Speed Racer

Local Persistent Volumes in Production

KubeCon EU, August 2020
Matt Schallert (Chronosphere)

Background

- Infrastructure engineer @ Chronosphere
  - Hosted metrics + monitoring
  - Large scale, high throughput
  - Built on M3
- Previously SRE @ Uber
  - In-house metrics team

chronosphere

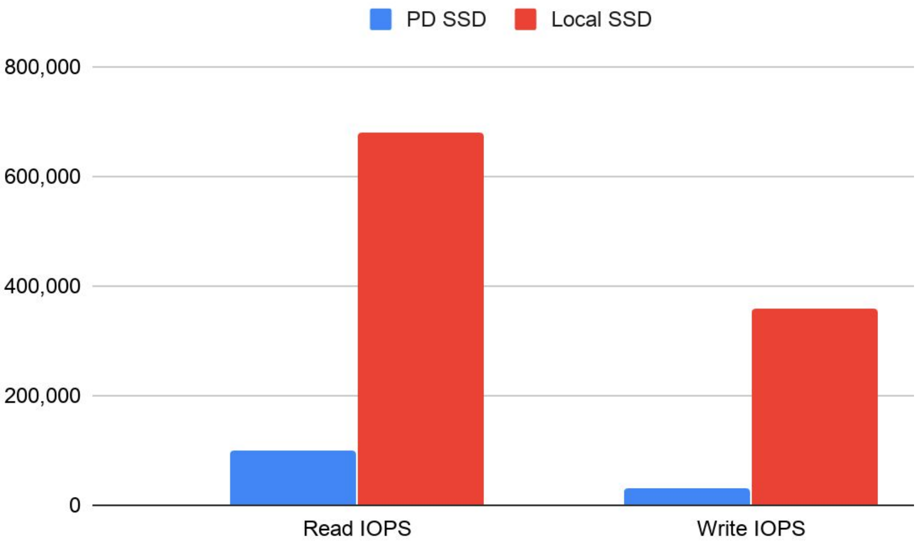
Local Persistent Volumes

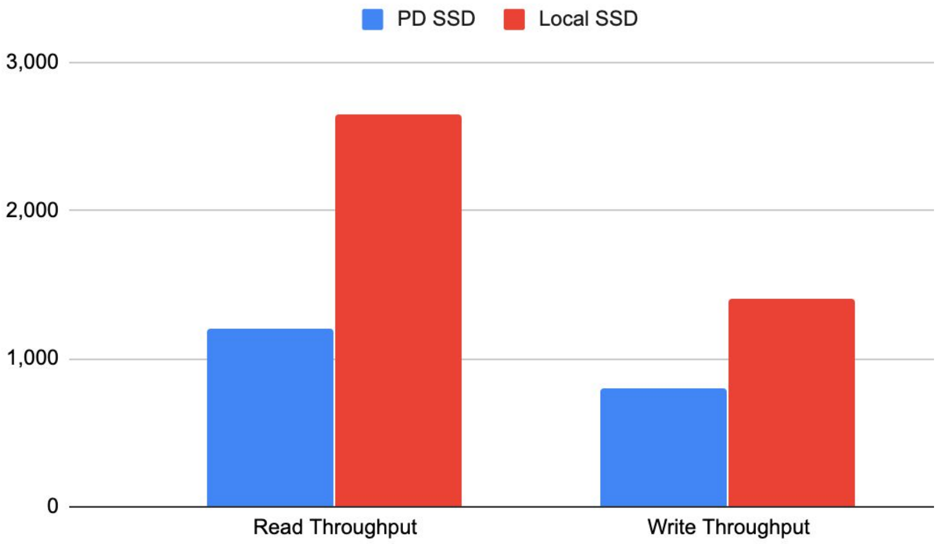
Local Persistent Volumes

Local Disks

- Offered in some form on most cloud providers
- Physical disks attached to host machine
  - Data persists for lifetime of instance
- Better performance, reduced cost
- Narrower use cases

Local Disk Performance

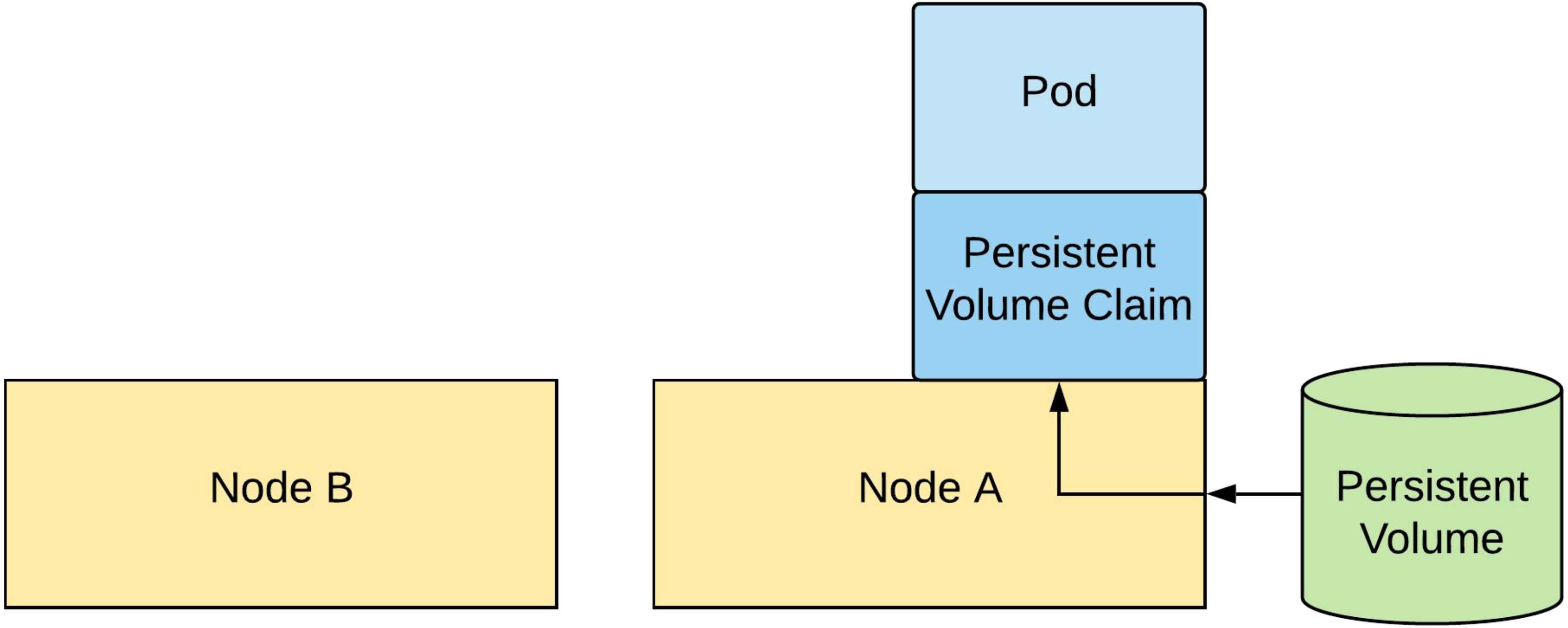

Read IOPS and Write IOPS


Read Throughput and Write Throughput (MB/s)

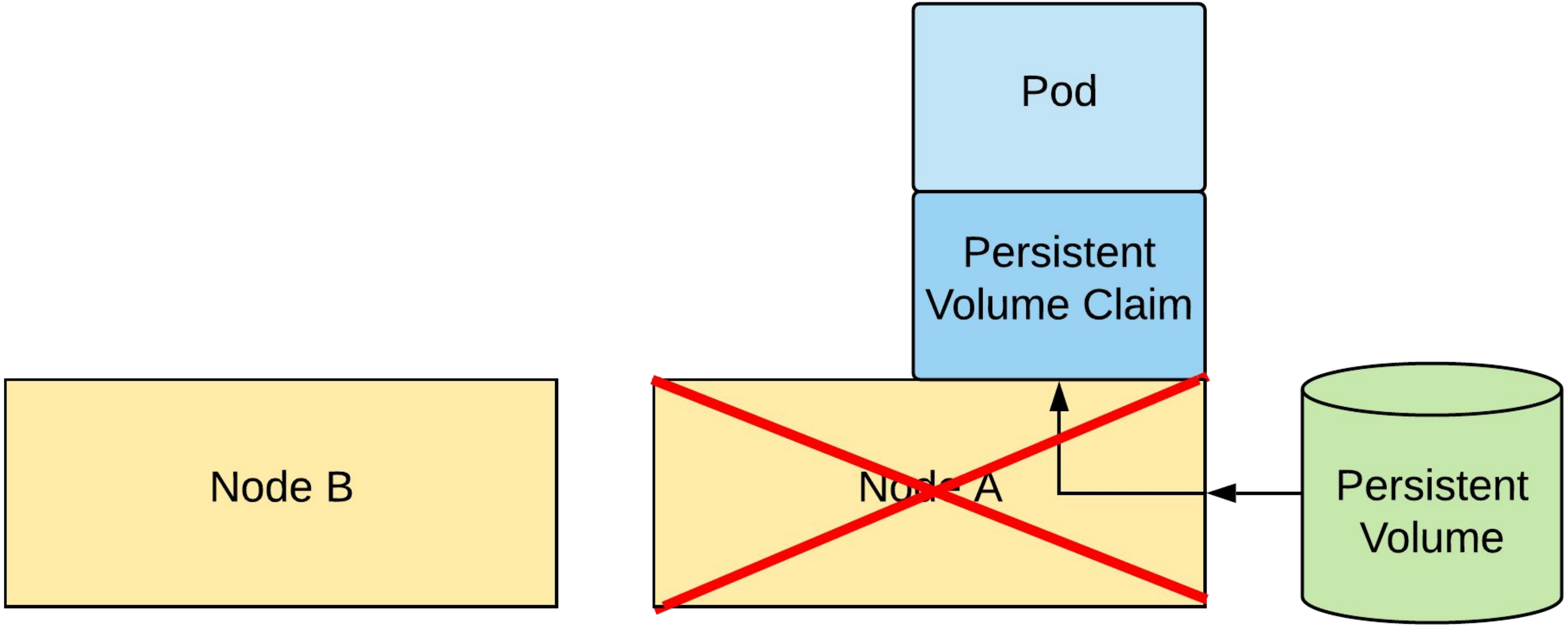
PD SSD	Local SSD
$0.17 / GB	$0.08 / GB

Local Persistent Volumes

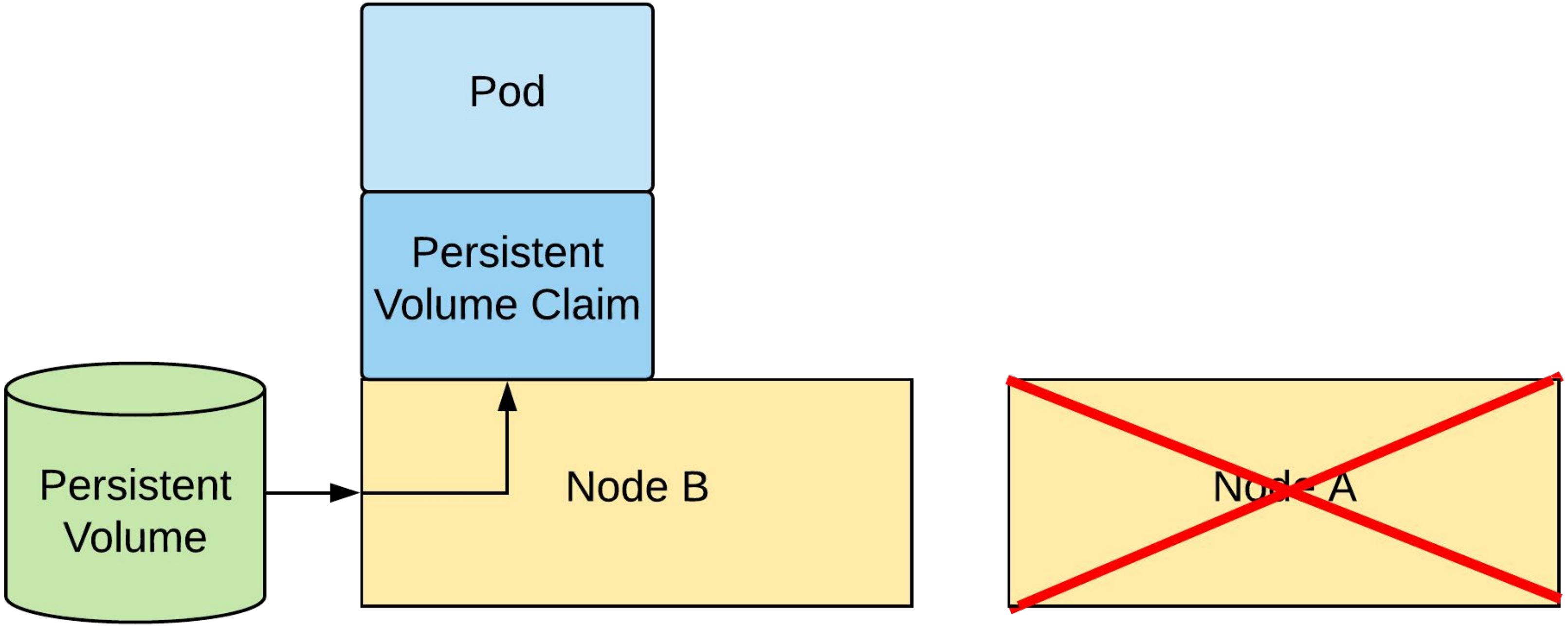
Persistent Volumes



Persistent Volumes

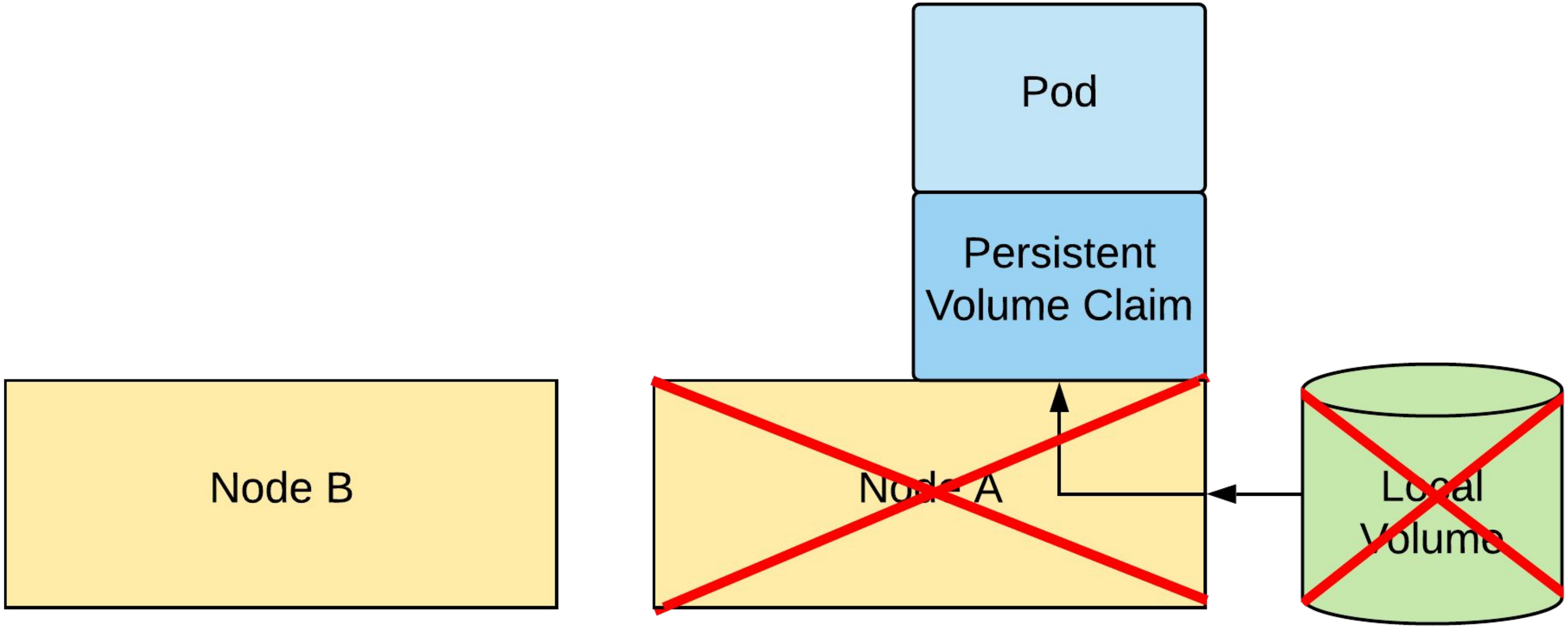


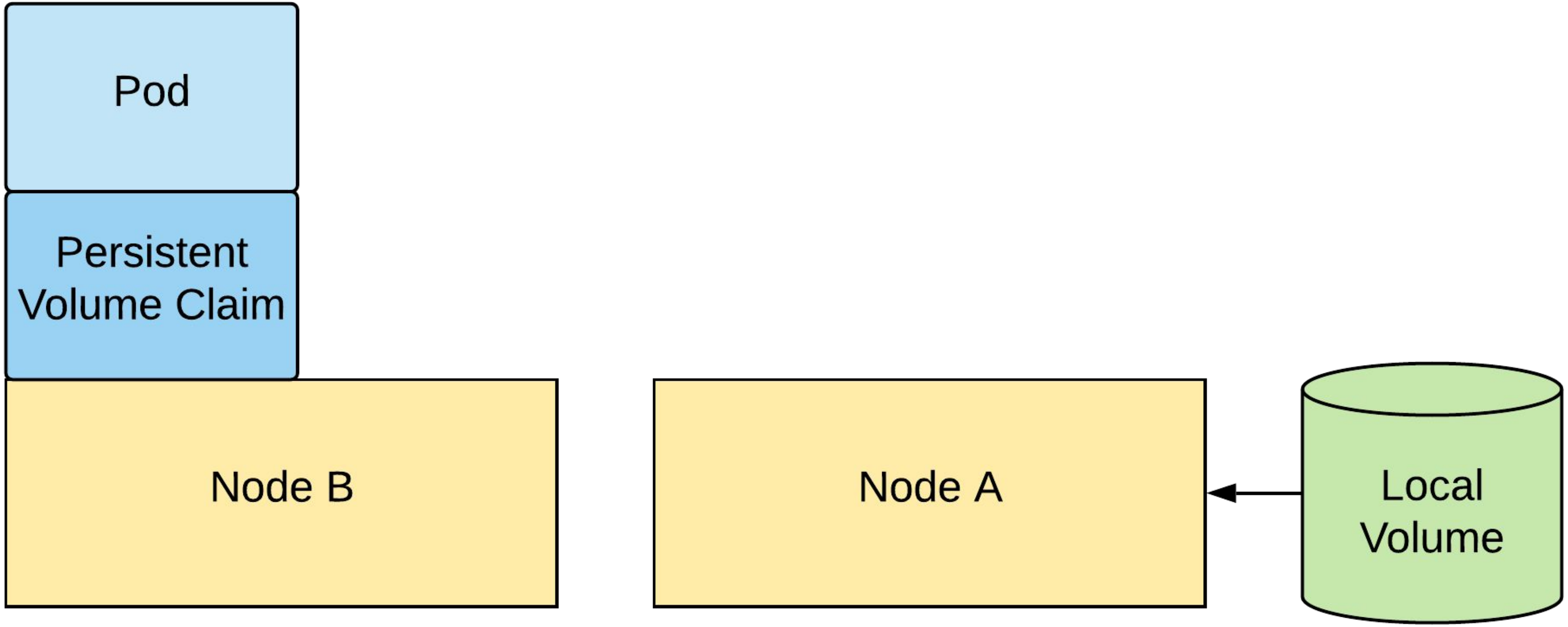
Persistent Volumes



Limitations of Persistent Volumes

- Assumption that storage could move with a pod from node to node
- Local SSDs break this assumption


Pod
Persistent Volume Claim
Node B
Node A
Local Volume


Pod
Persistent Volume Claim
Node B
Node A
Local Volume

How Local Volumes Work

- More general solution: “topology-aware volume provisioning”
  - Wait for pod to be scheduled before creating PVC + PV
  - Helps with multi-zone remote storage as well
- PVs created with “nodeAffinity”

Day One Operations

- sig-storage-local-static-provisioner
- Mount disks, point provisioner at path
- PVs created in cluster, provisioner handles lifecycle
  - Wiping disks before mounting, after delete, etc.

```yaml
apiVersion: storage.k8s.io/v1
kind: StorageClass
metadata:
  name: local
provisioner: kubernetes.io/no-provisioner
volumeBindingMode: WaitForFirstConsumer
```

```yaml
apiVersion: apps/v1
kind: StatefulSet
...
volumeClaimTemplates:
- metadata:
    name: data-volume
  spec:
    storageClassName: local
```

And That’s It!

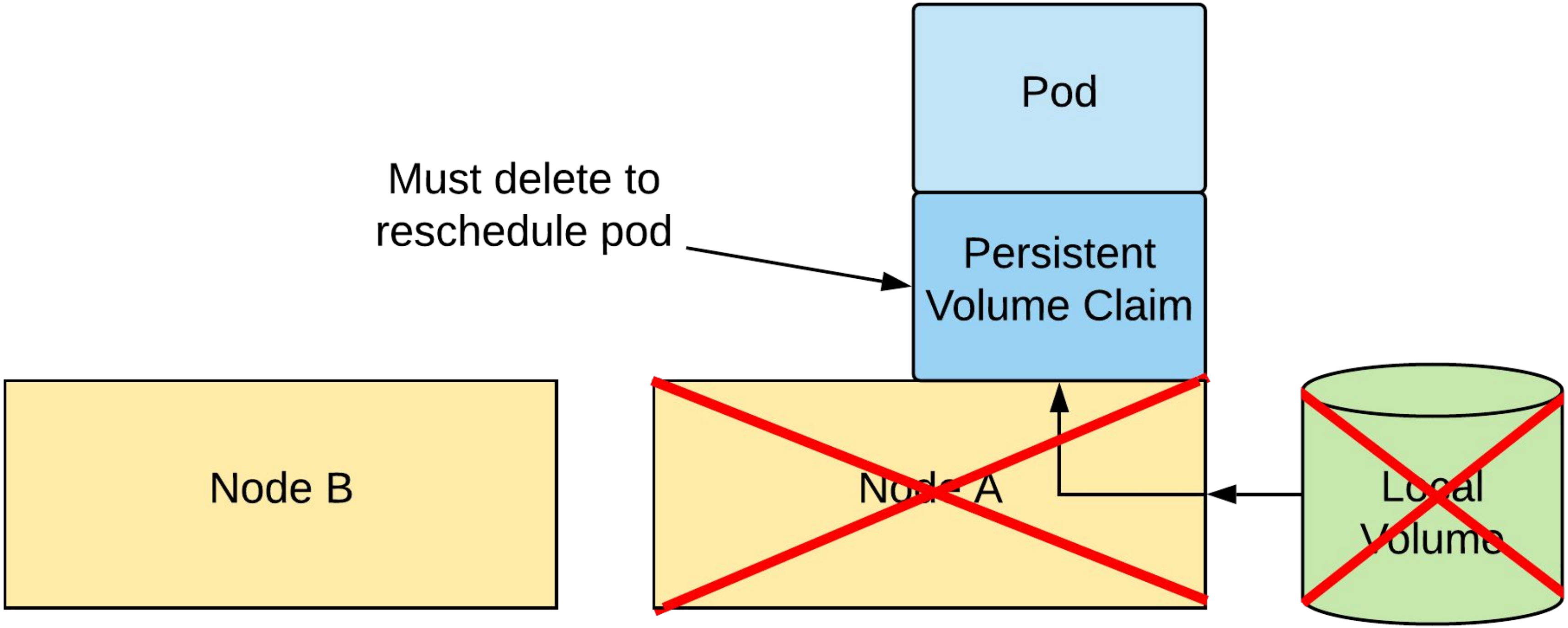
~~And That’s It!~~

Just Kidding

Day Two Operations

- Local Volumes != remote PVs!
- Different behavior in day-to-day operations
  - Node Failures
  - Backup & Restore
  - Cluster Operations
- Disk prep

Node Failures



@mattschallert

Node Failures

- PV remains attached to pod unless binding is explicitly broken
- If PV no longer exists (node failure), must delete PVC and Pod
- Pod rescheduled with new (empty) Local Volume
- Operators can automate

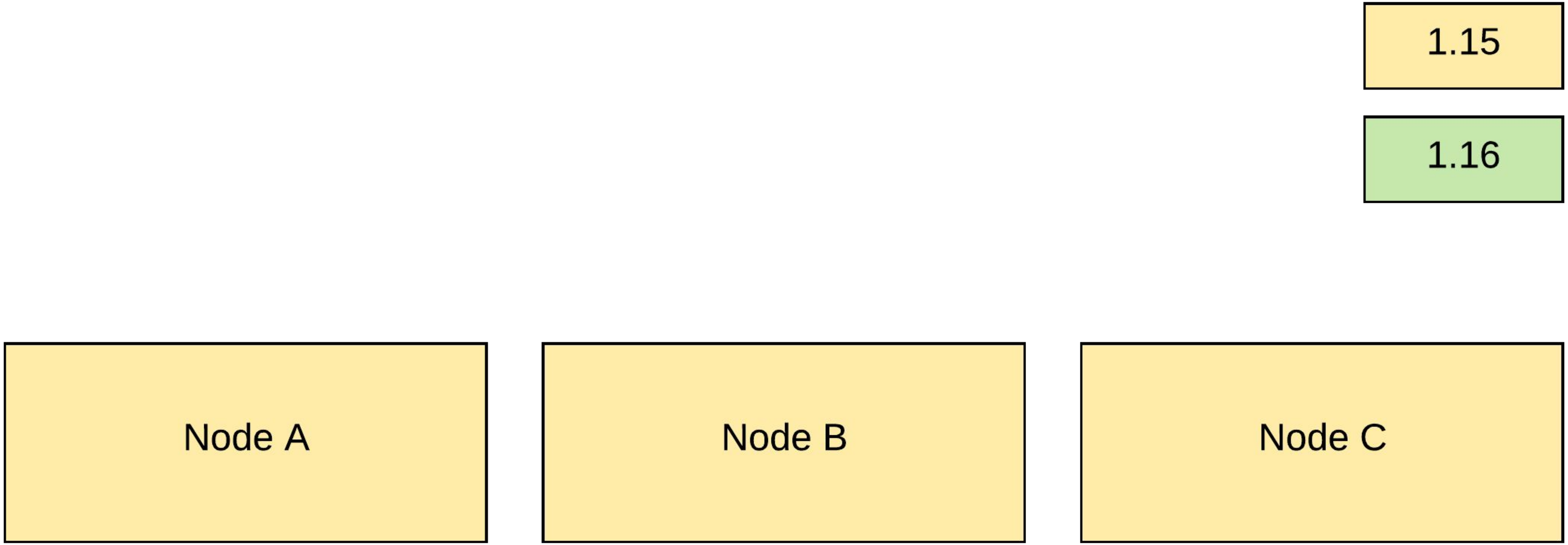
Backup & Restore

- Remote disks: snapshot + restore
- No snapshot support for local disks
- Minimal to no guarantees of local disk availability
- Sidecar pattern
  - Copy files from disk to object store

Cluster Operations

- Upgrades swap out nodes in a node pool or replace a node pool
  - Loss of all local disks!
  - Pods will be stuck (can’t bind to old volume)
- Preemptively evacuate node pools

Node Upgrades



1.15

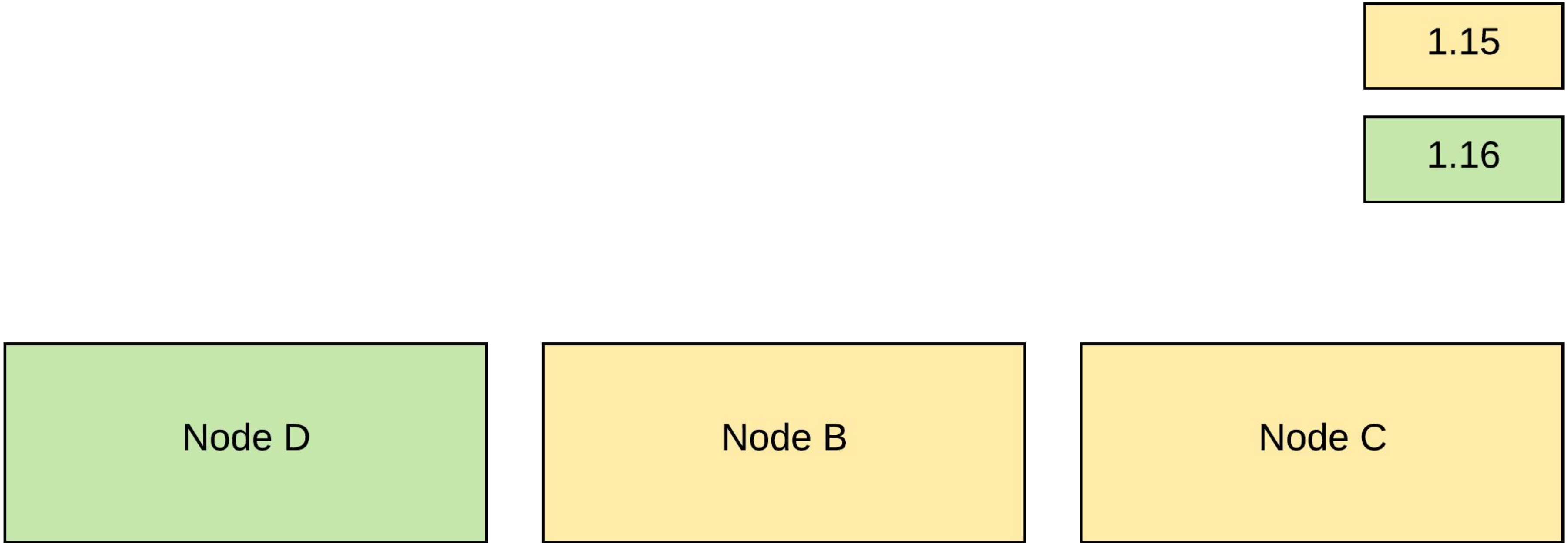
1.16

Node A

Node B

Node C

Node Upgrades



1.15

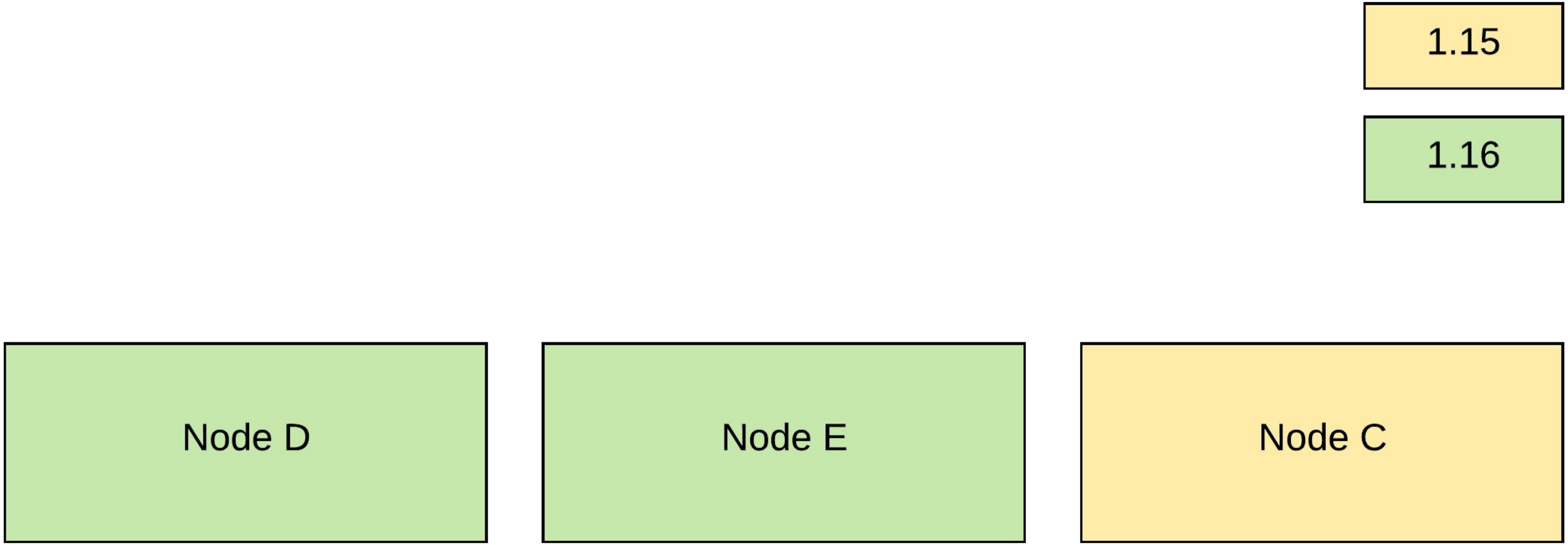
1.16

Node D

Node B

Node C

Node Upgrades



1.15

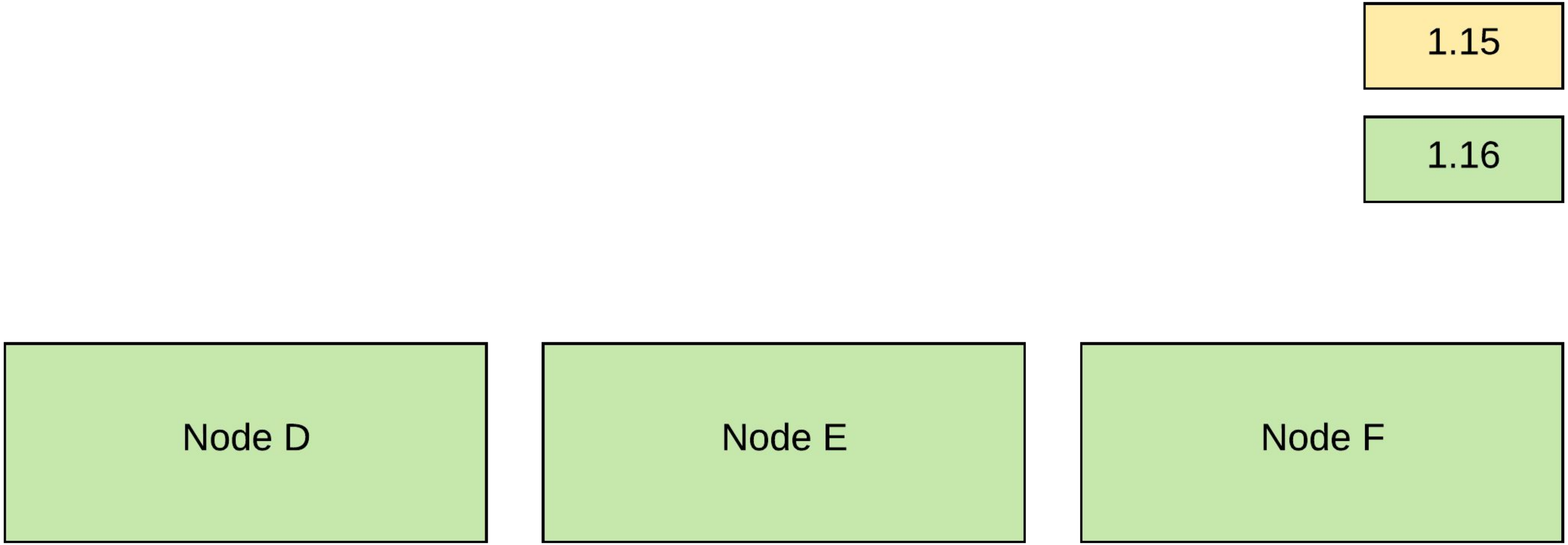
1.16

Node D

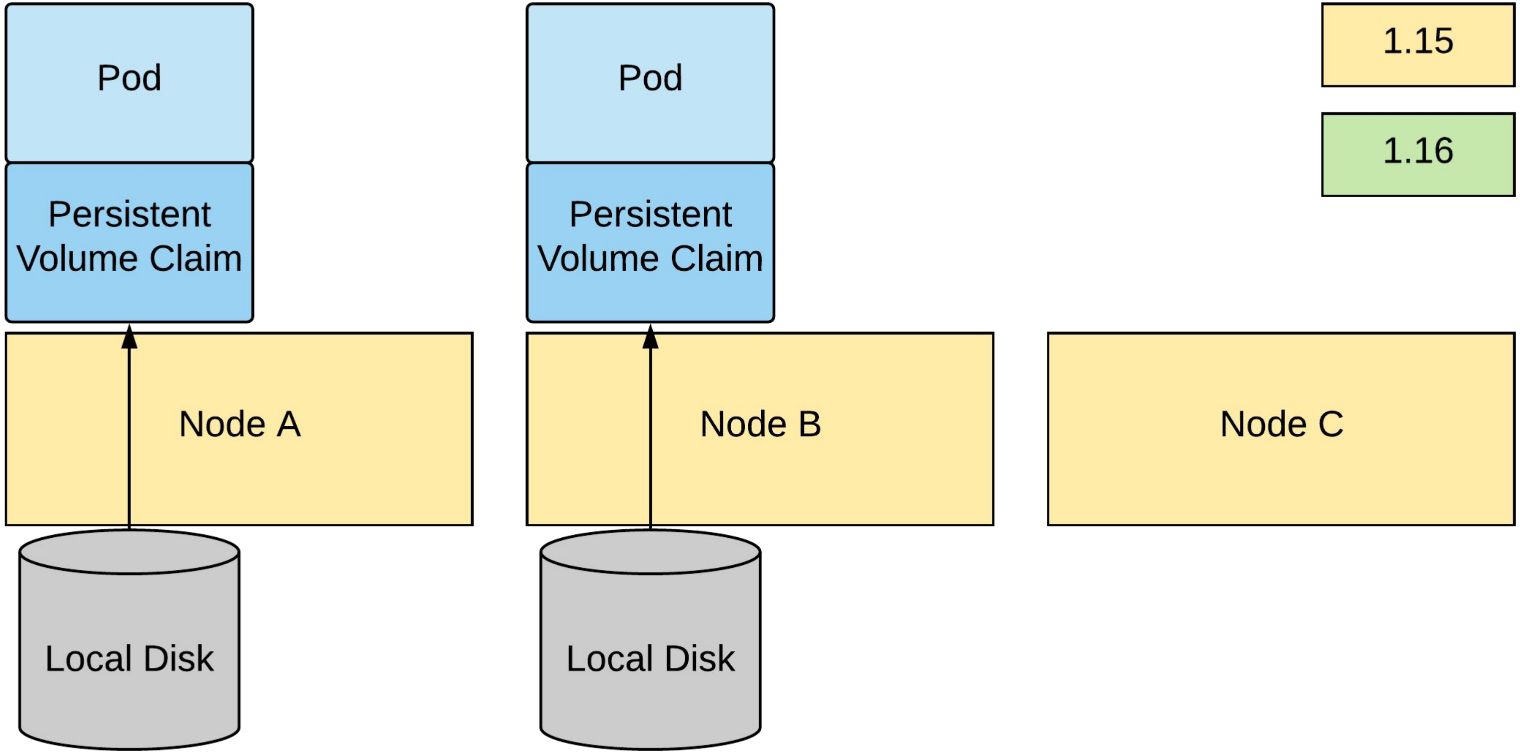
Node E

Node C

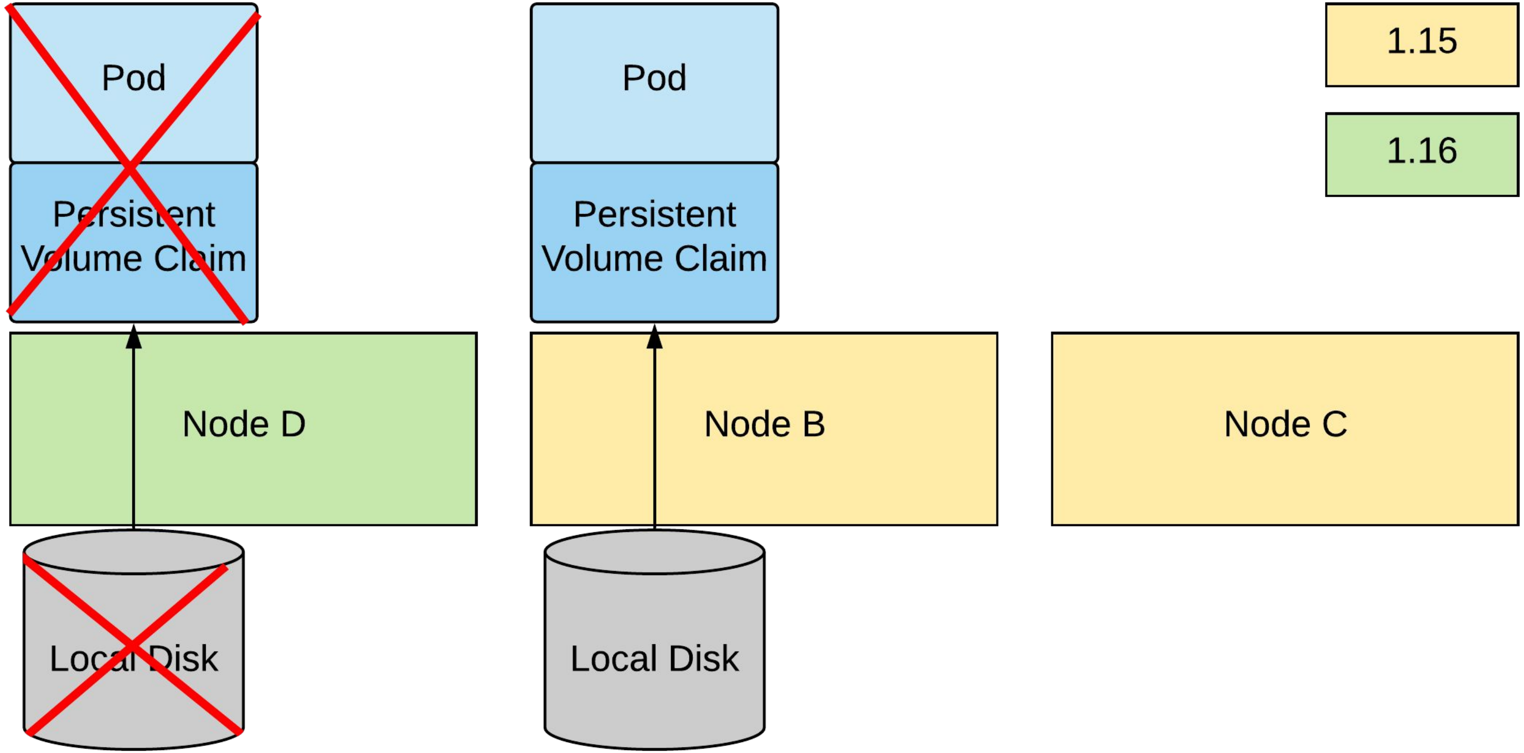
Node Upgrades



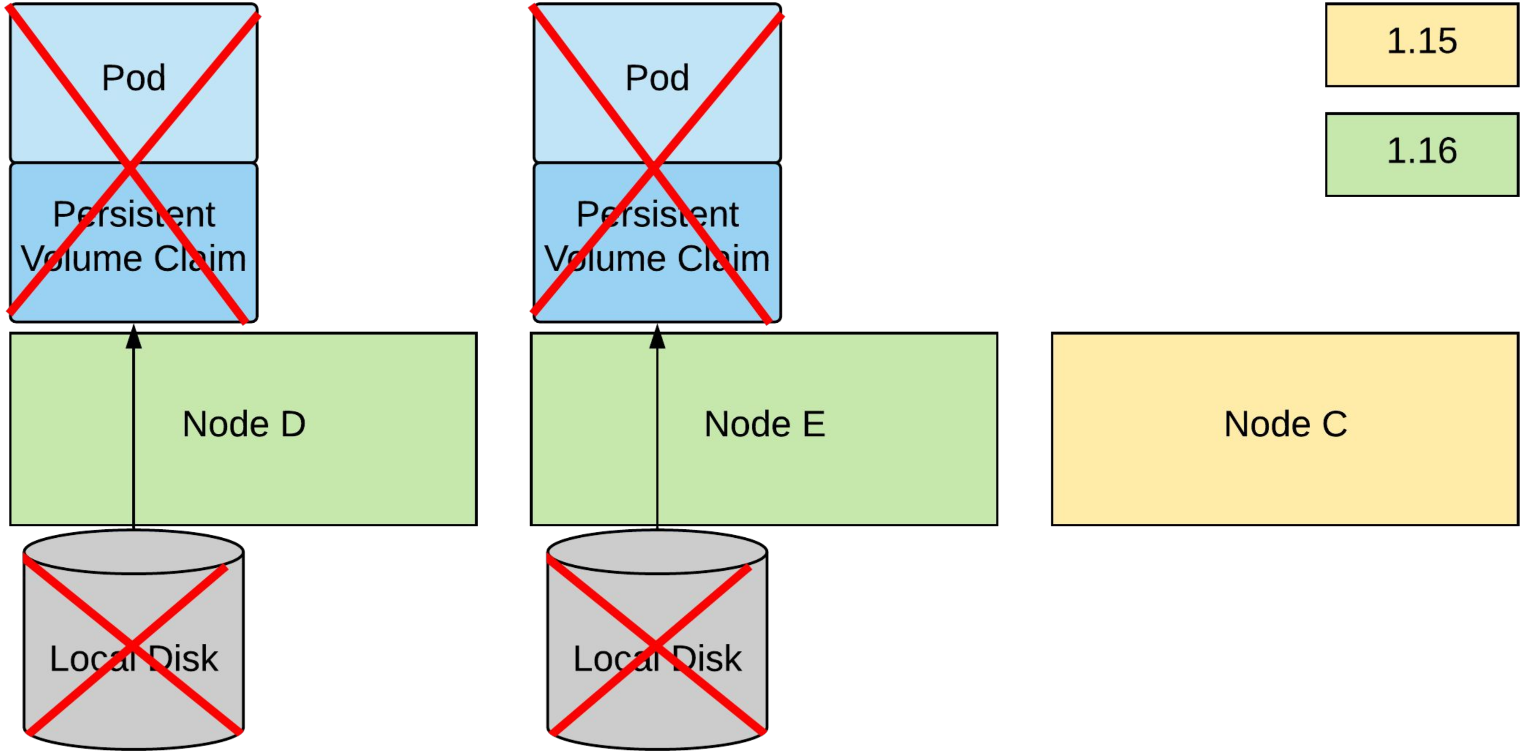
Node Upgrades: Local Volumes



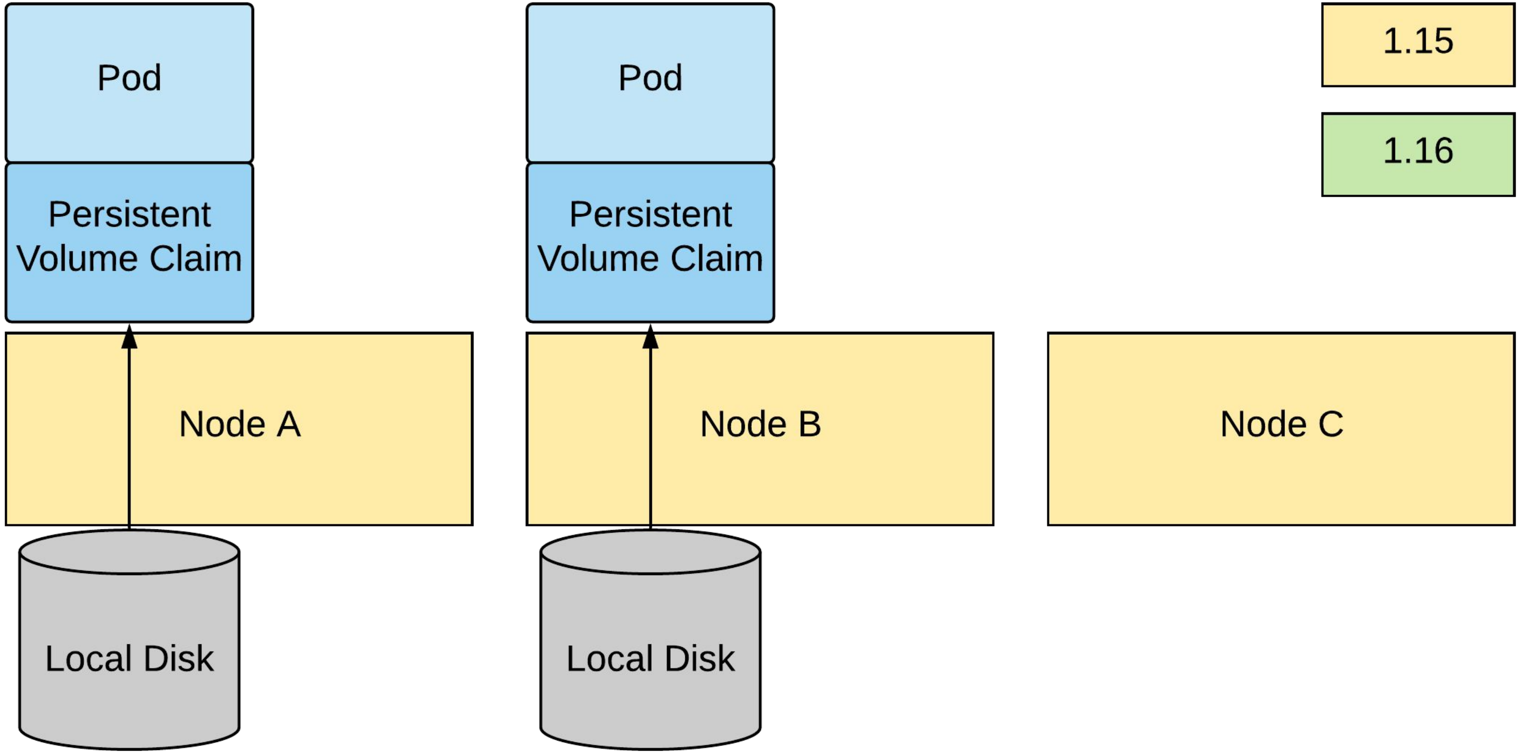
Node Upgrades: Local Volumes



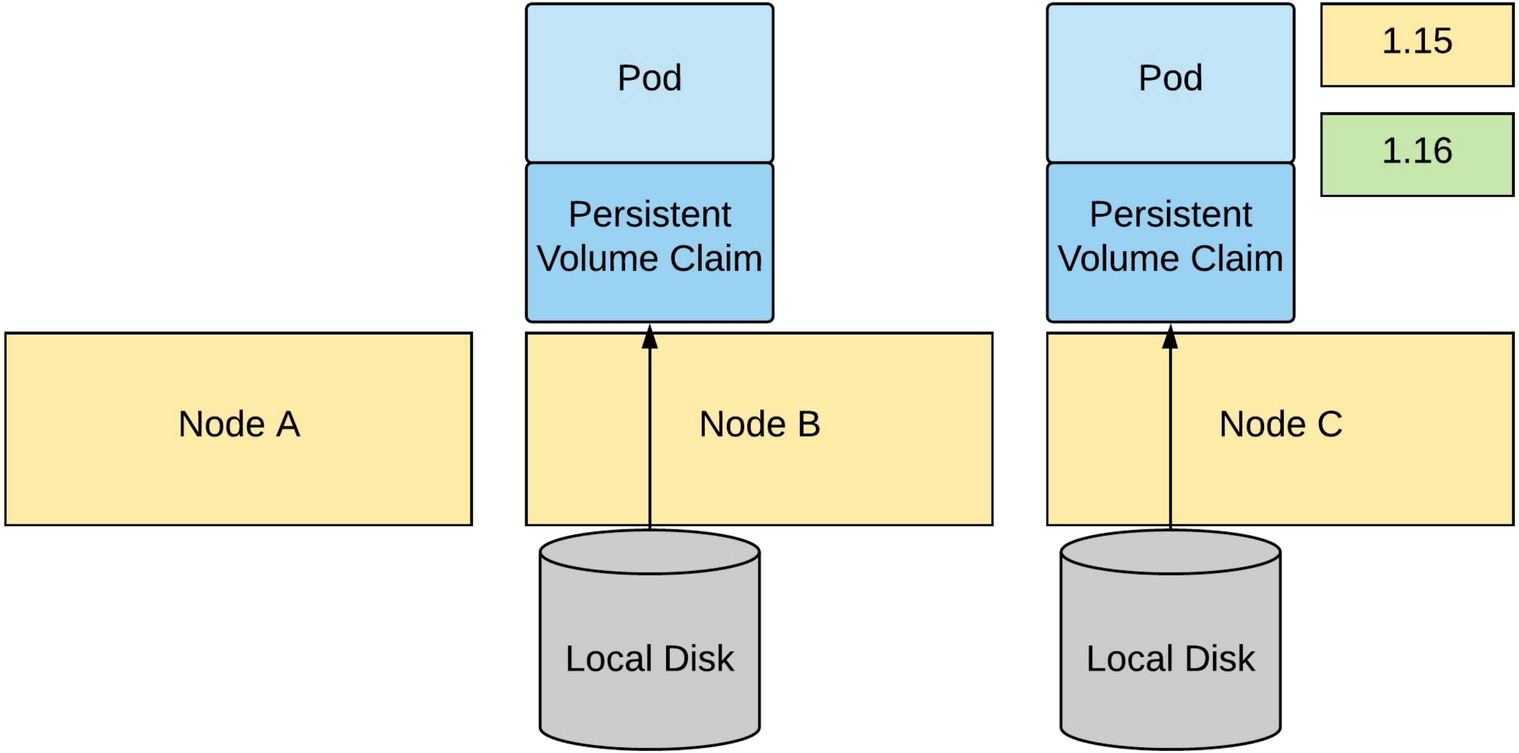
Node Upgrades: Local Volumes



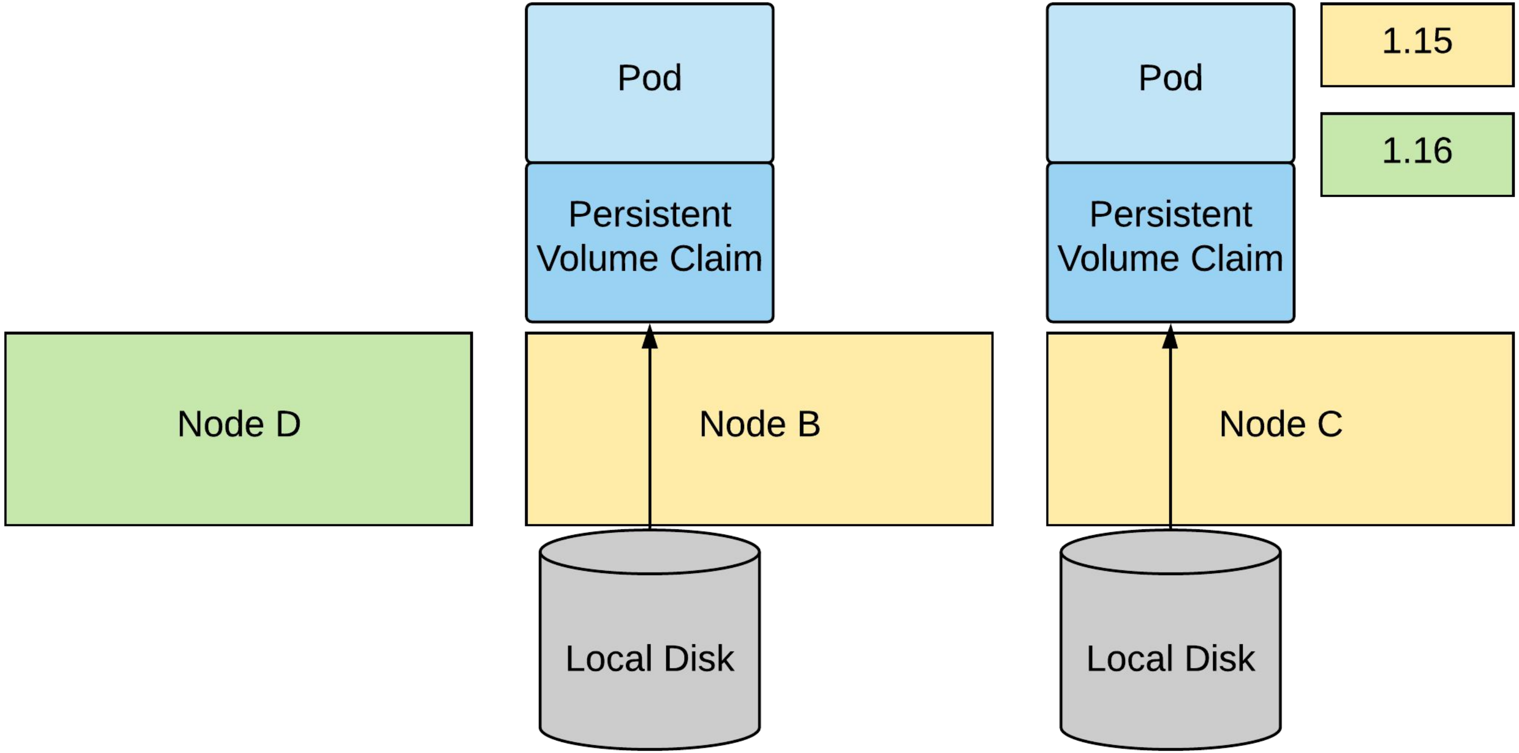
Node Upgrades: Local Volumes



Node Upgrades: Local Volumes



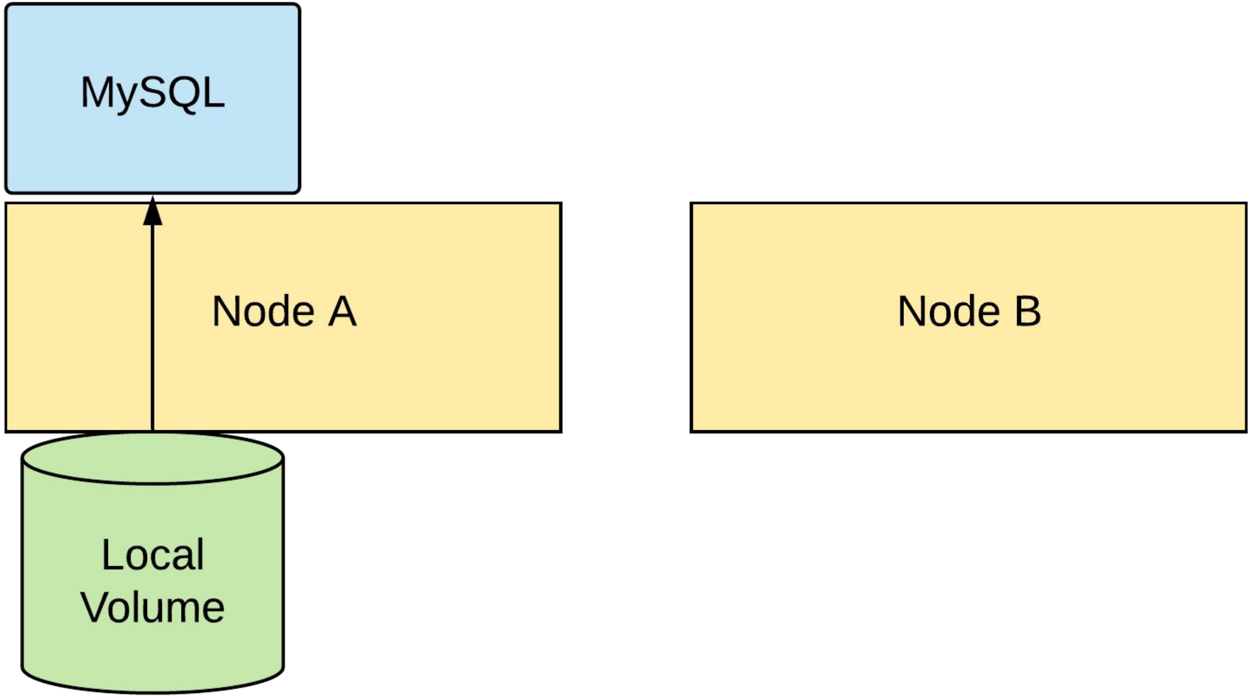
Node Upgrades: Local Volumes



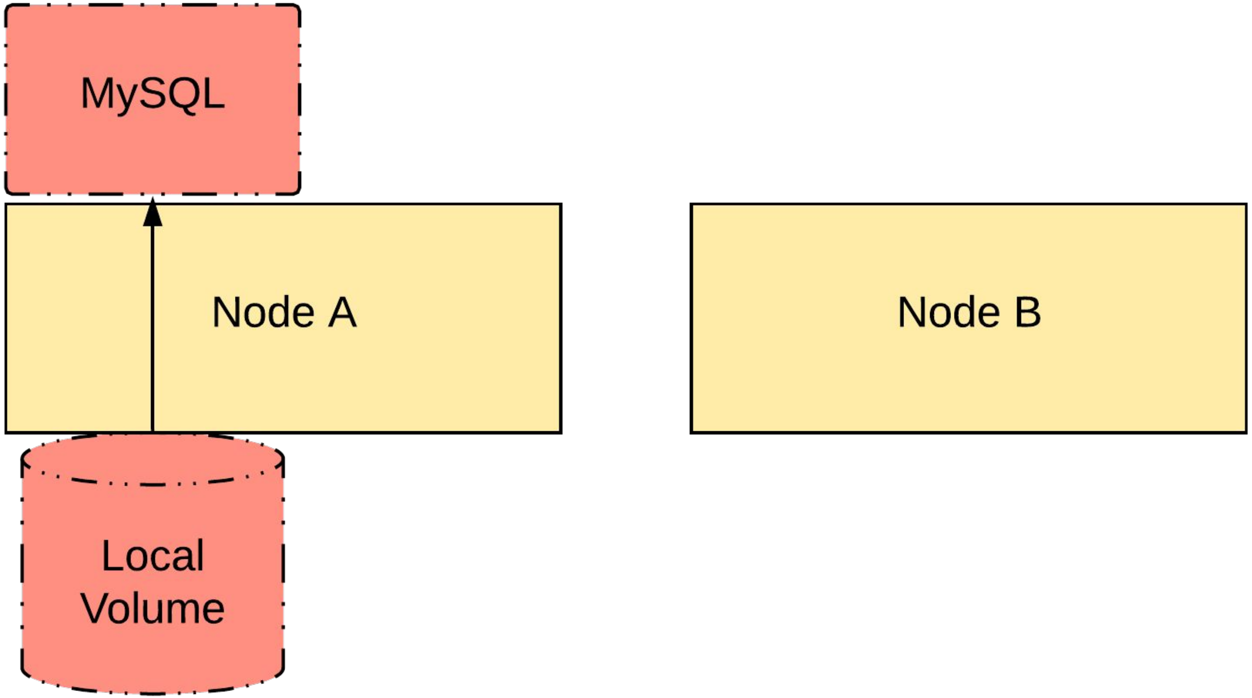
@mattschallert

Safely Using Local Volumes

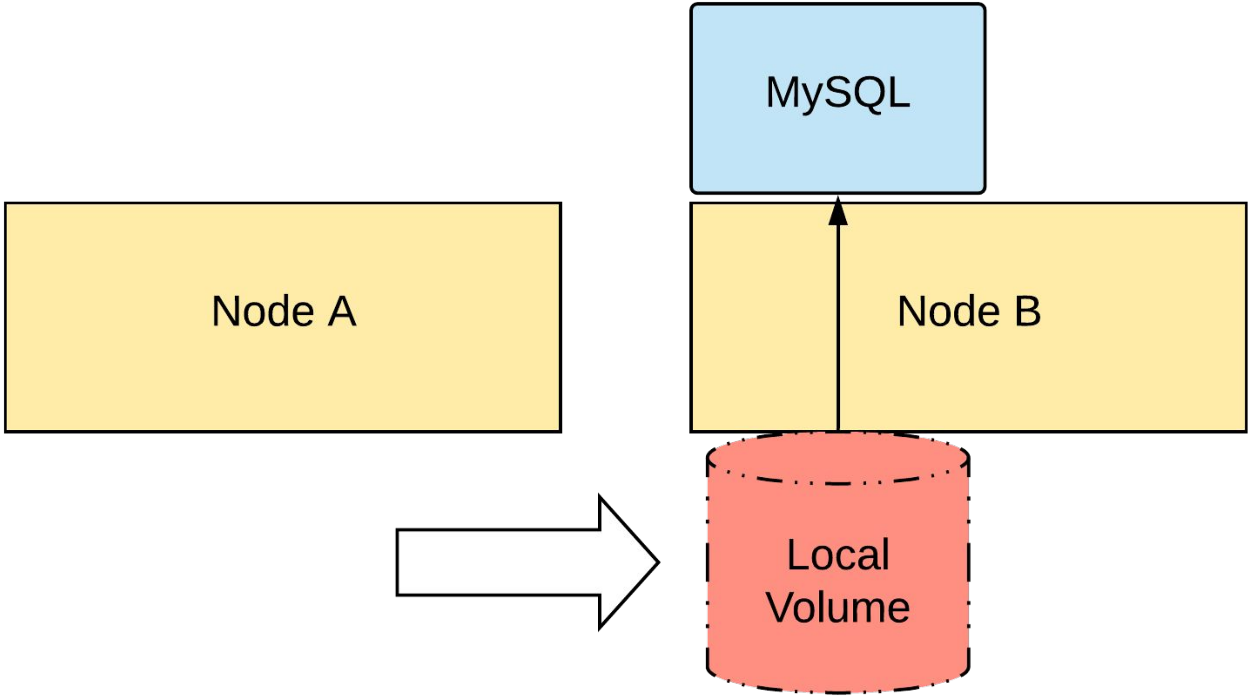
Anti-Pattern: Single Primary



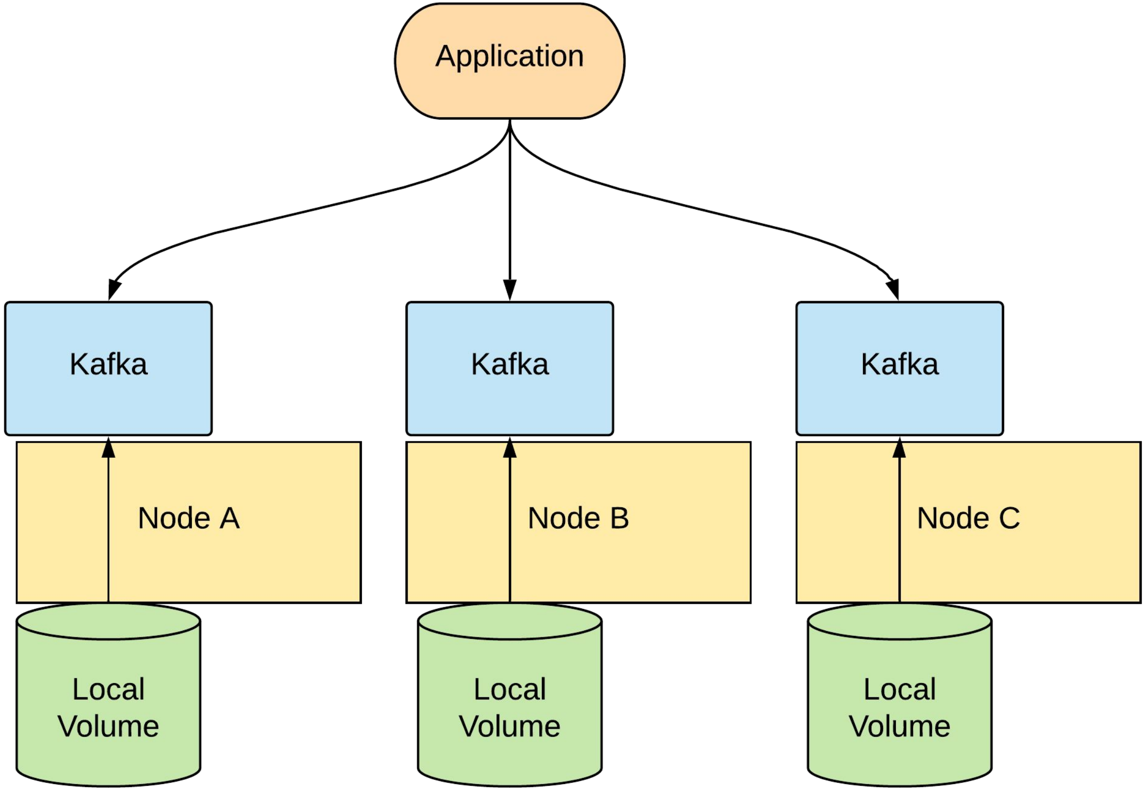
Anti-Pattern: Single Primary



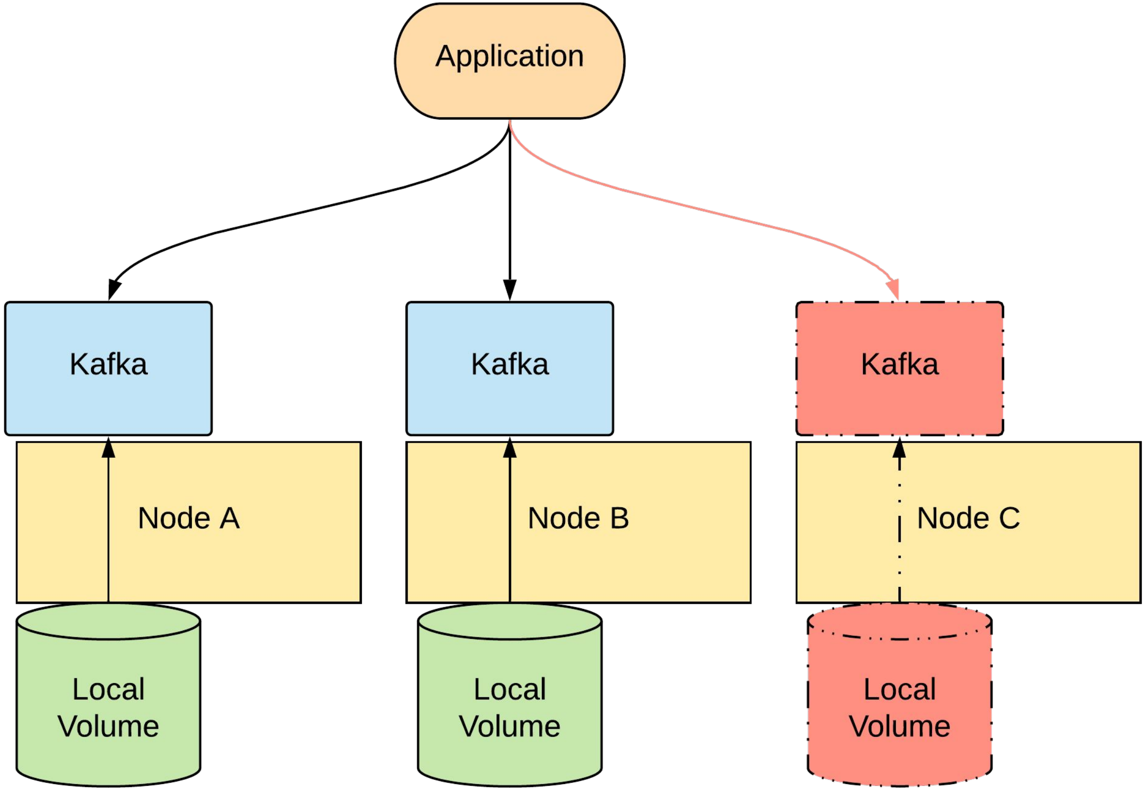
Anti-Pattern: Single Primary



Better Fit: Replicated Data



Better Fit: Replicated Data



Application

Kafka

Kafka

Kafka

Node A

Node B

Node C

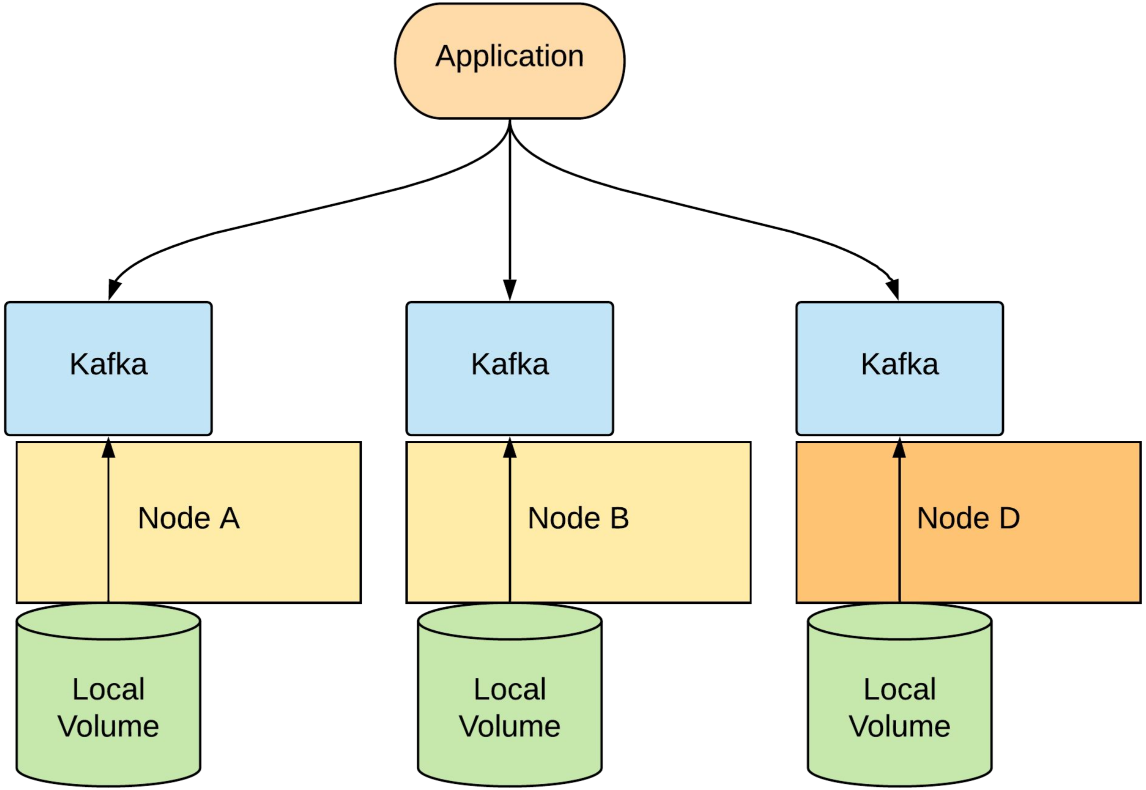
Local Volume

Local Volume

Local Volume

@mattschallert

Better Fit: Replicated Data



Use Cases

- Single primary DB instance? Probably not
- Replicated, fault-tolerant databases
- Local ephemeral cache
  - Data processing pipelines
  - Split between local disk + remote PV

Summary

- Consider impact on your operational practices
- Start with fault-tolerant or ephemeral use cases
- Read docs on provider-specific behavior
- Best practices in local-static-provisioner docs

Resources

- kubernetes.io/blog/2019/04/04/kubernetes-1.14-local-persistent-volumes-ga/
- kubernetes.io/blog/2018/04/13/local-persistent-volumes-beta/
- kubernetes.io/docs/concepts/storage/storage-classes/#volume-binding-mode
- kubernetes.io/blog/2018/10/11/topology-aware-volume-provisioning-in-kubernetes
- github.com/kubernetes-sigs/sig-storage-local-static-provisioner
- github.com/brunsgaard/eks-nvme-ssd-provisioner

Thank You! (+ Q&A)

- chronosphere.io
- Virtual booth
    - Let’s chat!

chronosphere