HPA: Replacing scale down forbidden window

author: jbartosik@
last modified: 2018-08-14

Context

Currently HPA uses scale down window. After HPA changes size of a deployment (either by scaling up or down) it will refrain for scale downs for the duration of the window.

Goals

A replacement should:

- Avoid frequent deployment size changes (initialization is costly, short break in a traffic shouldn't result in quick scale down and dropping requests when traffic increases back),
- Allow reasonably fast scale downs (to avoid unnecessarily allocating resources to pods that are not needed).

Method

- Let's call recommendations as calculated by HPA now "raw recommendations",
- When HPA calculates a raw recommendation it will record it (with a timestamp),
    - It will also erase any recommendations it doesn't need any more.
- Instead of applying a raw recommendation HPA will apply max of raw recommendations it calculated in last $duraton.

Testing

All scenarios will start with deployment size 100, target CPU utilization 60%. I'll execute all the scenarios against a cluster running a released K8s version and against cluster using HPA with my changes. Scenarios I want to check (with expected results):

- No load, some CPU used by background activity. Pods use 30% CPU. Modified HPA should scale down more quickly.
- Spiky activity. 0% CPU usage in odd minutes, 100% cpu in odd minutes. Modified HPA should not scale down, original HPA expected to make big changes to size.
- Small spikes in activity. 0% CPU usage on odd minutes, 50% on even minutes. Modified HPa should scale down at a rate similar to the first scenario. Original HPA expected to scale down rapidly.