Moab

What is the difference between node locked gres and global node gres?

Issue: What is the difference between node locked gres and global node gres?

Symptom:

If Moab is setup to have a gres on a specif node, such as a file-system, then those greses are tied directly to the task count of the job.

Example:

NODECFG[node001] GRES=ramdisk:6

When submitting a job as follows: 'msub -l walltime=60 -l nodes=1:ppn=4 -l other=rmdisk:6', Moab will miltiply the ''tast count' by the gres, other or software parameter. So in this case 4 * 6 = 24. The 24 become the total gres needed to satisfy the jobs need. In this case the node only has a totla of 6 and will never be able to satisfy the job. The checkjob outputs that the job is blocked on a gres.

Node locked generic resources reside on the same requirement and task definition as the processors. It has to be this way in order to ensure that the processors and the generic resource are satisfied by the same node. If you break up the processor requirement and the generic resource requirement there is no guarantee that they will each be given the same node. Tasks are atomic and moab assumes node locked generic resources are part of the same task definition and requirement as the processors in order to guarantee they are satisfied by the same node.

Because they are on the same requirement the generic resource request inherits the taskcount of the processor request. This has limitations, for example, you can't request an arbitrary amount of the generic resource as it either has to be the same as the taskcount or a multiple of the taskcount.

Unfortunately, because of the way Moab is architected to scheduler tasks atomically this is something that cannot be changed.
In this case.

In the caset of **GLOBAL** node gres, the gres is shared between all node as a pool and the task multiplyer is on its own request. In other words there is a request for the global node of '-l nodes=1:ppn=1 -l gres=6' and a request for ' -l nodes=1:ppn=4' for the compute node.

Solution: Consider submitting a job so that the job uses an appropriate gres. In this case '-l nodel=1:ppn=2 -l otherramdisk:3'. You can also rais the node gres on the node so that the total can be satisfied.

Unique solution ID: #1074
Author: Jason Booth
Last update: 2015-10-13 21:58