Contribution ID: **2** Type: **not specified**

K8s autoscaling based on custom metrics. Two examples of application: CMSWEB and HTCondor in the CMS Analysis Facility@INFN

Tuesday 1 December 2020 09:50 (20 minutes)

At the moment, Kubernetes only supports horizontal pod autoscaling based on predefined pod metrics (CPU and memory usage). Therefore, in order to achieve an actually green elastic cloud model (optimizing resource usage) a key point is to integrate this tool with autoscaling solutions based on custom metrics, and this requires the usage of third-party elements.
In this work we show the horizontal pod autoscaling based on custom metrics: in this workflow metrics are collected by a Prometheus server, and are then manipulated and made available to k8s-native Horizontal Pod Autoscaler (HPA) resources.
We show how we apply the presented feature to two HEP-related use cases: in the first one this solution is applied to CMSWEB (i.e. CMS web services) infrastructure, in the second one it is used to enhance elasticity of an analysis facility prototype on INFN-Cloud, with the automatic scaling of HTCondor instances.

Authors: SPIGA, Daniele (Universita e INFN, Perugia (IT)); CIANGOTTINI, Diego (INFN, Perugia (IT)); TEDESCHI, Tommaso (Universita e INFN, Perugia (IT)); KUZNETSOV, Valentin Y (Cornell University (US))

Presenter: TEDESCHI, Tommaso (Universita e INFN, Perugia (IT))

Session Classification: block 1 - presentations