Title: [YARN-1963] Support priorities across applications within the same queue

Authors: Sunil Govind, Vinod Kumar Vavilapalli

Last modified: November 11 2014

1.1 Preamble

Many a time, it is needed to execute some YARN applications at higher priority, regardless of other applications that are already running in a YARN cluster. The existing mechanism for enabling such a use-case in YARN's CapacityScheduler is by creating multiple queues and making users submit applications to higher priority and lower priority queues separately, each potentially setup with appropriate capacities. It is desirable to enable users of YARN clusters to define a priority for an application at its submission time. YARN-1963 is focused on handling such application-priorities as described in this document.

1.2 Problem statement

Using existing features in YARN's Capacity Scheduler, to support different priorities among a queue's application, one can do the following:

a. Configure multiple queues and define resources to each queue based on the priority.
b. User submits applications to different queues based on the priority define for that queue.
c. Access control permission can be set to each queue based on demand.

This is an inflexible way to enable priorities for users and can be made easier by letting priorities be specified for Application at submission time itself. Different priority applications can run in same queue and can use the resources of that queue based on the application-priority.

1.3 Requirements

- Users should be able to set priorities for each of their applications. These priorities should be respected in every leaf-queue.
- Priorities have significance across users. Higher priority implies scheduling importance irrespective of the user.
- Other scheduling constraints within a leaf-queue like user-limits, submission-time (FIFO

behavior) should meaningfully interact with priorities

- A high priority application from a user who already hit-limits should continue to wait in the queue.
- Applications with same priority should be ordered by their submission-times

- ACLs on priorities – Access Control Lists should exist on priorities to avoid users from getting incentivized to submit apps all with the highest priority.
- Admins should be able to enable/disable this feature for the entire scheduler as well as per queue.

1.4 Proposal

YARN-1963 defines priority for an application and makes the application run in the cluster as per its assigned priority.

Solution for this proposal has mainly three parts,

- **Setting the priority to an application:** Application-priority can be set via the submission API.
- **Scheduling applications based on Priority**: Based on the application-priority, a higher priority application's requests will be scheduled first from Resource Manager. Once all these requests of higher priority applications are served, then lower priority application requests will get served from Resource Manager.
- **Access Control Lists**: User must be granted permission to run a specific priority application. This can be taken as access control configuration for each priority level.

1.5 Details

1.5.1 How can End User submit an application by specifying priority?

1. Priority can be set via below methods:
   - **With API**: Below API can be used to set the Application priority.

     application.setApplicationPriority(ApplicationPriority priority)
   - **With command line parameter:** While submitting an application using yarn command, **yarn.app.priority** can be used to set the priority of the application.
2. ApplicationPriority is a continuum with their domain space as a set of integers.

Administrators can define known priority labels and then map them to a corresponding integer. Scheduler is completely agnostic to the priority labels, it only acts on the integer values. Admins can also define a default priority. A default definition looks like below:

Priority Label	**Mapped Value**
very_high	5
High	4
Normal	3
Low	2
very_low	1

End user can use this priority label to set the corresponding priority for the application.

1.5.2 How Admin can configure priority labels?

1. Configure priority label for the cluster:

Priority labels can be defined by Admin and labels can be configured through configuration parameter named **yarn.application.priority-label**.

As mentioned earlier, priority labels will be mapped to its respective integer value only in scheduler. For all other uses, priority label will be used as String itself. A sample configuration is given as below.

yarn.application.priority-label=high:3,medium:2,low:1

Optionally the integer mapping is not mandatory to provide. Resource Manager will take the same in the order of priority from lower to higher if the mapping is not specified.

2. Configure priority label per queue level:

Admin can define a max priority label per queue level if required. By default, all priority labels defined in cluster level will be applicable in queue level also, but priority labels higher than the max priority label cannot be used in the specific queue.

Sample configuration for capacity scheduler will be like below.

yarn.scheduler.root.<queue_name>.max_priority_label=high

A default priority label also can be configured per queue level. This will help for those applications which are submitted to the queue without specifying any priority. All such applications can run into this category of priority label.

yarn.scheduler.root.<queue_name>.max_priority_label=medium

Note: Queue Level priority label configurations can be configured separately for Capacity Scheduler and Fair Scheduler for now.

1.5.3 APIs or CLI support needed for Application Priority

1. Query Priority Labels under different criteria:

Admins and End users will require few sets of API's (also REST support) for priority labels.

- API to provide a list of complete priority labels in a given queue
- API to provide a list of priority label for a user in a given queue

CLI should support above mentioned query result and it can be made as a yarn application command.

2. Support for add/modify/delete operation on Application priority label

Application priority labels are configured in xml and it will be preloaded with configured or system defined integer mapping during startup. Dynamic support is required to change the below configurations even at run time with API/REST and Admin commands.

- Add or Delete Priority labels configured for a given queue
- Change default priority label for a given queue

Note: Deletion of priority label will first ensure all application running under that priority is completed, then label will be removed from queue.

3. Change Priority of a submitted application at runtime

As per few use cases, it will be very convenient for a user or admin to change the priority of an application during runtime.

- A User, who submitted an application with a given priority, should be able to change priority during runtime with the help of an API.
- Admin also will have privilege to change the priority of an application during time with help of API as well as CLI (admin command)

Note:

- The User or Admin, who is changing the priority of an application, must have proper ACL privileges for the successful change of priority.

- Once priority is changed for a running application, all new pending requests from RM will be granted with updated priority level.

This part will be taken up in separate JIRA and will be tracked as soon as the initial framework is ready.

1.5.4 Scheduler side support for Application-priority:

Scheduler changes for Application priority have to be done separately for Fair and Capacity schedulers. A brief design approach will be as follows.

In each queue, there can be many applications submitted with different priorities. In such cases, Scheduler will try allocating resources to an application that has higher priority than others and then for next lower priority application and then so on.

- By changing the application-comparator in Capacity and Fair Scheduler, always a higher priority application can be fetched and then used for later resource allocation. If there are multiple applications at same Application-priority level, then comparison will be based on submission-timestamp as happening normally now.
- During APP_ATTEMPT_ADDED event handling in both Capacity and Fair Scheduler, the application-priority can be fetched from Application Submission context. Application-priority can then set to the scheduler app level and can be used later.

Note: Interaction with scheduling parameters

- User-limit: When priority comes into picture for each application, it will not be fair to schedule resources in a uniform manner for all application in a queue with respect to user limits.

  It will be better to consider user-limit per priority level, and user-limit can be applied to those applications which fall in to specific priority level. Also this check can be started from highest priority applications and iteratively it can come down to lowest priority applications.
- User-limit-factor: Same explanation applicable for user-limit-factor also.

1.5.5 Access Control Lists for Application-priority:

User-level access permission to use different application-priorities can be configured and

controlled. This is to avoid situations where all users try running VERY_HIGH applications in the cluster and thus degrading the value of priorities.

- Access Control Lists can be set per priority-label within each queue.
- Separate ACL configurations are required for Fair and Capacity schedulers.

1. How ACL's can be configured per queue level?

Below is an example configuration that can be added in capacity scheduler configuration file for each Queue level.

yarn.scheduler.capacity.root.<queue_name>.<priority_label>.acl=user1,user2

Under each configuration, a set of users and groups can be configured and only those users/groups should be able to run that specified priority applications.

Note: Similar configuration can be kept under fair scheduler configuration file to define ACLs per queue.

2. Behavior together with queue acls:

If Queue level ACLs are configured, then this priority-based acl configuration should be a subset of that configuration – a user should have both queue-acl and priority-acl to be able to submit applications at that priority in that queue.

Note: Under both scenarios mentioned above, such application which are not able to meet the ACL criteria can be marked as Rejected. There is also an alternative to fall back to default priority label of queue and continue submitting application. This can be confirmed.

1.5.6 View Application-priority from RM Web UI:

Priority can be displayed in RM Web UI. This will help in getting the details of the submitted application-priority.

Few Points of Note:

1. **MAPREDUCE-314 Starvation problem**: Priority inversion problem can be solved by using below way.
    a. Applications Master's head room will be modified to reflect the priority change.

Based on this input from Resource manager, respective application masters can act on the change.

2. **Preemption:** Without preemption, if a lower priority application is using full Queue's capacity, then a higher priority application submitted in the same Queue has to wait. This wait period will depend on how fast the lower priority application will be completed. During this time Capacity Scheduler will ensure that resource allocation will not be happening for lower priority application further.

   Priority based preemption policy can help in this regard, a sub JIRA will be opened to track this.