Moab
Job starvation and/or deferral

Problem:

Jobs with priority reservations at or near the top of the queue, which should run (especially large jobs) get skipped over, remain idle in the eligible queue, or change to deferred status and move to the blocked queue.

Affected Versions:

All.

Solution:

Scenario 1: temporary resource spikes cause Moab to occasionally delay job starts. You may see something like this in checkjob output:

Message[0] 15 nodes unavailable to start reserved job after 30 seconds (reserved node n1234 is in state 'Idle' - check node)

To resolve this, try setting NODEAVAILABILITYPOLICY DEDICATED in moab.cfg. After restarting Moab with this policy in effect, the scheduler will ignore reports of compute node usage spikes, and rely much more on its own accounting of cluster resources when determining where and when to run jobs. This may result in jobs running on nodes that truly do have elevated load, thereby decreasing job performance, but should help with deferrals due to things like transient load spikes reported by Torque. Adaptive still recommends putting other policies and monitoring tools (e.g., via nagios and logstash) in place to increase administrator awareness of node health outside of Moab, as well as the Warewulf Project's Node Health Check script:
http://go.lbl.gov/nhc

Also, Moab has various timing parameters related to retrying delayed and deferred jobs: DEFERSTARTCOUNT, DEFERTIME, RESERVATIONRETRYTIME, NODEFAILURERESERVETIME, and JOBRETRYTIME. For example: if you'd like to have the system assume that conditions leading to job deferral will self-resolve quickly, you might try increasing DEFERSTARTCOUNT (default == 1) and lowering DEFERTIME (default 1:00:00).

Scenario 2: a large job looks like it should be able to run, since it gets a job reservation with the highest priority; however, smaller jobs continue to backfill, jumping over the large job. In this scenario, the cluster never drains enough nodes for the large job to run.

To resolve this, either set "RSVSEARCHALGO WIDE" in moab.cfg, and then recycle, or modify the job with:
"mjobctl -m flags=widersvsearchalgo ". This is a known issue. For historical reasons, Adaptive has elected not to make this the default behavior, but long-term observation demonstrates the safety of this parameter. If this works, you can "set it and forget it", leaving it in moab.cfg for good.

If neither of these solutions help, you'll want to report the incident in a support ticket.

Unique solution ID: #1007
Author: Rick McKay
Last update: 2016-12-28 19:27