Container Native Improvements

- IBM Storage Scale Days 2024
- March 5-7, 2024 | Stuttgart Marriott Hotel Sindelfingen

- Stefan Roth
- Alexander Wolf-Reber

Disclaimer

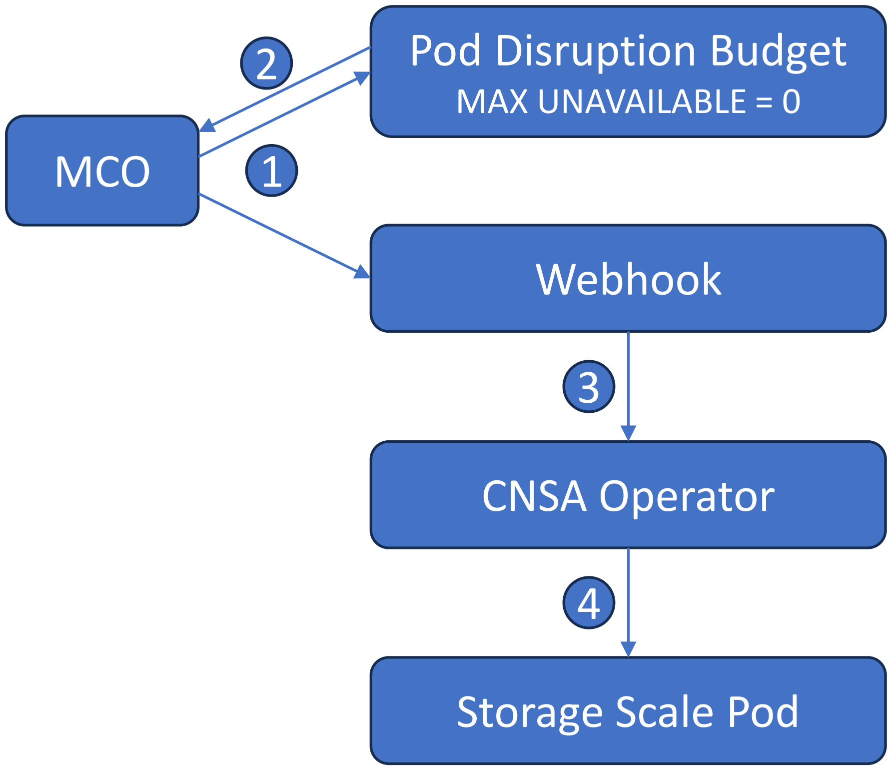
Storage Scale

- IBM's statements regarding its plans, directions, and intent are subject to change or withdrawal without notice at IBM's sole discretion. Information regarding potential future products is intended to outline our general product direction and it should not be relied on in making a purchasing decision. The information mentioned regarding potential future products is not a commitment, promise, or legal obligation to deliver any material, code, or functionality. The development, release, and timing of any future features or functionality described for our products remains at our sole discretion.
- IBM reserves the right to change product specifications and offerings at any time without notice. This publication could include technical inaccuracies or typographical errors. References herein to IBM products and services do not imply that IBM intends to make them available in all countries.

IBM Storage Scale Days 2024

This breakout session introduces some new functions that will be introduces with CNSA 5.2.0.

5.1.7 / 5.1.9 - Pod Disruption Budget

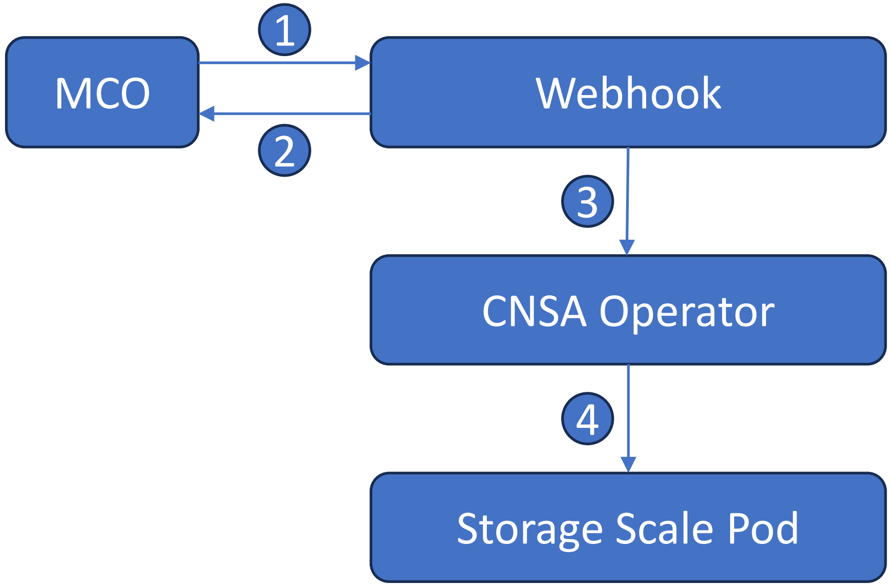


1. MCO evicts core pod
2. PodDistruptionBudget prevents pod eviction
3. Webhook sends eviction request to operator
4. Operator deletes Storage Scale pod

Confusion

- „This PDB blocks MCO upgrade for no reason."
- „Should I delete the PDB ?"
- „Why can the Storage Scale pod not be evicted ?"
- No user-friendly response for failed eviction in MCO logs

5.2.0 – No Pod Disruption Budget



1. MCO evicts core pod
2. Webhook prevents pod eviction
3. Webhook sends eviction request to operator
4. Operator deletes Storage Scale pod

Advantages

- No confusing Pod Disruption Budget
- User-friendly resonse from webhook

```
Error: admission webhook "veviction.scale.spectrum.ibm.com" denied the request:
Eviction of pod worker2 acknowledged. Waiting for operator to remove the pod.

To check progress:
oc describe daemon -n ibm-spectrum-scale

If progress appears to be halted, see the troubleshooting section in the user
guide: https://www.ibm.com/docs/en/scalecontainernative.
```

Node Cordons

- Which node are cordoned ?
- Did the operator cordon the node ?
- Did someone else cordon the node ?

```
$ oc describe daemon –n ibm-spectrum-scale
…
Status:
…
 Cordon And Drain:
    Nodes Cordoned By Operator:
    Nodes Cordoned By Others:  worker2.example.com
…
```

```
$ oc get node
NAME                    STATUS                      ROLES
worker0.example.com    Ready                       worker
worker1.example.com    Ready                       worker
worker2.example.com    Ready,SchedulingDisabled    worker
worker3.example.com    Ready                       worker
worker4.example.com    Ready                       worker
worker5.example.com    Ready                       worker
```

Node Drains

- Is the operator draining a node ?
- Which pods is the operator currently evicting ?
- Did the operator fail to evict pods ?
- Why did eviction of pods fail ?

```
$ oc describe daemon –n ibm-spectrum-scale
…
Status:
…
  Nodes Draining:
    Node:                  worker9.example.com
      Ongoing Pod Evictions:
           openshift-ingress/router-default-7f595df5c-f7r76
           openshift-monitoring/prometheus-operator-admission-webhook-767dd47d44-654zs
           openshift-operator-lifecycle-manager/collect-profiles-28367415-drxrw
           test-namespace/test-pod
      Pod Evictions Failed:
           test-namespace/test-pod
…
Events:
 Normal    NodeDrainFailed      3m9s    Daemon       Failed to evict pod test-namespace/test-pod on node worker9.example.com. Reason:
Cannot evict pod as it would violate the pod's disruption budget.
```

Pod Deletion

- Has the operator planned to delete Storage Scale pods ?
- Why does the operator want to delete Storage Scale pods ?
- Why can the operator currently not delete Storage Scale pods ?

```
$ oc describe daemon –n ibm-spectrum-scale
…
Status:
…
  Pods Unsafe To Delete:
    Delete Unsafe Reason:  Other pod worker1 is already terminating.
    Pods:                  worker0, worker2, worker6, worker7, worker8
  Pods Waiting To Be Deleted:
    Delete Reason:   6 pods will restart because configuration parameter 'PrefetchPct' changed to '35'.
    Pods:            worker0, worker1, worker2, worker6, worker7, worker8
    Delete Reason:   Eviction of pod worker1 requested by machine-config-controller.
    Pods:            worker1
…
```

Local File System - SNC

- Create local file system in a **Shared Nothing Cluster (SNC)**
- Create a LocalDisk CR for each disk or volume
    - Specify the node where disks or volume exists
    - Specify the device path
- Create a Filesystem CR
    - Type is local
    - Specify a list of disks
- CNSA 5.2.0 tech preview

```yaml
apiVersion: scale.spectrum.ibm.com/v1beta1
kind: LocalDisk
metadata:
  name: worker0-sdb
spec:
  node: worker0.example.com
  device: /dev/sdb
```

```yaml
apiVersion: scale.spectrum.ibm.com/v1beta1
kind: Filesystem
metadata:
  name: local-sample
spec:
  local:
    blockSize: 4M
    replication: 3-way
    pools:
    - name: system
      disks:
        - worker0-sdb
        - worker0-sdc
        - worker1-sdb
        - worker1-sdc
        - worker2-sdb
        - worker2-sdc
```

Local File System

- SNC limitations
  - Storage efficiency (3-way replication)
    - having to do 3 copies requires paying 3 to get 1
  - Network overhead
    - Writing flow of 500 MB/s to disk from a POD causes sending 1000 MB/s through network for other two replica nodes
  - Replication overhead
    - Whenever a failure group went offline we need to sync missed writes once it comes back. This is all orchestrated by the CNSA operator but will take time
- We consider support **shared disks** in future release (not in CNSA 5.2.0)
  - Shared disks can be more storage efficient with 2-way replication or no replication with external RAID
  - No network overhead if replication disabled
- We consider supporting CSI block drivers in future release
  - Just specify desired fs capacity, CNSA operator autonomically request block volumes to serve filesystem

Local File System – Interrest

Is there interest on local file systems ?

a) SNC file system ?

b) Shared disks file system ?

c) Shared disks in „all nodes share disks“ mode (SAN)?

K8s Limits

- Why are memory limits important for CNSA?
  - When a container goes past the memory limits it will get killed by the kubelet. This would pull the fs out underneath the application pods and might result in data loss.
  - In the gpfs container there runs mmfsd and a few other daemons. We can control their memory consumption fairly well. But we also run mm-cmds there like mmrunpolicy or gpfs.snap which may consume temporarily excess amounts of memory. We do no want to get killed in response to that.
- What have we done in the past?
  - Prior to 5.1.9.1 we set the limit to the total node memory. This had a flaw since worker nodes running in VMs could dynamically change their total memory which resulted in a rolling pod restart.
  - In 5.1.9.1 we set the limit to a very large number which derailed some management tools for OCP.
  - In 5.2.0 we default for twice the request and allow to override in the cluster CR to a larger number. So, if you experience kills of the core pods you can increase the safety margin.
- What about CPU?
  - CPU is not that critical, since no container gets ever killed. Instead k8s throttles the container to the limit.
  - It does make sense to have the limit higher than the request so that Scale can burst CPU when there are idle cycles to grab.

Thank you for using

Storage Scale

Storage Scale System