Troubleshooting Jobs on Odyssey

Paul Edmon, PhD
ITC Research Computing Associate

Bob Freeman, PhD
Research & Education Facilitator
XSEDE Campus Champion
• Tackle PEND, FAIL, and slow performance issues
• Highlight different approaches to troubleshooting
• Arm you with useful SLURM and Unix commands

• Enable you to “Work smarter, better, faster”
PEND Problems

Why are you pending?

`squeue -u USERNAME -t PD`

`squeue -u USERNAME -t PD -o "%.18i %.9P %.8j %.8u %.2t %.10M %.6D %R %C"`

```
[bfreeman@rclogin07 ~]$ squeue -t PD | tail -n 20
 25793703_1 serial_re 1MUI-A_P ncheron PD 0:00 1 (Priority) 20
 25793729_4 serial_re 1MUI-B_P ncheron PD 0:00 1 (Priority) 20
 25793813_1 serial_re 10HR_Pro ncheron PD 0:00 1 (None) 20
 25793936_9 serial_re 2AQU_B_P ncheron PD 0:00 1 (None) 20
 25820727_7 serial_re CHEMBL18 ncheron PD 0:00 1 (Resources) 16
 25793863_6 serial_re 2AQU_A_P ncheron PD 0:00 1 (Priority) 20
 25794027_18 serial_re 4HLA_A_P ncheron PD 0:00 1 (Priority) 20
 25877428_47-50 shakgpu Banjanac ncheron PD 0:00 1 (Resources) 32
 25877480_1-50 shakgpu Banjanac ncheron PD 0:00 1 (Priority) 32
 25877531_1-50 shakgpu Banjanac ncheron PD 0:00 1 (Priority) 32
 25877583_1-50 shakgpu Banjanac ncheron PD 0:00 1 (Priority) 32
 25877635_1-50 shakgpu Banjanac ncheron PD 0:00 1 (Priority) 32
 25877688_1-50 shakgpu Banjanac ncheron PD 0:00 1 (Priority) 32
 25877742_1-50 shakgpu Banjanac ncheron PD 0:00 1 (Priority) 32
 25877794_1-50 shakgpu Banjanac ncheron PD 0:00 1 (Priority) 32
 25885657_50 shakhnovi 1HXW_Pro ncheron PD 0:00 1 (Resources) 20
 25902404_29-50 unrestric OMAx_Myr rmfernan PD 0:00 1 (Resources) 1
 25889308 vogelsber p3_30_13 ptorrey PD 0:00 1 (Resources) 64
 25890147 vogelsber p3_100_1 ptorrey PD 0:00 1 (Priority) 64
 25900499 vogelsber Aq-H-5_r kannan PD 0:00 2 (Dependency) 128
```
## PEND Problems

Why are you pending? Most common reasons are...

<table>
<thead>
<tr>
<th>Reason</th>
<th>Solution</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>Scheduler hasn’t gotten around to you yet. Hang tight...</td>
</tr>
<tr>
<td>Resources</td>
<td>The job is waiting for resources to become available</td>
</tr>
<tr>
<td>Priority</td>
<td>One or more higher priority jobs exist for this partition or reservation. Priority is influenced by age, job size, partition, QoS, and Fairshare.</td>
</tr>
<tr>
<td>Dependency</td>
<td>This job is waiting for a dependent job to complete (--dependency)</td>
</tr>
</tbody>
</table>
PEND Problems: Resources

- What did you request?
- What are the parameters of your SLURM submission script?
- `scontrol show jobid -dd JOBID`
What is your priority in the scheduling queue?

- `showq-slurm -U`
- `showq-slurm -p PARTITION -o`
PEND Problems: Fair-share

• What is a fair-share score?
  Score assigned to each lab that affects the scheduling priority of jobs. A user’s current and past usage is considered when determining the scheduling of job execution.

• How is it calculated?
  Score is based on usage and shares:
  \(0 \leq \text{Usage} \leq 1\), represents your proportional use of Odyssey
  \(\text{Shares} \sim = \text{slices of a pie, or the part of Odyssey that is yours}\)
  Premise: \(\text{Usage} = \text{Shares}\), you’ve hit your fairshare target

  \[
  \text{Fairshare Factor} = 2^{-\text{Usage/Shares}} \quad 0 < \text{FF} < 1
  \]

  When usage increases, FF decreases
  When usage decreases, FF increases
**PEND Problems: Fair-share**

- **What is my score?**

  `sshare -u USERNAME`

  ![Command Output]

  ![Command Output with FairShare Highlighted]

  ![Command Output with User Specific Output Highlighted]
FAIL Problems

• Determining the root cause is highly dependent on your ability to trace & document the problem

• Do you have log/error files?
  
  #SBATCH -o job.stdout.txt
  #SBATCH -e job.stderr.txt

• Did you use the -e parameter to get an error file?
  
  If omitted, STDERR is redirected to -o file or slum-JOBID.out

• Is the path non-existent?
  
  The path to place the output files must exist! Otherwise, SLURM won’t know what to do and will do nothing but FAIL.

• What else is going on in your SLURM submission script?
FAIL Problems

• Look at the fail trail in any log and/or error files.
• *Nota bene!* The last error listed may not be the root cause!
FAIL Problems

segmentation faults
shared/static library errors

- Are the correct software packages and versions loaded?
- Are there version conflicts?
- Are too many things loaded in your `.bashrc`?
Poor performance

• Much harder to diagnose!
• Transient issues:
  • other jobs on the node
  • network
  • storage
  • bad node
• Software install issue
• Code compile/optimization problem
**Poor Performance**

Slow run times? Most common reasons are...

<table>
<thead>
<tr>
<th>Reason</th>
<th>Diagnostics &amp; Potential solutions</th>
</tr>
</thead>
</table>
| SLURM submission errors | - Check --mem/--mem-per-cpu  
                          - Check cores (-n) + nodes (-N)  
                          - Try running job interactively               |
| Overloaded Storage    | - Is the mount point broken?  
                          - Is the storage being hammered? Don’t run out of home/lab directories.  
                          - Is one file being accessed by hundreds of jobs? Make copies.  
                          - Too many files in one directory?  
                          - Check status.rc                              |
| Overloaded Node       | - Too much Input/Output for the given network interface  
                          - Overuse of CPUs relative to SLURM request  
                          - Check status.rc; use squeue and top             |
| Sick Node             | - Memory errors  
                          - /scratch filled up  
                          - use lsload to look at node diagnostics          |
| Sick Code             | - Monitor/trace program execution with strace  
                          - Look at code performance with perf                |
Poor performance: Nagios Compute

Check the state of the systems at https://status.rc.fas.harvard.edu

Heavy load, but normal

Heavy load, oversubscribed
Poor performance: Nagios Storage

Includes storage like /n/regal and /n/holyscratch
Poor performance: Node health

Check the health of the compute node (abused by other jobs?)

- need to figure out which compute node your job is on
  
  `squeue -j JOBID`

- quick check using [https://status.rc.fas.harvard.edu](https://status.rc.fas.harvard.edu)

- look at load using `lsload`

- look at jobs on node using `squeue`

- log in to node and profile
  
  - `top`
  
  - `strace`
  
  - `perftrace`
Poor performance: Node health

Use `lsload` to examine the utilization of individual nodes

```
[bfreeman@rclogin05 ~]$ lsload

<table>
<thead>
<tr>
<th>Hostname</th>
<th>Cores</th>
<th>InUse</th>
<th>Ratio</th>
<th>Load</th>
<th>Mem</th>
<th>Alloc</th>
<th>State</th>
</tr>
</thead>
<tbody>
<tr>
<td>aag01</td>
<td>64</td>
<td>64</td>
<td>100.0</td>
<td>63.68</td>
<td>258</td>
<td>208</td>
<td>ALLOCATED</td>
</tr>
<tr>
<td>aag02</td>
<td>64</td>
<td>64</td>
<td>100.0</td>
<td>64.20</td>
<td>258</td>
<td>184</td>
<td>ALLOCATED</td>
</tr>
<tr>
<td>aag03</td>
<td>64</td>
<td>56</td>
<td>87.5</td>
<td>58.89</td>
<td>258</td>
<td>120</td>
<td>MIXED+COMPLETING</td>
</tr>
<tr>
<td>aag04</td>
<td>64</td>
<td>56</td>
<td>87.5</td>
<td>62.11</td>
<td>258</td>
<td>172</td>
<td>MIXED+COMPLETING</td>
</tr>
</tbody>
</table>

holy2a01201  64  64  100.0  64.00  264  181  ALLOCATED
holy2a01202  64  64  100.0  54.34  264  250  ALLOCATED
holy2a01203  64  62  96.9  63.45  264  210  MIXED
holy2a01204  64   0   0.0  0.20  264   0   IDLE+DRAIN /var/local/mem_error/holy2a01204_2
  2014-11-23 mem_error exists [root@2014-12-02T20:24:01]    ERROR Health check failed: Memory Check failed!
holy2a01205  64  64  100.0  49.60  264  132  ALLOCATED
holy2a01206  64  64  100.0  64.14  264  192  ALLOCATED
holy2a01207  64  64  100.0  64.32  264  186  ALLOCATED
holy2a01208  64  32  50.0  95.49  264  210  MIXED
holy2a02101  64  25  39.1  18.07  264  55   MIXED
holy2a02102  64  23  35.9  19.13  264  96   MIXED
holy2a02103  64  51  79.7  30.10  264  119  MIXED
holy2a02104  64  61  95.3  55.40  264  150  MIXED
holy2a02105  64  27  42.2  19.09  264  15   MIXED+DRAIN Closed to replace DIMM - MRE 12/1/

2014 [methier@2014-12-01T12:49:20]

holy2a02106  64  47  73.4  48.92  264  124  MIXED
holy2a02107  64  45  70.3  45.33  264  71   MIXED
holy2a02108  64  56  87.5  56.90  264  201  MIXED
holy2a02201  64  64  100.0  66.14  264  114  ALLOCATED
holy2a02202  64  64  100.0  66.34  264  256  ALLOCATED
holy2a02203  64  61  95.3  56.34  264  111  MIXED
holy2a02204  64  34  53.1  33.79  264  21   MIXED
holy2a02205  64  58  90.6  50.30  264  162  MIXED
```
Poor performance

What other jobs are running on my node?

`squeue -w NODENAME`

e.g. Ensure that the # of CPUs requested = # CPUs used
Poor performance

Log in to node and top

ssh USERNAME@NODENAME
top

```
top - 10:16:12 up 57 days, 22:49, 1 user, load average: 123.95, 123.25, 122.67
Tasks: 1410 total, 25 running, 1384 sleeping, 0 stopped, 1 zombie
Cpu(s): 57.7%us, 5.4%sy, 0.0%ni, 36.9%id, 0.0%wa, 0.0%hi, 0.0%si, 0.0%st
Mem: 264498560k total, 116480296k used, 148018264k free, 404116k buffers
Swap: 8388600k total, 1172k used, 8387428k free, 38103800k cached

<table>
<thead>
<tr>
<th>PID</th>
<th>USER</th>
<th>PR</th>
<th>NI</th>
<th>VIRT</th>
<th>RES</th>
<th>SHRS</th>
<th>S</th>
<th>%CPU</th>
<th>%MEM</th>
<th>TIME+</th>
<th>COMMAND</th>
</tr>
</thead>
<tbody>
<tr>
<td>11608</td>
<td></td>
<td>20</td>
<td>0</td>
<td>71.2g</td>
<td>47g</td>
<td>1064</td>
<td>S</td>
<td>1654.2</td>
<td>19.0</td>
<td>16617:17</td>
<td>segemehl.x</td>
</tr>
<tr>
<td>10613</td>
<td></td>
<td>20</td>
<td>0</td>
<td>5427m</td>
<td>5.1g</td>
<td>8208</td>
<td>R</td>
<td>100.0</td>
<td>2.0</td>
<td>266:24.42</td>
<td>R</td>
</tr>
<tr>
<td>10639</td>
<td></td>
<td>20</td>
<td>0</td>
<td>5770m</td>
<td>5.4g</td>
<td>8192</td>
<td>R</td>
<td>100.0</td>
<td>2.1</td>
<td>267:51.29</td>
<td>R</td>
</tr>
<tr>
<td>10651</td>
<td></td>
<td>20</td>
<td>0</td>
<td>5866m</td>
<td>5.5g</td>
<td>8224</td>
<td>R</td>
<td>100.0</td>
<td>2.2</td>
<td>267:17.18</td>
<td>R</td>
</tr>
<tr>
<td>19870</td>
<td></td>
<td>20</td>
<td>0</td>
<td>389m</td>
<td>259m</td>
<td>5644</td>
<td>R</td>
<td>100.5</td>
<td>0.1</td>
<td>118:21.74</td>
<td>ko</td>
</tr>
<tr>
<td>19871</td>
<td></td>
<td>20</td>
<td>0</td>
<td>316m</td>
<td>187m</td>
<td>6276</td>
<td>R</td>
<td>100.5</td>
<td>0.1</td>
<td>118:26.03</td>
<td>ko</td>
</tr>
<tr>
<td>19872</td>
<td></td>
<td>20</td>
<td>0</td>
<td>317m</td>
<td>188m</td>
<td>6296</td>
<td>R</td>
<td>100.5</td>
<td>0.1</td>
<td>118:24.72</td>
<td>ko</td>
</tr>
<tr>
<td>19873</td>
<td></td>
<td>20</td>
<td>0</td>
<td>316m</td>
<td>186m</td>
<td>5532</td>
<td>R</td>
<td>100.5</td>
<td>0.1</td>
<td>118:25.13</td>
<td>ko</td>
</tr>
<tr>
<td>19874</td>
<td></td>
<td>20</td>
<td>0</td>
<td>316m</td>
<td>186m</td>
<td>6272</td>
<td>R</td>
<td>100.5</td>
<td>0.1</td>
<td>118:24.53</td>
<td>ko</td>
</tr>
<tr>
<td>19875</td>
<td></td>
<td>20</td>
<td>0</td>
<td>316m</td>
<td>187m</td>
<td>6936</td>
<td>R</td>
<td>100.5</td>
<td>0.1</td>
<td>118:25.58</td>
<td>ko</td>
</tr>
<tr>
<td>19876</td>
<td></td>
<td>20</td>
<td>0</td>
<td>316m</td>
<td>188m</td>
<td>6916</td>
<td>R</td>
<td>100.5</td>
<td>0.1</td>
<td>118:25.19</td>
<td>ko</td>
</tr>
<tr>
<td>19877</td>
<td></td>
<td>20</td>
<td>0</td>
<td>316m</td>
<td>187m</td>
<td>6902</td>
<td>R</td>
<td>100.5</td>
<td>0.1</td>
<td>118:25.35</td>
<td>ko</td>
</tr>
<tr>
<td>19878</td>
<td></td>
<td>20</td>
<td>0</td>
<td>316m</td>
<td>185m</td>
<td>6264</td>
<td>R</td>
<td>100.5</td>
<td>0.1</td>
<td>118:25.68</td>
<td>ko</td>
</tr>
<tr>
<td>19879</td>
<td></td>
<td>20</td>
<td>0</td>
<td>316m</td>
<td>188m</td>
<td>6944</td>
<td>R</td>
<td>100.5</td>
<td>0.1</td>
<td>118:25.67</td>
<td>ko</td>
</tr>
</tbody>
</table>
```
Poor performance

Log in to node and monitor code execution

ssh USERNAME@NODENAME
ps aux | grep USERNAME
(obtain relevant process ID)
strace -p PROCESSID

```
Odelia:~ bfreeman$ ssh holy2a01208
[bfreeman@holy2a01208 ~]$ strace -p 29175
Process 29175 attached - interrupt to quit
select(1, [0], NULL, NULL, {2, 430607}) = 0 (Timeout)
fcntl(0, F_SETFL, O_RDONLY|O_NONBLOCK|O_LARGEFILE) = 0
read(0, 0x7ffe977125c, 1) = -1 EAGAIN (Resource temporarily unavailable)
ioctl(0, TCFLSH, 0) = 0
fcntl(0, F_SETFL, O_RDONLY|O_LARGEFILE) = 0
open("/proc", O_RDONLY|O_RDONLY|O_NONBLOCK|O_DIRECTORY|O_CLOEXEC) = 8
getdents(8, /* 1338 entries */ , 32760) = 32760
stat("/proc/1", {st_mode=S_IFDIR|0555, st_size=0, ...}) = 0
open("/proc/1/stat", 0_RDONLY) = 9
read(9, "1 (init) S 0 1 1 0 -1 4202752 28", ..., 1023) = 205
close(9) = 0
open("/proc/1/statm", 0_RDONLY) = 9
read(9, "5359 394 316 35 0 103 0\n", 1023) = 24
close(9) = 0
stat("/proc/2", {st_mode=S_IFDIR|0555, st_size=0, ...}) = 0
open("/proc/2/stat", 0_RDONLY) = 9
read(9, "2 (kthread) S 0 0 0 0 -1 214961", ..., 1023) = 158
close(9) = 0
```
Please talk to your peers, and
We wish you success in your research!

http://rc.fas.harvard.edu
https://portal.rc.fas.harvard.edu
rchelp@fas.harvard.edu
@fasrc

Harvard Informatics
@harvardifx