Introduction

https://issues.apache.org/jira/browse/YARN-4314

AM Container wait-time: When a client submits an application, there is a time period when the AM attempt is in an unassigned state. The time difference between AM scheduled time and AM start time is called AM container wait time. Individual attempts store AM container wait-time for this particular attempt. The app level AM container wait time is aggregated over all the attempts.

Container wait-time: When an attempt is in running state, It asks for containers with various capabilities. The time for which the AM waits for allocation of containers it called container wait time. It's calculated as the summation of number_of_containers multiply by the time elapsed to get those containers. This is also stored for an attempt. The app level container wait-time is summed over all the attempts.

Design

AM container wait-time: At attempt level, we have the attempt started time. When an attempt is in AMContainerAllocatedTransition, scheduled time is set to the current time. Scheduled time is subtracted from the start time to get the AM container wait time. The delta time is updated in the queue metrics and attempt metrics.

Container wait-time: AM asks containers of different capabilities from RM as a part of the AM-RM communication. Let's assume that the total number of currently pending containers are **n** and the last updated time for wait time is **t1,** and the current time is **t2.** Delta additive wait time will be **n*(t2-t1).** This wait time will be added to queue metrics and attempt metrics.

Individual attempt's metrics are added to get the app level metrics.

Adding wait-time in the queues: Both metrics are present for an attempt. Whenever they are updated, the delta time is also added to the corresponding queues. Queue metrics are the part of the JMX metrics which are emitted as they calculated.

Application Recovery

Both container and AM container wait time are stored in the RMState store as a part of the attempt data. So whenever the RM restarts or switched over, these metrics can be recovered. The state will only be saved when there is a state transition. For running attempts, the state data will not be saved. Upon recovery the wait times for all the in-flight apps will be set to -1 to show that the correct wait time of this application cannot be known.

Queue change of an application

AM container wait time is calculated by subtracting start time from scheduled time. In the case of queue movement, we will have to add the delta time period to the current queue and reset the start time. We can't reset the start time as it is a constant. To solve that there is a temporary variable called last_start_time which will be set to current time the case where we need to reset the start time. On the other hand, container-wait-time is not affected as it is added to the queue in every heartbeat.

Few scenario:

- If an app is in NEW, SUBMITTED or SCHEDULED state: The AM container wait time will be added to the current queue, and the last_start_time will be set to current time.
- If an app is in any other state: The am container wait time should not be added.

Salient features/Limitations:

1. All containers are considered equal in the terms of wait time regardless of capabilities they have asked for. This means currently there is no way to know, the wait time of containers requesting different capabilities.
2. RM restarts or switching over makes wait time for in-flight apps invalid.
3. AM container wait time is not reflected unless it is scheduled.
4. Both metrics are recoverable if an application is finished.