



KubeCon



CloudNativeCon

Europe 2019

Improving Availability for Stateful Applications

Michelle Au

Software Engineer, Google

Agenda



KubeCon



CloudNativeCon

Europe 2019

Persistent storage options

Building highly available stateful applications

- Failure domain spreading
- Demo
- Pod downtime and recovery



KubeCon



CloudNativeCon

Europe 2019

Persistent Storage Options

Supported Storage Systems



KubeCon



CloudNativeCon

Europe 2019

In-tree Drivers

- <https://kubernetes.io/docs/concepts/storage/#types-of-volumes>
- Over 15!

CSI Drivers

- <https://kubernetes-csi.github.io/docs/drivers.html>
- Over 35!

Wide range of characteristics

- Local vs remote, cloud vs appliance vs software-defined, distributed vs hyper-converged, etc.

Storage Characteristics



KubeCon



CloudNativeCon

Europe 2019

Accessibility

- At what granularity does your app have to be co-located with storage?

Availability

- At what granularity is storage still available during an outage?

Durability

- Under what conditions could my data be lost?

Access Mode

- How many nodes can access the volume concurrently?

Storage Characteristics



KubeCon



CloudNativeCon

Europe 2019

Performance

- Read/write/mixed IOPS and throughput

Cost

- Including operation, maintenance

Examples



KubeCon



CloudNativeCon

Europe 2019

Example	Accessibility	Availability	Durability	Access Mode	Performance	Cost
Local disk	Single node	Single node	Single disk*	Single node	Best	\$

* Most cloud local disks are not durable beyond VM

Examples



KubeCon



CloudNativeCon

Europe 2019

Example	Accessibility	Availability	Durability	Access Mode	Performance	Cost
Local disk	Single node	Single node	Single disk*	Single node	Best	\$
Cloud disk	Single zone	Single zone	3x	Single node	Better	\$\$

* Most cloud local disks are not durable beyond VM

Examples



KubeCon



CloudNativeCon

Europe 2019

Example	Accessibility	Availability	Durability	Access Mode	Performance	Cost
Local disk	Single node	Single node	Single disk*	Single node	Best	\$
Cloud disk	Single zone	Single zone	3x	Single node	Better	\$\$
Replicated cloud disk	Multi zone	Multi zone	3x	Single node	Good	\$\$\$

* Most cloud local disks are not durable beyond VM

Examples



KubeCon



CloudNativeCon

Europe 2019

Example	Accessibility	Availability	Durability	Access Mode	Performance	Cost
Local disk	Single node	Single node	Single disk*	Single node	Best	\$
Cloud disk	Single zone	Single zone	3x	Single node	Better	\$\$
Replicated cloud disk	Multi zone	Multi zone	3x	Single node	Good	\$\$\$
Single NFS	Global	Single server	Varies	Multi node	Good	\$\$\$

* Most cloud local disks are not durable beyond VM

Examples



KubeCon



CloudNativeCon

Europe 2019

Example	Accessibility	Availability	Durability	Access Mode	Performance	Cost
Local disk	Single node	Single node	Single disk*	Single node	Best	\$
Cloud disk	Single zone	Single zone	3x	Single node	Better	\$\$
Replicated cloud disk	Multi zone	Multi zone	3x	Single node	Good	\$\$\$
Single NFS	Global	Single server	Varies	Multi node	Good	\$\$\$
Scaleout/HA Filer	Global	Global	Varies	Multi node	Varies	\$\$\$\$

* Most cloud local disks are not durable beyond VM



KubeCon



CloudNativeCon

Europe 2019

Building Highly-Available Stateful Applications

Pod Anti-Affinity



KubeCon



CloudNativeCon

Europe 2019

Spread replicas across failure domains

```
affinity:
```

```
  podAntiAffinity:
```

```
    requiredDuringSchedulingIgnoredDuringExecution:
```

```
    - topologyKey: failure-domain.beta.kubernetes.io/zone
```

```
      labelSelector:
```

```
        matchExpressions:
```

```
        - key: app
```

```
          operator: In
```

```
          values:
```

```
          - my-app
```

Model 1: App has no redundancy



KubeCon



CloudNativeCon

Europe 2019

Need high availability at storage layer

- Multi-writer
- Globally accessible and available
- Example: Scaleout/HA filer

Model 1: Deployment

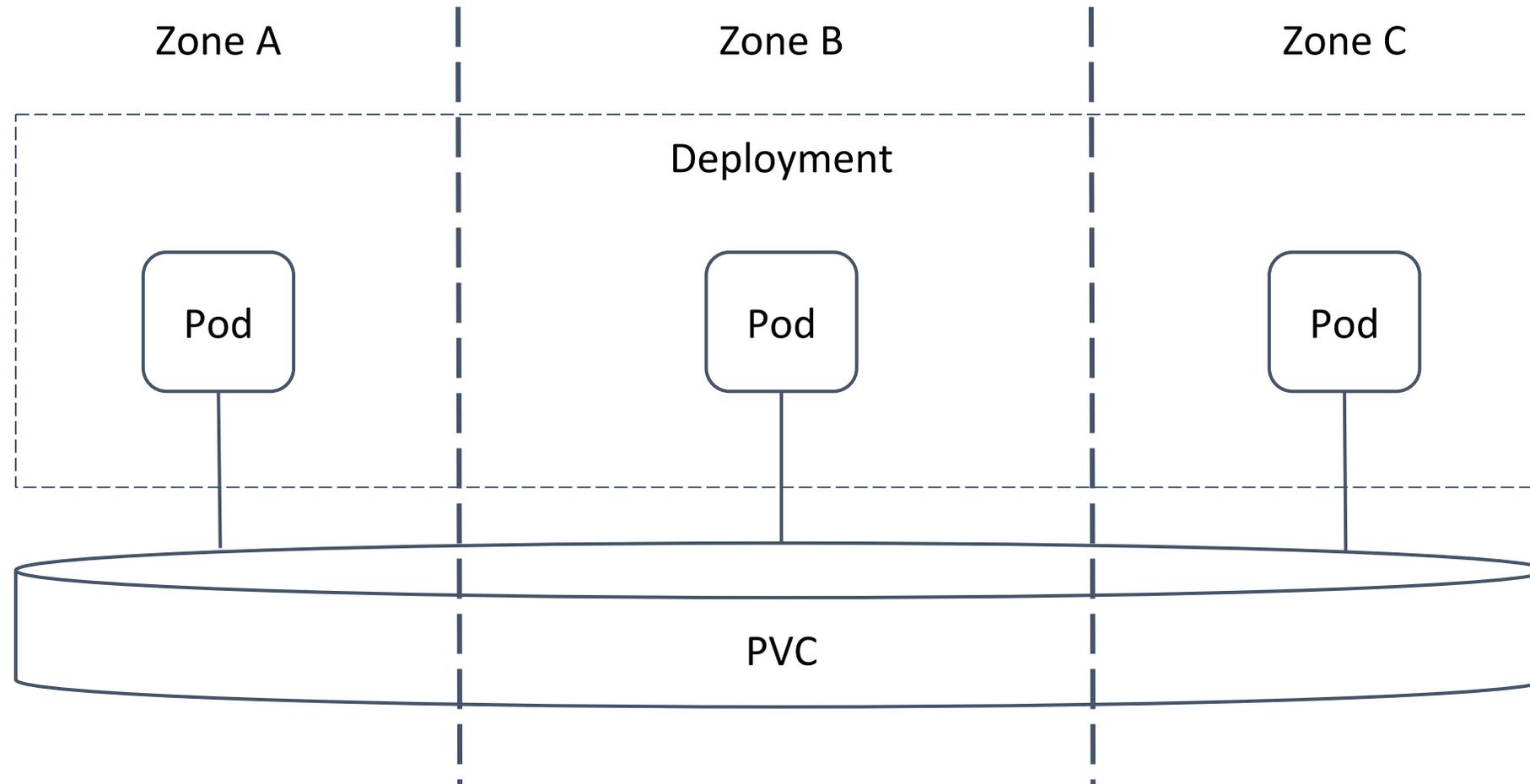


KubeCon



CloudNativeCon

Europe 2019



Model 1: Deployment

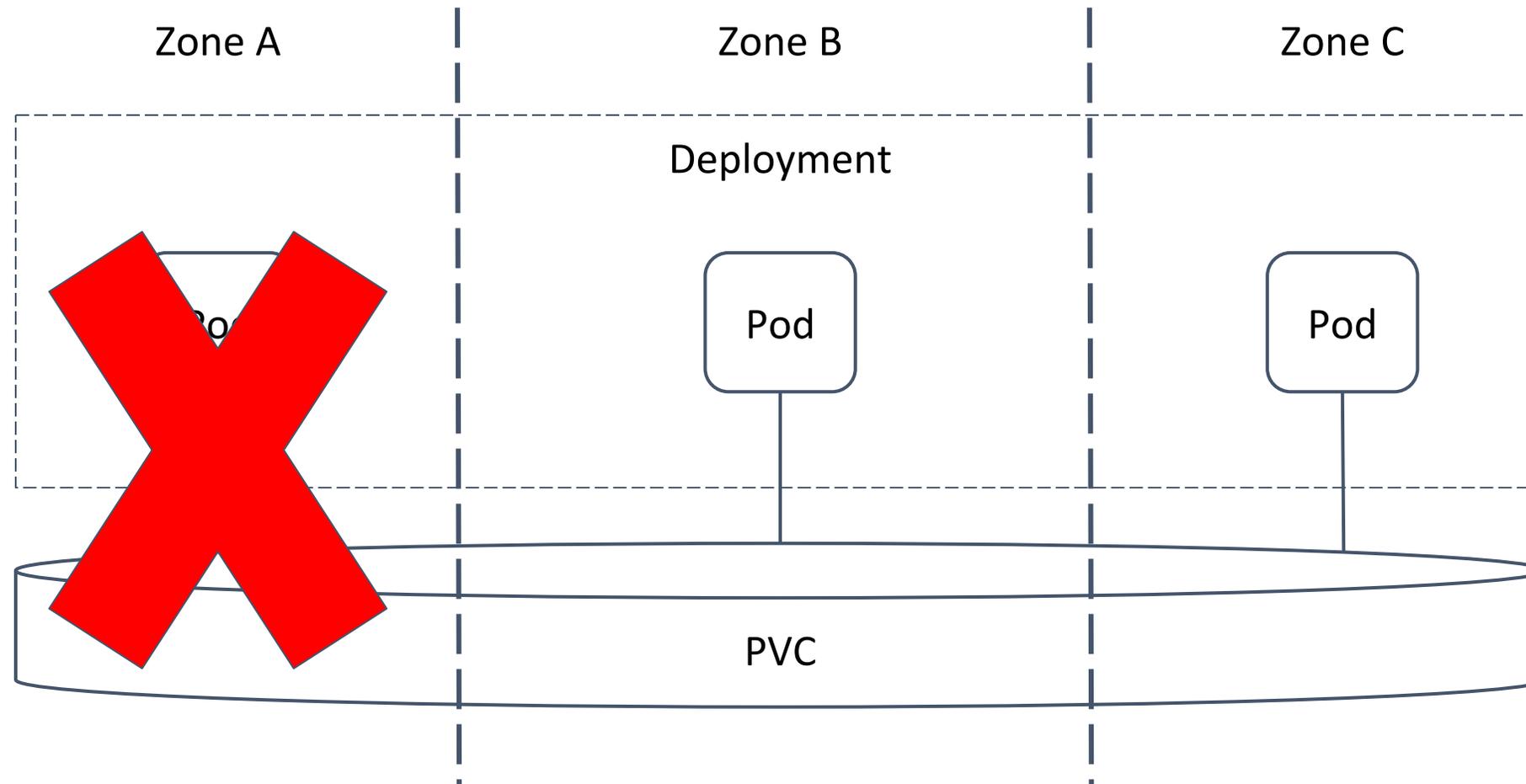


KubeCon



CloudNativeCon

Europe 2019



Model 2: App has redundancy



KubeCon



CloudNativeCon

Europe 2019

Distributed apps: shard and replicate data between pods

- Example: Cassandra, MongoDB

Do not need high-availability at storage layer

- Single writer
- Non-global accessibility and availability
- Example: Local disks, cloud disks

Model 2: StatefulSet

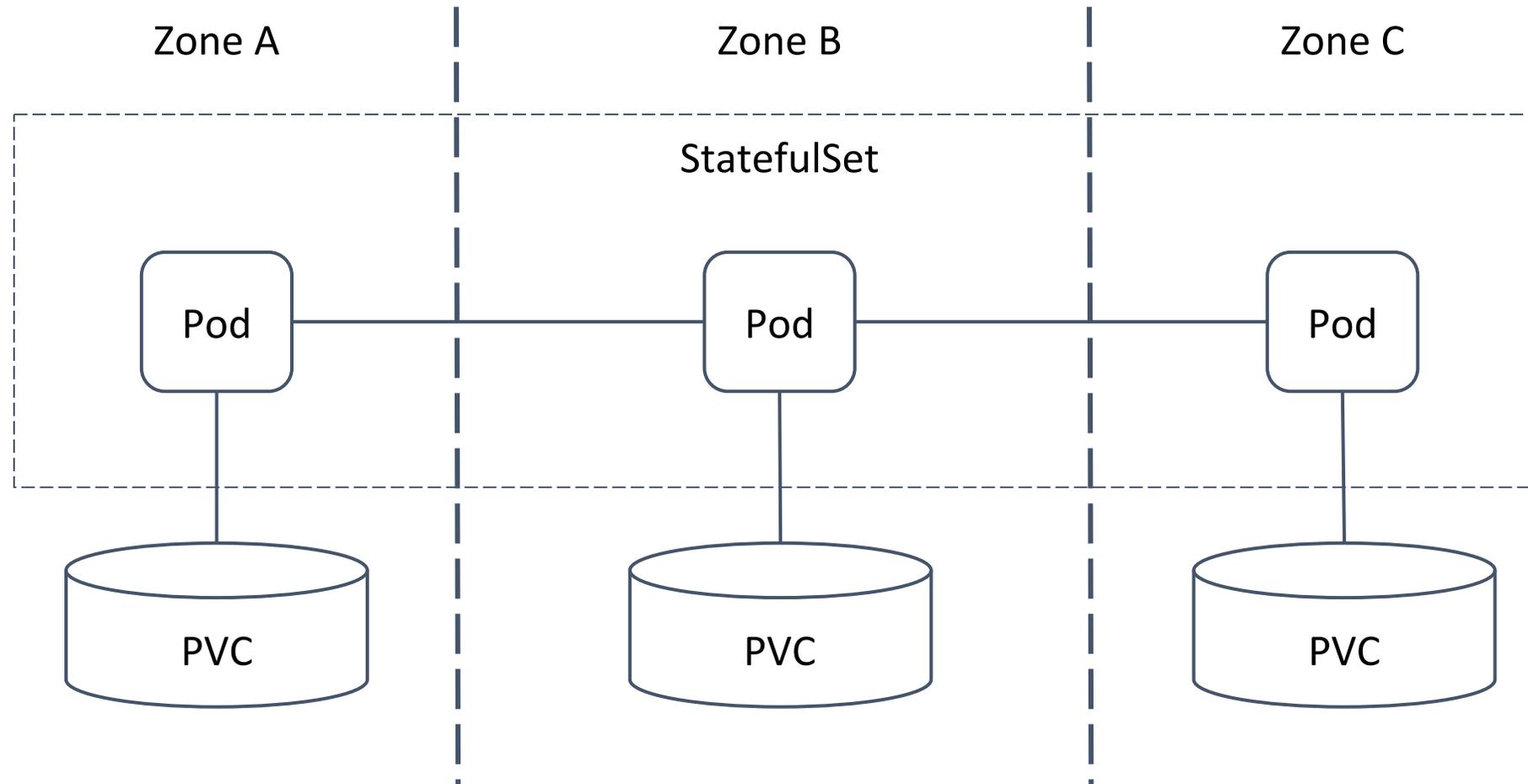


KubeCon



CloudNativeCon

Europe 2019



Model 2: StatefulSet

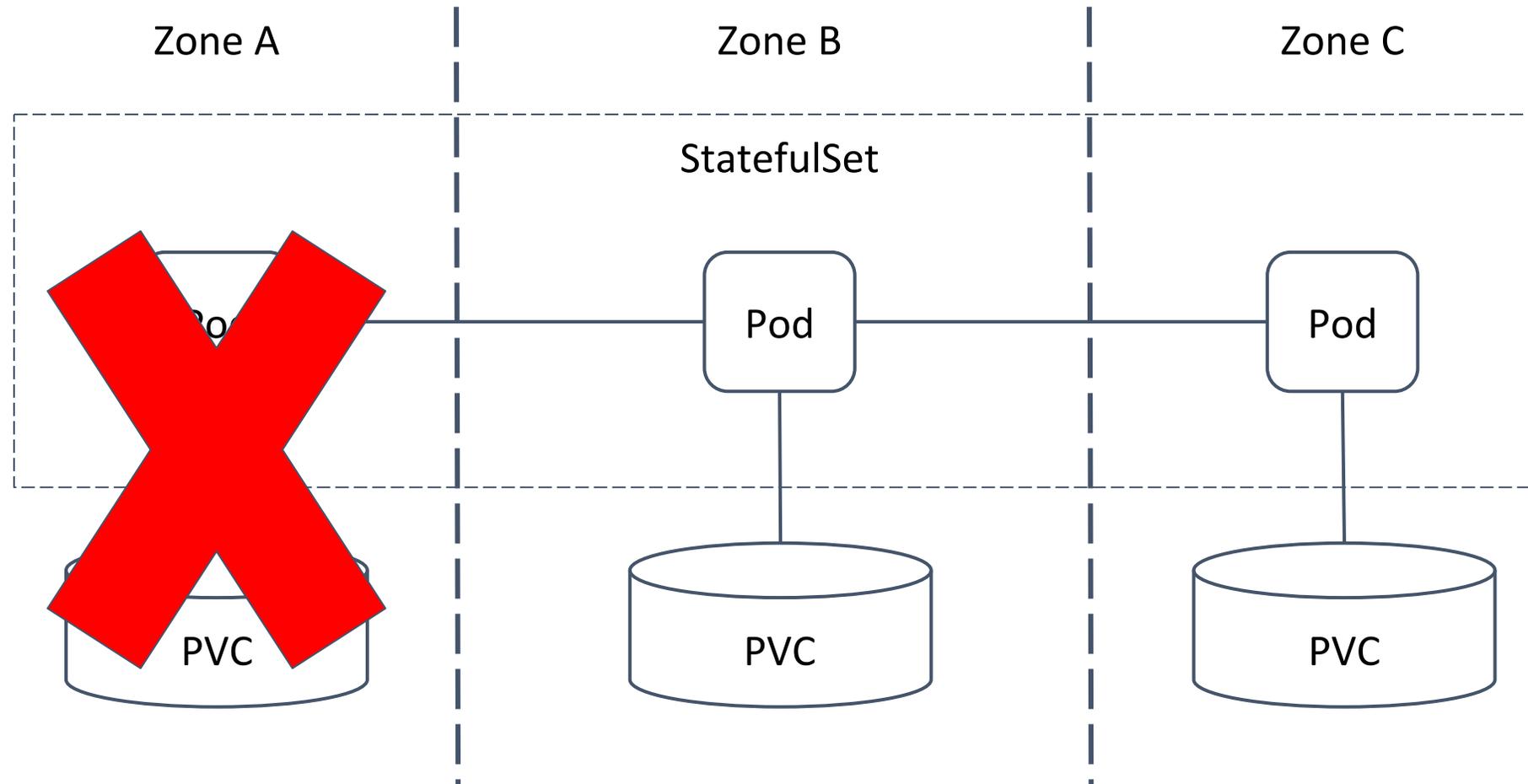


KubeCon



CloudNativeCon

Europe 2019



Volume Topology



KubeCon



CloudNativeCon

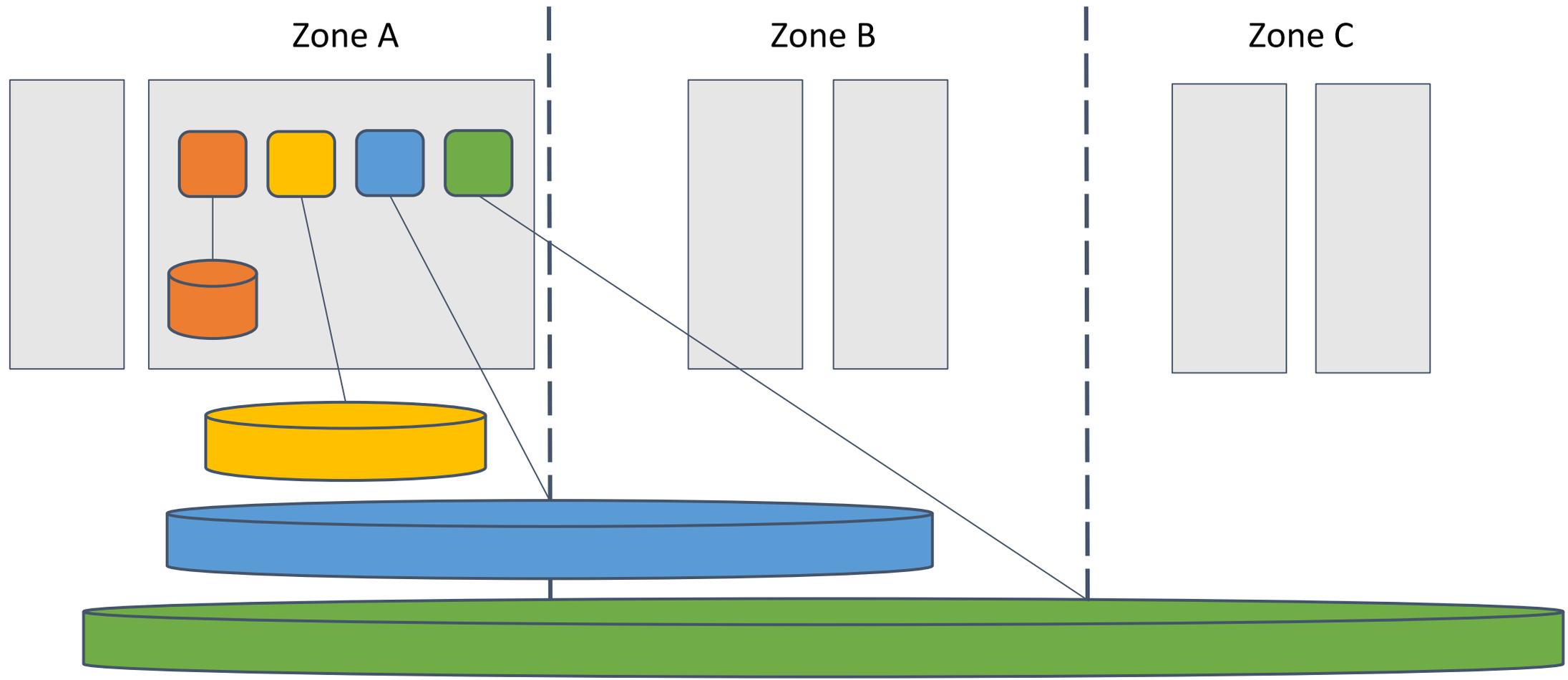
Europe 2019

Scheduler understands volume accessibility constraints

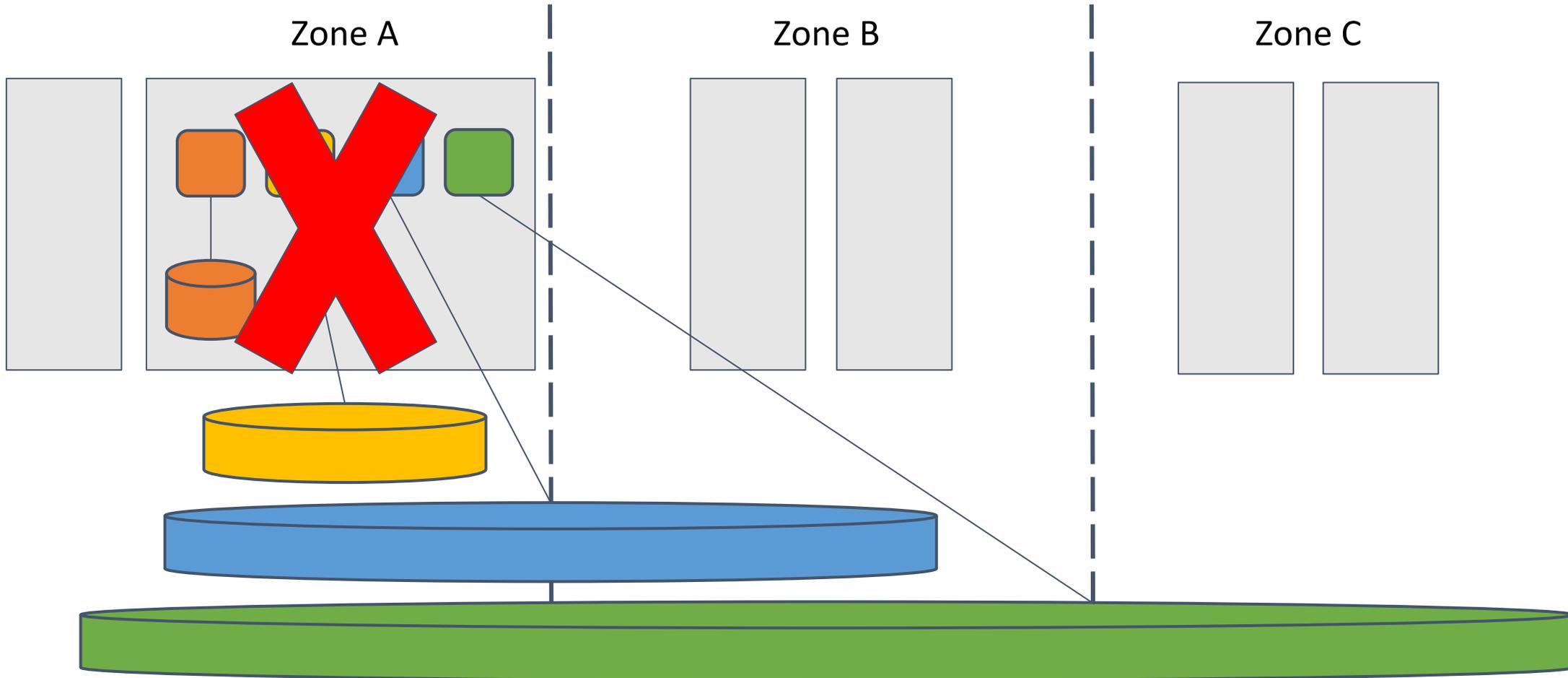
- No user configuration needed
- Storage driver provides topology

Auto-scale replicas and dynamically provision volumes across zones (except local)

Demo



Demo



Demo

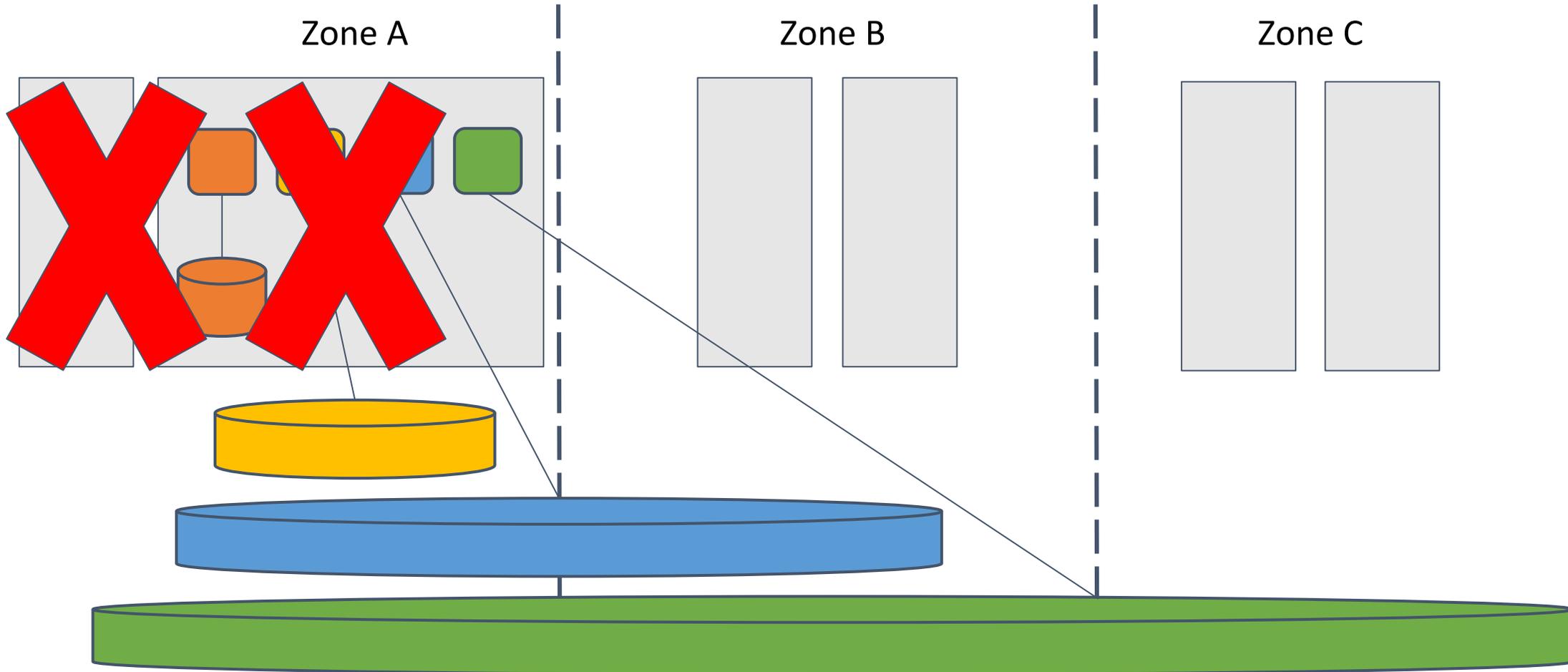


KubeCon



CloudNativeCon

Europe 2019



Downtime



KubeCon



CloudNativeCon

Europe 2019

Time to detect failure
+
Time to replace pod

StatefulSet Caveat



KubeCon



CloudNativeCon

Europe 2019

Stateful applications may require exactly-once semantics

- Two containers cannot write to the same volume

During split brain, replacement Pod cannot be started

- Node fencing can help

StatefulSet pod recovery can be long

- Minutes: automated
- Hours: manual

Summary



KubeCon



CloudNativeCon

Europe 2019

Kubernetes features for high-availability

- Volume topology, pod anti-affinity, node taints

Stateful application models

- Deployment vs Statefulset
- Storage redundancy vs application redundancy

Design for redundancy and account for downtime

Additional Resources



KubeCon



CloudNativeCon

Europe 2019

[Deployments](#) and [StatefulSets](#)

[Pod anti-affinity](#)

[Volume topology blog post](#)

[Node taints and tolerations](#)

[Node fencing discussions](#)

Get Involved



KubeCon



CloudNativeCon

Europe 2019

Kubernetes Special Interest Groups (SIGs)

- [sig-storage](#), [sig-apps](#), [sig-node](#), [sig-scheduling](#)
- Community meetings, slack

Me

- Github/Slack: msau42
- Twitter: [_msau42_](#)

Questions?



KubeCon



CloudNativeCon

Europe 2019



KubeCon



CloudNativeCon

Europe 2019