Moab

How can I make sure jobs run within a single infiniband switch?

Problem:

MPI jobs should run within a single infiniband switch if at all possible, to maximize bandwidth between nodes.

Solution:

Using Moab nodesets, this can can be implemented easily.

First you will need to add a property or feature to each node in the pbs_server nodes file, to let Moab know what switch each node is connected to - we also add "all" to make sure we can reference all nodes in the system, eg:

server_priv/nodes:

node00 np=8 switcha all

node01 np=8 switcha all

node02 np=8 switcha all

node03 np=8 switcha all

node04 np=8 switchb all

node05 np=8 switchb all

node06 np=8 switchb all

node07 np=8 switchb all

Using these features we can put the following in moab.cfg:

NODESETPOLICY FIRSTOF

NODESETATTRIBUTE FEATURE

NODESETISOPTIONAL FALSE

NODESETLIST switcha,switchb,all

With this configuration, Moab will try to run a job on the nodes connected to switcha first, then switchb and if the job does not fit in either of those, it tries to run the job across all nodes on all switches.

What if I want to force a single users jobs to always run within a single switch?

In this case, we can reuse the configuration from above and add the following job template to moab.cfg:

JOBCFG[single.min] USER=fred

JOBCFG[single.set] NODESET=FIRSTOF:FEATURE:switcha,switchb

JOBMATCHCFG[island] JMIN=single.min JSET=single.set

Using this configuration, user fred's jobs will be forced to always run within a single switch.

What if I don't want jobs to go beyond one switch, ever?

You can simply leave out "all" from the NODESETLIST, like this:

NODESETLIST switcha,switchb

What if I want to allow jobs to span switches in general, but provide a way for users to request the job to only run on one switch?

In this case, we add another feature to all nodes, here I used singleswitch:

server_priv/nodes:

node00 np=8 switcha,singleswitch,all

....

....

We can keep the NODESET* settings from earlier, but add the following job template:

JOBCFG[single.min] RFEATURES=singleswitch

JOBCFG[single.set] NODESET=FIRSTOF:FEATURE:switcha,switchb

JOBMATCHCFG[island] JMIN=single.min JSET=single.set

With this configuration, the user can request the job to run within a single switch using:

msub -l nodes=2:ppn=8:singleswitch

Documentation link:

http://docs.adaptivecomputing.com/9-1-2/MWM/moab.htm#topics/moabWorkloadManager/topics/optimization/nodesetoverview.html

Moab

Author: Michael Aronsen
Last update: 2018-02-21 13:16