KubeCon | CloudNativeCon

North America 2024

KubeCon | CloudNativeCon

North America 2024

Longhorn: Introduction and Deep Dive

Phan Le

Senior Software Engineer, SUSE

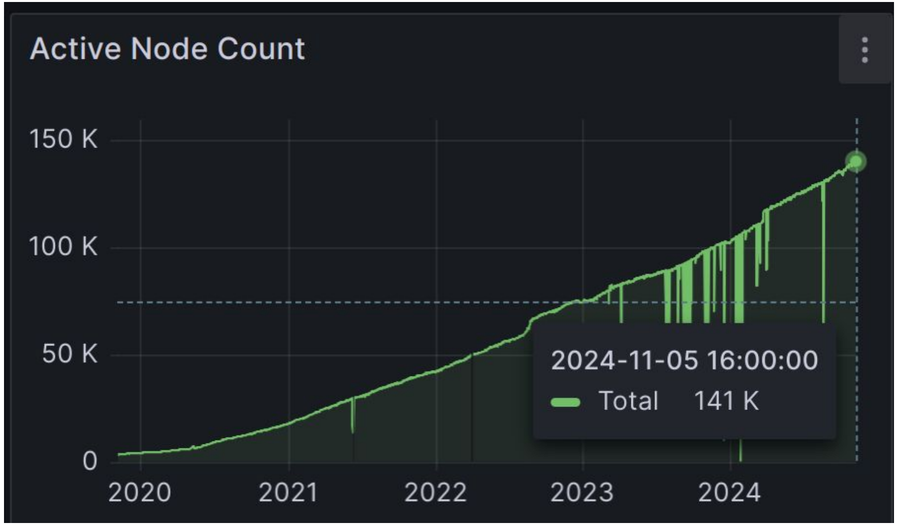
LONGHORN

One click to add persistent storage support to any Kubernetes cluster

Open Source
Distributed Storage Software for Kubernetes
https://longhorn.io/

Community Adoption

- 141,000+ worldwide live node count reached and counting!
- 44%+ growth year over year



Active Node Count

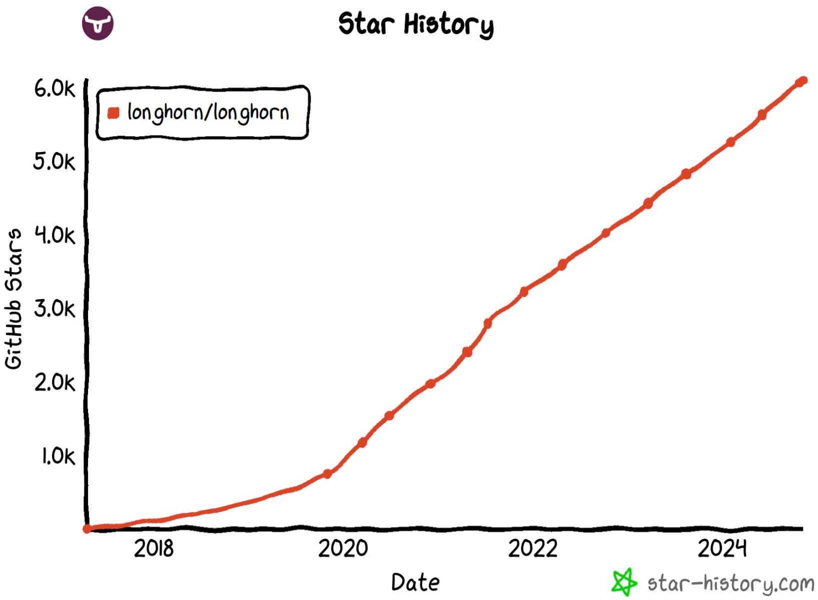
2024-11-05 16:00:00

Total 141 K

Metrics available at https://metrics.longhorn.io/

Source code: https://github.com/longhorn/upgrade-responder

- 3,700+ users in the Slack channels
- GitHub star: 6.1k



Overview

Use case

- Reliable, scalable storage solution for stateful workloads in Kubernetes clusters
- Hyper-converged solution. Run on the same cluster using the local disks and provide replicated storage for pods.

Reliability

- Crash consistent
- Multiple layers of protection against data loss, including built-in snapshot and backup support

Usability

- One click installation
- Polished user experience

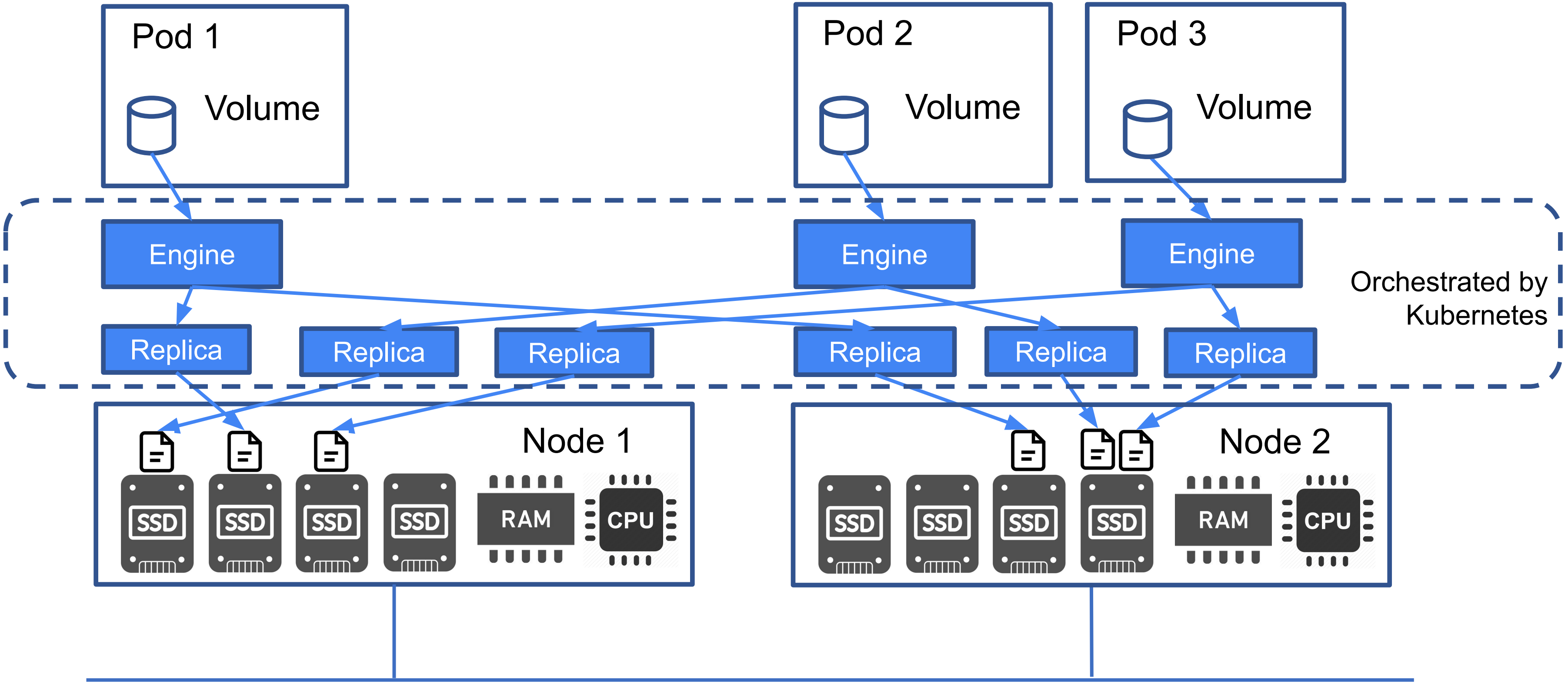
Maintainability

- Easy to understand
- Easy to recover even in the worst-case scenario
- Upgrade without interrupting the workload

What Longhorn Supports

- **Kubernetes Persistent Volume Support**
  - Block, FS volumes
  - RWO, RWX
- **CSI Protocol Support**
  - Volume Provision, Attachment, Snapshot, Clone, Restore, Expansion
- **Volume Capabilities**
  - Thin provisioning
  - Snapshot, CoW
  - Trim, Expand
  - Live upgrade, migration
- **IO Performance**
  - Data Locality, Strict local Volume
  - v1 & v2 Data Engines
- **Intuitive UI**
- **Storage, Storage Topology**
  - v1 & v2 Longhorn disk
  - Disk/Node/Zone replica scheduling anti-affinity
  - Storage tag
- **Data Protection**
  - Data replication
  - Data encryption in transit & at rest
  - Bit-rot protection
- **Data Services**
  - In-cluster snapshot & revert
  - Out-of-cluster backup & restore
  - Disaster recovery volume
- **Space Usage Management**
  - Space efficiency for Snapshot
  - Backup compression

Architecture - Engine



Latest Feature Release v1.7

- Data Reliability and Integrity
  - Support Periodic and On-Demand Full Backups
  - High Availability of Backing Images
- Resilience
  - RWX Volumes Fast Failover
- Scheduling
  - Volume Locality for RWX Volumes
  - Auto-Balance Pressured Disks
- Networking
  - Storage Network Support for RWX Volumes
- Longhorn CLI

V2 Data Engine new features:

- Support disk drivers
  - AIO, NVMe
- Online Replica Rebuilding
- Volume Operations
  - Filesystem trim

Roadmap - v1.8 (Q1 2025)

V2 Data Engine

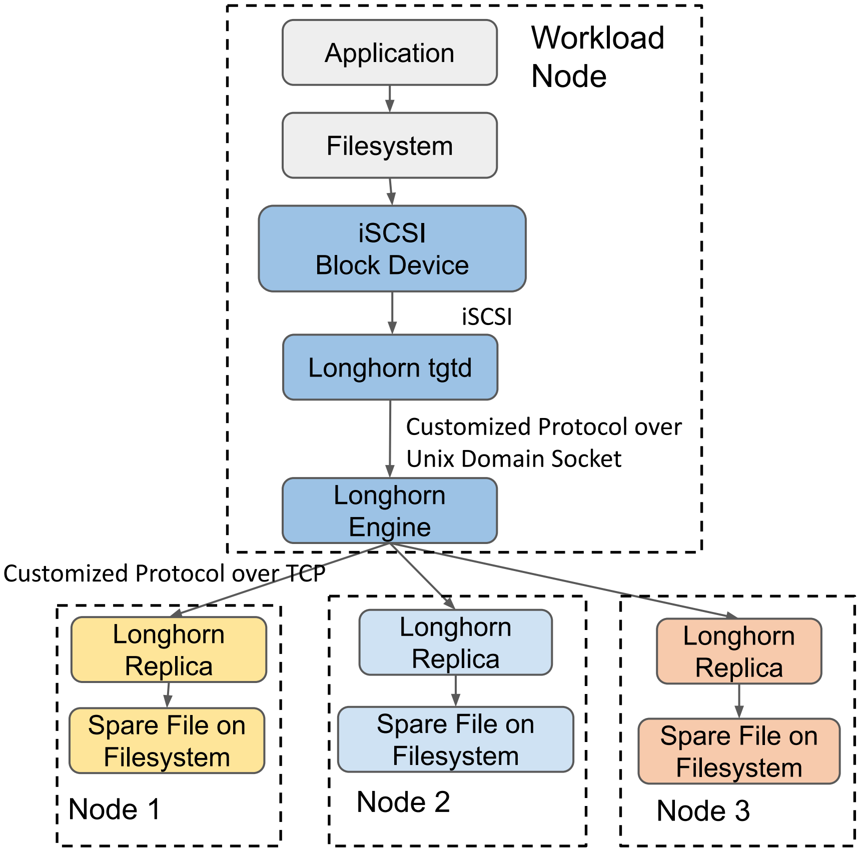
- Performance
  - Configurable CPU Cores
  - Dynamic Scheduler
- Replica Rebuilding
  - Online Rebuild Improvement
  - Snapshot Checksum
  - Delta Replica Rebuilding:
- Data Recovery
  - Auto Salvage
  - Disaster Recovery Volume
- Volume Live Upgrade
- Live Migration
- Backing Image
- Volume Expansion

Other Features:

- Multiple Backup Stores
- Longhorn CLI Commands for Operations

Longhorn Engine v1 vs v2

Longhorn Engine v1



Workload Node

Application

Filesystem

iSCSI Block Device

iSCSI

Longhorn tgtd

Customized Protocol over Unix Domain Socket

Longhorn Engine

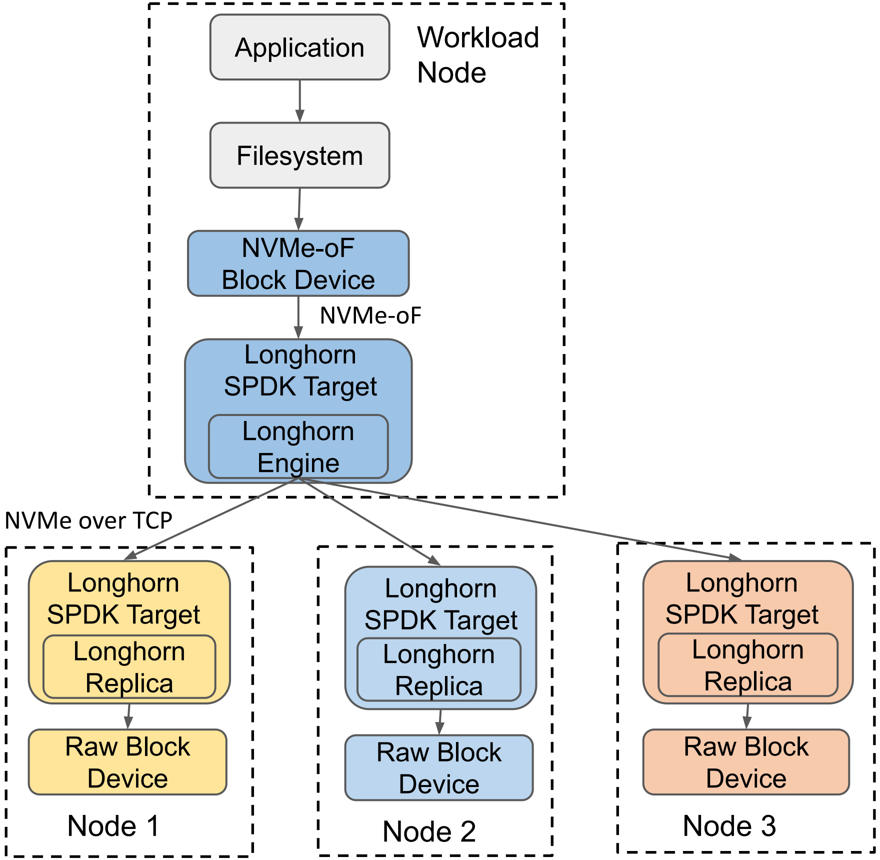
Customized Protocol over TCP

Node 1: Longhorn Replica → Spare File on Filesystem

Node 2: Longhorn Replica → Spare File on Filesystem

Node 3: Longhorn Replica → Spare File on Filesystem

Longhorn Engine v2



Workload Node

Application

Filesystem

NVMe-oF Block Device

NVMe-oF

Longhorn SPDK Target

Longhorn Engine

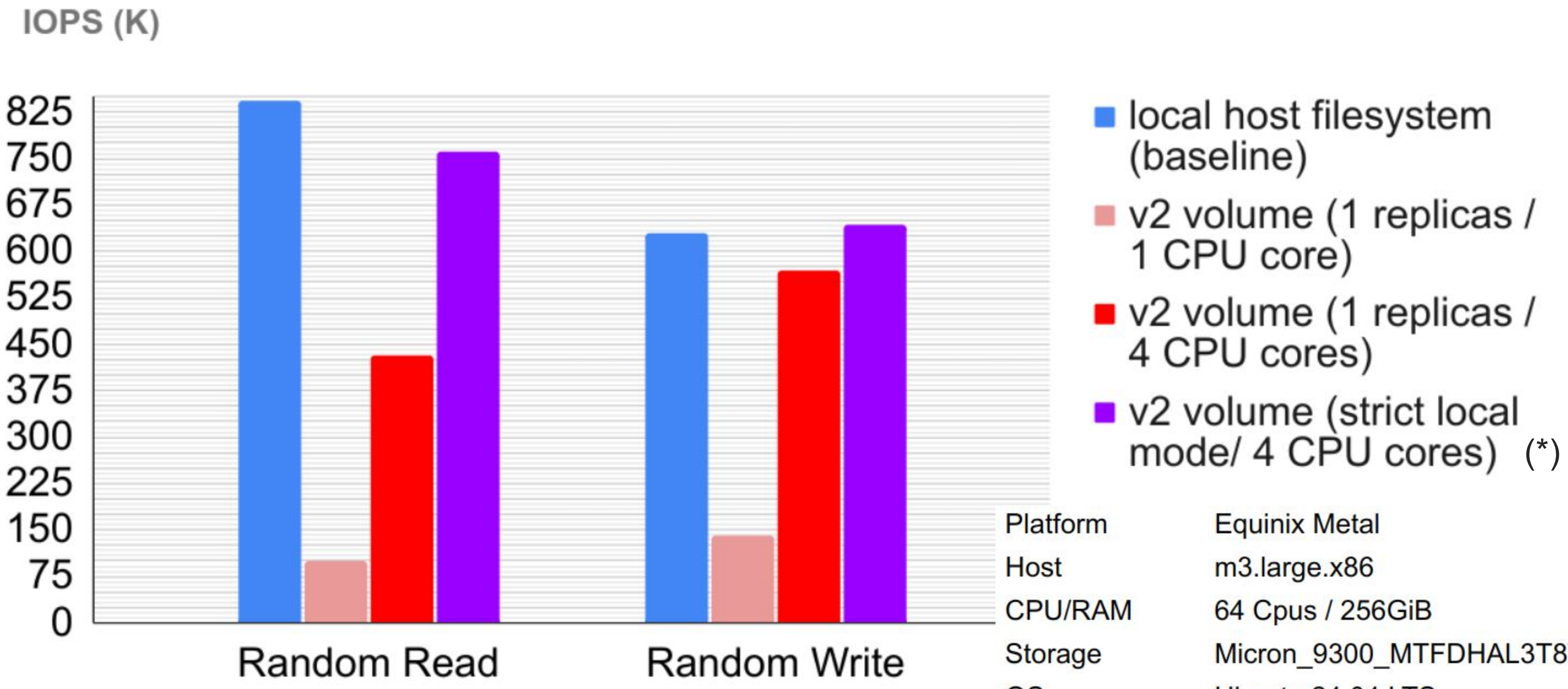
NVMe over TCP

Node 1: Longhorn SPDK Target, Longhorn Replica → Raw Block Device

Node 2: Longhorn SPDK Target, Longhorn Replica → Raw Block Device

Node 3: Longhorn SPDK Target, Longhorn Replica → Raw Block Device

V2 Engine Performance - IOPs

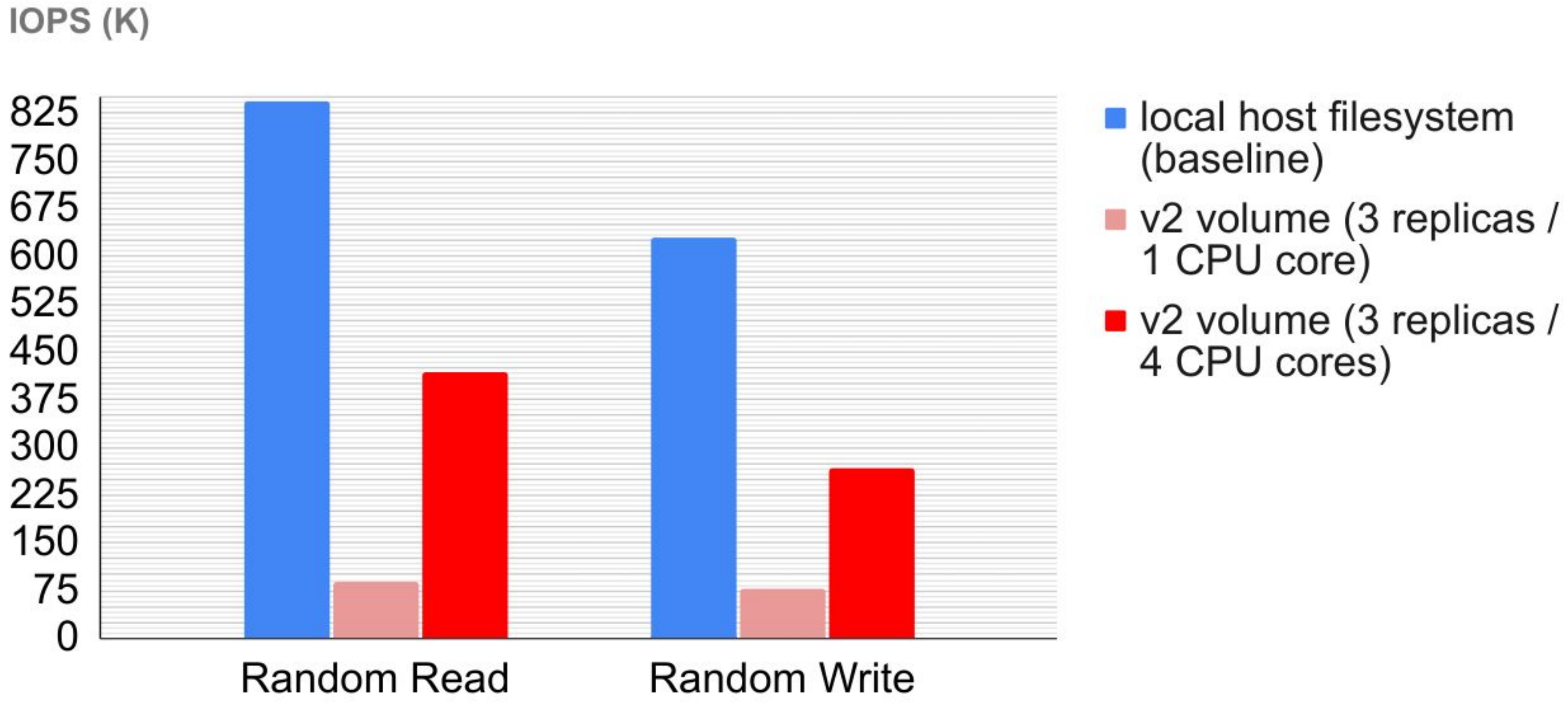


Platform	Equinix Metal
Host	m3.large.x86
CPU/RAM	64 Cpus / 256GiB
Storage	Micron_9300_MTFDHAL3T8TDP
OS	Ubuntu 24.04 LTS
Kernel	6.8.0-45-generic

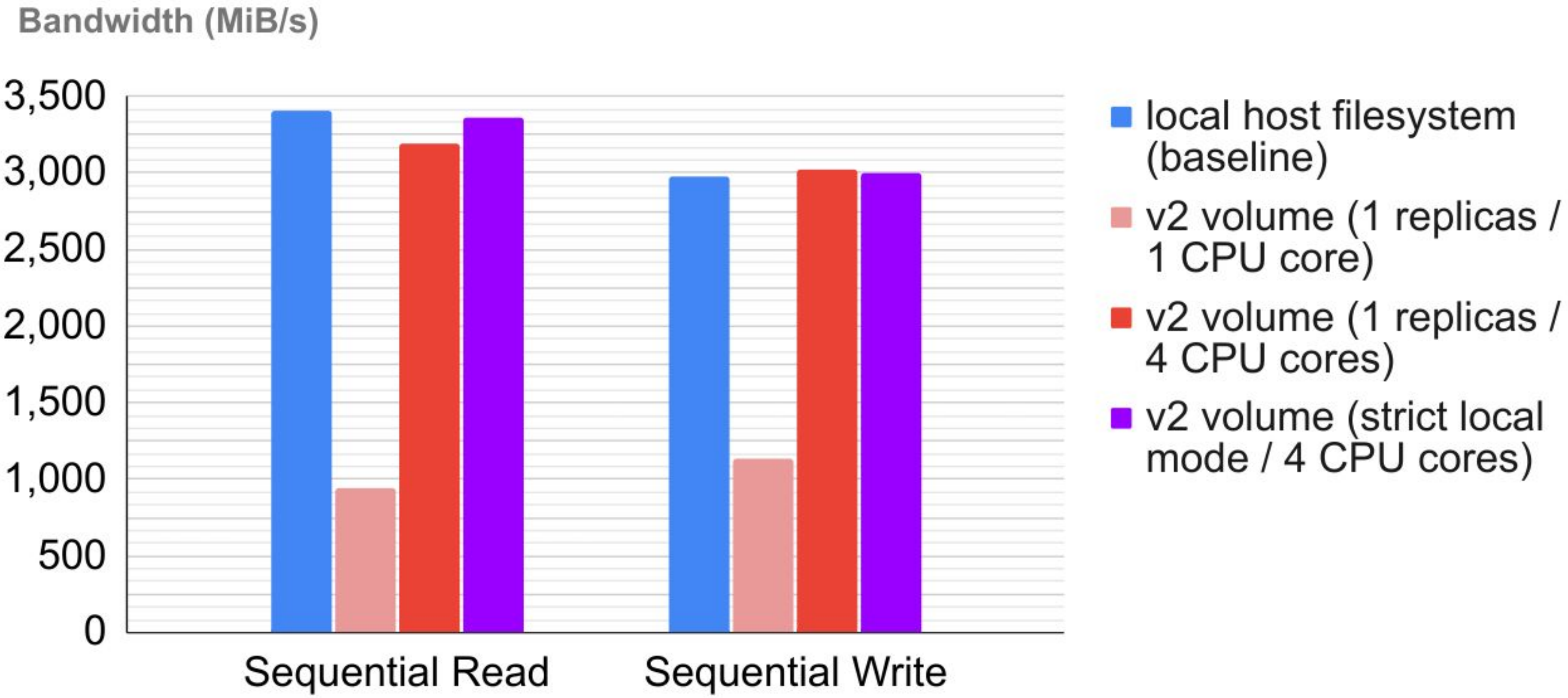
SPDK is using NVMe driver

* Note that strict local mode is coming

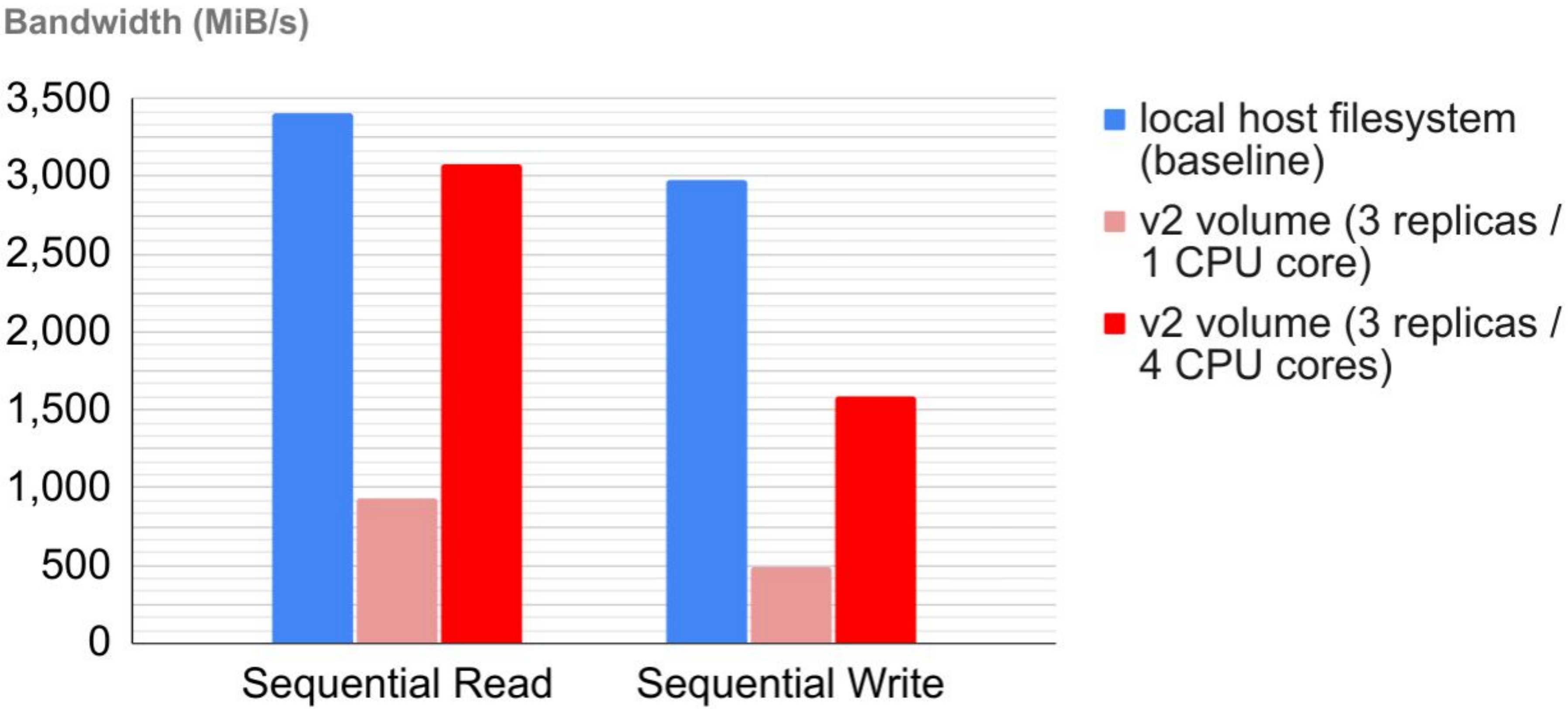
V2 Engine Performance - IOPs - 3 replicas



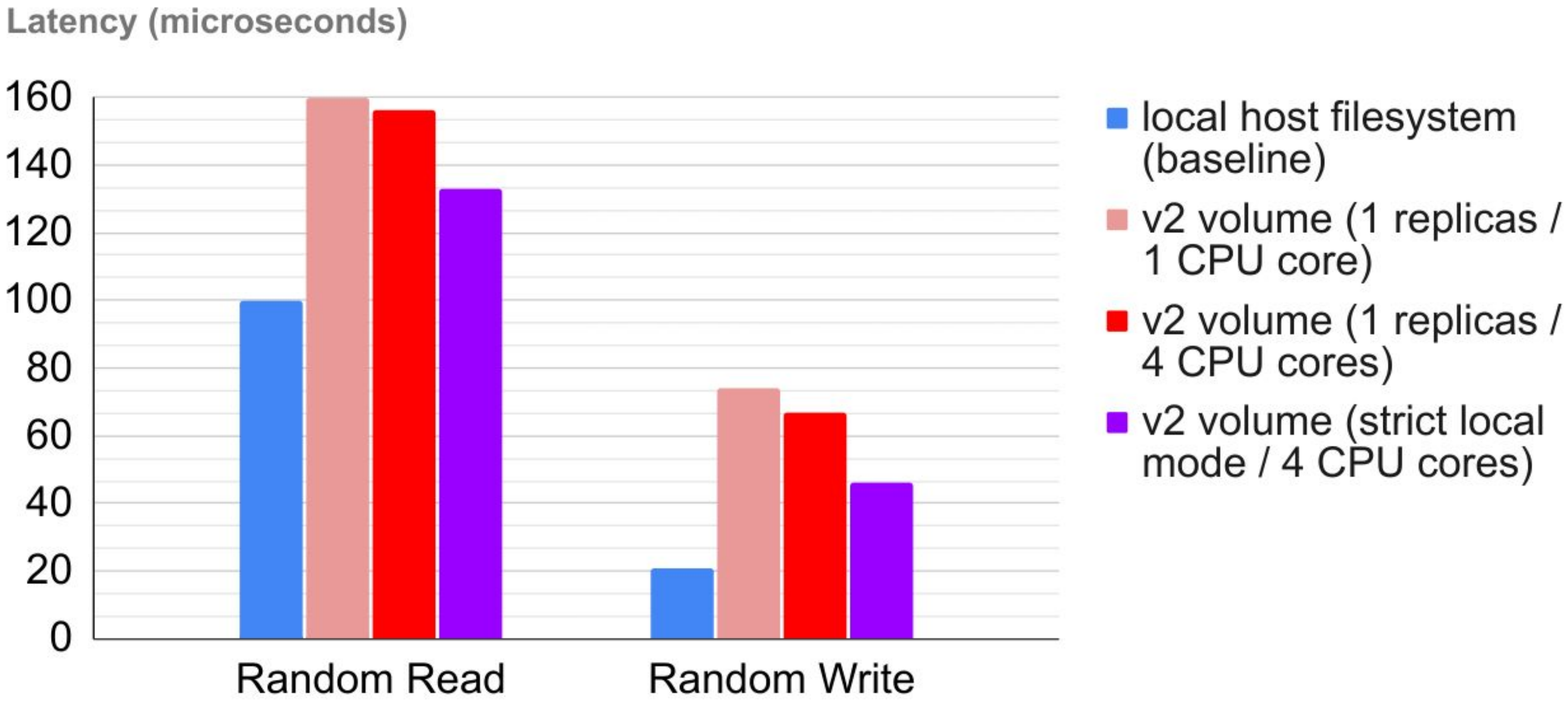
V2 Engine Performance - Bandwidth



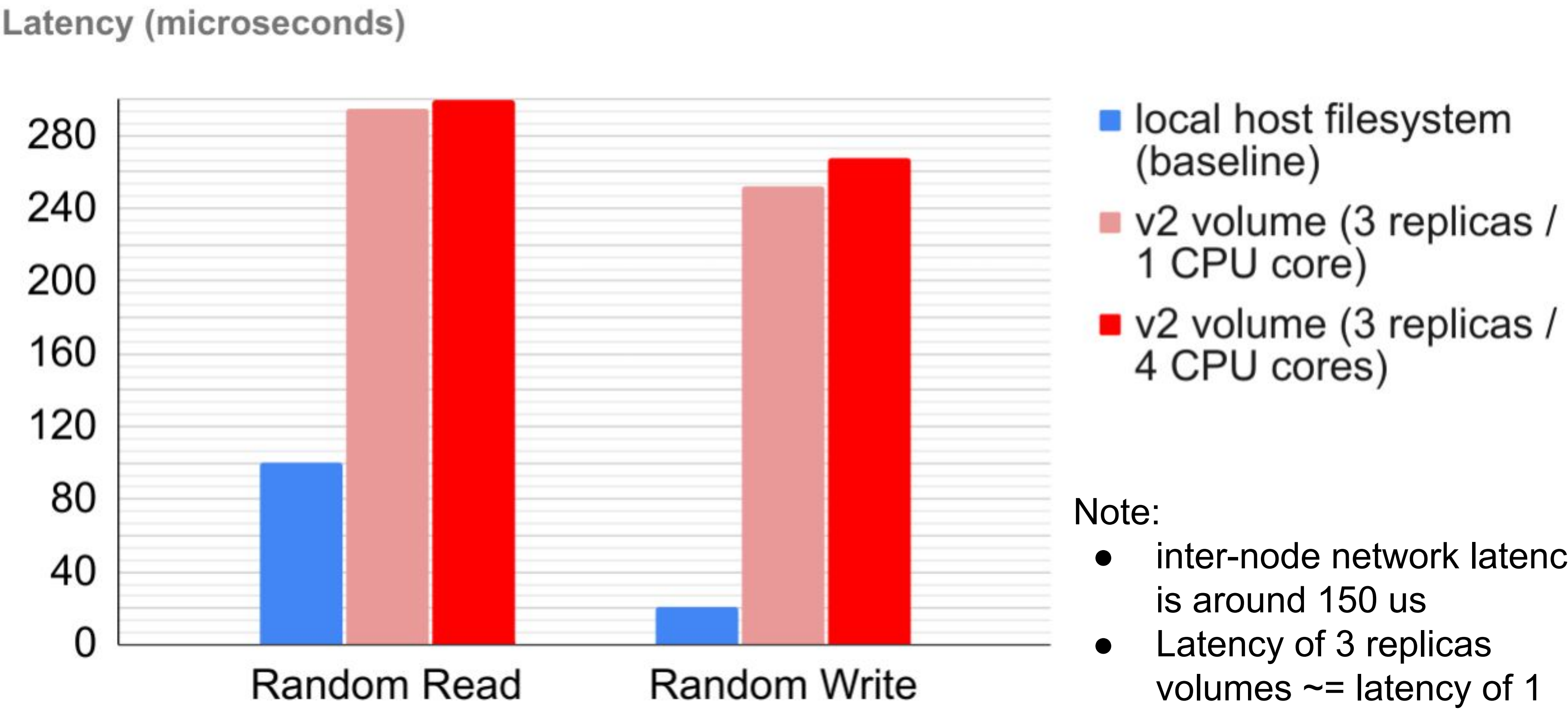
V2 Engine Performance - Bandwidth - 3 replicas



V2 Engine Performance - Latency



V2 Engine Performance - Latency - 3 replicas



Note:

- inter-node network latency is around 150 us
- Latency of 3 replicas volumes ~= latency of 1 replica + network latency

KubeCon CloudNativeCon

North America 2024

V2 Engine Demo

- Adding NVMe disk
- Create a SC with v2 engine
- Create a PVC
- Deploy a workload pod
- Run fio benchmark inside workload pod
- Take a snapshot and a backup for Longhorn volume

Please provide us feedbacks at

Q&A

https://kccncna2024.sched.com/event/1hoxZ