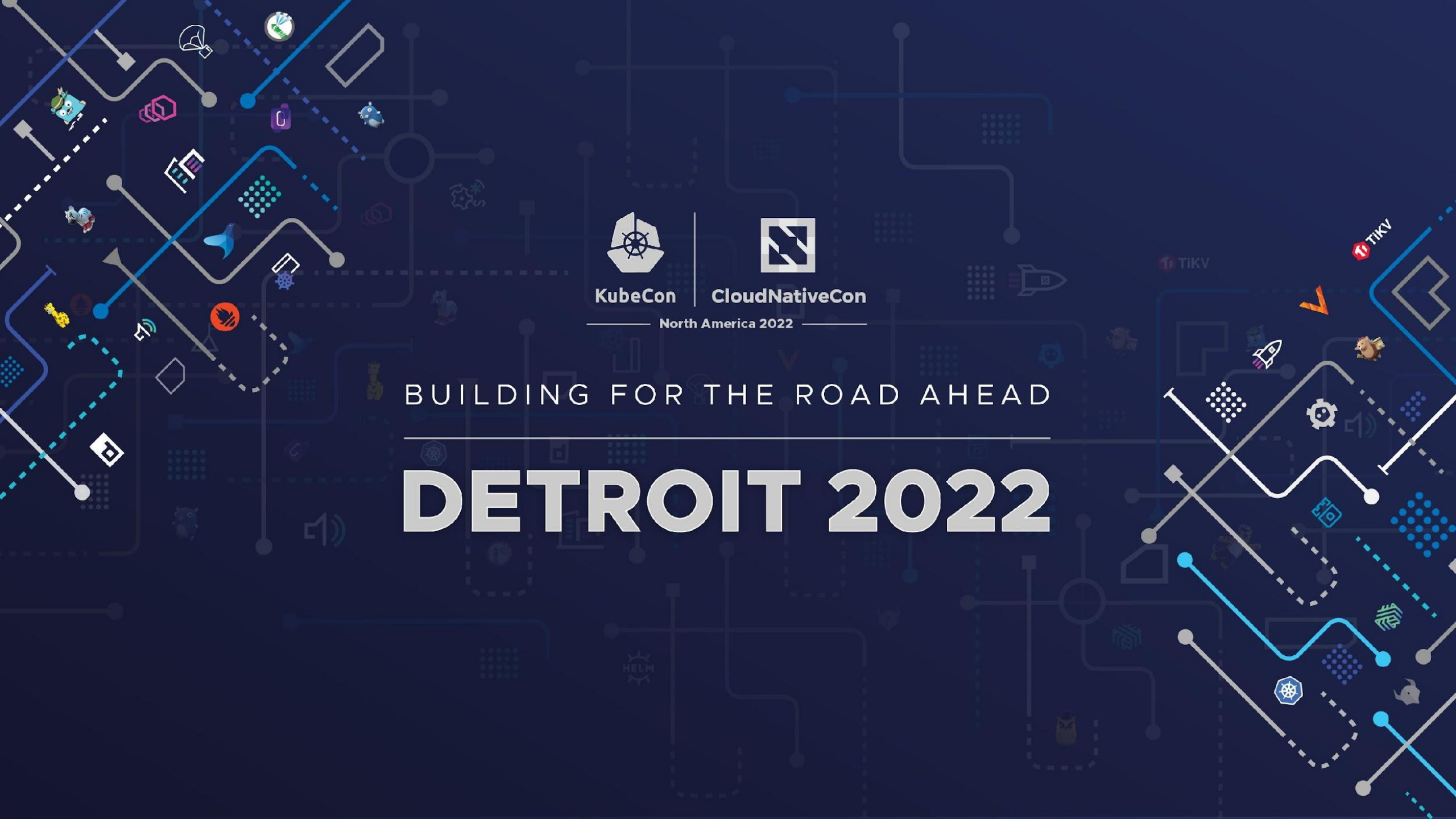

KubeCon
CloudNativeCon
North America 2022
BUILDING FOR THE ROAD AHEAD
DETROIT 2022
TiKV
TiKV
HELM

BUILDING FOR THE ROAD AHEAD

DETROIT 2022

Storage Wars

Seán C McCord

Who am I?

- Seán C McCord
  - former Principal Architect at Sidero Labs, maker of Talos Linux (and now Omni!)
  - original author of containerization of Ceph used by RedHat and Rook
  - seeker of distributed storage solutions for over 25 years
  - contributor to many open source projects over many years, across many fields
  - gainfully unemployed 😉

This talk is...

- NOT a comparison of any cloud provider systems
- NOT a comparison of vendor CSIs
- NOT benchmark fest

- an Overview of the main open source storage solutions available for Kubernetes
- a guide to the Key Criteria to decide which is right for you

This talk is...

Plan:

1. Types of Storage
2. Location
3. Characteristics of Storage
4. Storage Interfaces
5. Contenders
6. Summary

Types of Storage

- Object stores
- Block stores
- Shared filesystems

Object Stores

- Key-Value database for data
- Based on web tech
  - Massively, widely readable
  - Network-native
  - Easily integrated and layered
- REST interfaces
- Writes, like web, more difficult
- Cannot use directly

Block Stores

- Present storage as block-oriented devices: disks
- Full control of filesystem and its tuning
- Direct map into Kubernetes PV
- Single pod attachment
- No one standard, but some protocols exist:
  - iSCSI
  - NVMEoF
- Can be made to offer the other types
  - MinIO for object storage
  - NFS for shared filesystem

Shared filesystems

- Presents files and directories across a number of nodes as a filesystem
- NFS
- Always locking problems
  - bottlenecks
  - contention locks
  - slow
- Least common denominator
- Easy to setup... and forget about

Location

- Clouds
  - single cloud vendor? Use their system
- In-Cluster vs Out-of-Cluster
  - Common manifests, Kubernetes for all
  - Greater portability, modularity
  - Danger
    - Storage is stateful
    - Data has value
    - Not easily or quickly replicated
    - Resource contention

Storage Characteristics

- Scalability
- Performance
- Cost

Characteristic: Scalability

- Traditional RAID
  - single controller
  - highly centralised
  - limited replication factors
- Standard SAS expanders
  - redundant controllers
  - highly centralised
  - limited tiering
- Storage clusters
  - eliminate single points of failure
  - horizontally scalable
  - faster as they grow
  - dynamic, fine-grained replication, topology

Characteristic: Performance

- Benchmarks misleading
  - drives themselves
  - controllers and interfaces
  - workload needs
  - unexpected scaling effects
  - test as precisely as possible
- Some systems slow down as they scale
- Others speed up with scale
- Still, there *are* architectural choices which influence real-world performance

Characteristic: Cost

- Disks, controllers... hardware
- Complexity of the system
- Maintenance - drives _will_ fail... often
- Growth / Scalability

Storage Interfaces

- iSCSI
  - old standard
  - used by many
  - open-iscsi
    - pre-container age
    - bad practises
- NVMEoF
  - new standard
  - cleaner, simpler, faster
  - clean containers
- Ceph
  - RBD
  - CephFS
- NFS

The Contenders

Contenders: Vendor adapters

- Majority of CSI providers
- Specific vendor hardware or service
- Just an adapter
- If you have that vendor, just use it

Contenders: Proprietary

- Black boxes
- No way to evaluate
- Examples (no particular order)
  - Hedvig: iSCSI, all three types
  - Kumoscale: NVMEoF, NVME only
  - StorageOS: size-limited freemium
  - StorPool: iSCSI pooling abstractor
  - onDat: generic storage adapter
  - PortWorx: size-limited freemium, claims performance

Contenders: Local Storage

- Pod - Node binding
- native Persistent Local Volumes
- TopoLVM: use LVM volumes

Contenders: Shared Filesystems

- NFS
- Gluster via Kadalu
  - in-cluster operator
  - aggregating shared fileystem
- CephFS

Contenders: Pooling/Aggregating

- Group and repackage for Kubernetes
- VDA (Virtual Disk Array)
  - NVMEoF
  - simple pooling aggregator
  - not much tooling
- MinIO
  - feature-rich object store
  - aggregates wide variety of storage backends
- LinStor
  - somewhat aggregative, somewhat replicative
  - pluggable providers for many things

Contenders: Storage Clusters

- OpenEBS family
  - most limited of storage clusters here
  - just block storage
  - limited replication and topology control
  - cStor
    - original engine
    - ZFS-based
    - iSCSI interface
    - rugged, tested, slow
  - Jiva
    - stepchild; upgraded iSCSI interface
  - Longhorn

Contenders: Storage Clusters

- OpenEBS family (continued
  - Longhorn
    - Rancher-sponsored, Rancher-focused
    - variously rewritten, but still iSCSI
  - Mayastor
    - shiny: rust, NVMEoF, even Nix
    - very new
      - 1.0 this year, breaking changes
      - simple replication only
      - history of docs problems
      - requires external etcd database

Contenders: Storage Clusters

- SeaweedFS
  - RADOS-based (like Ceph)
  - Simpler, more focused re-envisionment of Ceph
  - Optimised for container workloads and small files
  - in- or out-of-cluster opteration
- Ceph
  - complex to start
  - scales well, faster with size
  - resource-intensive
  - immensely tunable
  - high topology awareness
  - rugged, tested, and highly fault-tolerant

Contenders: Storage Clusters

- Rook/Ceph
  - Operator for Ceph (and others)
  - Ceph administration = easy
  - Trades some control for automation

Summary: Comparison Table

Name	Supported Types	Administrative Complexity	Scalability	Reliability	Performance
Ceph	OBF	Hard 	3	3	2
Kadalu/Gluster	F	Medium 	2	2	2
Linstor	B	Easy 	2	2	3
Longhorn	B	Medium 	2	2	2
MinIO	O	Easy 	2	2	2
NFS	F	Easy 	1	1	1
OpenEBS/cStor	B	Medium 	2	3	2
OpenEBS/Mayastor	B	Medium 	2	1	3
Rook/Ceph	OBF	Easy 	3	3	2
SeaweedFS	OBF	Medium 	2	2	2
TopoLVM	B	Easy 	1	2	3
VDA	B	Hard 	2	2	2

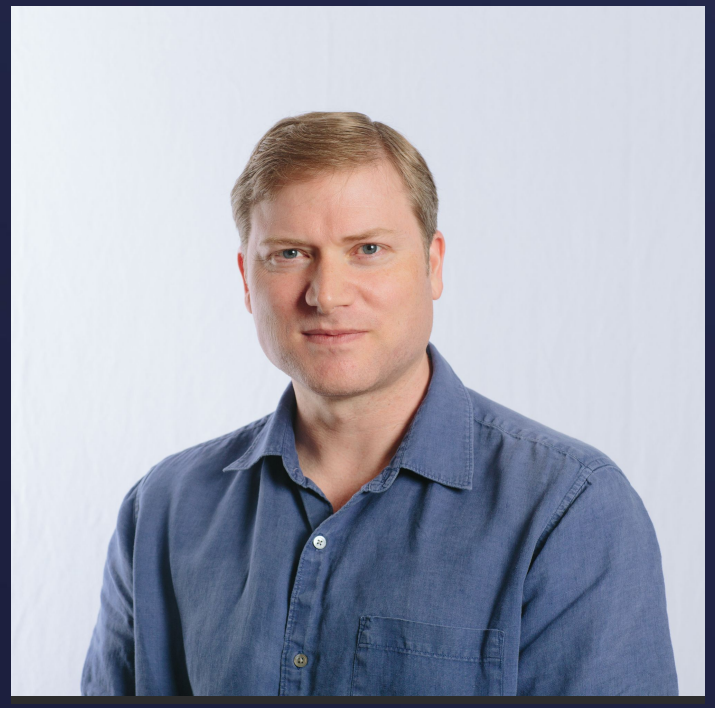
Executive Summary

- Pay someone to handle it: `PortWorx`
- Nothing fancy, just store it: `Linstor`
- Need control or scaling, but Ceph is scary:
  - **Performance** over *Ruggedness*: `OpenEBS/Mayastor`
  - **Ruggedness** over *Performance*: `OpenEBS/cStor`
- Best features, scaling, and fault tolerance
  - **Stability** over all: `Ceph`
  - *Otherwise*: `Rook/Ceph`

Storage Wars

KubeCon | CloudNativeCon
North America 2022
BUILDING FOR THE ROAD AHEAD
DETROIT 2022

KubeCon | CloudNativeCon
North America 2022
BUILDING FOR THE ROAD AHEAD
DETROIT 2022



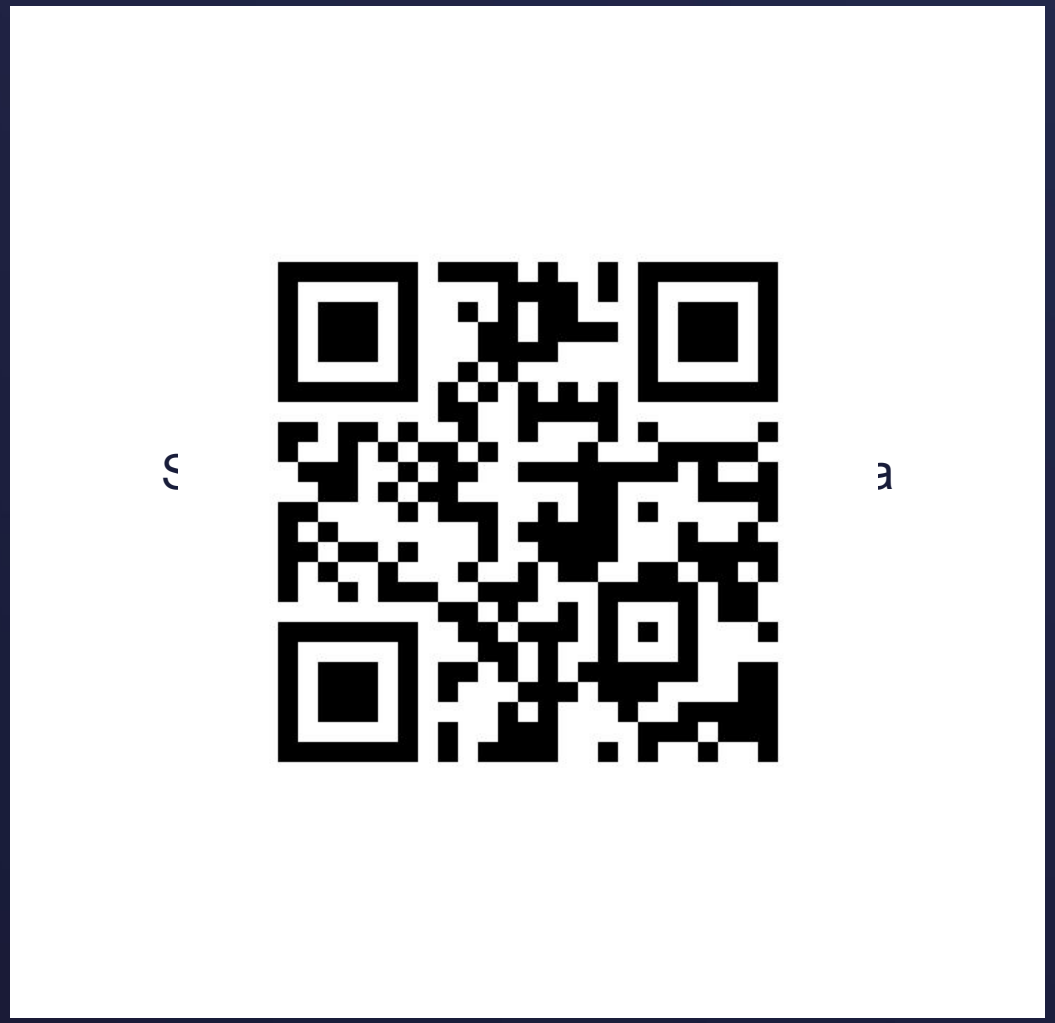
October 24-28, 2021

Seán C McCord
CTO, *CyCore Systems*

KubeCon CloudNativeCon North America 2022

BUILDING FOR THE ROAD AHEAD

DETROIT 2022



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