Summary

Optimize the loop iteration period (static 100 ms) in the Reconciler and Desired State of the World (DSW) by increasing the sleep period when no changes are detected. As a result, the Kubelet will listen to gRPC events stream from the CRI implementation.
(Initial issue:
https://github.com/kubernetes/kubernetes/issues/126049)

Motivation

In volume manager :

A. desired state of the world function : populates the desired state to the cache
   a. findAndAddNewPods
   b. findAndRemoveDeletedPods

B. Reconciler function : reconciles the state
   a. reconstructVolumes : tries to reconstruct the actual state of world by scanning all pods' volume directories from the disk
   b. unmountVolumes
   c. mountOrAttachVolumes
   d. unmountDetachDevices
   e. cleanOrphanVolumes
   f. ……

Adapt the DSWP and Reconciler run loop period based on the events emitted by the CRI .

Goal

Reduce unnecessary work during inactivity.

Non-Goal

Completely remove the batch loop.

Proposals

Proposal 1 : https://github.com/kubernetes/kubernetes/pull/126450

: the PR allows users to customize or override the loop period configuration using the kubelet conf file :

Reason/suggestion (**sig node**) : move to event-based approach: https://github.com/kubernetes/kubernetes/issues/126049#issuecomment-2278659439

Proposal 2: https://github.com/kubernetes/kubernetes/pull/126668 : This proposal increases the timer without the event-based approach. If a change is detected, the function resets the sleep period. However, this PR will likely be closed since changes are detected late.

Proposal 3 : switch to event-based approach.

Design Details

Triggering the existing implementation based on the event type :

- CONTAINER_CREATED_EVENT
- CONTAINER_DELETED_EVENT

Gradually increase after the third execution (to no impact the existing retry logic) (, e.g., +100ms on each iteration) the sleep period to a maximum (e.g., 1 second). If any event is detected, reset the interval back to the default value (100ms).

Risks and Mitigations

- A bug or an issue on the event-based approach implementation: a flag will needed to activate the feature (alpha initially).
- CRI bug on the event system :
    - Impact: Will take more time to mount/unmount volumes (1 second max instead of 100 ms).
    -