KubeCon, Seattle F2F Meeting

We are planning to meet at KubeCon in Seattle. We have a conference available at the)

Location: Municipal Room (Level B)

Zoom Link: https://zoom.us/j/614261834

Date

The meeting is scheduled for 12/12/2018 between 9AM to 12noon. Breakfast and beverages will be served from 8AM. The meeting can be extended to 1PM, if needed.

Agenda

- Core functions for group snapshots (Xing, Jing)
    - Execution Hook
    - https://docs.google.com/presentation/d/1H2dV-pDf9qExtxPF-ybYjUv7oHx2SAT1ESW1_Zto4YE/edit#
- CSI inline volume discussion
    - https://github.com/kubernetes-csi/drivers/issues/85
    - https://github.com/kubernetes/kubernetes/pull/68232
- Populator/DataSources sync (jgriffith)
- PVC Namespace transfer updates (jgriffith)
- Readiness Gate (Xing, Jing, jgriffith???)
- AttachDetachController if kubelet is unavailable (status about issue #65392)
- Snapshot topology (Xing, Jing)
- Migrating/upgrading from a dynamic provisioner to a CSI plugin
    - Without changes to kubernetes, the experience is likely to be poor
        - Either old volumes never get new functionality, or
        - PVCs could be upgraded to CSI in a semi-disruptive way
    - Implementing a PV swap (or rebind) in core Kubernetes is one option that would significantly improve things
    - Another alternative is to implement a more complex scheme for mapping of “old” PV types to CSI plugins such that vendor-provided CSI plugin could take responsibility for a subset of PVs
- Common iscsi\fc\mount libs. When CSI driver will be able to use it for production?

- CSI operator (Jan)

Attendees

1. Sambasiva Bandarupalli <samba@diamanti.com>
2. Abhay Kumar Singh <abhay@diamanti.com>
3. Chakri Nelluri <chakri@diamanti.com>
4. Shilpa Mayanna <shilpa@diamanti.com>
5. Vaibhav Kamra <vkamra@kasten.io>
6. Tom Manville <tom@kasten.io>
7. Sagy Volkov, Red Hat <sagyvolkov@gmail.com>
8. Vangelis Koukis, Arrikto <vkoukis@arrikto.com>
9. John Griffith, <john.griffith8@gmail.com>
10. Ardalan Kangarlou <Ardalan.Kangarlou@netapp.com>
11. Ben Swartzlander <Ben.Swartzlander@netapp.com>
12. Jose Rivera <jarrpa@redhat.com>
13. Saad Ali <saadali@google.com>
14. Serguei Bezverkhi <sbezverk@cisco.com>
15. Shailesh Mittal <sm@datera.io>
16. Xing Yang <xingyang105@gmail.com>
17. Simon Croome <simon.croome@storageos.com>
18. Matt Smith <mss@datera.io>
19. Bradley Childs <bchilds@redhat.com>
20. Murali Balcha <Murali.balcha@trilio.io>
21. Yash Desai <desaiy@vmware.com>
22. Adam Litke <alitke@redhat.com>
23. Deep Debroy <ddebroy@docker.com>
24. Shay Berman <bshay@il.ibm.com>
25. Ryan Wallner <ryan.wallner@portworx.com>
26. Kevin Fox <Kevin.Fox@pnnl.gov>
27. Subodh Mathur <subodh@nutanix.com>
28. Hemant Kumar <hekumar@redhat.com>
29. Michelle Au <msau@google.com>
30. Jan Šafránek <jsafrane@redhat.com>
31. Masaki Kimura <masaki.kimura@hitachivantara.com>
32. Sandeep Pissay <ssrinivas@vmware.com>
33. Luay Alem <luay.alem@gmail.com>
34. Mathusan Selvarajah <mselvara@redhat.com>
35. Luis Pabon luis@portworx.com
36. Julian Hjortshoj Julian.hjortshoj@dell.com
37. Sergey Kornfeld serge.kornfeld@huawei.com
38. Michael Adam obnox@redhat.com

39. John Mulligan jmulliga@redhat.com
40. Humble Chirammal hchiramm@redhat.com
41. Gregory Touretsky gtouretsky@infinidat.com
42. Kiran Mova kiran.mova@openebs.io
43. Goutham Ravi gouthampravi@gmail.com
44. Matthew Wong mawong@redhat.com
45. Rena Shah renash@microsoft.com
46.

Meeting Notes

Snapshot (Slides)

We are planning to add execution hook and snapshot group of volumes (whole application).
There is a proposal for application snapshots but sig-arch has some concerns.
CSI calls for freeze and unfreeze file system is needed.

- File system freeze and unfreeze hooks aren't enough, we need something for the application too.
- For retries there should be 2 parameters, one that says how many times and the other at what time interval.
- quiesce and unquiesce are application level hooks.
- Freezing file system requires privileged permissions. If we do that with CSI node RPC, it should already has those permissions to format the filesystems.
- We could need both quiesce and file system freeze. It depends on the use case.
- Pod and application readiness could be independent of each other. We could use liveness checks to decide if application could server requests while quiesced.
- Can CSI file system freeze be called before quiesce or after? Xing - it should be called after.
- GroupSnapshot interacts with **pod disruption budget** and needs to be explored.
- How will quiesce/unquiesce work when volume is shared between multiple pods? It is possible that it may be too hard to get it right in this case and hence we may want to disable quiescing in that case.
- Is it possible to quiesce different paths in same volumes when volume is shared between different application? Xing - there should be a way to filter based on PVC.
- Publish/Unpublish calls can not be made when File system is frozen. These details needs to be flexed out. (CSI spec needs more details)
- We need to consider recovery of frozen file system.

- Restoring snapshots for stateful sets is trickier because it requires either modification to stateful set controller itself or a new CRD to manage restore or having GroupSnapshot saving info of its source.
- Snapshots and restoring volume from snapshot needs to take into account topology consideration of both volume and snapshot.
- Need to make quiesce/freeze optional (both hook and CSI RPC calls)
- Scheduler does not know how to handle delayed topology aware provisioning of volumes from snapshot.
- For most cloudproviders snapshots are typically regional but volumes are zonal.
- QA:
    - Do the ExecutionHooks be executed via side-car?
        - FS quiesce requires high security level context.
        - NodeFreezeFilesystem could be executed by csi node driver
        - side cars may be required to handle timeout options
    - What happens if quiesce takes lot of time?
        - will the quiesce impact the pod readiness
        - there can be a change in the readiness/liveness logic
        - controlled by the application developers
    - Can I disable snapshots on some volumes?
        - 2 PVCs used in same volume - 1 for data and 1 for logs
        - How to mention don't snapshot logs volumes
        - PVC should have the information - that will then be passed to the node plugin
    - Can the quiescing be optional in the snapshot - say application consistent snapshot requires quiescing but not in the scheduled snapshot
    - There are still open questions under how to handle error cases. How to handle cases like Freeze has happened and is unable to unfreeze. How publish/unpublish - recovery work in case of errors.
    - Request for revert snapshot feature from the field

Inline Volumes for CSI(Design PR)

- Inline volumes are volumes that can used without PV and PVC.
- Inline volumes for CSI allows implementing something like ephemeral volumes.
    - Something like data images created from Dockerfiles.
- Supporting attachments for inline CSI volumes is something we found to be problematic.
- Admins can blacklist via pod security policies.
- Namespace propagation is a problem.

Volume Readliness (Slides)

- Volume Readliness is a field on PVC to indicate if volume is ready for attach/detach. Only when a volume is ReadyToUse it can be attached or detached or mount/unmount.
- The Readiness Gates also represent the conditions
- Detailed proposal - https://docs.google.com/document/d/1JsEA9h3mH-MTG3QGHlpMw-t3ZfSECFGxB4yLQbfUnjU/edit#
- Jan - Why can't we use taints and tolerations for achieving same thing, rather than introducing new concepts.

-

Migrating/upgrading from dynamic provisioner to a CSI plugin

- Updates
    - Default csi implementation like NFS etc., does very basic operations like publish/unpublish
    - Existing systems already have volumes created with older provisioners. Once the CSI driver has been included into the cluster, how to make use of the new capabilities built into the new CSI driver for old volumes:
        - Rebind is one option. (Jan said he can fix the rebind command and document how to do it. That will require the manual removal of known Finalizers)
        - Another option is to add an API that will pass the volume to CSI drivers to identify if any of them own that up. Currently there is no option to pass the volume id (that was not created from non-CSI)
        - Volume created in 1.9 will not get the features added for volumes in 1.14 - could this approach be used?
        - Allow the application developer to migrate. This would require a admin access. For the developer - PVC is the only interface and connecting to PV etc., are taken care automatically.
        -
    - During migration from intree NFS to specific NFS might also require Rebind

Common iSCSI/FC CSI

- The alpha versions are available to use.
- iSCSI - should it be containerized?
    - The initial idea is to use the go-binding to make use of the iSCSI drivers
    - What are the issues:
        - iscsid is not namespaced
            - Run iscsid on the host. Iscsiadm inside container talks to the host - can result in incompatible iscsid - causing crashes
            - Run iscsid in the container - needs access to the system (privileged access). But if container crashes - will leave out stale entries.
    - Need to work with open-iscsi to make it namespace compatible. A kernel patch is open.

PVC Namespace transfer updates (https://github.com/kubernetes/enhancements/pull/643)

- Add some notes to the PR with a generic API that can help with namespace transfer for all kinds of objects.
- Need to consider cases like - how to transfer related objects that are not associated with namespace.
- Annotations are error prone. Need to convert to Spec parameters

Populator/DataSources sync

- Make use of the DataSource, which is currently used for clone.
- Adds dataSource spec to the PVC
- Need to work on the type checking for the dataSource
- Should there be generic CRD DataSource or have specific sources like HTTPSource
    - Keeping them seperate allow for different drivers to be written.
- Needs help with reviews pushing forward the PRs.
- Requires the ReadinessGate to be prototype this feature.
- Flow:
    - Create the Volume

    - Watch the PVC being attached, kick in the process of populating by creating a container in the same namespace as the application pod that is using the volume.

CSI Operator

- Does all the plumbing work like creating the service account, deployments etc., for installing the csi operator
    - RBAC rules
    - Images that need to be used for side cars
    - Mount points for side cars
    - Helps with upgrading the csi driver or side car images
- Works currently with OpenShift / K8s 12 and uses OpenShift CI. This can be moved to the csi repository.
- Admins will be interacting with the Operator CR.
- Remove the storageClassTemplate
- Side cars may not follow the K8s version and more tied into csi driver version.