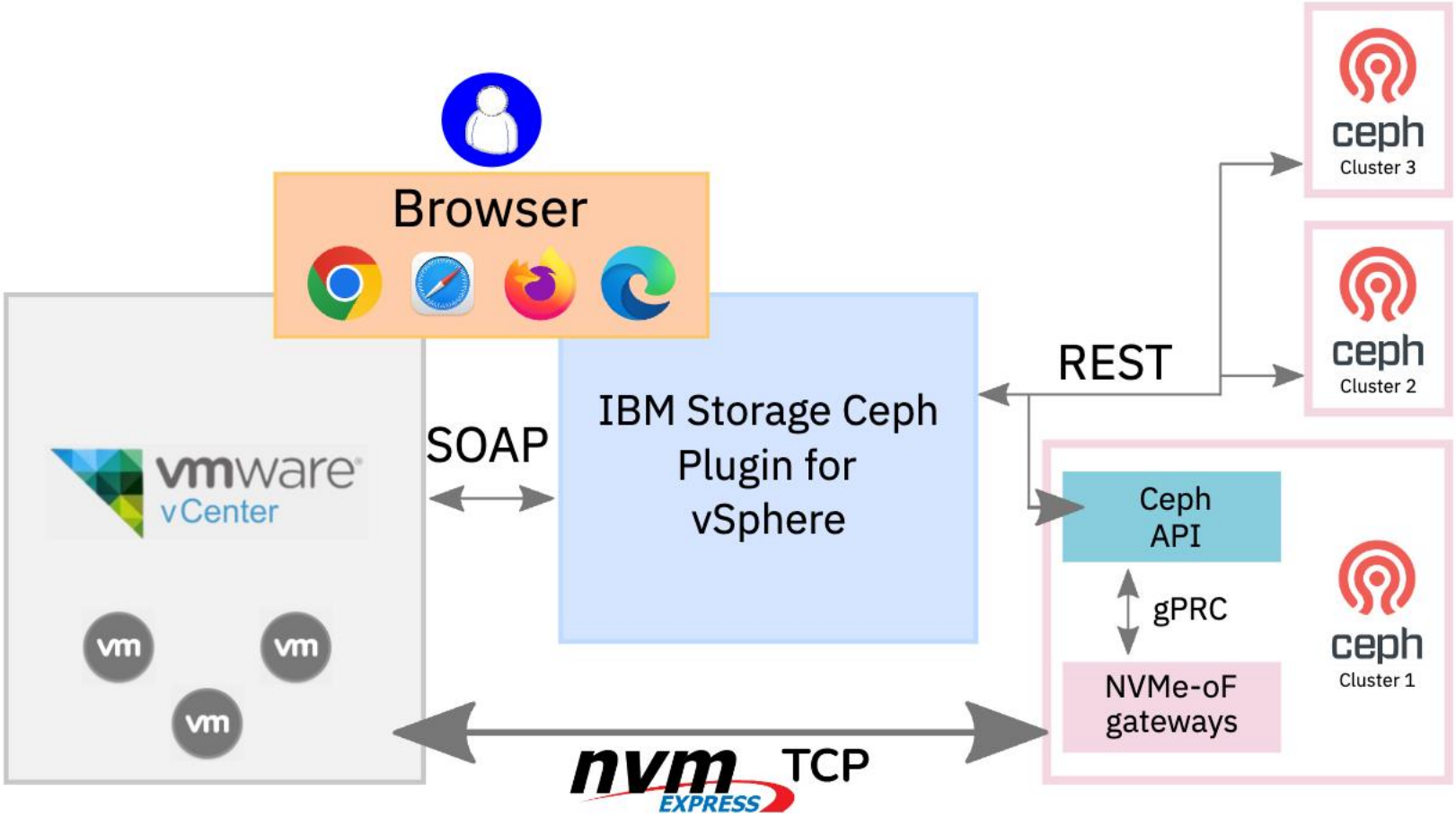
⚠ **Important:** An NVMe-oF gateway can only be part of one gateway group and should never be part of two or more gateway groups.

ⓘ **Note:** The RHEL and ESXi initiators can have smaller NVMe over Fabric namespace discovery imits. Confirm all discovery limits with your software vendor and version.

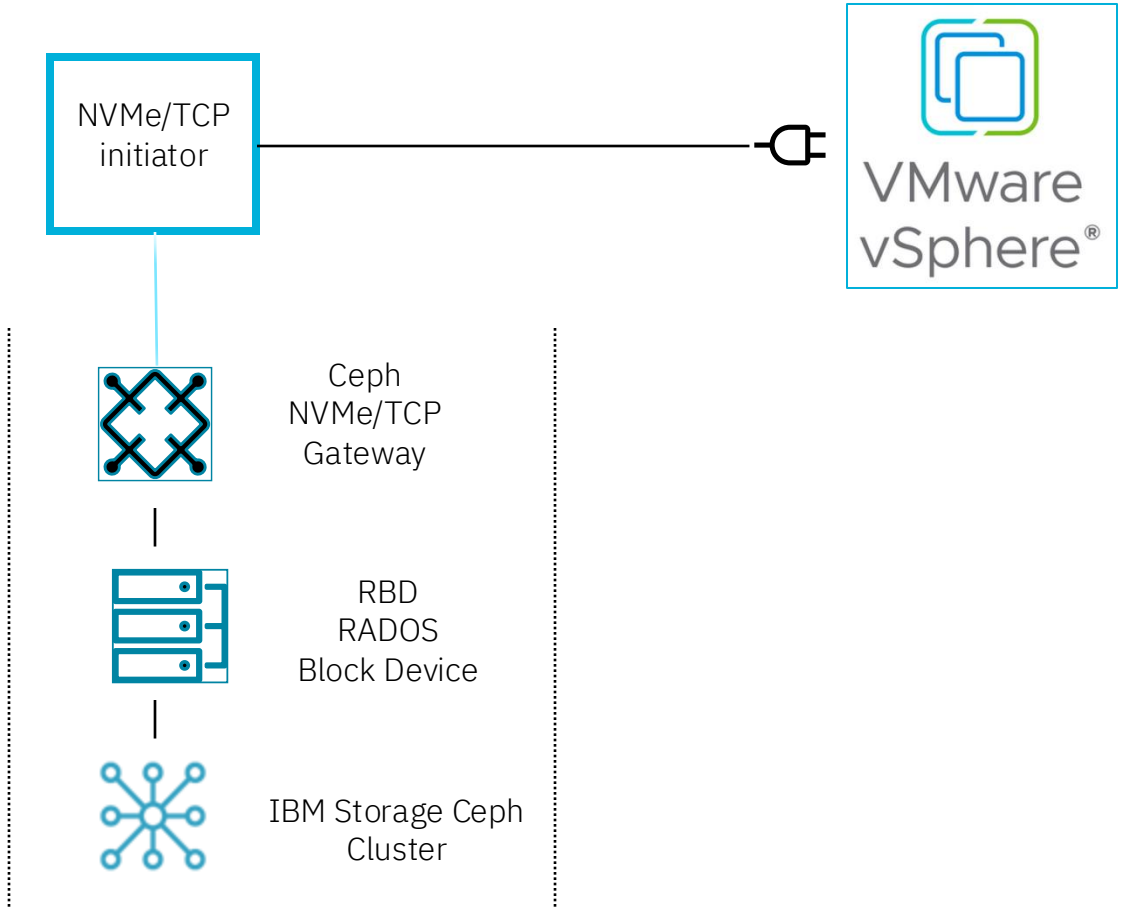


vSphere plug-in for IBM Storage Ceph NVMe/TCP

The IBM Storage Ceph Plugin for vSphere allows management and control of storage volumes from within the VMware vSphere™ Client. The plug-in and associated workflow is integrated into the vSphere user interface and allows for provisioning and management of VMFS datastores across multiple storage systems



VMware vSphere plug-in enhancements



VMware vSphere plug-in

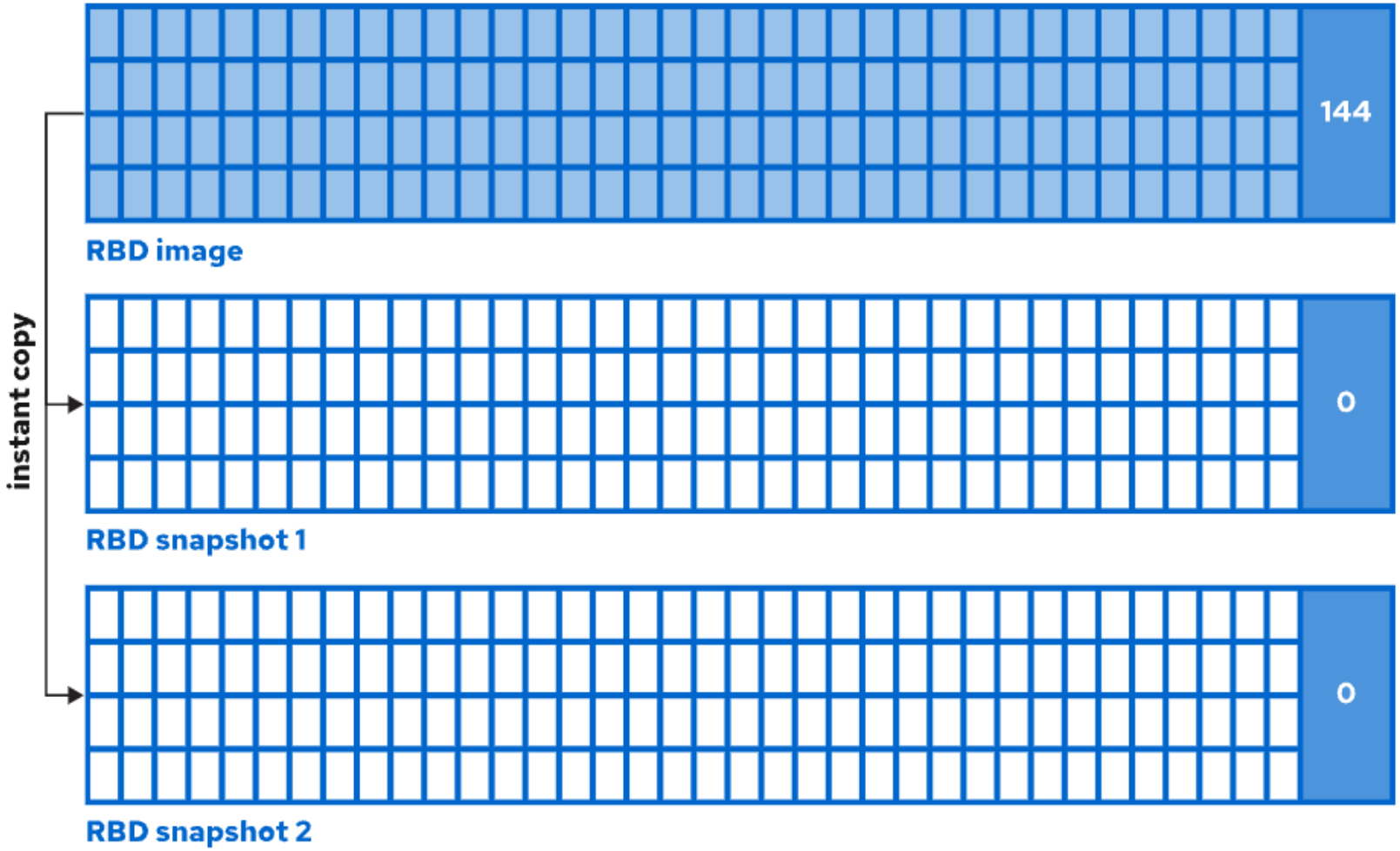
- Multi NVMe/TCP Gateway support
- Multi IBM Storage Ceph Cluster support
- Software upgrades
- Snapshot Management (Tech Preview)

IBM Storage Ceph enhancements

- Create non-default IBM Storage Ceph RBD Pool
- Create non-default IBM Storage Ceph RBD subsystem.
- VMware 7 update 3 and higher or vSphere 8.0 to support NVMe over TCP

Ceph RBD snapshots

RBD snapshots are read-only copies of an RBD image created at a particular time.



Ceph RBD snapshot algorithm

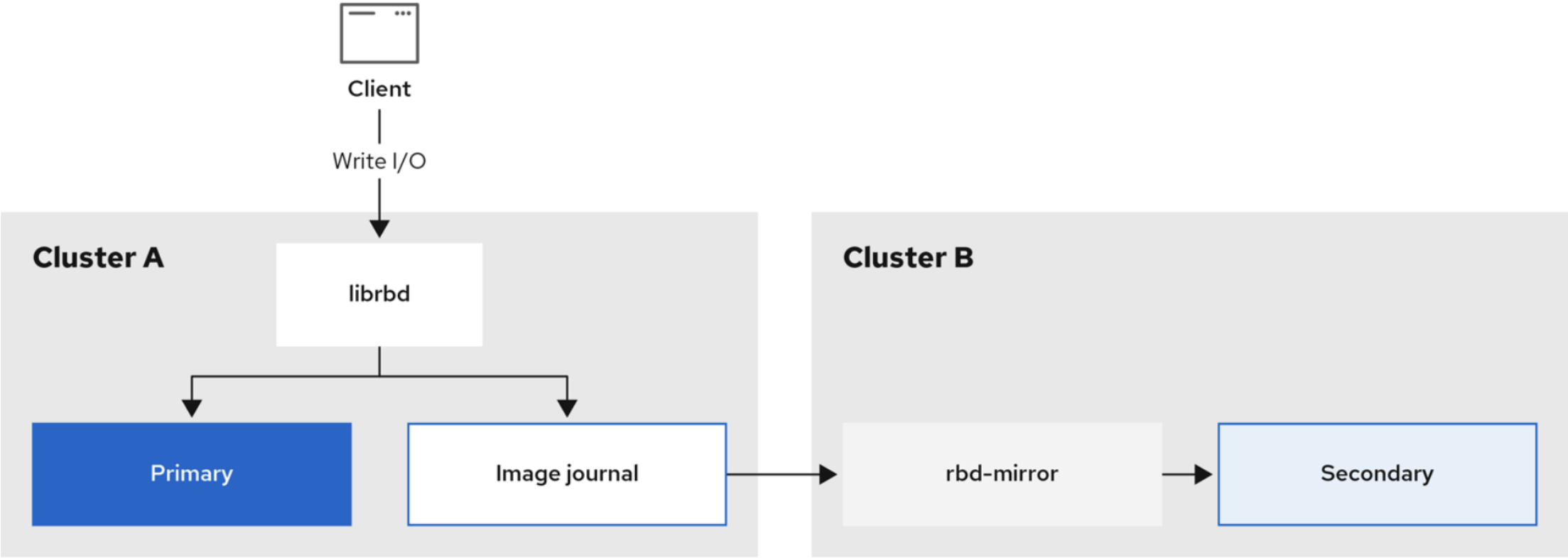
The snapshot COW procedure operates at the *object* level, regardless of the size of the write I/O request made to the RBD image. If you write a single byte to an RBD image that has a snapshot, then Ceph copies the entire affected object from the RBD image into the snapshot area.

Ceph RBD clones – Day 3

RBD *clones* are read/write copies of an RBD image that use a protected RBD snapshot as a base. An RBD clone can also be flattened, which converts it into an RBD image independent of its source.

Ceph RBD Mirroring – Day 3

RBD *clones* are read/write copies of an RBD image that use a protected RBD snapshot as a base. An RBD clone can also be flattened, which converts it into an RBD image independent of its source.



Ceph NVMe/TCP principles of operation – day 3

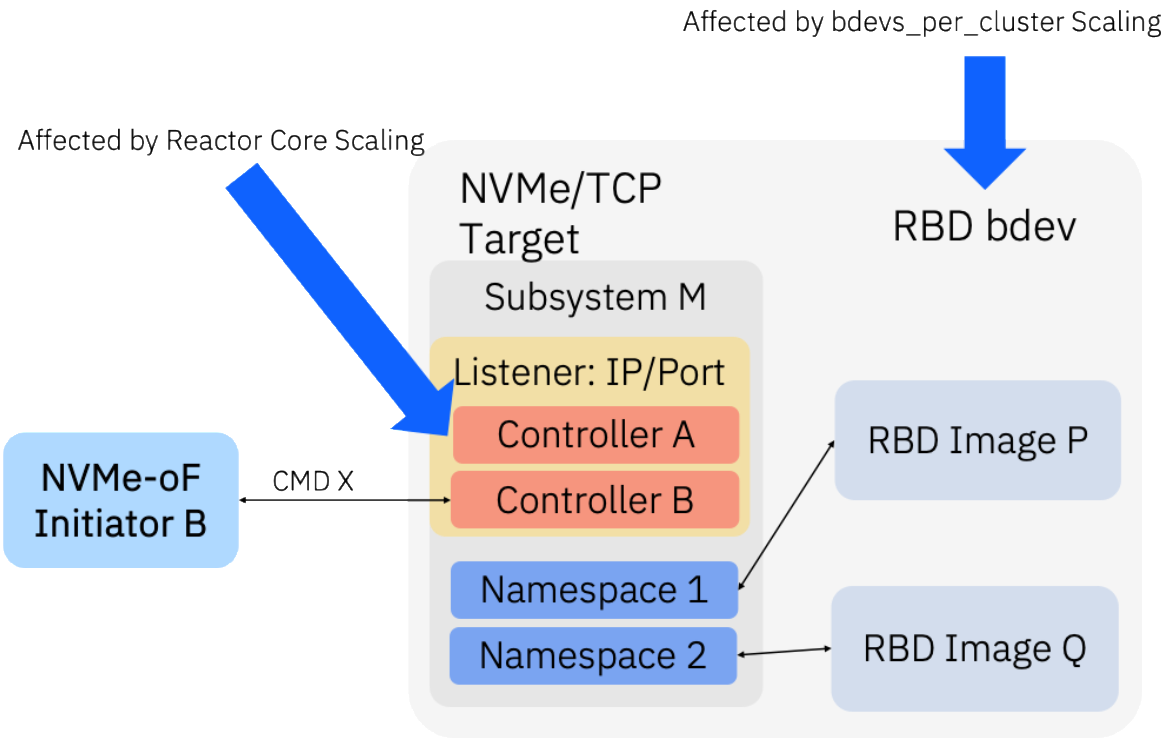
Ceph NVMe/TCP principles of operation – *Day 3*

NVMe-oF Commands

- RBD backend in SPDK
  maps NVMe operations to RBD API

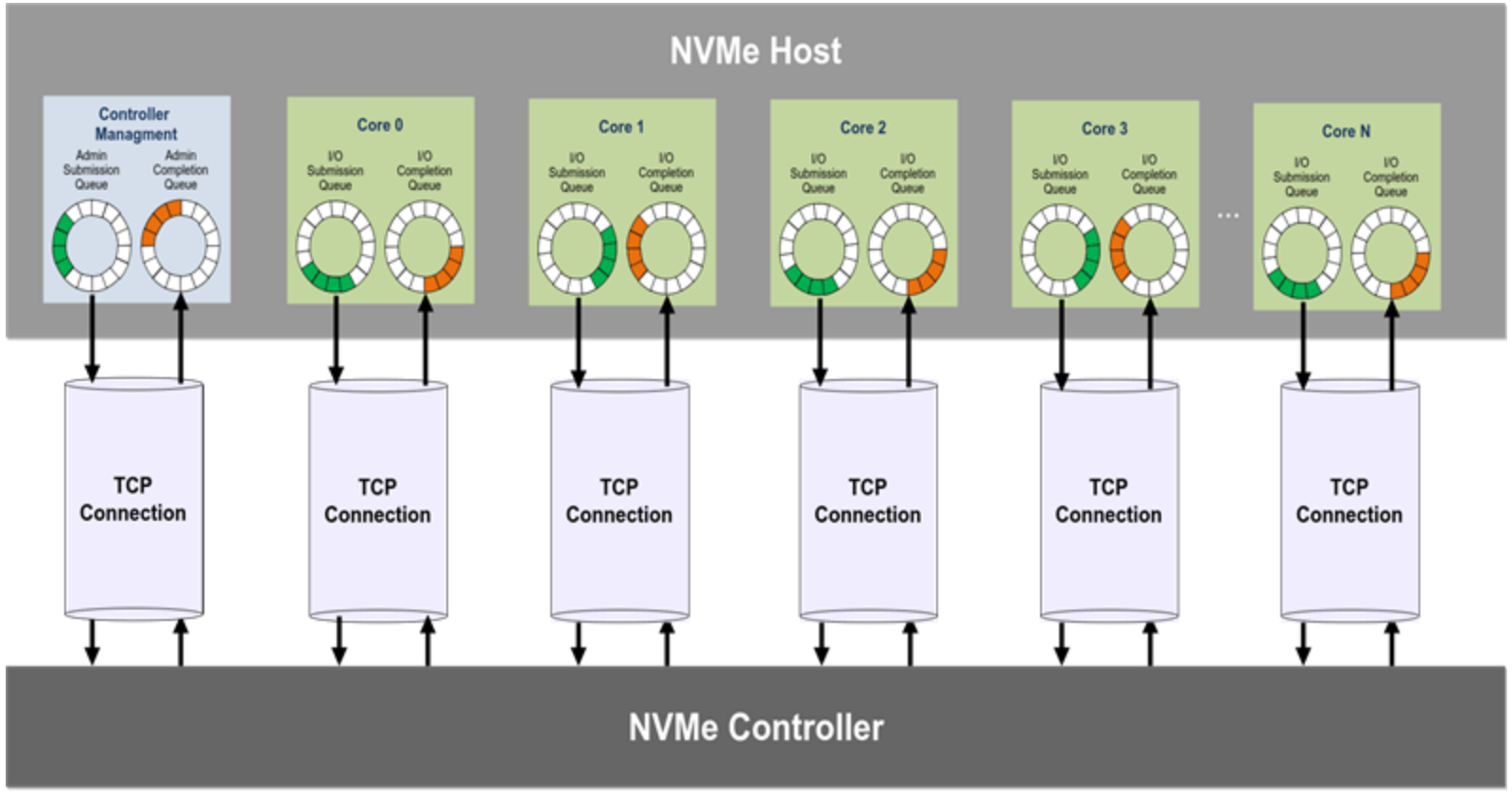
- Natively supported
  - Read
  - Write
  - Unmap
  - Flush
  - Write zeroes
  - Compare and write

- Emulated
  - Compare
  - Copy
  - Abort**



NVMe client connects and controllers (Reactor core association model)

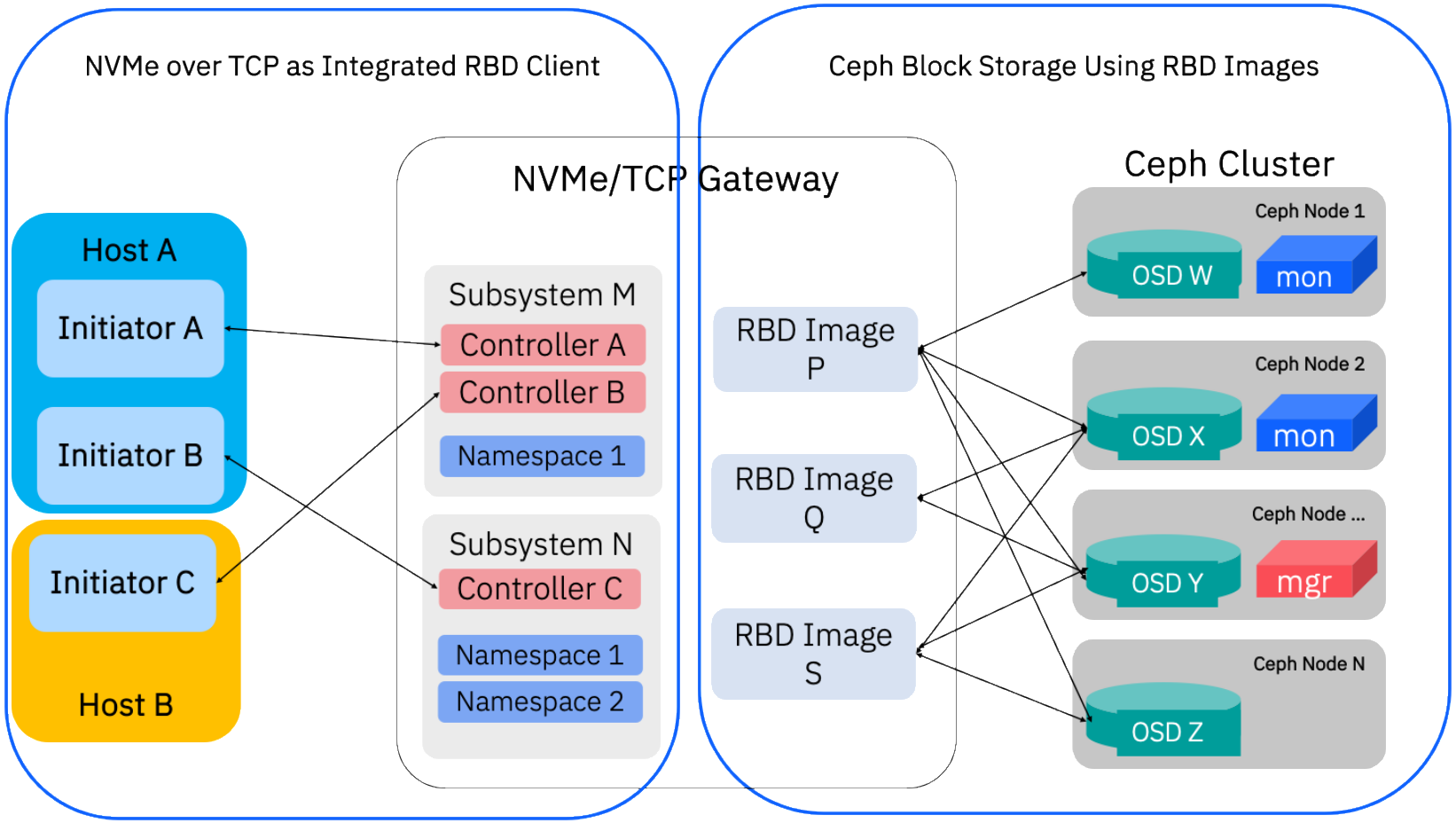
NVMe Controllers



- *Ephemeral*: For each client session, a unique and ephemeral controller is associated
- Connection binding is performed in NVMe-oF connection time
- No controller-wide sequencing or reassembly constraints
- No shared state across NVMe queues and TCP

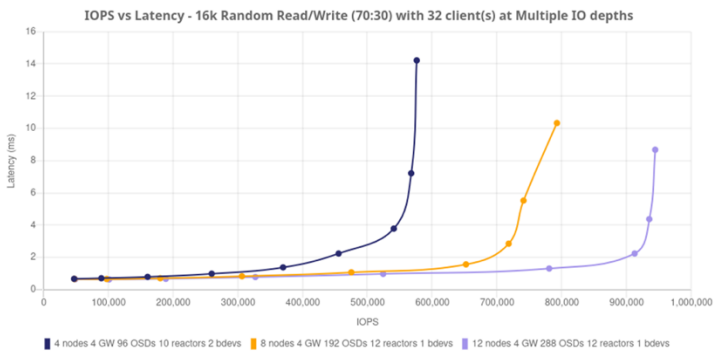
End to end NVMe over Fabrics ecosystem

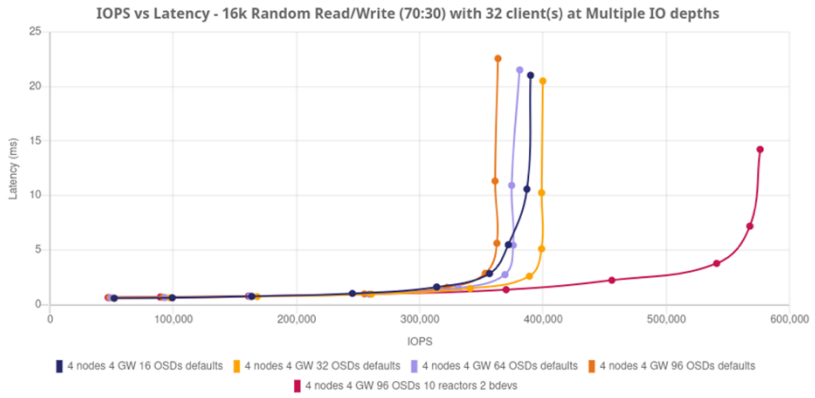
- Gateway is the Ceph NVMe/TCP service running on at least 2 nodes
- Namespace is mapped to an RBD image
- Subsystem is a logical grouping of Namespaces
- Initiator is the client software that gets a Controller within the Gateway (for the duration of the session)

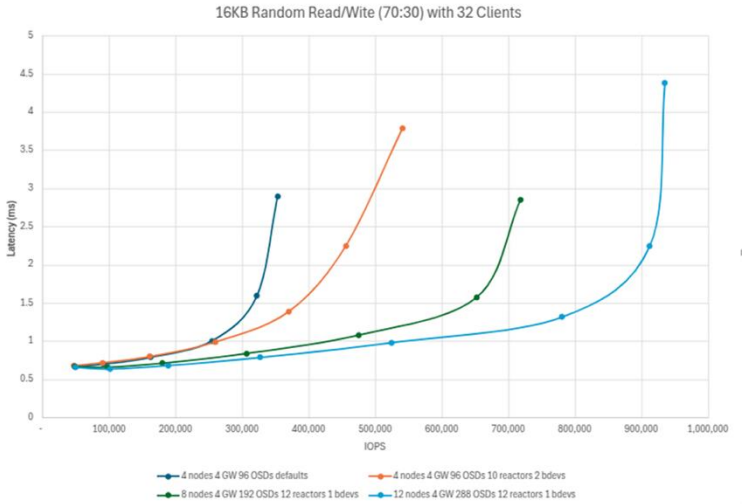


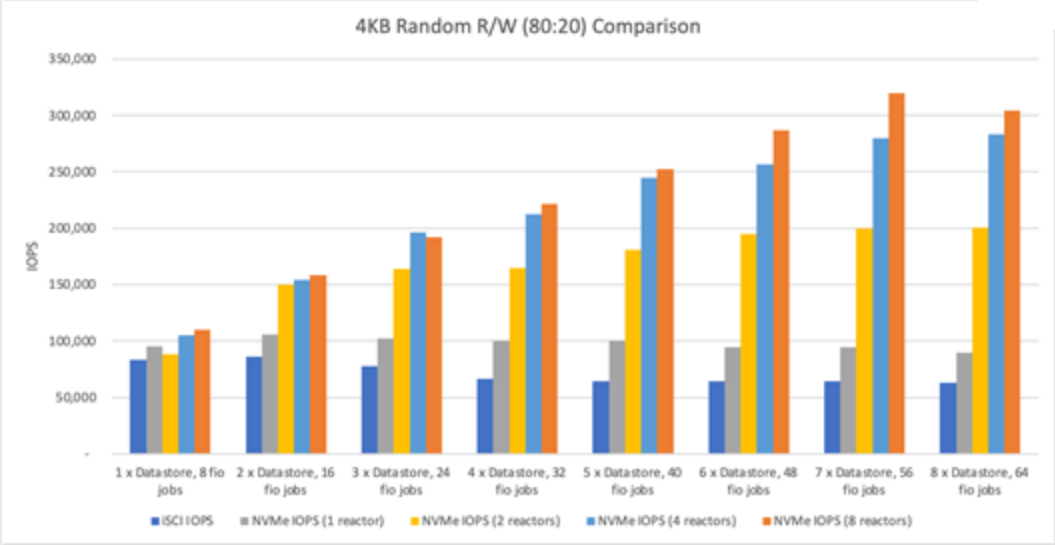
What about performance?

https://community.ibm.com/community/user/blogs/mike-burkhart/2024/12/20/ibm-storage-ceph-71-performance

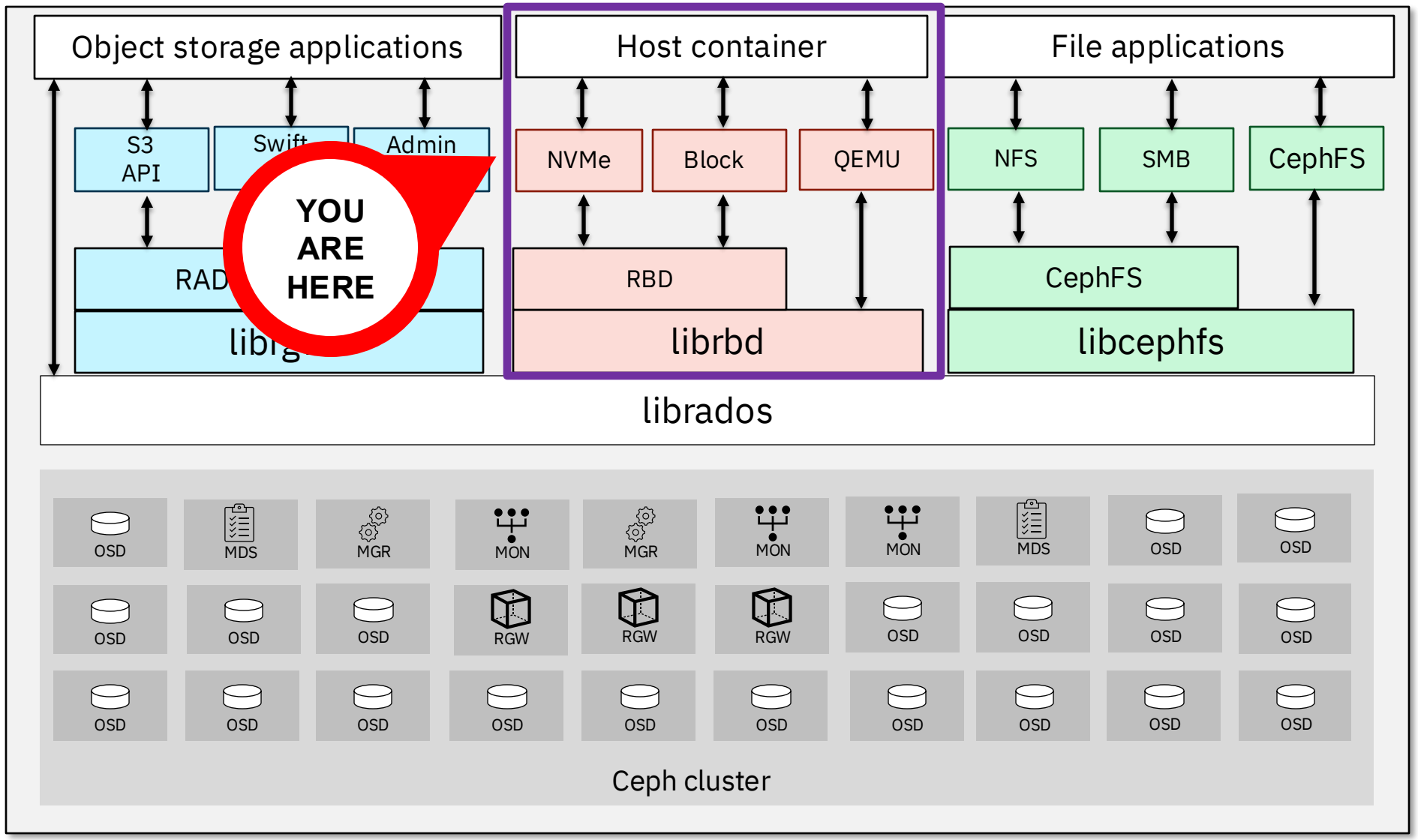








Building out an IBM Storage Ceph ecosystem for RBD and NVMe block services



What's next?

- S3 Bucket lifecycle, tiering and transitions
- IBM Deep Archive (S3 GLACIER)
- Ceph snapshots, clones, mirrors?
- Ceph management and monitoring
- Replication for RBD, CephFS, RGW

- And. . . .

IN CONCLUSION

Where to get help within minutes for IBMers

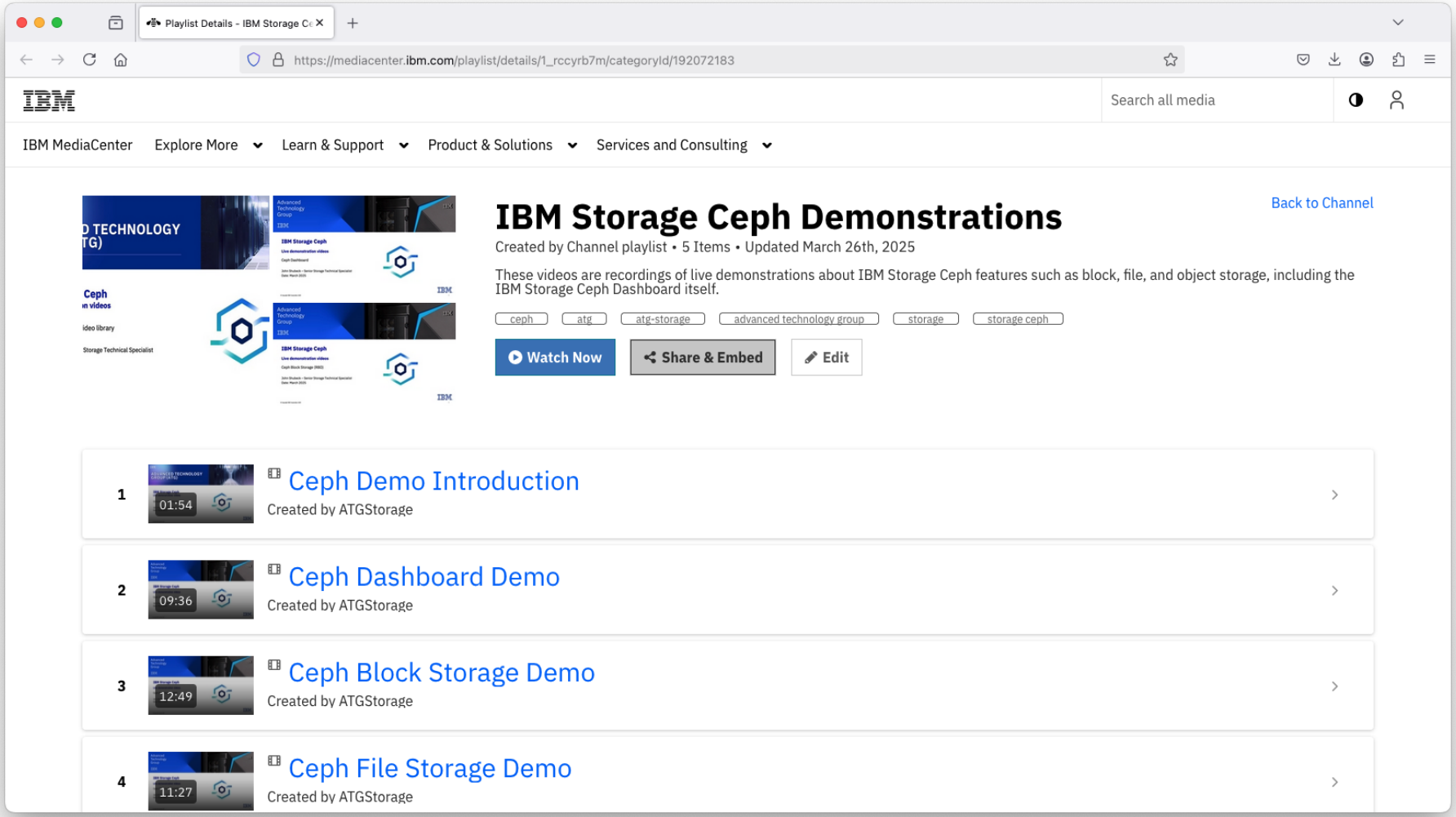
#ceph-help

Where to get Ceph Community help within minutes for Anybody!

#ceph

New IBM Storage Ceph demonstrations in IBM Mediacenter

DELIVERED



https://mediacenter.ibm.com/playlist/details/1_rccyrb7m/categoryId/192072183

Accelerate with ATG Survey

Please take a moment to share your feedback with our team!

You can access this 6-question survey via Menti.com with code 5151 0447 or

Direct link https://www.menti.com/alhsf3bgvxu6
Or

QR Code

Thank you!