Moab

How to configure scheduling leeway for transient load

Problem: Moab indicates that nodes are Idle in mdiag -n, checknode, showq, etc., but in reality they have higher load, and jobs will not run there. This leads to confusion.

Solution: By default, Moab takes load into account when scheduling jobs (NODEAVAILABILITYPOLICY COMBINED). However, while Moab node states correspond to those reported up from Torque, they don't necessarily indicate availability for scheduling. When reported processor utilization of any specific compute node exceeds 0.01 ("loadave=..." in pbsnodes) at the start of an iteration, Moab will mark that node as Busy. This can happen with transient load and, of course, on systems not entirely dedicated to running workload only under Moab/Torque.

This article suggests a couple of possible options for cluster administrators to address this situation.

1) Configure NODEMAXLOAD in moab.cfg to permit some transient load.

Example:

NODEMAXLOAD 0.75

With this setting, Moab will to continue scheduling jobs on nodes with load as high as 0.75, and will mark the node as busy once the load reaches that threshold

The NODEMAXLOAD documentation states:

"Specifies the maximum CPU load on an idle or running node. If the node's load reaches or exceeds this value, Moab will mark the node busy."

http://docs.adaptivecomputing.com/9-0-3/MWM/help.htm#topics/moabWorkloadManager/topics/appendices/a.aparameters.html#nodemaxload

Notes:

- NODEMAXLOAD is an absolute "core usage" value, not a percentage.
- NODEMAXLOAD is a global setting. To limit the scope of this setting, or to set a different value for specific nodes, like so:

NODECFG[node123] MAXLOAD=0.85

- To negate the global setting on specific nodes, set MAXNODE to -1.
- Keep in mind: because of inherent delays between Moab, pbs_server, and pbs_mom processes running on the nodes, you should not treat the output of Moab client commands as real-time information. (However, for many commands, you may include "--blocking" to bypass cached data and get the

most up-to-date information possible.)

2) Deploy the LNBL Node Health Check script configured with "check_ps_loadavg".

This will mark the nodes as down (initially in Torque, and subsequently in Moab) until the load drops below the specified threshold. Refer to the project page here:

https://github.com/mej/nhc

This page includes all the documentation.

3) Configure Moab to ignore transient load altogether by setting NODEAVAILABILITYPOLICY DEDICATED:PROCS.

http://docs.adaptivecomputing.com/9-1-1/MWM/help.htm#topics/moabWorkloadManager/topics/appendices/a.aparameters.html#nodeavailabilitypolicy

This instructs Moab to ignore load when scheduling, and only take its own tracking of configured & scheduled resources into account. If using this type of configuration, it is best to include additional, external method for monitoring node health (such as #2 above) that automatically performs an action to deal with the node.

See also:

Job starvation and/or deferral

Why will my job not start when there is no other job on the compute resource but CPU usage on that node is high?

Unique solution ID: #1169
Author: Rick McKay
Last update: 2017-05-18 01:39