
KubeCon
CloudNativeCon
Europe 2023
TiKV

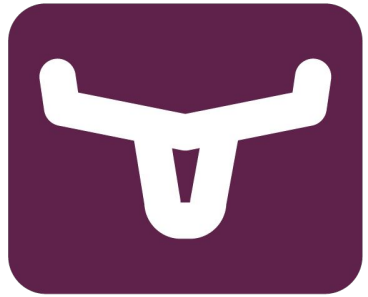
Longhorn - Intro, Deep Dive, Q & A

David Ko

Senior Engineering Manager at SUSE

Agenda

- What is Longhorn
- Feature List
- Momentum/Community, Story/Roadmap
- Releases
- How Longhorn Works
  - Control Plane
  - Data Plane
  - Snapshot, Backup, Replica Rebuilding
  - Disaster Recovery
  - Volume Live Migration
- What is Next?



LONGHORN

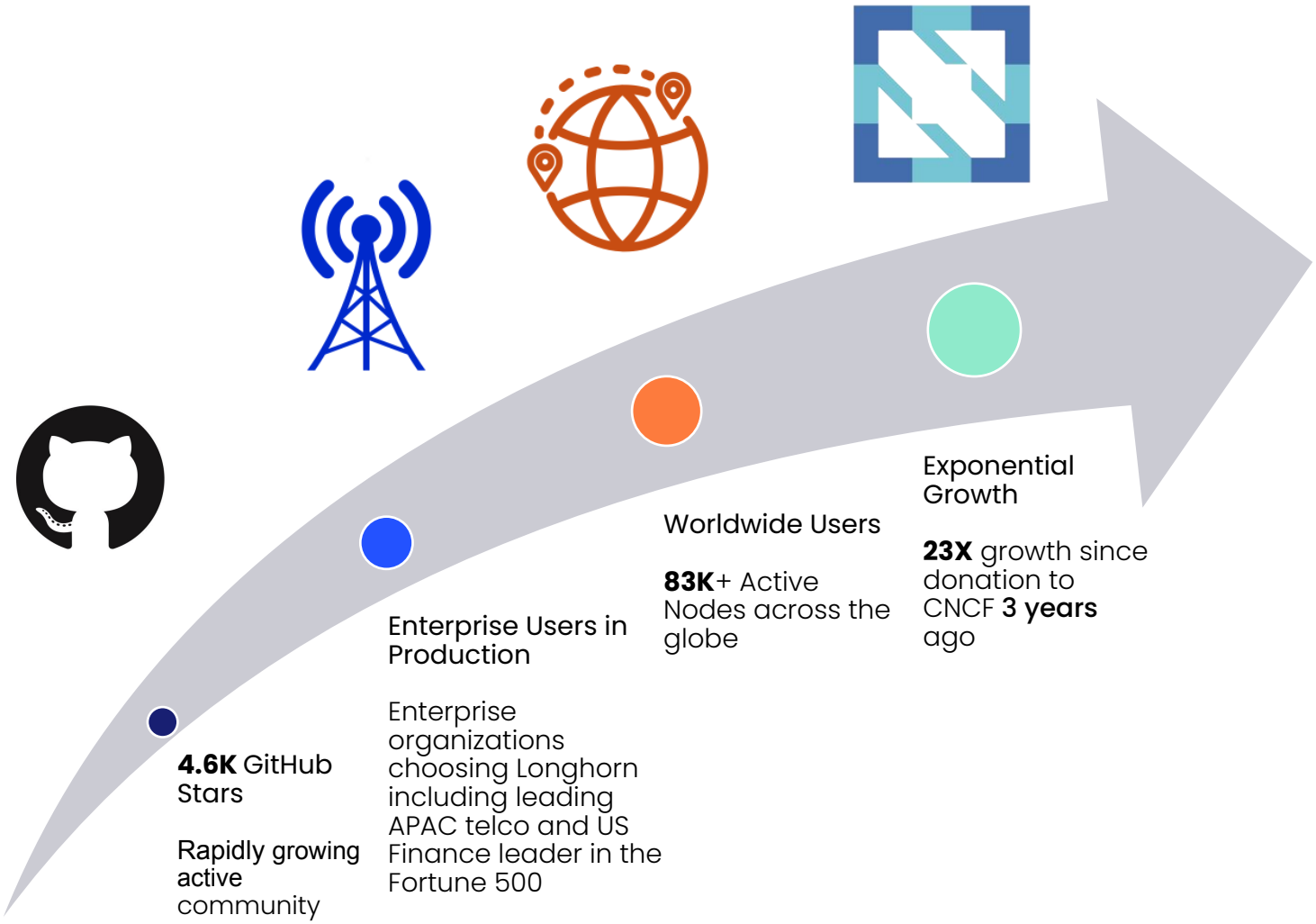
What is Longhorn

- Highly available, software-defined persistent block storage for Kubernetes
- Lightweight, reliable, and easy-to-use
- Adds persistent volume support to any certified K8s cluster.
- Storage Agnostic – any ext4/xfs filesystem can be added to a Longhorn cluster
- NFS and S3 compatible (backup store)
- Kubernetes-first design implemented in CRDs and controller pattern
- Open source and owned by the CNCF

Feature List

- Enterprise-grade distributed block storage software for Kubernetes
- Volume thin-provisioning
- Volume snapshot & revert
- Volume backup & restore
- Volume clone & expansion
- Volume encryption in-transit and at-rest
- Auto replica rebalancing & Cross-zone replication
- Storage Tag for node and disk selection
- Cross-cluster disaster recovery volume with defined RTO and RPO
- Non-disruptive live volume upgrade
- Recurring jobs (snapshot/backup)
- Block/FS volume types
- RWO/RWX access modes
- AMD64/ARM64/s390x arches
- Intuitive UI
- …

Longhorn Momentum



4.6K GitHub Stars

Rapidly growing active community

Enterprise Users in Production

Enterprise organizations choosing Longhorn including leading APAC telco and US Finance leader in the Fortune 500

Worldwide Users

83K+ Active Nodes across the globe

Exponential Growth

23X growth since donation to CNCF **3 years** ago

Longhorn Story, road to 1.3, 1.4 & 1.5*

KubeCon CloudNativeCon Europe 2023

'14

'16

Open sourced

Re-written in Go,
built for Docker

'18

Longhorn 0.3

Support S3 for backup
Basic K8s support

'19

Longhorn 0.6

DR Volume
Helm chart

CNCF Sandbox

'20

Longhorn 1.0

Volume resizing
Automatic recover faulted volume
Air gap
First GA Release

'21

Longhorn 1.1

RWX support
ARM64 Support
Prometheus support

'21

Longhorn 1.2

Encryption at Rest and In Transit
Policy Based Recurring Job

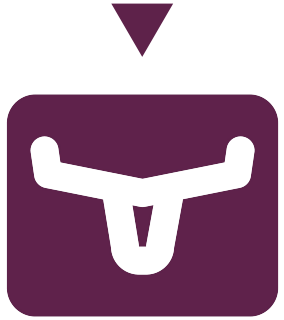
CNCF Incubating

Kubernetes Native API support (v1beta2) with Admission and Conversion Webhooks

Storage Network

Public Cloud Managed Kubernetes Cluster Support

Space Efficient Snapshot Purge



Longhorn 1.3
June 2022

CSI Snapshot for Longhorn Snapshot

Backing Image Download

Secured Communication among Control/Data Plane components

Upgrade Path Enhancement via Version Filtering

Longhorn Story, road to 1.4

KubeCon CloudNativeCon Europe 2023

'14

'16

Open sourced

Re-written in Go,
built for Docker

'18

Longhorn 0.3

Support S3 for backup
Basic K8s support

'19

Longhorn 0.6

DR Volume
Helm chart

CNCF Sandbox

'20

Longhorn 1.0

Volume resizing
Automatic recover faulted volume
Air gap
First GA Release

'21

Longhorn 1.1

RWX support
ARM64 Support
Prometheus support

'21

Longhorn 1.2

Encryption at Rest and In Transit
Policy Based Recurring Job

CNCF Incubating

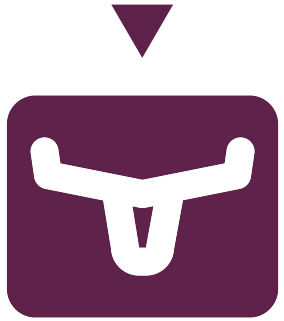
Reclaim/Shrink/Trim Volume Support

RWX Volume GA

Longhorn System Backup & Restore

Snapshot Checksum - Data Consistency

Bit-Rot Protection



Longhorn 1.4
December 2022

K8s 1.25 Support

ARM64 GA

Support Bundle Enhancement

Online Volume Expansion

Local Volume - Data Locality Strict-Local

Longhorn Story, road to 1.5*

KubeCon CloudNativeCon Europe 2023

'14

'16

Open sourced

Re-written in Go,
built for Docker

'18

Longhorn 0.3

Support S3 for backup
Basic K8s support

'19

Longhorn 0.6

DR Volume
Helm chart

CNCF Sandbox

'20

Longhorn 1.0

Volume resizing
Automatic recover faulted volume
Air gap
First GA Release

'21

Longhorn 1.1

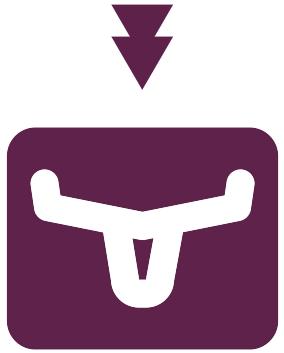
RWX support
ARM64 Support
Prometheus support

'21

Longhorn 1.2

Encryption at Rest and In Transit
Policy Based Recurring Job

CNCF Incubating



Longhorn 1.5*
June 2023

Next SPDK Data Engine - Experimental

Cluster Autoscaler GA

Instance Manager Engine & Replica Consolidation

Automatic FS Trim & RWX volume Trim

Upgrade Path Enforcement & Downgrade Prevention

Backing Image Management via VolumeSnapshot

CIFS & Azure BlobStore Backup Store

Kubernetes Drain Awareness & Policy

New Longhorn VolumeAttachment

Recurring Jobs - Snapshot Delete & Cleanup

Longhorn Releases

Active Maintained Branches

- 1.3 and 1.4

Upcoming Releases

- 1.3.3 – *April, 2023 - released today!*
- 1.4.2 – *May, 2023*
- 1.5.0 – *June, 2023*

How Longhorn Works

Control Plane

- Kubernetes Controller + CR

Data Plane

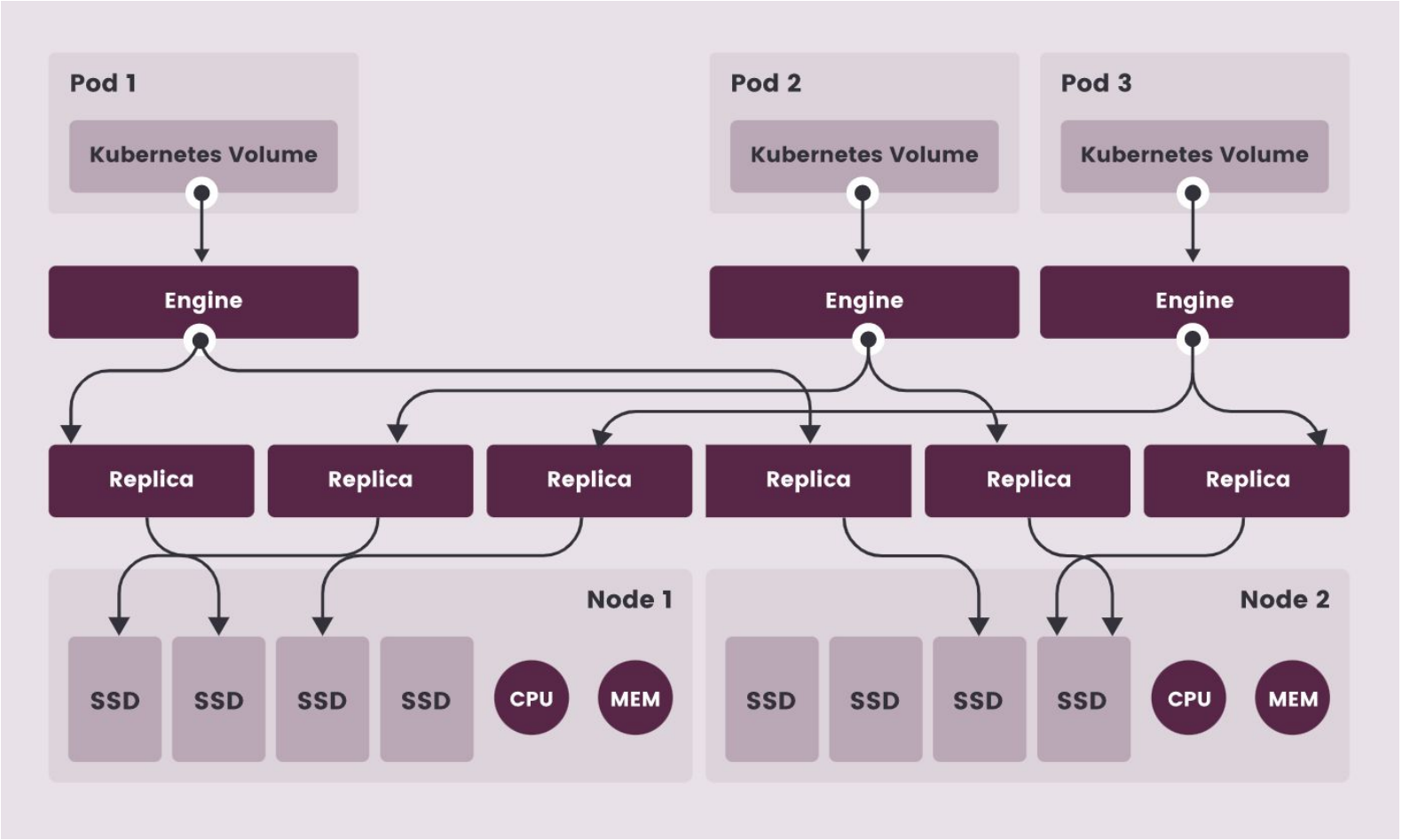
- Volume Frontend (iSCSI)
- Volume (Engine)
- Volume replica (Replica)

Volume Lifecycle

- CSI
- PVC/PV

Data Placement

- Longhorn disk (FS on host)



Longhorn Engine & Replica

- Volume Frontend (iSCSI)
  - open-iscsi (iSCSI initiator)
- Volume (Engine)
  - longhorn tgt (iSCSI target)
  - longhorn engine (TCP data server and volume controller)
- Volume replica (Replica)
  - longhorn replica (local/remote TCP data server and replica controller)
  - Data operation (snapshot, rebuild, coalesce/merge, prune, purge, backup, etc)

Longhorn Engine & Replica - SPDK (1.5*)

SPDK - Software Performance Development Kit

- Used in high performance cloud applications
- Has a generic “block device” application layer with many different implementations, easy to implement new block devices
- Has support for exposing block devices for remote block devices: iSCSI and NVMe over Fabrics
- Has a logical volume feature which stores data in a series of sparse snapshots
- Designed for asynchronous programming
- Uses memory pools to minimize memory allocation

Longhorn Engine & Replica - SPDK (1.5*)

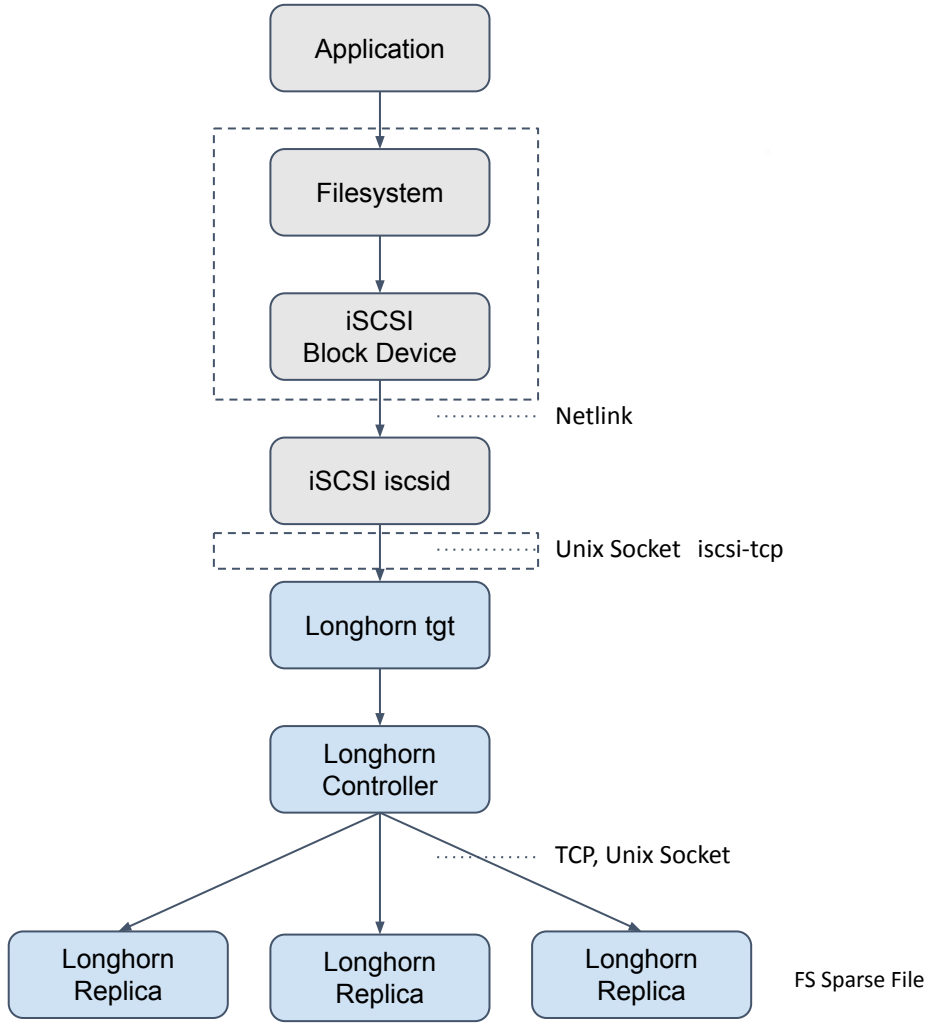
- Volume Frontend (NVMe-oF)
  - nvme-cli (NVMe-oF initiator)
- Volume (Engine)
  - SPDK tgt (NVMe-oF target)
  - longhorn engine (longhorn SPDK bdev)
- Volume replica (Replica)
  - longhorn replica (local/remote SPDK logical volume exposed by NMVe-oF target)
  - Data operation (snapshot, rebuild, coalesce/merge, prune, purge, backup, etc)

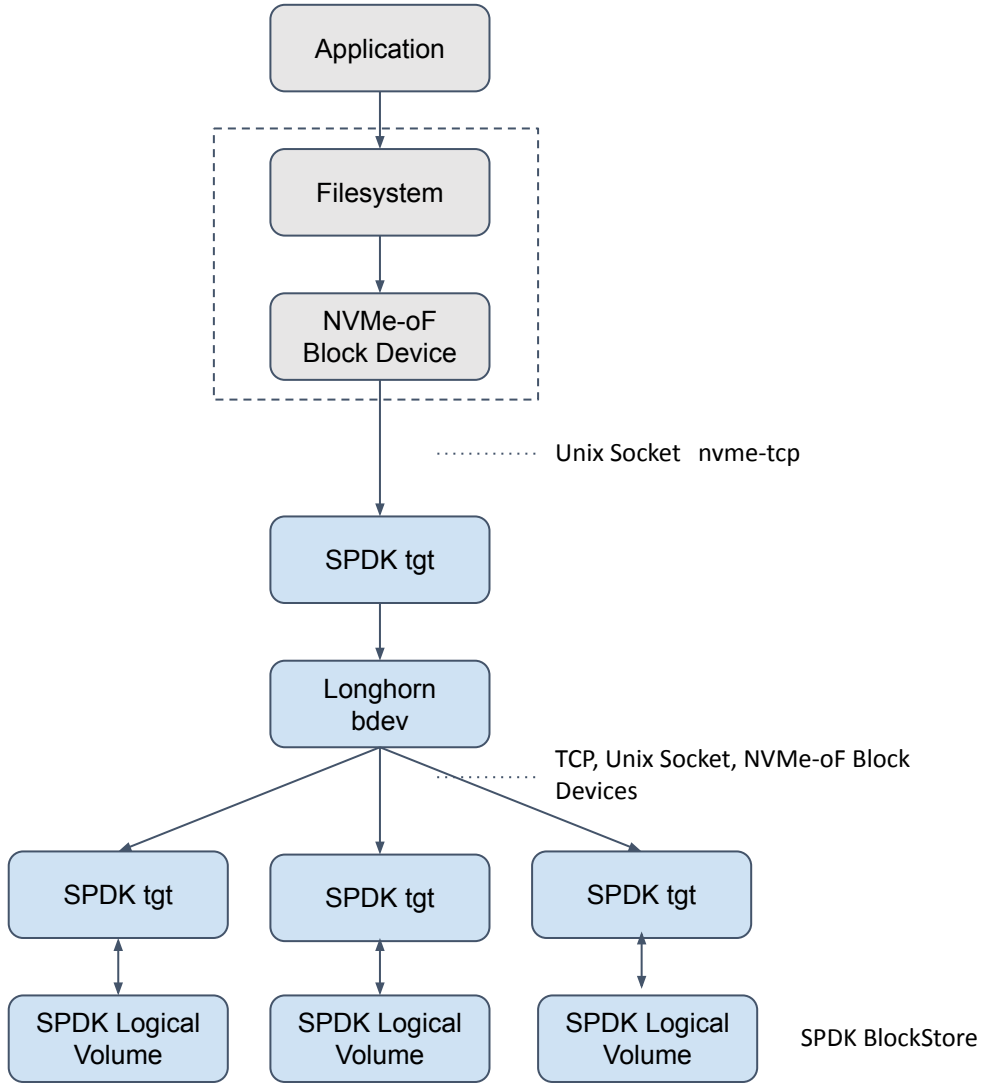
Multiple Data Engines

SPDK:

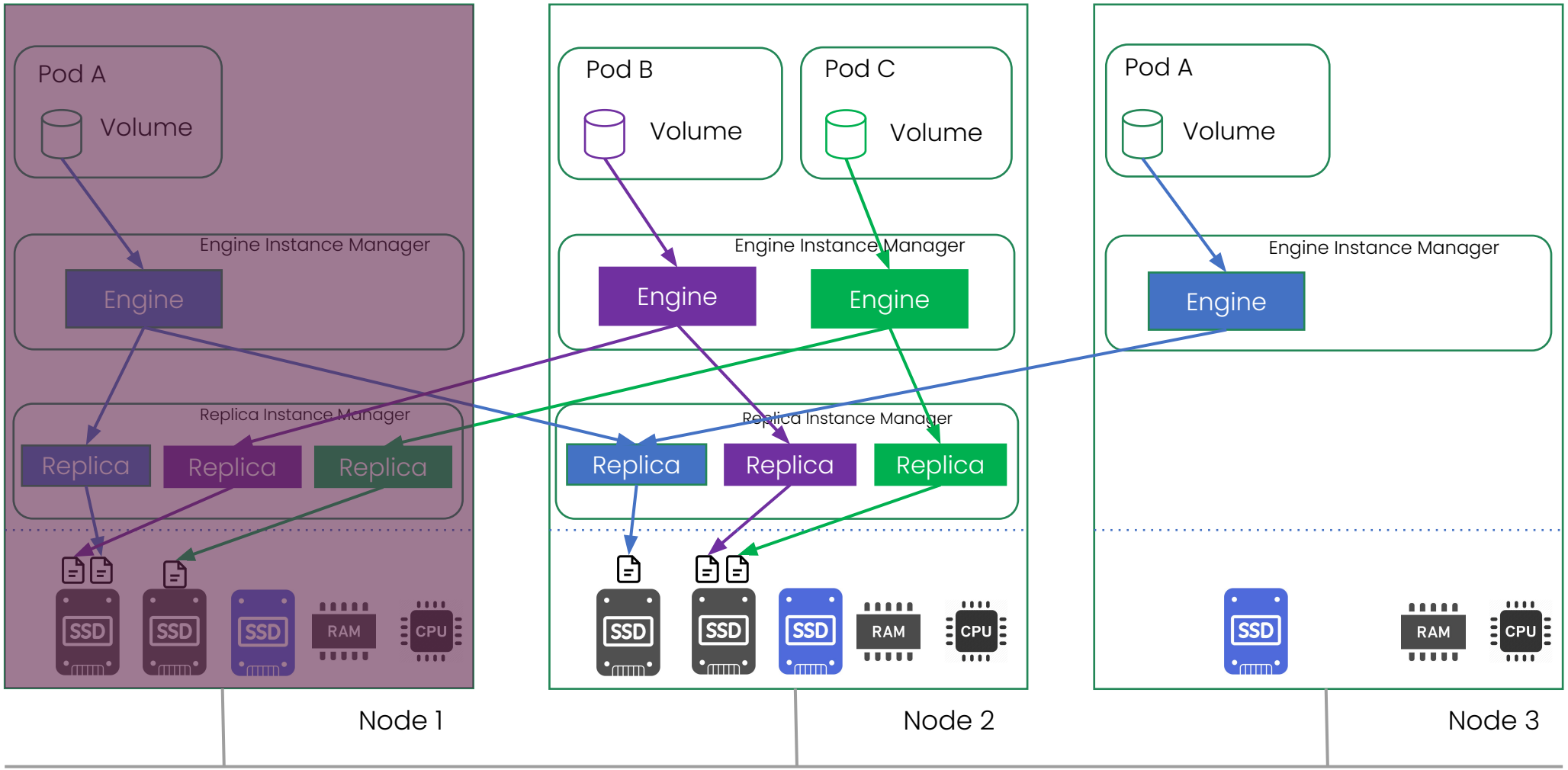
- **1.5 Experimental**
- **1.6 Feature Parity**

Longhorn Current & Future Engine Data Paths

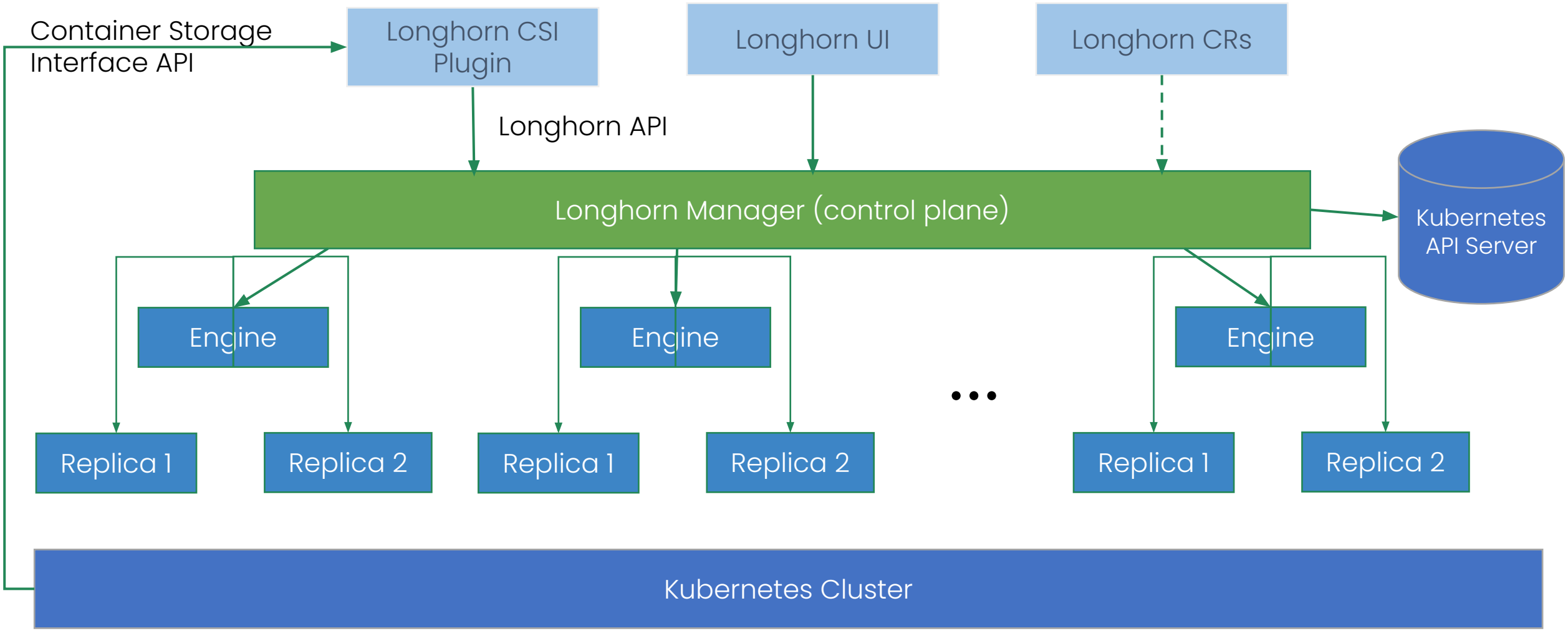




Longhorn Engine & Replica - Resilience & Failover

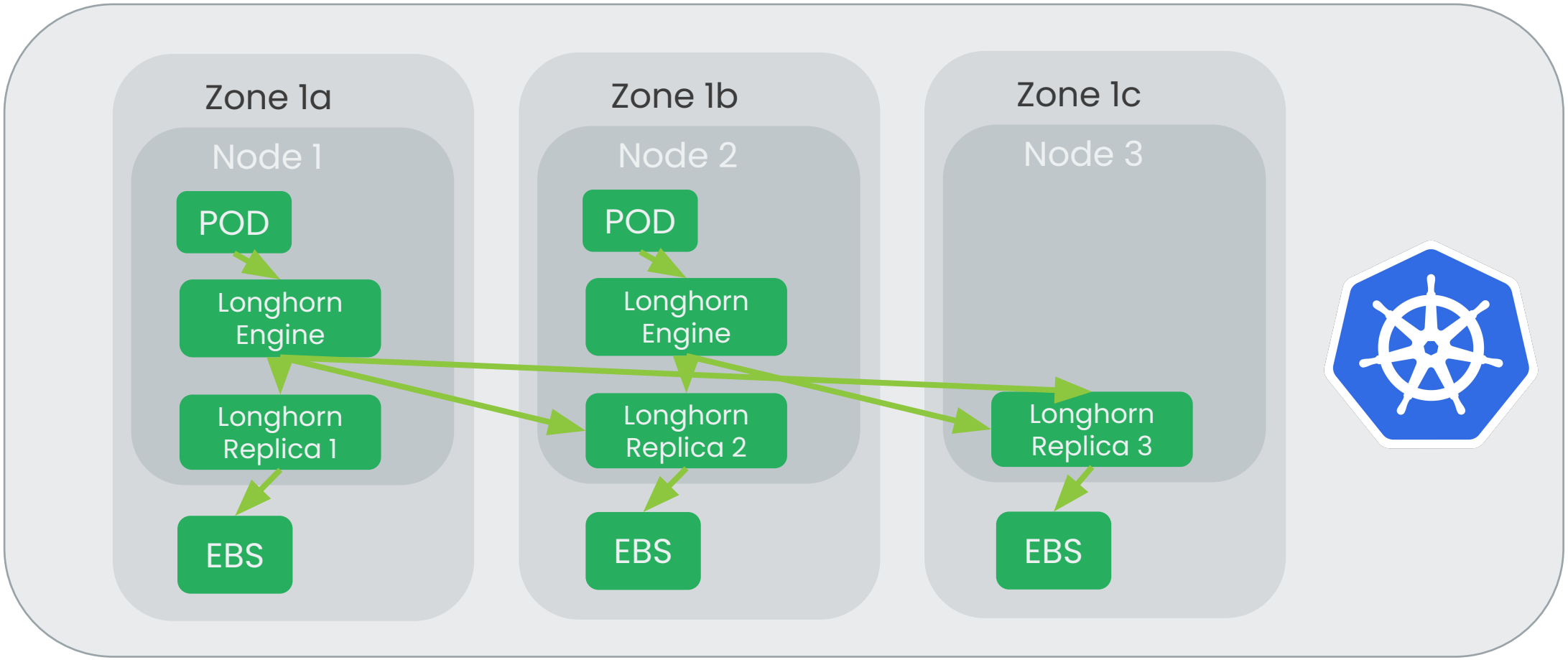


Longhorn Manager - Control Plane



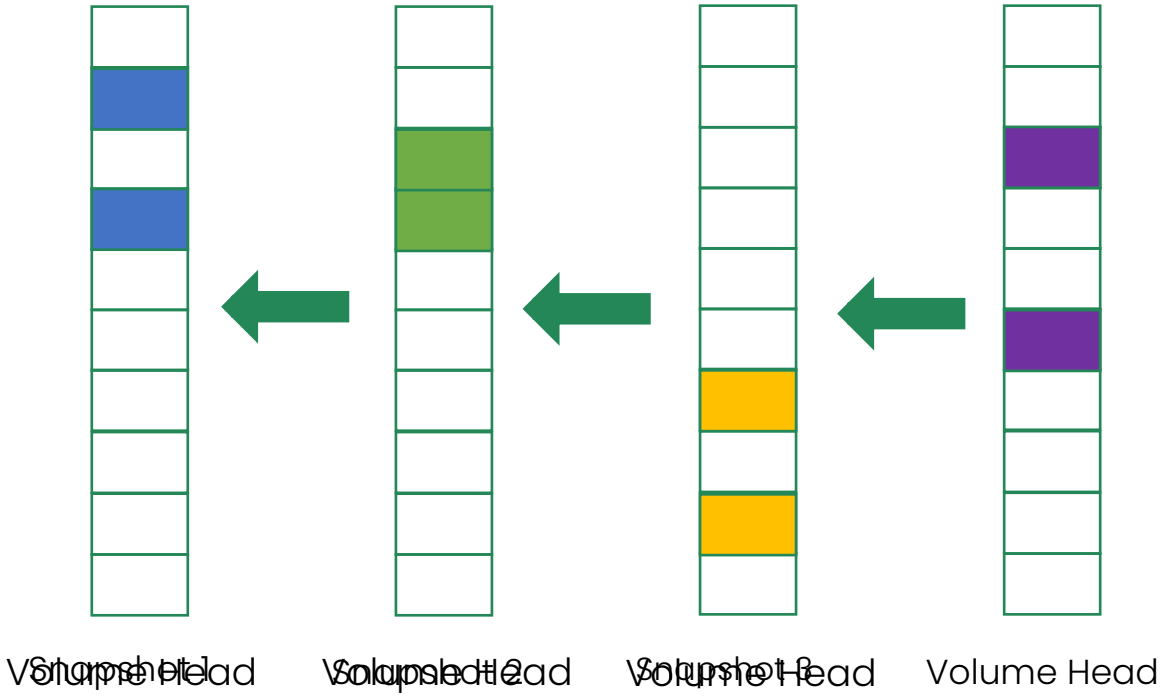
Longhorn Volume HA

Longhorn provides high availability block device across the availability zone

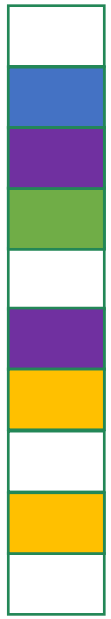


Longhorn Volume Snapshot

Snapshot Chain



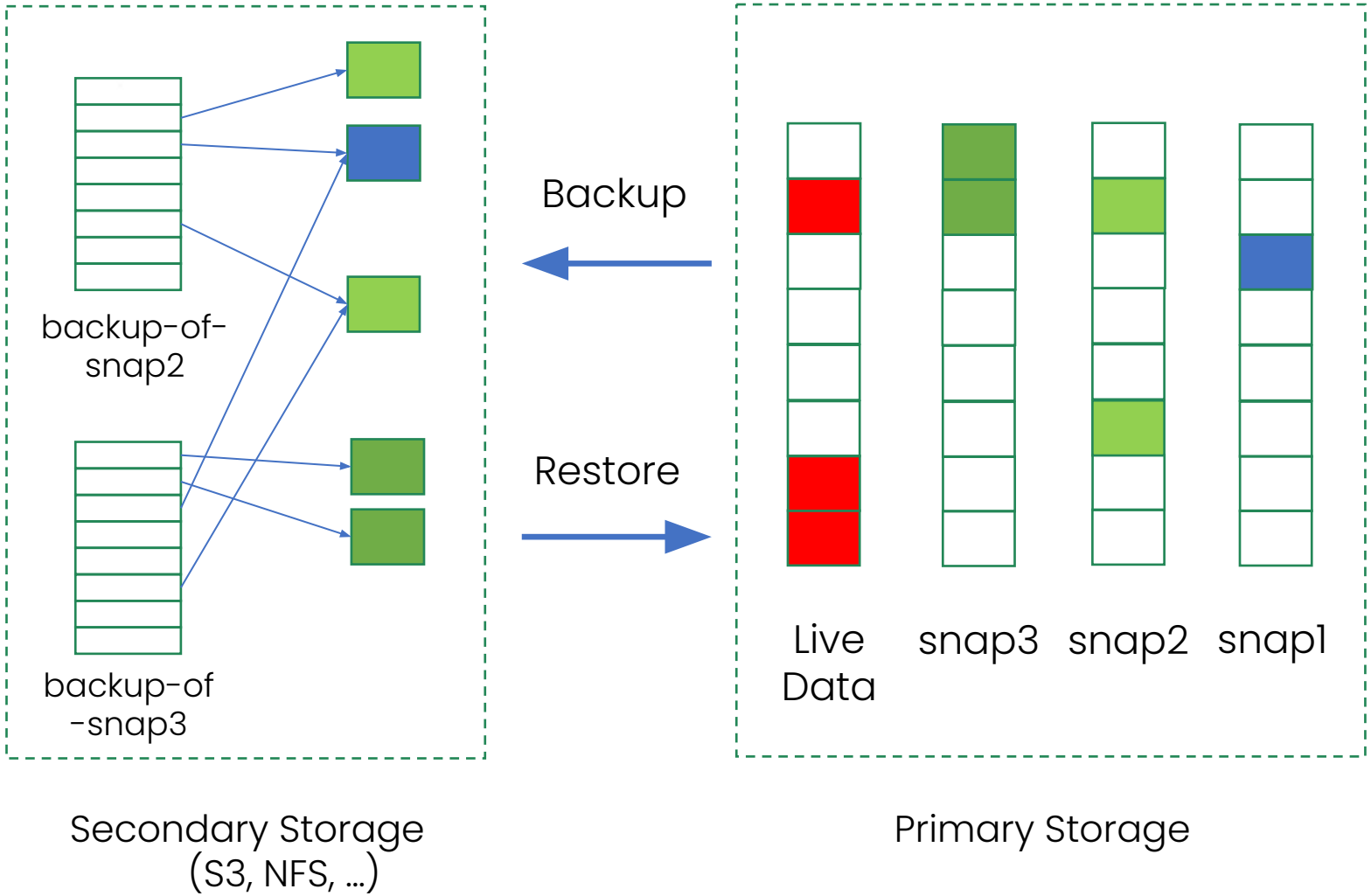
Volume Content



- Block Size: 4k
- Based on Linux Sparse File
- Read: lazily fill up a read index
- Write: always to the volume head, update read index as well

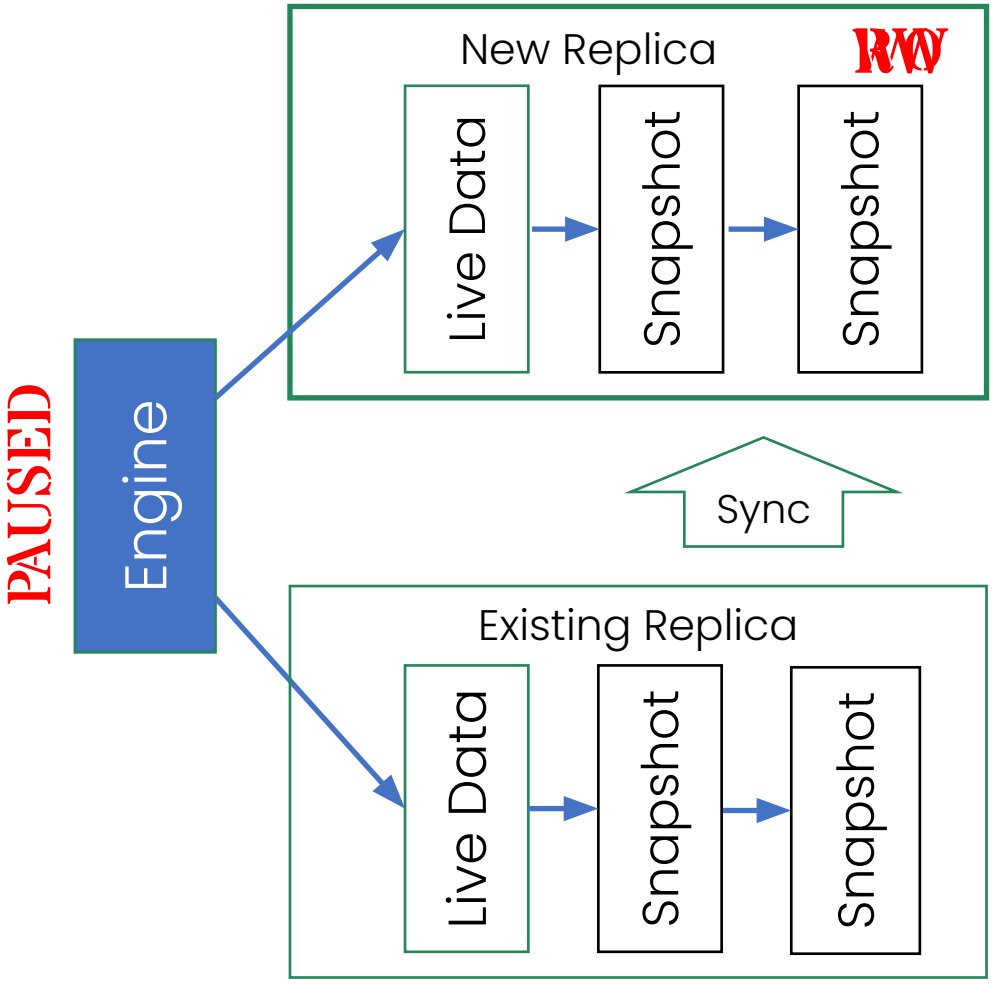
Longhorn Volume Backup

- AWS EBS-style backup
- Only changed blocks are copied
- 2M block size

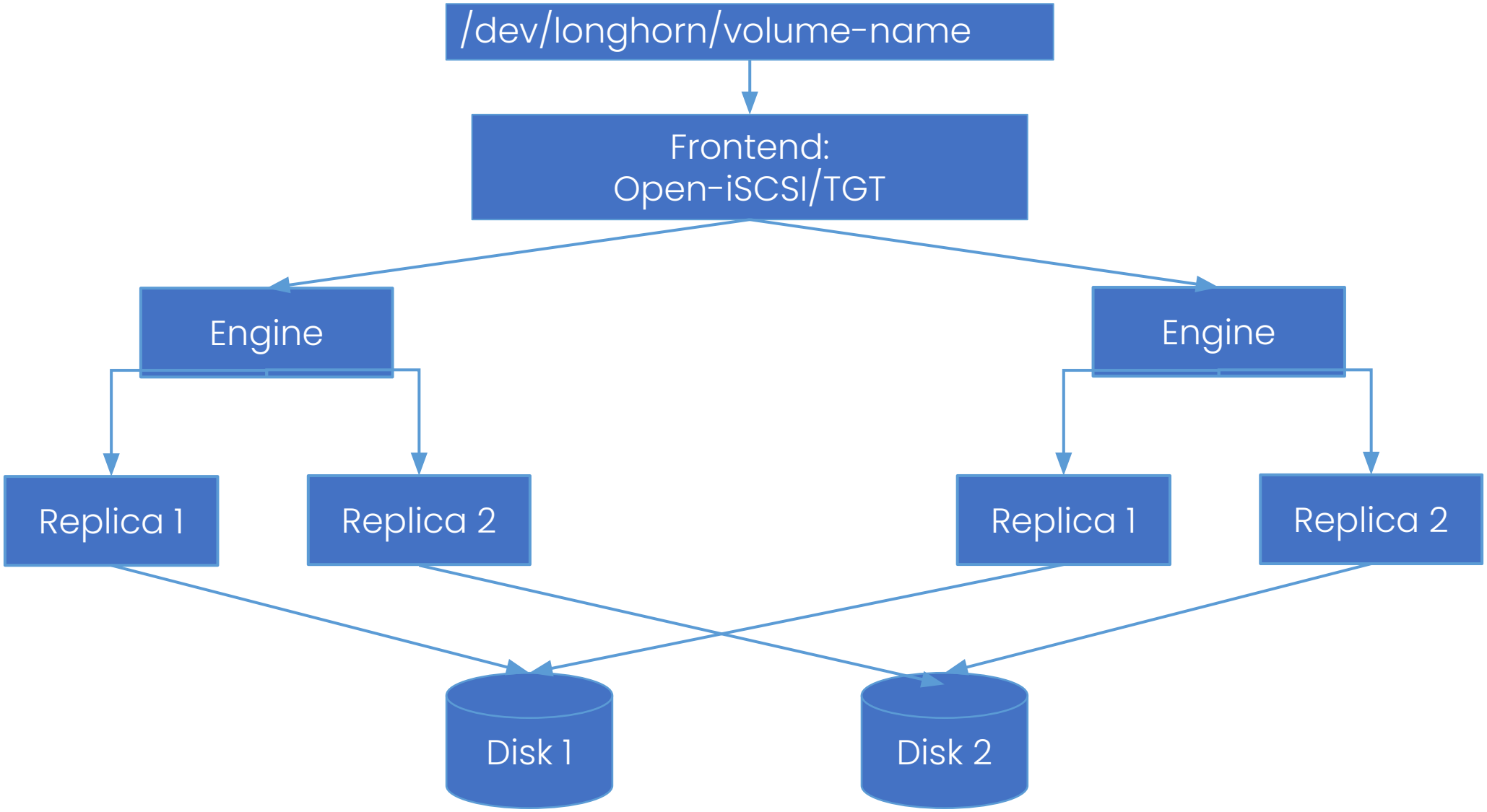


Longhorn Volume Replica Rebuilding

1. Pause engine
2. Take snapshot of existing replica
3. Add new replica in WO mode
4. Unpause engine
5. Sync snapshots
6. Set new replica to RW

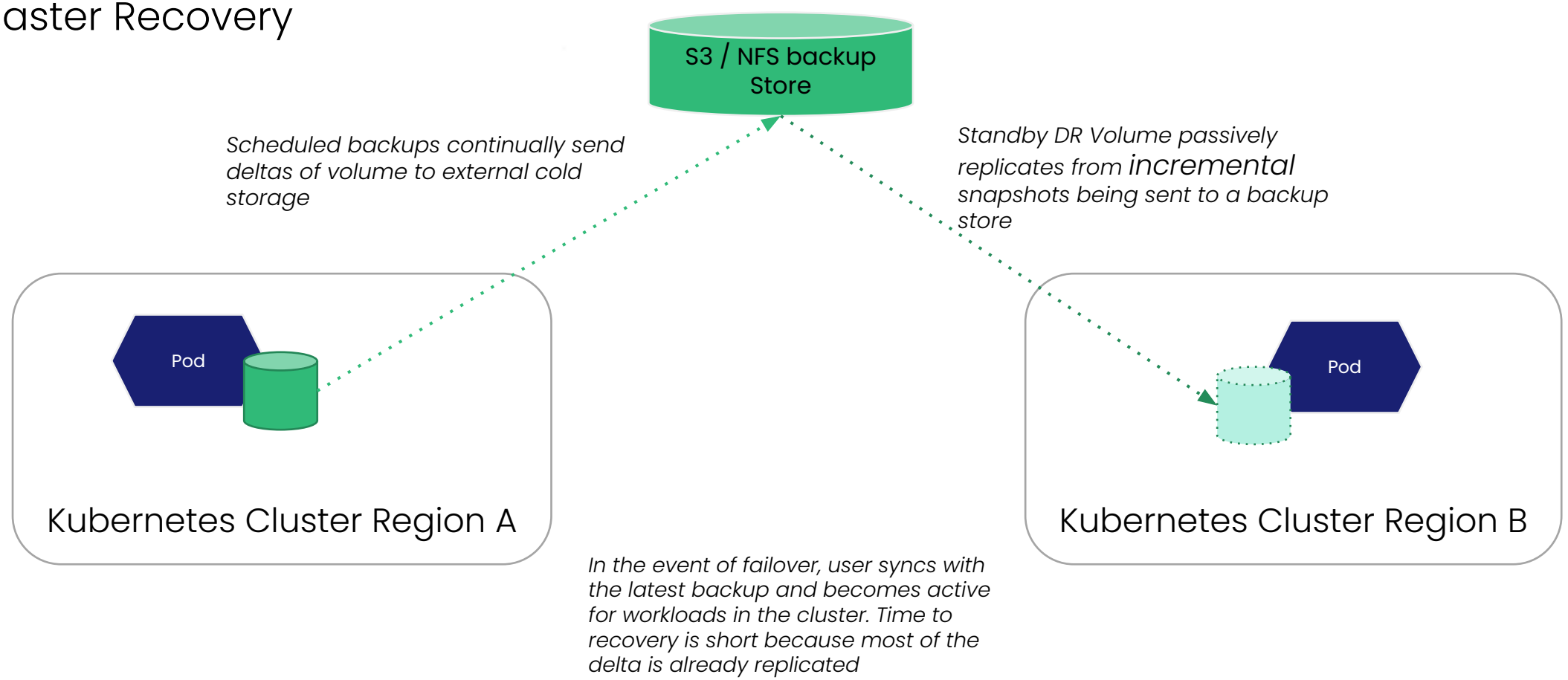


Longhorn Volume Live Migration



Longhorn Disaster Recovery

Multi-Cluster, Multi-site Disaster Recovery



Performance Benchmark between v1.4 and v1.5*

- Test Environment
  - Cloud provider: Equinix Metal
  - Machine: m3.small.x86 (Intel Xeon 8 cores/16 threads Processor, 5.1 GHz)
  - Storage: 1 SSD (Micron_5300_MTFD)
  - Network throughput between nodes: 15 Gbps

- Test Methodology
  - Uses the Longhorn developed kbench utility
  - Uses the fio command to test IOPS, bandwidth, and latency
  - Tested one replica using raw disk, Longhorn 1.4, and Longhorn 1.5*
  - Tested three replicas with Longhorn 1.4 and Longhorn 1.5*

1 Replica Performance Comparison

Disk

	Read	Write
IOPS Random	46,266	80,324
IOPS Sequential	45,658	96,055
Bandwidth Random I/O (KiB/sec)	454,321	454,156
Bandwidth Sequential I/O KiB/sec)	445,128	460,035
Latency Random I/O (ns)	165,637	44,131
Latency Sequential I/O (ns)	66,303	47,132

Longhorn v1.4 (strict-local)

	Read	Write
IOPS Random	36,682	22,430
IOPS Sequential	23,685	38,169
Bandwidth Random I/O (KiB/sec)	310,098	416,286
Bandwidth Sequential I/O KiB/sec)	441,751	429,559
Latency Random I/O (ns)	388,195	220,571
Latency Sequential I/O (ns)	250,138	215,796

Longhorn v1.5 (SPDK)*

	Read	Write
IOPS Random	89,665	79,689
IOPS Sequential	68,157	91,274
Bandwidth Random I/O (KiB/sec)	449,018	460,535
Bandwidth Sequential I/O KiB/sec)	450,726	457,083
Latency Random I/O (ns)	143,188	47,727
Latency Sequential I/O (ns)	62,306	47,734

3 Replicas Performance Comparison

KubeCon CloudNativeCon Europe 2023

Longhorn v1.4

	Read	Write
IOPS Random	40,423	15,979
IOPS Sequential	54,040	28,178
Bandwidth Random I/O (KiB/sec)	607,119	354,477
Bandwidth Sequential I/O KiB/sec)	1,096,812	264,512
Latency Random I/O (ns)	711,105	338,918
Latency Sequential I/O (ns)	732,359	341,736

Longhorn v1.5 (SPDK)*

	Read	Write
IOPS Random	90,313	68,708
IOPS Sequential	62,766	70,002
Bandwidth Random I/O (KiB/sec)	417,650	458,761
Bandwidth Sequential I/O KiB/sec)	455,303	448,301
Latency Random I/O (ns)	142,474	81,121
Latency Sequential I/O (ns)	62,984	81,001

What is Next?

Goals

- IO Performance
  - Longhorn SPDK Data Engine (1.5 experimental, 1.6 feature parity)
  - Local Volume Passthrough (1.6)
  - Performed frontend for Longhorn iSCSI Data Engine (1.6)
- Object Storage Interface
  - S3 object storage volume type (1.6 experimental)
- Run Anywhere
  - Cloud, Constrained Environment, Edge


KubeCon
CloudNativeCon
Europe 2023
Thank You 🙏
Q & A 🙋

Please scan the QR Code above
to leave feedback on this session