



KubeCon



CloudNativeCon

Europe 2026

#KubeCon #CloudNativeCon

Rook: Storage for Kubernetes

Deepika Upadhyay, Clyso

Artem Torubarov, Clyso

Niels de Vos, IBM



Agenda



KubeCon



CloudNativeCon

Europe 2026

- Introduction to Rook and Ceph
- Cluster Topology & Maintenance
- Production User Study
- Ceph-CSI Driver
- Erasure Coding





KubeCon



CloudNativeCon

Europe 2026

Introduction to Rook



What is Rook?



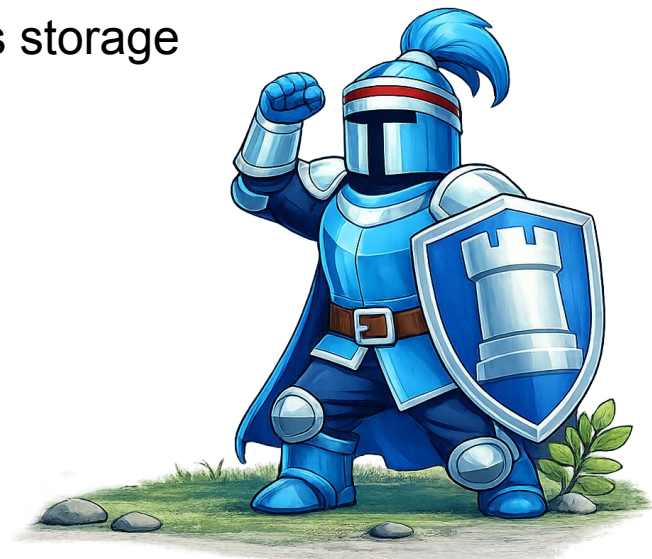
KubeCon



CloudNativeCon

Europe 2026

- Brings Ceph storage into your Kubernetes cluster
- Manages Ceph storage with an operator and CRDs
- Automated deployment, configuration, upgrades
- Allows apps to consume storage like any other K8s storage
 - Storage Classes, PVCs
- Open Source (Apache 2.0)
- Happy 10th Birthday!
 - Created November 2016



Architectural Layers



KubeCon

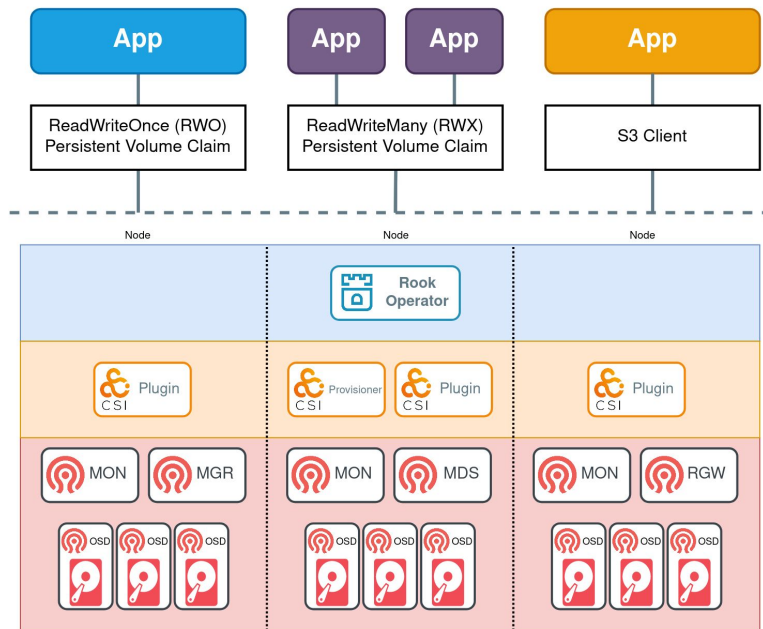


CloudNativeCon

Europe 2026

- Rook
 - Operator **deploys** and **manages** Ceph
- CSI
 - Ceph CSI driver dynamically **provisions** and **mounts** storage to user application pods
- Ceph
 - **Data Layer**

Rook Architecture



Ceph: Open-source, distributed Enterprise storage platform



KubeCon



CloudNativeCon

Europe 2026

✓ **Block** (RWO)
Ceph RBD

✓ **File** (RWX)
CephFS

✓ **Object** (S3)
Ceph RGW

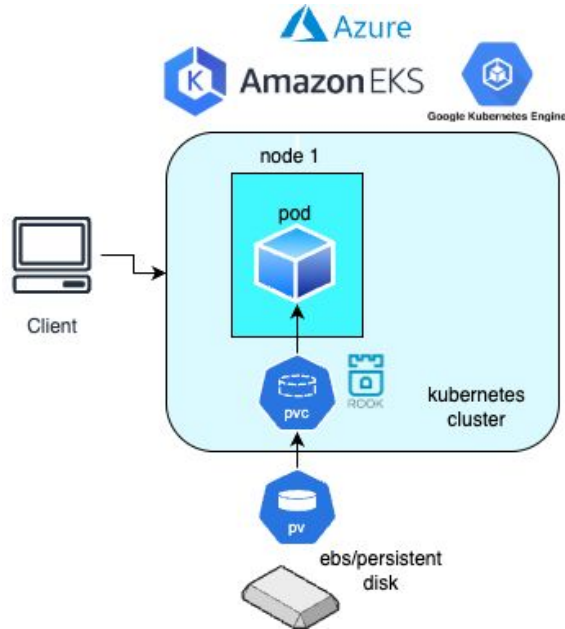
All-in-One Storage Solution for Kubernetes



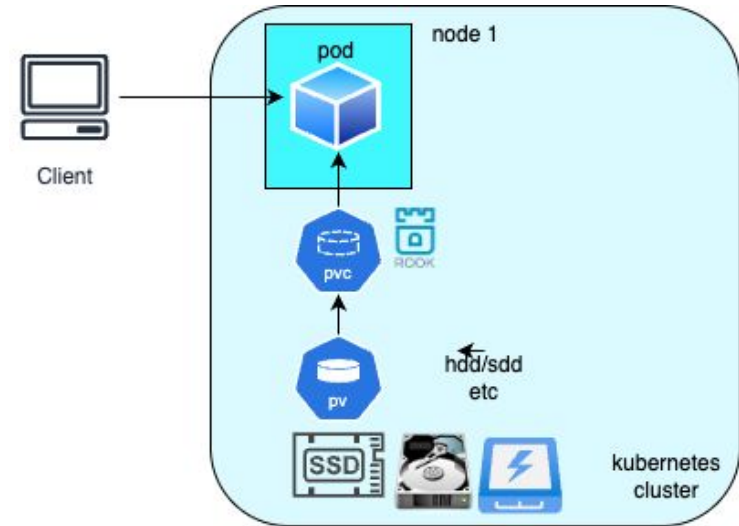
Installation Environments

Anywhere Kubernetes runs!

Installation Environments



Running in the Cloud → (EBS, Persistent Disks, etc.) for flexibility



On-Premises/Bare Metal → Use local SSDs/HDDs for performance & control.

Installation Environments

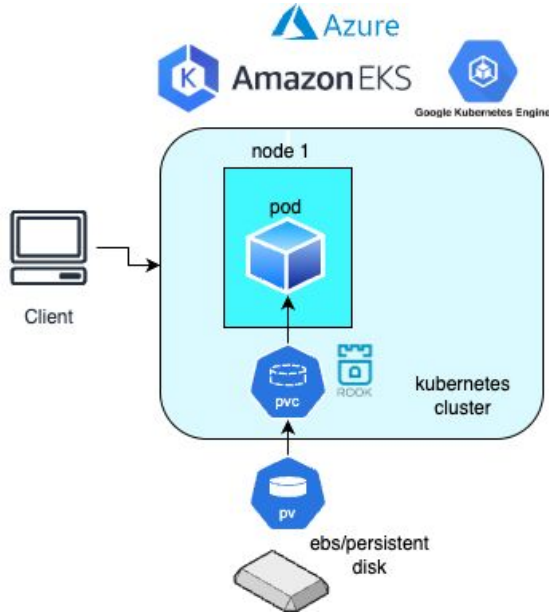


KubeCon

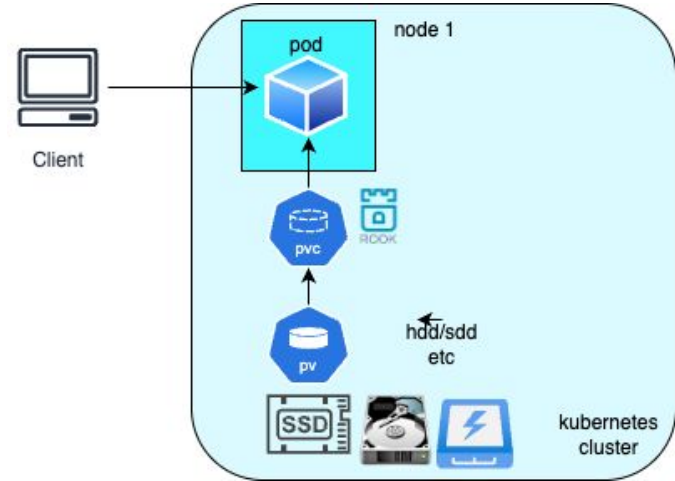


CloudNativeCon

Europe 2026



baremetal/vm



 **Hybrid/Multi-Cloud** → Mix & match for resilience.

Why choose Rook



KubeCon



CloudNativeCon

Europe 2026

Challenge

 **Fear of data loss:** No Storage Across AZs

 Limited PVs Per Node,
Poor Performance of Small PVs

 Lack of Cross-Cloud Support

How Rook Helps

 Replicates & distributes storage across AZs

 unlimited scaling with dynamic and optimized data placement

 Unified storage layer across multiple cloud providers



Add capacity without downtime

- Ceph is designed to scale up and out



Thin provisioning = Efficient storage use

- Adding storage capacity is independent from PVCs



KubeCon



CloudNativeCon

Europe 2026

Recent Features

- Enabled by default with Rook in v1.18+
- Provides configuration via CRDs for the Ceph-CSI drivers
 - RBD, CephFS, NFS, NVMe-oF
- Rook v1.20 (Coming ~May 2026)
 - All Ceph-CSI configuration will be done via CRs instead of with Rook operator settings

NVMe-oF Gateway (experimental)



KubeCon



CloudNativeCon

Europe 2026

- NVMe over Fabrics
 - Allows Ceph RBD(Block) volumes to be accessed via the NVMe/TCP protocol
- External clients outside the cluster can connect to Ceph block storage using **standard NVMe-oF initiators**
- Delivers low-latency, high-throughput block storage access over existing network infrastructure.
- Experimental in Rook v1.19 (with Ceph Tentacle v20)
 - We need your feedback!



KubeCon



CloudNativeCon

Europe 2026

Cluster Maintenance



Cluster Maintenance



KubeCon



CloudNativeCon

Europe 2026

- What is the impact of maintenance operations?
 - Can the cluster maintain availability during maintenance?
 - What severity of outages can the cluster survive?
- Topology is crucial to understand for reliability and durability



Topology: 3 replicas across failure domains



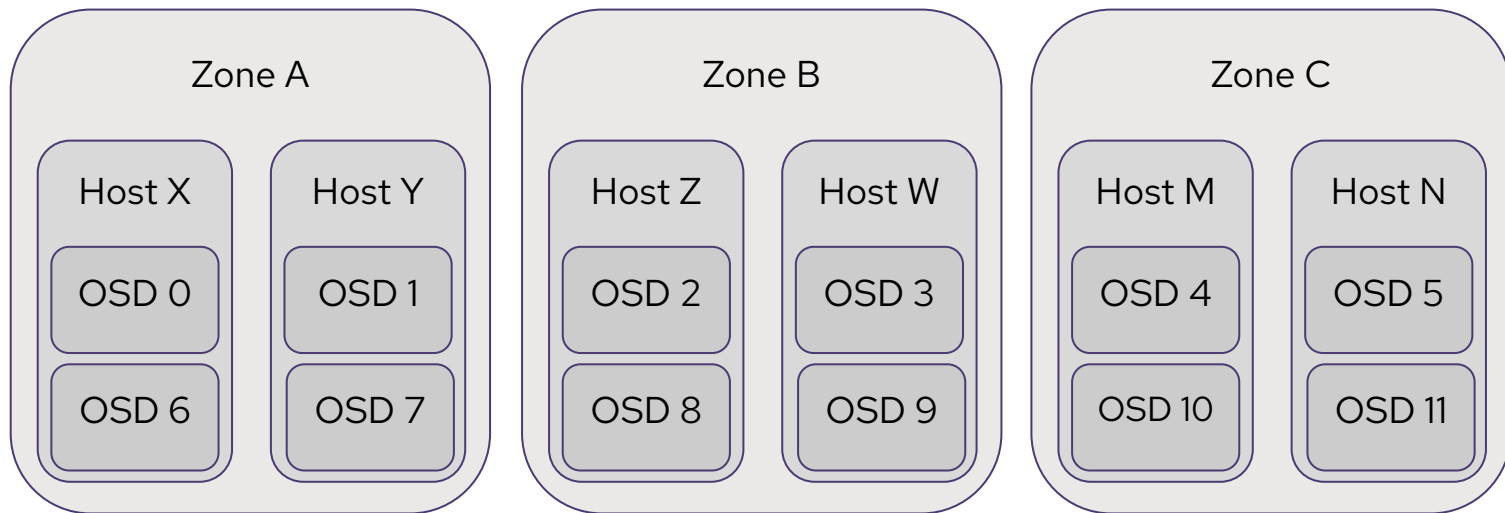
KubeCon



CloudNativeCon

Europe 2026

Zone Topology



Availability: One failure domain can be down



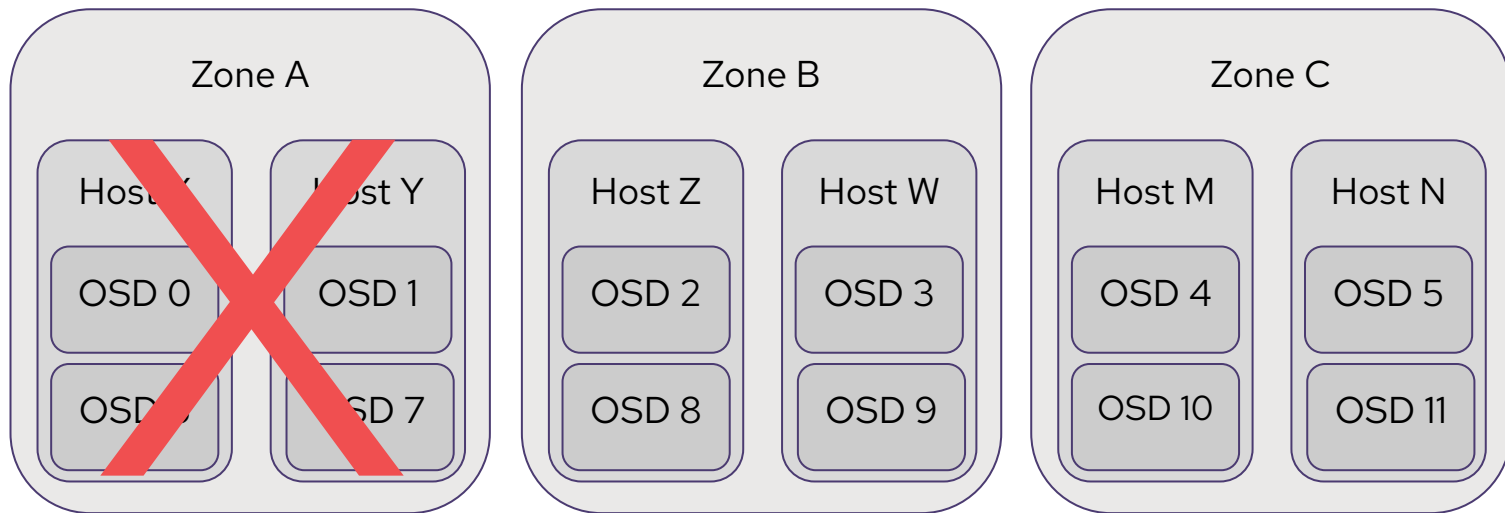
KubeCon



CloudNativeCon

Europe 2026

One zone goes down, cluster is fully online



Durability: 2 failure domains can be lost



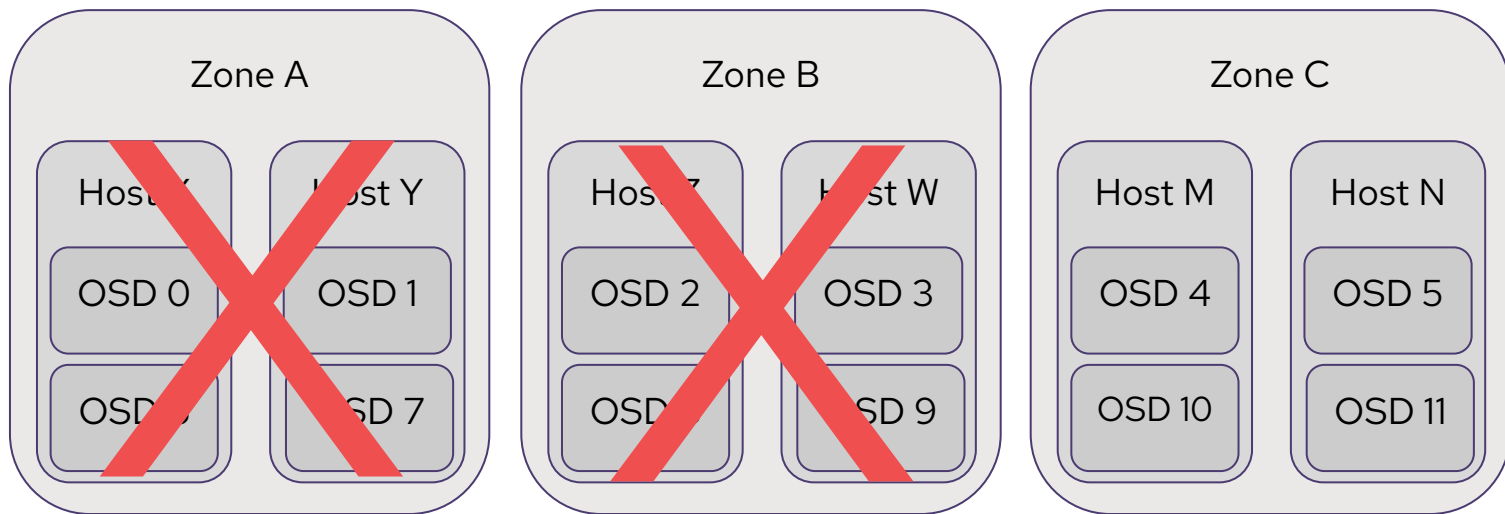
KubeCon



CloudNativeCon

Europe 2026

Two zones lost, data is still safe although offline



Node Maintenance



KubeCon



CloudNativeCon

Europe 2026

- Pod Disruption Budgets (PDBs) ensure storage remains available during node maintenance
 - Topology aware
- Rook upgrades ensure only one failure domain is affected to ensure no downtime

Kubectl Plugin



KubeCon



CloudNativeCon

Europe 2026

- CLI tool for troubleshooting
 - Cluster status
 - DR status
- Management operations that do not fit desired-state patterns
 - Ceph commands (toolbox replacement)
 - Restore mon quorum
 - Remove OSDs
 - Mon and OSD maintenance operations
 - Restore CRs accidentally marked for deletion
- What would you like to see automated?



KubeCon



CloudNativeCon

Europe 2026

Production User Deployment



Project Scope



KubeCon

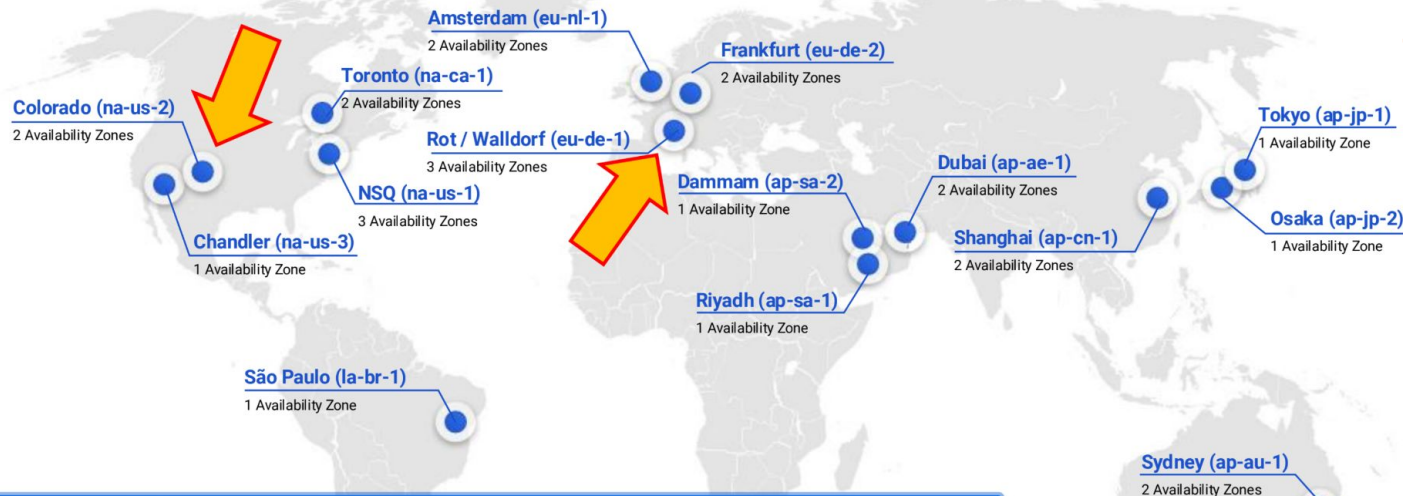


CloudNativeCon

Europe 2026

SAP Cloud Infrastructure: Overview

Global distribution of our physically isolated data centers



15 Regions
26 Availability Zones

>

170k VMs
500k Containers

20 PiB Memory
120 PiB Storage
3.2 Tb/s SAP Backbone

Storage:

- Openstack SWIFT
- Proprietary NAS

EU Sovereign Cloud



KubeCon



CloudNativeCon

Europe 2026

NeoNephos

Foundation **in formation**



Europe



Finanziert von der
Europäischen Union
NextGenerationEU

Gefördert durch:



Bundesministerium
für Wirtschaft
und Klimaschutz

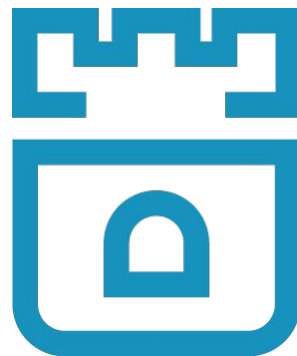
aufgrund eines Beschlusses
des Deutschen Bundestages



Storage:



ceph



The Fleet



KubeCon



CloudNativeCon

Europe 2026

Cluster Status

 First Region Live:	May 2024
 Live Regions:	10
 Storage Nodes:	251
 OSDs:	~2,800
 Raw Capacity:	~37 PiB
 Target:	30 regions, 120 PB

Node Hardware:

 Dell PowerEdge R7615
 AMD EPYC 64-core
 384 GB RAM
 16x 14TB NVMe
 100GbE

Performance Envelope



KubeCon



CloudNativeCon

Europe 2026

- **Workload:** 2M objects (4 KB each), 20 K6 clients, single bucket.
- **Method:** Ramping load over 30 minutes until p90 latency exceeded 500 ms.

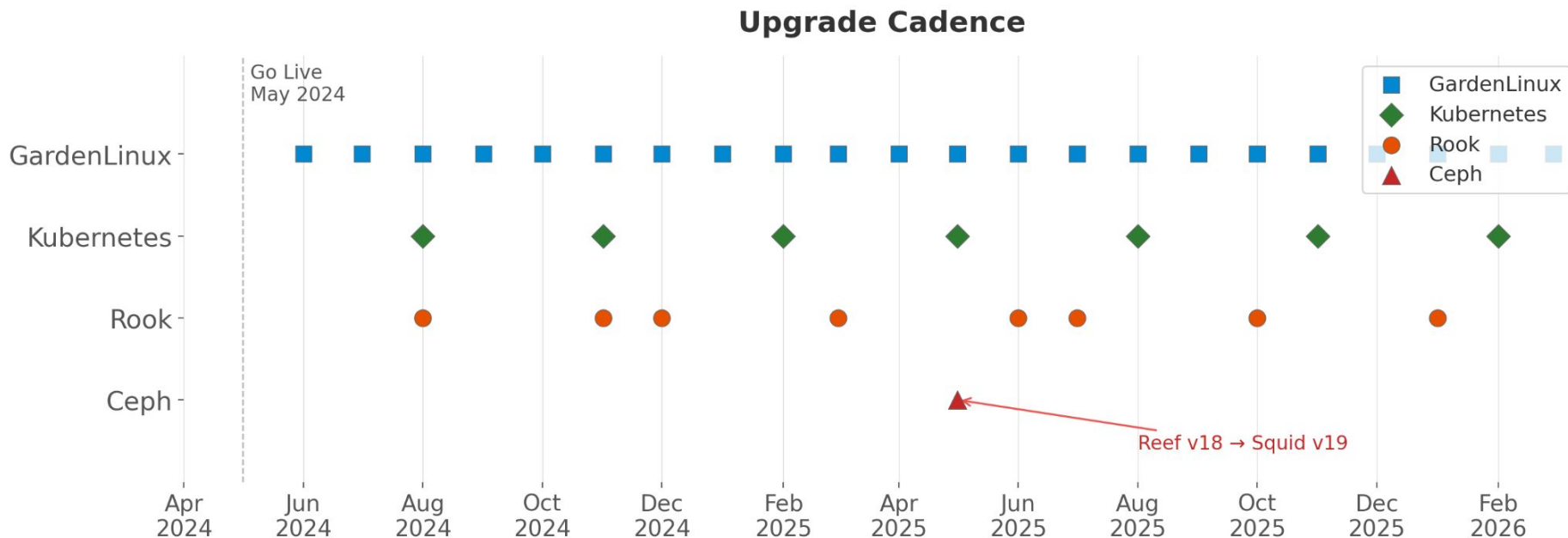
~90K GET/s	saturation begins
~170K GET/s	breaking point



Day 1: Infrastructure as a Code

- GitOps
- Base Blueprint as Helm Chart
- Region Overlay: CephBlockPools, RGW placement rules.

Upgrades: Zero Customer-Facing Downtime



Text version in Rook Blog



KubeCon



CloudNativeCon

Europe 2026





KubeCon



CloudNativeCon

Europe 2026

Ceph-CSI Driver

- **Ceph Container Storage Interface (Ceph-CSI)** project hosts **CSI Drivers** for Ceph
- Following are the different drivers that currently exist under **CephCSI** project:
 - **RBD Driver:** Rados Block Device (RBD)
 - **CephFS Driver:** Ceph Filesystem (CephFS)
 - **NFS Driver:** Network Filesystem (CephNFS)
 - **(new) NVMe-oF Driver:** RBD over NVMe-oF

Volume Mode recommendations



KubeCon



CloudNativeCon

Europe 2026

In-cluster workload usage:

- RBD
 - RWX BlockMode: Virtual Machines with KubeVirt, supports Live Migration
 - RWO Filesystem: ext4 formatted, use for databases and small file workloads
- CephFS
 - RWX Filesystem mode: simultaneous multi node access to files

Usage by external systems (or limited kernel driver availability):

- NFS
 - CephFS volumes exposed over NFS
- NVMe-oF
 - RBD volumes exposed over NVMe-oF

Single container-image that supports all different drivers

RADOS OMAP Usage for state and checkpointing

Ceph-CSI extensively makes use of **RADOS OMAP** to store state information of volumes and snapshots.

Go-Ceph: Go bindings for Ceph APIs

The **Go-Ceph** project is a collection of API bindings that support the use of native Ceph APIs, which are C language functions, in Go.

Standard Ceph-CSI Features



KubeCon



CloudNativeCon

Europe 2026

Kubernetes CSI features:

- PVC expansion
- Snapshotting and cloning of PVCs
- Topology aware
 - Read from the OSD nearest the client
- Create and restore from Snapshots, clone volumes
- Ephemeral volumes

Non Kubernetes-native:

- PVC encryption
 - LUKS (RBD), fscrypt (ext4, CephFS)
 - Various Key Management Systems (KMS) such as Vault, Azure, IBM HPCS, KMIP etc is supported.

Recently added Ceph-CSI Features



KubeCon



CloudNativeCon

Europe 2026

VolumeGroup Snapshots

Label PVCs to create consistency groups

RBD Block Usage Metrics

Size of the RBD-image, and number of blocks allocated

RBD Change Block Tracking support

Required for differential backups by comparing two snapshots

Auto Fencing support

Prevent data corruption, block misbehaving clients

CSI-Addons



KubeCon



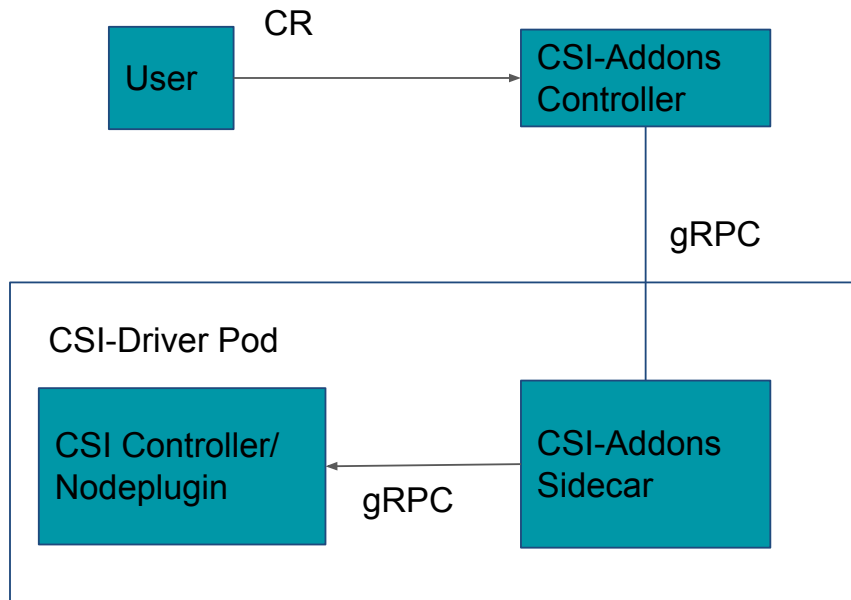
CloudNativeCon

Europe 2026

[CSI-Addons](#) hosts the extensions to the [CSI specification](#) that provides advanced storage operations.

Various components involved:

- CSI-Addons Controller
- CSI-Addons Sidecar
- CSI-Driver



CSI-Addons features in Ceph-CSI



KubeCon



CloudNativeCon

Europe 2026

Reclaim Space Operation:

- **Reclaim Space** operation executes ``rbd sparsify`` on images and ``fstrim`` on filesystem mode volumes

Network Fence/Class Operation

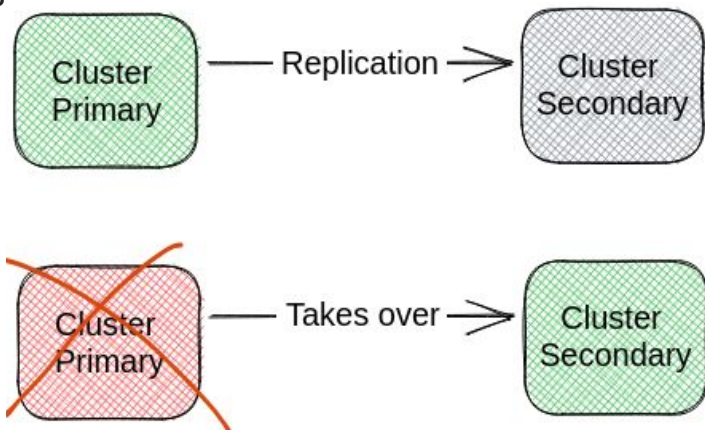
- **Network Fence Class** operation to advertise the client IP on the node required for Network Fencing
- **Network Fence** operation provides an API for blocking a list of given CIDR IP ranges
- This plays a critical role in **Metro Disaster Recovery** and **Node-Loss** scenarios

Encryption Key Rotation:

- Rotation of Key Encryption Keys(KEKs) for encrypted volumes.

Volume/VolumeGroup Replication Operation

- **Volume/VolumeGroup Replication** operation provides common and reusable APIs for storage disaster recovery
- It allows enabling/disabling mirroring and changing state(primary/secondary) of rbd mirrored group/images
- The volume/volume-group replication operation automates rbd-mirroring, allowing **promote**, **demote**, **resync** and get volume replication information operations on rbd group/images
- This plays a critical role in **Regional Disaster Recovery**.





KubeCon



CloudNativeCon

Europe 2026

Erasure Coding with Rook



Why Erasure Coding?



KubeCon



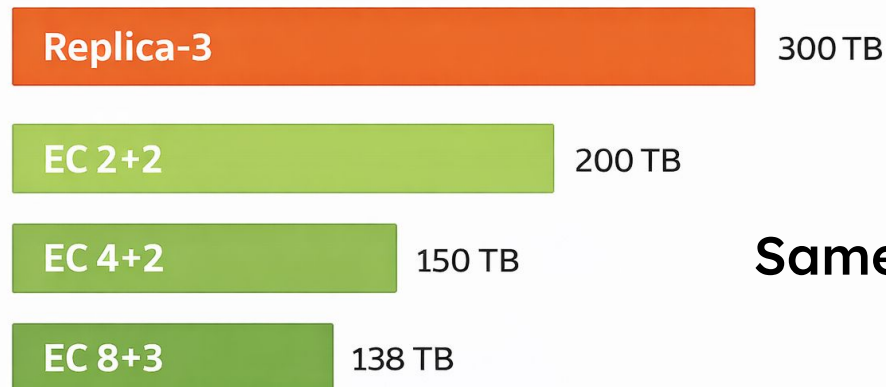
CloudNativeCon

Europe 2026

Why store 3 copies of your data when math can do the job?

- Most Ceph clusters use replica-3 for durability → 3× storage required
- Erasure Coding provides similar durability with significantly lower storage overhead

Raw Capacity Needed for 100TB Data



Same durability, ~50% less storage
(EC 4+2 vs Replica-3)

How Ceph Erasure Coding works



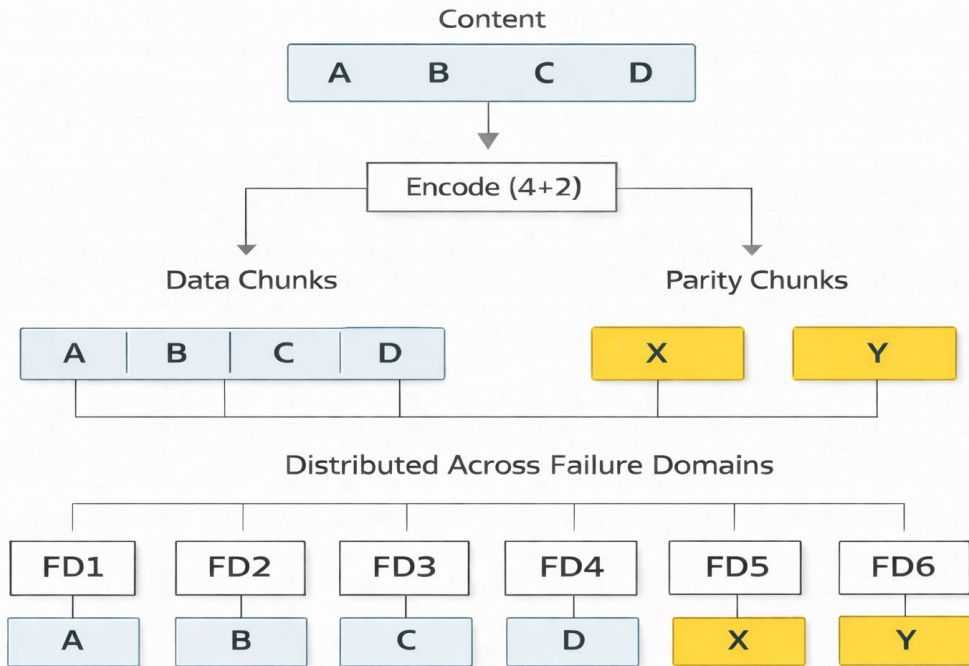
KubeCon



CloudNativeCon

Europe 2026

Data is split into chunks and extra parity chunks are generated for recovery.



EC Profile: 4+2

$k = 4$ data chunks

$m = 2$ coding chunks

Minimum failure domains
required = $k + m = 6$

Failure tolerance = $m = 2$

Setting up Erasure Coding with Rook



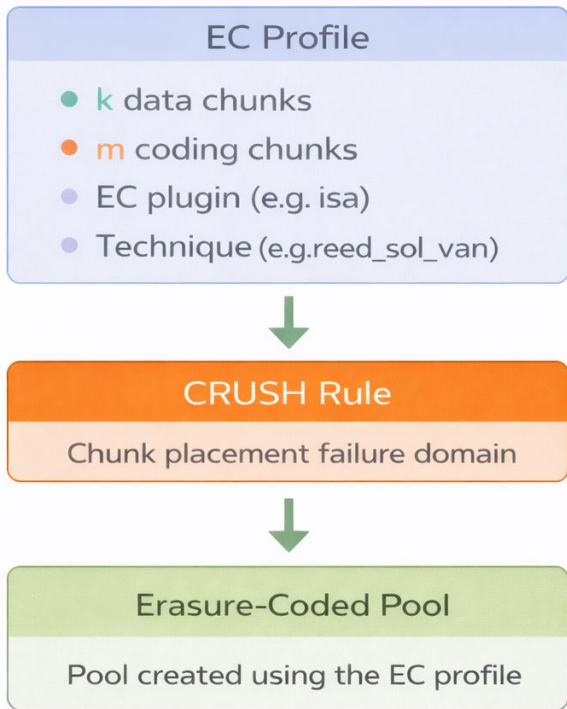
KubeCon



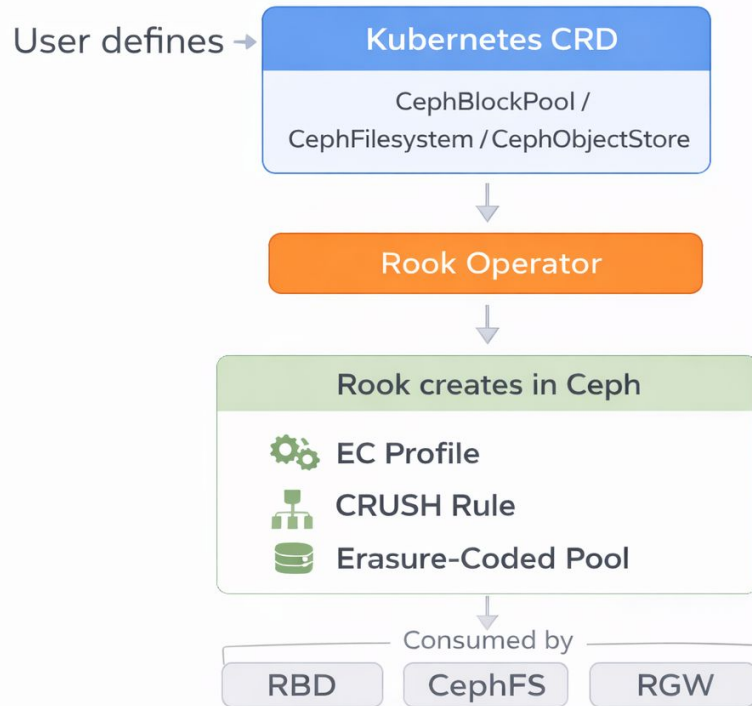
CloudNativeCon

Europe 2026

Ceph Erasure Coding Requires



Rook Simplifies this by



Things to Consider Before Using EC



KubeCon



CloudNativeCon

Europe 2026

- **Performance trade-off**
 - Higher **CPU usage** due to chunking & parity computation
 - Slower writes than replication
 - Not ideal for small or latency-sensitive workloads
- **Choose K+M carefully**
 - Determines capacity efficiency and failure tolerance
 - Requires **K + M failure domains**, Cannot be changed later
 - Use **M $\geq$ 2** for production
- **Workload suitability**
 - **Ideal:** archives, backups, media, ML datasets, cold storage
 - **Confirm benchmarks:** databases, VM disks, transactional systems

“Fast EC”: Ceph Tentacle v20



KubeCon



CloudNativeCon

Europe 2026

Ceph v20 introduced major performance improvements for EC pools

- **Partial Reads**
 - Reads only the required shards instead of reconstructing the full stripe, Significantly improving small-read latency.
- **Partial Writes**
 - Update only the affected chunks instead of rewriting the entire stripe saving on drive operations
- **Parity Delta Writes**
 - Parity shards updated using delta calculations rather than full recomputation, Reducing CPU usage and network overhead
- **Larger Striping**
 - default stripe_unit 16KB improves block and file perf

FastEC vs (old) EC vs Replica-3

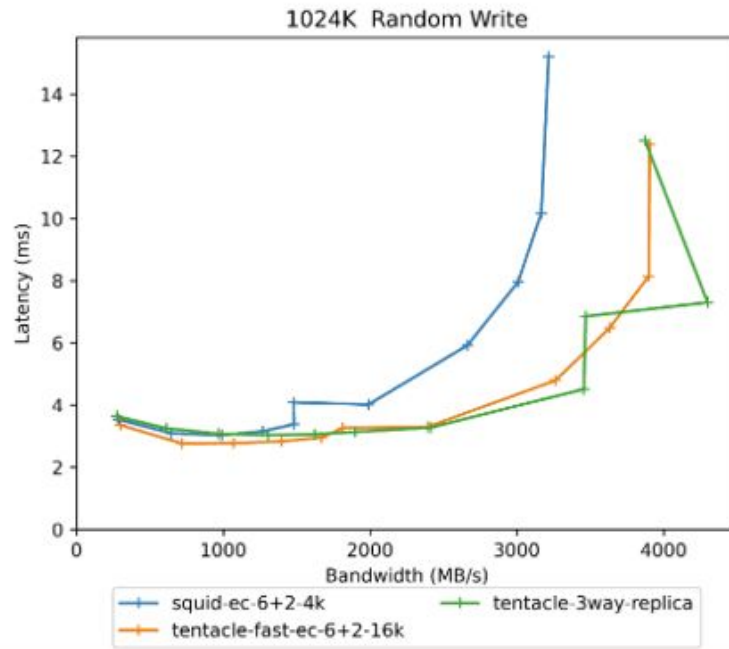
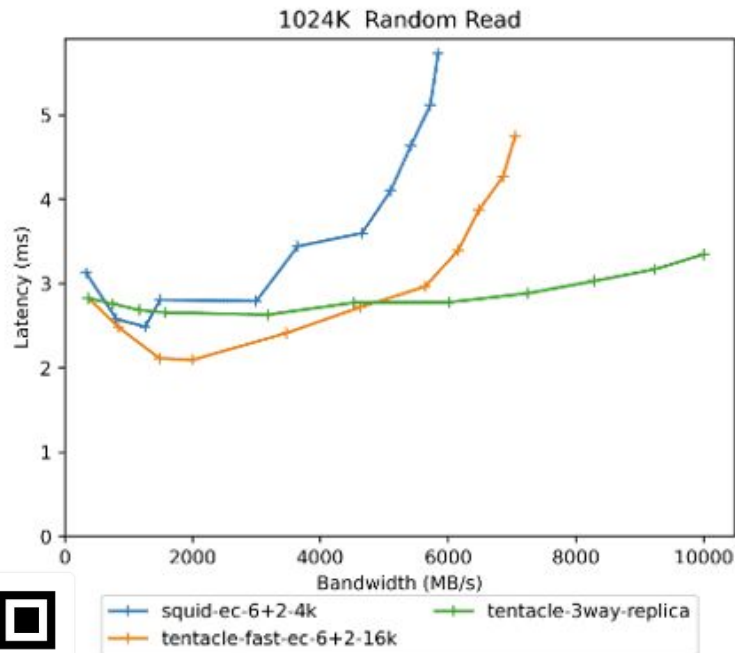


KubeCon



CloudNativeCon

Europe 2026



With Large Read & Write the EC perf is almost same as replica-3



Configuring Erasure Coding



KubeCon



CloudNativeCon

Europe 2026

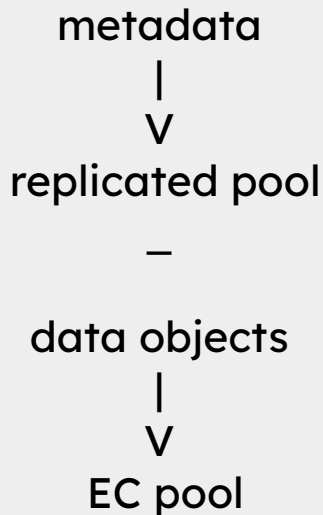
Erasure-coded pools do not support OMAP. Metadata stored by RBD, CephFS and Object requires a separate replicated metadata pool.

Metadata Pool

```
apiVersion: ceph.rook.io/v1
kind: CephBlockPool
metadata:
  name: replicated-metadata-pool
  namespace: rook-ceph
spec:
  failureDomain: host
  replicated:
    size: 3
```

EC Data Pool

```
apiVersion: ceph.rook.io/v1
kind: CephBlockPool
metadata:
  name: ec-data-pool
  namespace: rook-ceph
spec:
  failureDomain: host
  erasureCoded:
    dataChunks: 4
    codingChunks: 2
  parameters:
    allow_ec_optimizations: "true"
```



Configuring EC: RBD StorageClass



KubeCon



CloudNativeCon

Europe 2026

StorageClass for Block EC Usage

```
apiVersion: storage.k8s.io/v1
kind: StorageClass
metadata:
  name: ec-block-sc
parameters:
  clusterID: rook-ceph
  ...
  pool: replicated-metadata-pool
  dataPool: ec-data-pool
provisioner: rook-ceph.rbd.csi.ceph.com
allowVolumeExpansion: true
```

- The **pool** parameter specifies the replicated metadata pool
- The **dataPool** parameter specifies the erasure-coded data pool

EC Features Coming Soon



KubeCon



CloudNativeCon

Europe 2026

Later this summer with Ceph v21 (Umbrella)

- **Direct Reads:**

Client talks directly to each OSD to retrieve the data, so read performance comes to the same level as replicated

- **OMap support:**

No more need for a separate replicated pool for metadata

- **Pool migration:**

Convert a pool from replication to EC, or replication to EC, or change the parameters of EC such as K & M values. Internally a new pool is created and data is migrated



KubeCon



CloudNativeCon

Europe 2026

Visit Us!

Booth: Project Rook | Kiosk P-14B | Hall 5, Gouda Zone

Tue: 10:15–14:40

Wed: 10:00–13:30

Thu: 10:00–12:00



KubeCon



CloudNativeCon

Europe 2026

Questions?

Website and Docs	https://rook.io
Slack	https://slack.rook.io
X	@rook_io
Project Pavilion	Come to the Rook booth!



KubeCon



CloudNativeCon

Europe 2026

Appendix



Rook Pods

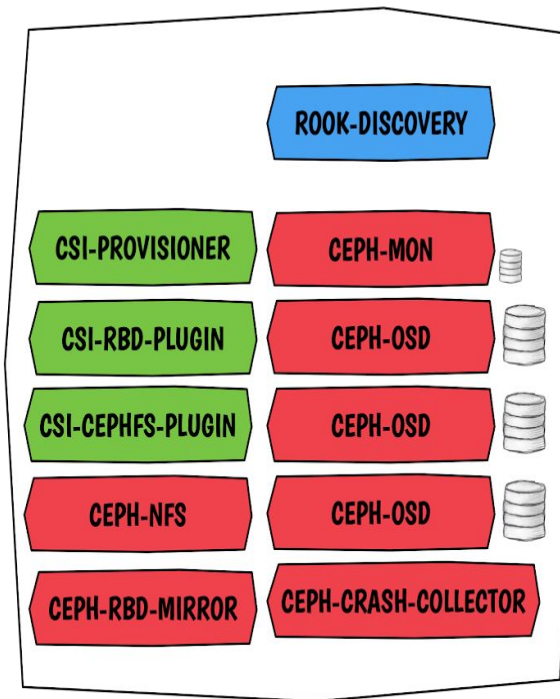
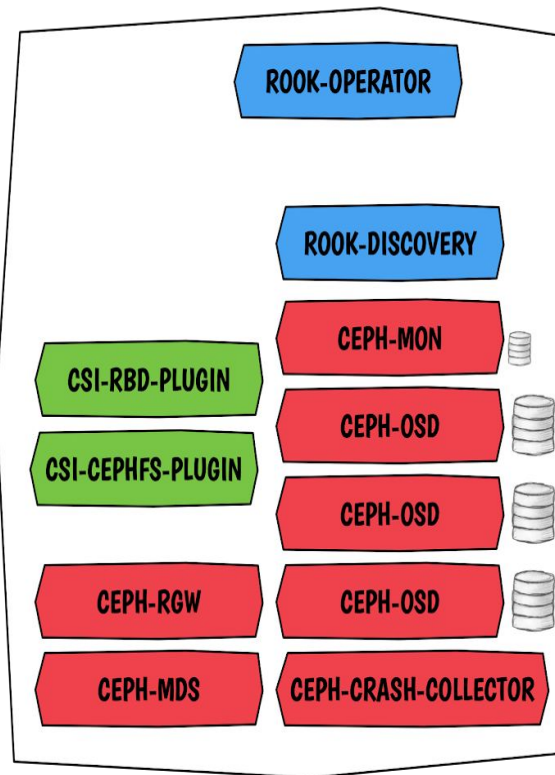
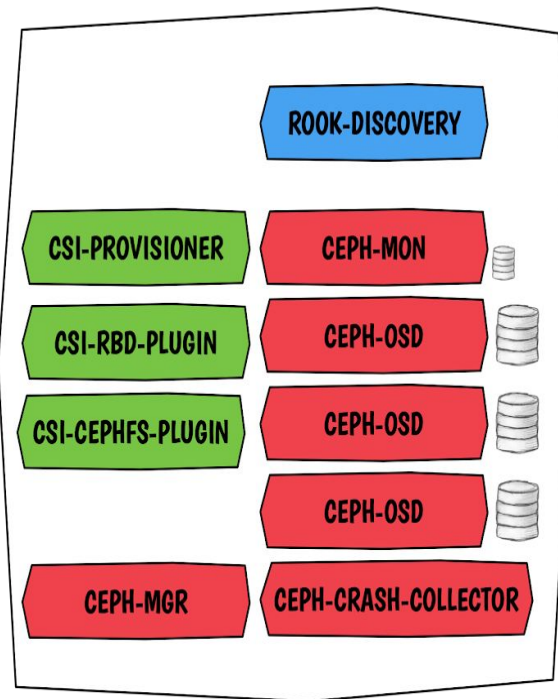


KubeCon



CloudNativeCon

Europe 2026



CSI Provisioning

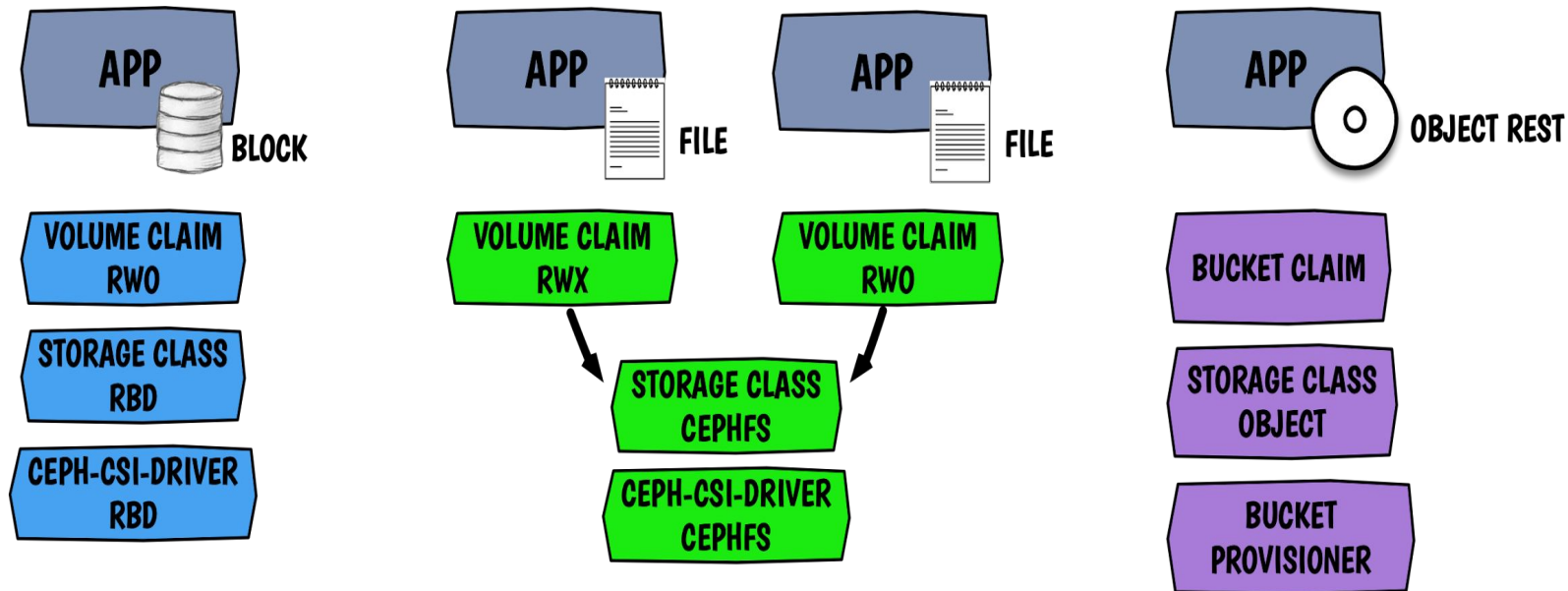


KubeCon



CloudNativeCon

Europe 2026



Ceph Data Path



KubeCon



CloudNativeCon

Europe 2026

