HNC: Make excluded namespaces configurable

yiqigao@google.com - Mar. 18, 2021 - **WORLD VIEWABLE**
https://bit.ly/hnc-excluded-namespaces
Status: Implemented. Approvers: [x]rjbez17, [x]adrianludwin
Please join kubernetes-wg-multitenancy@googlegroups.com to comment

Background

Currently the excluded namespaces list is hard-coded, it would be nice to have it configurable. More importantly, when HNC admission controllers are not responsive, e.g. not ready yet, the operations in the excluded namespaces may be blocked by our FAIL CLOSE webhooks, which is not ideal.

See issue #374 and #1023 for more details.

Goal

Users will be able to exclude namespaces before installing HNC by:

- Setting `excluded-namespace` container args in the `config/manager/manager.yaml`:

```
containers:
- command:
  - /manager
  args:
  # ... other args ...
  - "--excluded-namespace=kube-system"
  - "--excluded-namespace=kube-public"
  - "--excluded-namespace=hnc-system"
  - "--excluded-namespace=kube-node-lease"
```

- Allow users to add the `hnc.x-k8s.io/excluded-namespace = true` label to the excluded namespaces that match those listed in the container args.

Detailed Design

We will add a `namespaceSelector` to our object webhook in the `ValidatingWebhookConfigurations` to apply the validation only on those namespaces *without* the `hnc.x-k8s.io/excluded-namespace` label. With that, when HNC or specifically the webhook services are down, operations in the excluded namespaces won't be affected.

Other webhooks on HNC CRs don't need this `namespaceSelector` because they are supposed to prevent abuse of the CRs in the excluded namespaces. Adding the `namespaceSelector` would deprive their capability.

As for the namespace webhook, unfortunately we found that the request to add a label filtered by the webhook's `namespaceSelector` won't get validated. In our case, if we set the webhook to only validate those namespaces without the excluded namespace label, people can add the excluded namespace label to any namespaces and bypass the webhook. Therefore, we cannot add the `namespaceSelector` to the namespace webhook.

Prevent abuse

To avoid abuse of the `hnc.x-k8s.io/excluded-namespace` label, our reconcilers will only respect the excluded namespaces listed in the container args and remove the label if the namespace shouldn't be excluded. The namespace webhook will *always* deny a request if it's trying to add the `hnc.x-k8s.io/excluded-namespace` label, except on the excluded namespaces themselves.

Default settings

We will have `kube-system`, `kube-public`, `hnc-system` and `kube-node-lease` listed in the `config/manager/manager.yaml` by default. Thus, they are the default excluded namespaces.

HNC will also add the `hnc.x-k8s.io/excluded-namespace` label to the `hnc-system` namespace by default, and users will have to add the label to the other excluded namespaces manually. The `hnc-system` namespace is created by HNC. Without setting this label by default on the `hnc-system` namespace, HNC cannot create secrets for webhook certs.

Discussions

This design contains two usability limitations:

- Users have to add labels manually to make it work, besides adding container args to the deployment configuration;
- The excluded namespaces are not configurable after installing HNC, e.g. when another component in the cluster creates the `cert-manager` namespace, we cannot exclude it unless we redeploy HNC.

It would be nice to make excluded namespaces configurable, for example, in the `HNCConfiguration` object. However, it's much harder than using a command-line parameter, and the need to configure it later after installation is low. In that case, the admins can either set all possible excluded namespaces container args before installing HNC and then add labels to the namespaces when they exist, or simply redeploy HNC with the new configuration.

As for automating the namespace labeling, if HNC never comes up, those namespaces would never get labeled, thus never getting excluded from HNC validating webhooks anyway. Besides, it would be nice for users to modify namespaces like `kube-system`, instead of HNC making that change without the users explicitly telling us to. It's more predictable and safer for HNC to fully ignore namespaces, rather than having a custom set of operations to perform on them, especially when it's not enough to ensure safety (when HNC doesn't come up).

Possible Future Extensions

We will implement the approach described in this doc with the container args in the HNC deployment configuration and the excluded namespace labels for now.

In the future, if there is a clear need to make the excluded namespaces configurable in the `HNCConfiguration` object, we can add it. We should continue to honour the excluded namespaces container args to make this change backward compatible.

If someone really wants to automate the excluded namespace labeling, we can have a discussion then.