YARN – 1042 Design Document

Author: cheersyang@hotmail.com

Problem It Solved

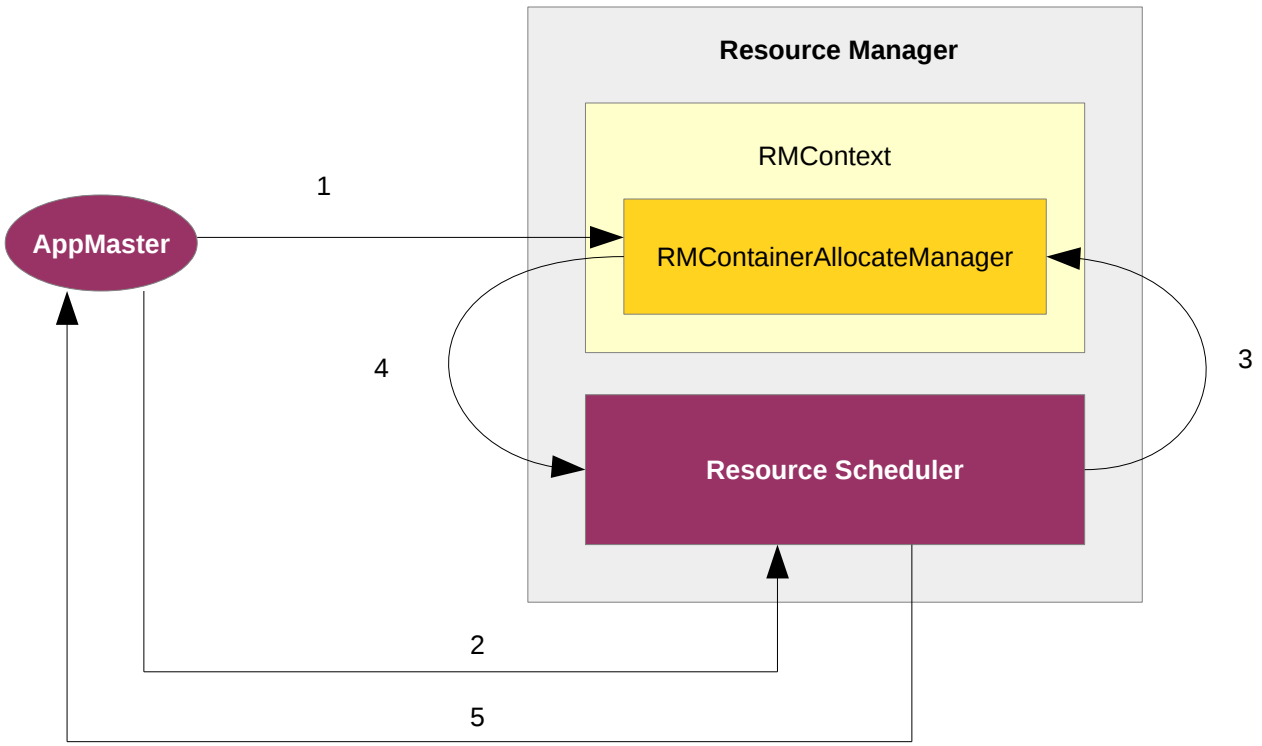
Add the ability to Resource Manager to apply certain allocation policies while assigning containers. A policy can be defined with following attributes

- Rule : AFFINITY, ANTI_AFFINITY, NO_PREFERENCE
- Scope : NODE, RACK
- maxTimeAwaitBeforeCompromise

The NO_PREFERENCE rule is the default value, RM will allocate the container to wherever resource scheduler wants; AFFINITY rule means the application prefers to group their containers on to the same node/rack; ANTI_AFFINITY rule means the application prefers to distribute their containers to different node/rack. Where Scope decides to apply the policy to node level or rack level.

The preference can be described as REQUIRED or PREFRRED, REQUIRED means the policy must be satisfied otherwise RM cannot assign a new container to the application, and PREFRRED means the policy can be, under some circumstances, compromised. In this version, we provided a time limit maxTimeAwaitBeforeCompromise. When it sets to be a value larger than 0, RM will wait for that time before assigning a container which will disobey the policy. In feature, it's possible to add more attributes to extend possibly compromised situations, such as time of failures, resource limits etc.

How It Works



Explaination

1. AppMaster registers an app attempt to RMContainerAllocateManager, when a ContainerAllocateRule is given, the application ID and the rule will be added into a map.
2. AppMaster sends AllocationRequest in a heartbeat to Resoure Manager asks for resource to execute the application.
3. The Resource Scheduler queries RMContainerAllocateManager to get the ContainerAllocateRule of this application.
4. The Resource Scheduler calls particular ContainerAllocateHandler while assigning containers to a Node Manager
5. The Resource Scheduler sends the AllocationResponse back to AppMaster, then AppMaster assigns tasks to run on these containers.
6. Repeat 2-6 in the application life cycle.

Major code changes

ApplicationSubmissionContext

The container allocate rule is a per-app configuration, add a new argument ContainerAllocateRule to the submission context, so AM can send it to RM through ApplicationClientProtocol when submitting the app attempt.

RMContext

When Resource Manager starts up, add an instance of RMContainerAllocateManager in its context.

RMContainerAllocateManager

RMContainerAllocateManager stores a map to track the mapping between an application and its container allocate rule. The map is updated when register or un-register an application.

AffinityContainerAllocateHandler and **AntiaffinityContainerAllocateHandler**

Both implement interface ContainerAllocateHandler, it is called by the Resource Scheduler, e.g CapacityScheduler, during it assigns containers to a scheduler node. If assign a container to a node would cause a conflict to the placement policy, the assignment will be rejected. In feature, we can add more handlers for more placement policies.

LeafQueue

Call a container allocate handler to evaluate if this assignment would against the given policy, injected to the process when capacity scheduler assigns containers.

TestContainerAllocateRule

Unit tests. It currently test node-affinity-required and node-anti-affinity-required policies.