Network Storage and Job Scheduler

I. Network storage

1．HOME=/home/`whoami`

The $HOME directory of each user shares an SSD cluster, which gives a total of ~ 2M iops with basic measures to maintain the QoS.

On each management node and compute node are mounted the same $HOME export,　thus providing a consistent view of directory structure. With the current settings, changes made under $HOME on one client should be visible on all other client nodes within 3 seconds. In other words, global consistency is not guaranteed within this period.

For each client node, the peak performance of the $HOME file system is ~ 1M iops and ~ 12GB/s throughput.

Because of the low ratio between the peak client iops (~ 1M) and the peak cluster iops (~ 2M), running tasks that stress the metadata servers on more than a few compute nodes will slow down the complete SSD cluster.

It is therefore appreciated to remove unnecessary and duplicate entries under $HOME from $PATH, $LD_LIBRARY_PATH, $PYTHON_PATH, and etc., keeping them as short as possible.

This is because a program will need to search for a number of files in these paths each time it starts. Lengthy paths translate to lengthy program starts, resulting in less publications at increased cost in ￥.

Best practice guide for Presto users -- For short observations, always process *more than 10 data files each time calling realfft or accelsearch, instead of spending a few seconds to traverse LD_LIBRARY_PATH to start the program, but then only a fraction of a second to process a single data file before exit.*

2．/home.low.iops/`whoami`

This is the storage cluster of mechanical hard disks, recommended for storage of user specific files accessed at low iops and medium throughput. Typical figures are no more than a few hundreds of iops and up to 1 ~ 2 GB/s throughput.

3．/data31

Storage cluster for observation data. Sub-directories for each project can be accessed with prior authorization.

II. Job scheduler

The Portable Batch System (PBS) is used for job scheduling. Available queues are cu_fat, cu_slim, and gpu.

Resource limits for each queue are:

<table>
<tr><th></th><th>cu_slim</th><th>cu_fat</th><th>gpu</th></tr>
<tr><td>CPU cores per node</td><td>4</td><td>80</td><td rowspan="2"></td></tr>
<tr><td>CPU cores per job</td><td>4</td><td>160</td></tr>
<tr><td>GPU cards per node</td><td colspan="2"></td><td>4</td></tr>
<tr><td>Number of running jobs</td><td colspan="3">3</td></tr>
<tr><td>Number of queued jobs</td><td colspan="3">2</td></tr>
<tr><td>Time limit (job life time)</td><td>5000h</td><td colspan="2">24h</td></tr>
</table>

Notes:

1. Job queues must be specified explicitly. The default queue is “batch”. Jobs submitted to “batch” will be queued forever.

2. The gpu queue is only available to users registered for GPU resource. The following PBS directive is needed for GPU job requests:

```
# PBS -Wx=GRES:gpu@2
```

where “gpu@2” indicates that two GPU cards are needed to run this job.

3. For more information, please refer to the manual of the “Portable Batch System” and relate documentation.