Intra-queue preemption : Scheduler impact analysis

Overview

Intra Queue preemption focuses to normalize various application's resource usage within a queue. Application Priority, User Limit etc will be considered to normalize application in the first phase development of Intra-Queue preemption module. This document focuses on checking the impact of intra-queue preemption against various scheduler features. Main intention is to call out certain features in scheduler and to announce that such feature will not be working well preemption.

Intra-Queue Preemption : Short Overview

Application Priority

As of today, existing inter queue preemption module identifies oldest containers to preempt based on a preemption iterator from FiFoOrderingPolicy. This ensures that apps are ordered first w.r.t its priority. Then containers are selected from within each application by checking each container's priority. Main intention is to select lowest priority container from a lowest priority application.

Hence if a high priority application has some outstanding demand, Intra-queue preemption module tries to get resources from lowest priority apps within same queue.

User Limit

Post YARN-2113, intra-queue preemption module could also normalize resources in a queue based on user limit as well. This capability helps underserved applications of a given user to get more resources from user's who are abusing resources above its resource limits. Strict enforcements are added to ensure that preemption will not bring down any user's resource usage under its user-limit.

Impact Analysis against Scheduler features

In ideal world, preemption module will be working as a feeder for scheduler. Preemption module will be able to predict how next scheduling may happen and based on this assumption, selection of container will be performed. Preemption module must avoid a case where few containers are preempted but scheduler again assigned containers to the same preempted apps itself.

a) vs Queue Capacity and User Limit

When resource demand from higher priority applications are considered during preemption calculation, intra-queue preemption module ensure that the demand from these high priority applications are genuine by considering Queue Capacity and its User limit. Preemption module will skip such app's demand if any such violations found.

b) vs Queue Priority

Queue priorities are considered during the selection of queues based on its resource usage. Hence Intra queue preemption does not have any impacts.

c) vs Node Locality Delay

An application with a demand of OFF_SWITCH containers has to wait for node locality delay or scheduling missed opportunities. Hence preemption of container from such apps may force a delay in completion. Ideally this case is fine, but could have some impacts as mentioned above.

d) vs Node Label

No impacts as preemption module considers label in its calculations.

e) vs AM Resource Limit

Preemption modules considers applications which are running to preempt containers. Also preemption module ensure that its takes demand from only active apps. Hence no impact.

f) vs AM Containers

Preemption module will skip AM containers for now. However we need to improve to consider some AM containers in case of urgent need. This has to be implemented.

g) vs Reserved containers

As of today, reserved containers are not considered in Intra-Queue preemption. This is to be done in next phase.

Open Discussion

Container selection for preemption is only based on application priority (submission time) and container priority. It is possible some applications container may be costly to preempt. However no such policies are available today to save containers based on a COST.