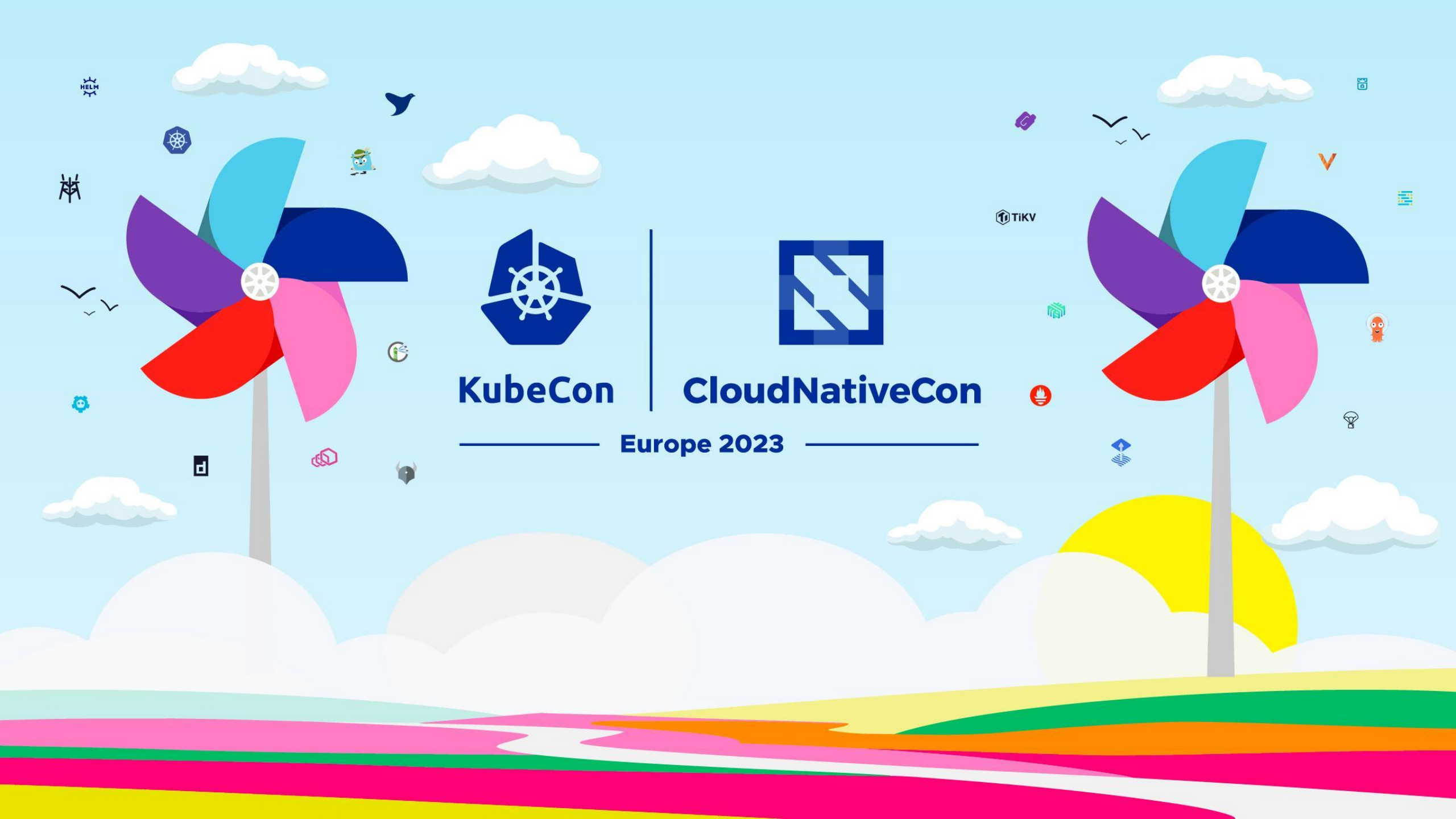
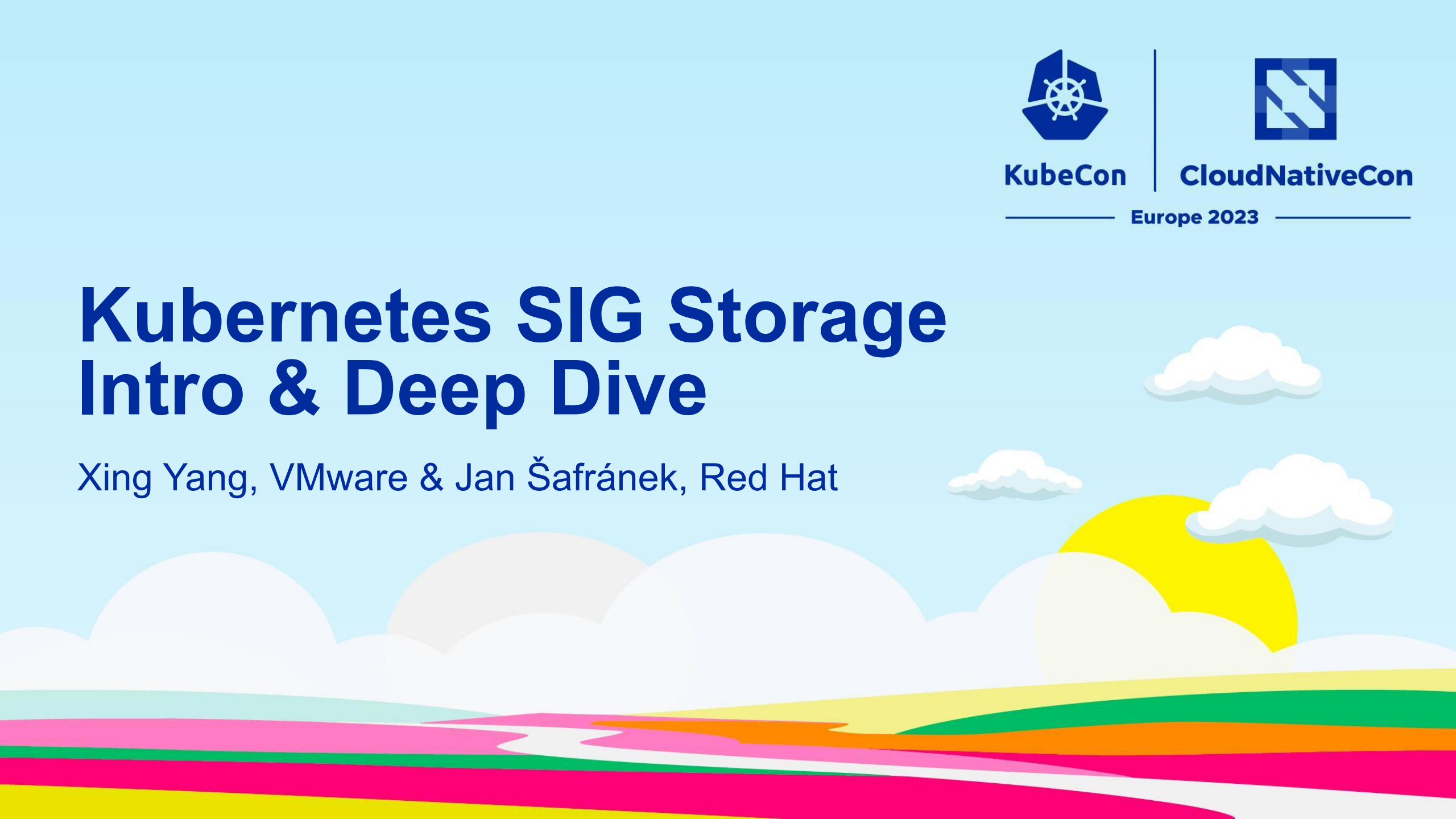

KubeCon
CloudNativeCon
Europe 2023
TiKV



Kubernetes SIG Storage Intro & Deep Dive

Xing Yang, VMware & Jan Šafránek, Red Hat

Agenda

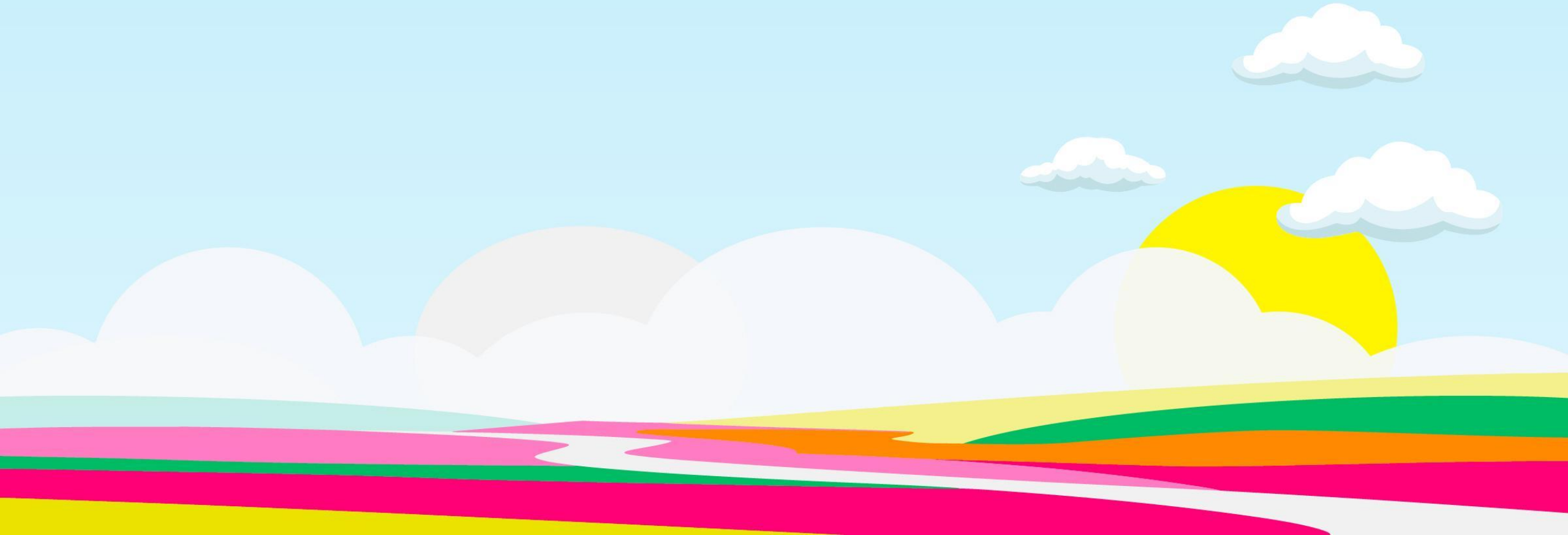
- Who we are
- What we did in 1.26
- What we did in 1.27
- Features in design/prototyping
- How to get involved
- Q & A

SIG-Storage

- Co-Chairs: Saad Ali (Google) and Xing Yang (VMware)
- Tech Leads: Michelle Au (Google) and Jan Šafránek (Red Hat)
- #sig-storage slack channel members: 5300+
- #csi slack channel members: 1400+
- #sig-storage-cosi slack channel members: 280+
- #csi-windows slack channel members: 230+
- SIG-Storage zoom meeting attendees: 25
- Unique approvers for SIG-owned packages: 30

What we do

- Defined in SIG Storage Charter
  - Persistent Volume Claims and Persistent Volumes
  - Storage Classes and Dynamic Provisioning
  - Kubernetes volume plugins
  - Secret, ConfigMap, DownwardAPI, Projected and EmptyDir Volumes (co-owned with SIG-Node)
  - Container Storage Interface (CSI)
    - CSI sidecars
    - Most CSI drivers are owned by SIG Cloud Provider or other community
  - Container Object Storage Interface (COSI)
    - Alpha in 1.25



What we did

What we did in 1.26

- GA
  - Delegate FSGroup to CSI Driver instead of Kubelet
  - On-going effort: CSI Migration
    - Azure File
    - vSphere
- Beta
  - Reconcile Default Storage Class Assignment
    - Allows existing PVCs without “storageClassName” updated to use the new default StorageClass
  - Non-graceful Node Shutdown

Non-graceful Node Shutdown

- Graceful node shutdown
  - “`ssh <node> poweroff`”
  - Will stop containers + unmount volumes
- Non-graceful Node Shutdown
  - Can’t ssh to node because of:
    - HW failure
    - Network lost
    - Kernel panic
    - …

Non-graceful Node Shutdown

- Beta in 1.26; introduced as Alpha in 1.24
- Expected workflow
  a. Enable NodeOutOfServiceVolumeDetach feature gate for kube-controller-manager.
  b. Detect a node is unhealthy.
  c. Shut down the node (IPMI, cloud API, …)
  d. Apply taint to the **shutdown** node:
     - `node.kubernetes.io/out-of-service: "NoExecute"`

-> All Pods are force-deleted + all volumes are detached *immediately*.

What we did in 1.26 (cont.)

- Alpha
  - Provision volumes from Cross-namespace snapshots
    - Allows volumes to be provisioned from a snapshot in a different namespace
    - Thursday, April 20 • 17:25 - 18:00:
      Across Kubernetes Namespace Boundaries: Your Volumes Can Be Shared Now!
      Masaki Kimura & Takafumi Takahashi, Hitachi

What we did in 1.27

- Beta
  - ReadWriteOncePod PersistentVolume Access Mode
  - Selinux Relabeling with Mount Options
    - Speeds up container startup by mounting volumes with the correct SELInux label
    - Beta Blog pending
  - Robust VolumeManager Reconstruction after kubelet restart
    - Allows kubelet to populate information about how existing volumes are mounted
  - Node Expand Secret
    - Supports per-PVC secrets for volume resizing during node side filesystem expansion
    - Beta Blog pending
  - On-going effort: CSI migration
    - RBD (Beta, off-by-default)
  - Prevent Unauthorised Volume Mode Conversion
    - Prevent unauthorised volume mode conversion
  - (SIG-Apps) Auto remove PVCs created by statefulset
    - Adds an option to allow PVCs created by StatefulSet to be removed automatically
    - Beta Blog pending

ReadWriteOncePod PV Access Mode

- Beta in 1.27; Beta Blog pending
- Existing PersistentVolume Access Mode in Kubernetes
  - RWO – ReadWriteOnce
  - ROX – ReadOnlyMany
  - RWX – ReadWriteMany
- New PersistentVolume Access Mode: RWOP - ReadWriteOncePod

PersistentVolume AccessMode	Driver Supports SINGLE_NODE_*_WRITER	Driver Does Not Support SINGLE_NODE_*_WRITER
ReadWriteOncePod	SINGLE_NODE_SINGLE_WRITER	Don't use ReadWriteOncePod if driver is incapable
ReadWriteOnce	SINGLE_NODE_MULTI_WRITER	SINGLE_NODE_WRITER (Existing behavior)

Volume Mode Conversion

- Beta in 1.27; Alpha Blog
- Prevents unauthorised volume mode conversion
- **Feature flag *prevent-volume-mode-conversion* will be enabled by default in snapshot-controller and csi-provisioner in releases for 1.28**
- API Changes
  - A *SourceVolumeMode* field in *VolumeSnapshotContent*
  - An annotation *AllowVolumeModeChange* on *VolumeSnapshotContent*
- Behaviour
  - **Reject volume mode change when rehydrating a volume from snapshot unless the *AllowVolumeModeChange* annotation has been set to true.**

What we did in 1.27 (cont.)

- Alpha
  - Volume Group Snapshot
    - Introduces a VolumeGroupSnapshot API to take a crash consistent snapshot of multiple volumes.
  - Alpha Blog pending

Volume Group Snapshot

- New Kubernetes APIs
  - VolumeGroupSnapshot
  - VolumeGroupSnapshotContent
  - VolumeGroupSnapshotClass
- CSI spec changes
  - New Group Controller Service
  - New Group Controller capability CREATE_DELETE_GET_VOLUME_GROUP_SNAPSHOT
  - New gRPC interfaces for create/delete/get volume group snapshot
- New controller logic in snapshot-controller and CSI snapshotter sidecar
  - For dynamic provisioning, set a label selector in VolumeGroupSnapshot spec for PVCs with the matching labels to be snapshotted together.

CSI Migration Schedule

Core CSI Migration is GA in 1.25

Driver	Alpha	Beta (in-tree deprecated)	Beta (on-by-default)	GA	Target "in-tree plugin" removal
OpenStack Cinder	1.14	1.18	1.21	1.24	1.26
Azure Disk	1.15	1.19	1.23	1.24	1.27
Azure File	1.15	1.21	1.24	1.26	1.30 (Target)
AWS EBS	1.14	1.17	1.23	1.25	1.27
GCE PD	1.14	1.17	1.23	1.25	1.28 (Target)
vSphere *	1.18	1.19	1.25	1.26	1.30 (Target)
Ceph RBD	1.23	1.27			
CephFS	1.28 (Target)				
Portworx	1.23	1.25			

* vSphere version < 7.0u2 is no longer supported for in-tree vSphere volume in 1.25+

In-Tree Storage Driver Removal

Driver	Deprecated	Code Removal
Flocker	1.22	1.25
GlusterFS	1.25	1.26
Quobyte	1.22	1.25
ScaleIO	1.16	1.22
StorageOS	1.22	1.25

Features in Design/Prototyping

- Changed block tracking
- Kubernetes Volume Provisioned IO
- Runtime Assisted Mounting
- (SIG-Apps) Volume Expansion for StatefulSets

How to get involved

How to Get Involved

KubeCon CloudNativeCon Europe 2023

- Start at the SIG Storage page:
  - https://github.com/kubernetes/community/tree/master/sig-storage
- Attend our meetings:
  - Bi-weekly meeting: 9am PT every second Thursday.
  - Issue triage: 10am PT every Wednesday.
- Mailing List:
  - kubernetes-sig-storage@googlegroups.com
- Slack channel:
  - #sig-storage
  - #csi
  - #sig-storage-cosi

***SIG Meet & Greet*: Friday, April 21 • 12:30 - 14:30**

Resources

- SIG Storage page
- Storage concepts
- CSI driver docs
- CSI spec
- CSI sample driver hostpath deployment example
- SIG Storage annual report


ubeCon
CloudNativeCon
Europe 2023
Please scan the QR Code above to
leave feedback on this session


KubeCon
CloudNativeCon
Europe 2023
Thank you