Job & Queue Management Tools

SciNet User Group

March 11, 2015

Outline

- Go over what tools are available to monitor jobs & the queue
- Solicit ideas for improvement

Standard tools

showq: what's in the queue?

```
$ showq

active jobs------------------------
JOBID              USERNAME      STATE PROCS   REMAINING            STARTTIME

27669964           atlaspt3    Running     1  1:22:59:46  Tue Mar 10 23:14:24
27669961           atlaspt3    Running     1  1:22:59:46  Tue Mar 10 23:14:24
27669973           atlaspt3    Running     1  1:22:59:46  Tue Mar 10 23:14:24
27669963           atlaspt3    Running     1  1:22:59:46  Tue Mar 10 23:14:24
27669972           atlaspt3    Running     1  1:22:59:46  Tue Mar 10 23:14:24
...
```

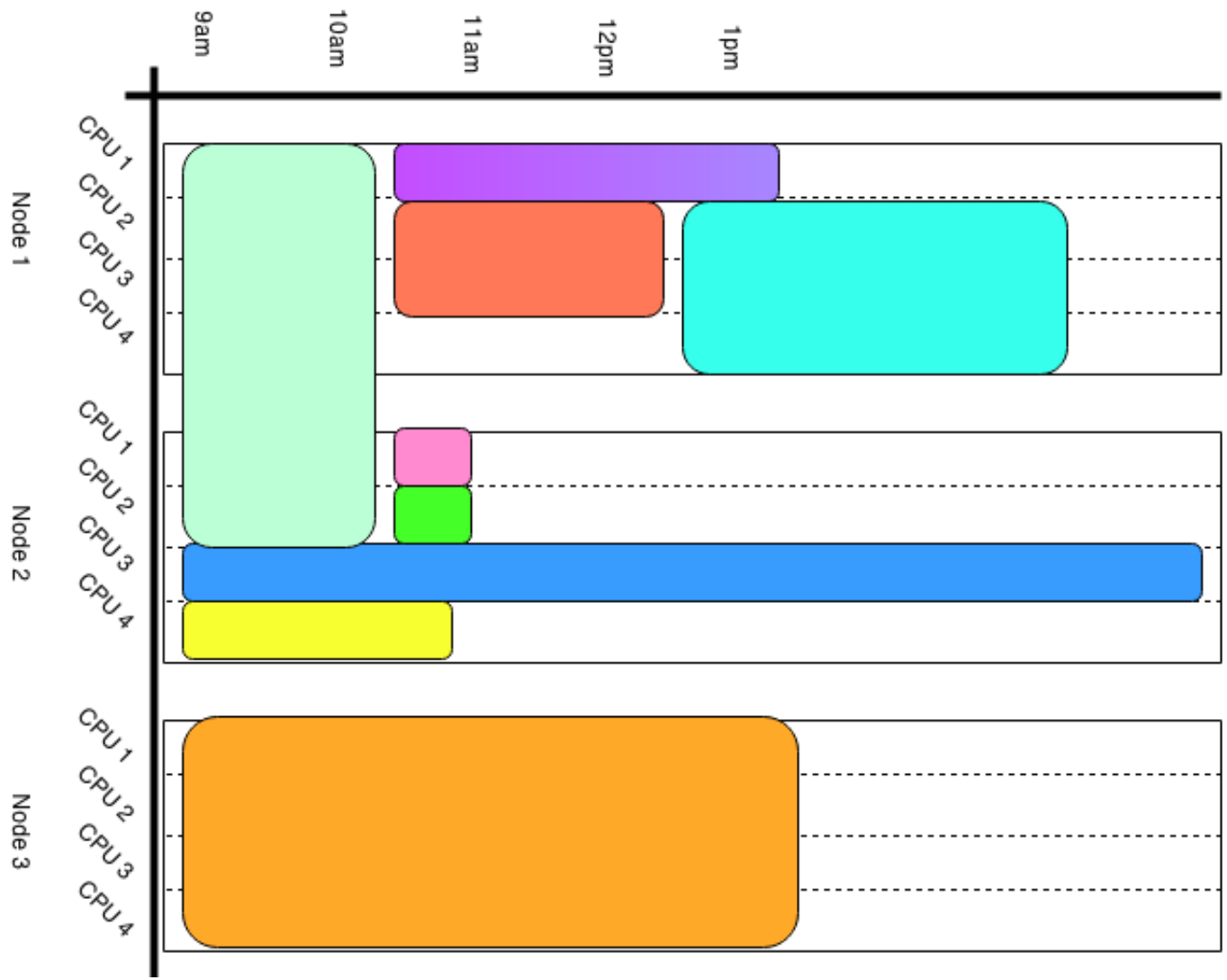
Jobs are sorted by priority.

Job states:

- Running: uh, running
- Eligible: waiting to be run
- Blocked: the job violates some condition:
  - 'Idle' means you have max jobs already in the queue
  - 'Hold' means its waiting for a dependent job

showbf: what can i run right now?

Sometimes gaps appear in the schedule, as nodes are reserved for large jobs.



showbf: what can i run right now?

If your job fits into a gap, it can run right away. These kind of gap-filling jobs are often referred to as 'backfill'.

```
$ showbf
Partition     Tasks  Nodes      Duration   StartOffset       StartDate
---------     -----  -----  ------------  ------------  --------------
ALL             146     23      00:36:26      00:00:00  10:46:48_03/09
ALL              18      7       1:35:40      00:00:00  10:46:48_03/09
ALL              10      6      INFINITY      00:00:00  10:46:48_03/09
torque          146     23      00:36:26      00:00:00  10:46:48_03/09
torque           18      7       1:35:40      00:00:00  10:46:48_03/09
torque           10      6      INFINITY      00:00:00  10:46:48_03/09
```

Any job with:

- nodes <= 23 and walltime <= 36m26s
- nodes <= 7 and walltime <= 1h35m40s
- nodes <= 6

will run immediately.

showstart: when will my job start?

```
$ showstart 12345678 -e all
job 12345678 requires 64 procs for 1:16:00:00

Estimated Rsv based start in                00:25:36 on Mon Mar  9 17:46:43
Estimated Rsv based completion in          1:16:25:36 on Wed Mar 11 09:46:43

Estimated Priority based start in             9:44:39 on Tue Mar 10 03:05:46
Estimated Priority based completion in     2:01:44:39 on Wed Mar 11 19:05:46

Estimated Historical based start in        2:10:26:29 on Thu Mar 12 03:47:36
Estimated Historical based completion in   4:02:26:29 on Fri Mar 13 19:47:36

Best Partition: DDR
```

- **Rsv**: earliest possible start time
- **Priority**: if no new jobs are added to the queue before starting
- **Historical**: estimate based on historical averages

showstats: average job properties

```
$ showstats -u user123
statistics initialized Fri Dec  5 09:42:20

         |------ Active ------|---------------------------------- Completed
user      Jobs Procs ProcHours   Jobs    %    PHReq    %    PHDed    %   FSTgt
user123      0     0      0.00    115 100.00 702.6K 100.00 228.0K 100.00 -----

         ... ---------------------------------|
user     ...   AvgXF  MaxXF  AvgQH  Effic  WCAcc
user123         0.51   1.56   1.28  60.71  44.03
```

- PHReq: Proc-hours requested
- PHDed: Proc-hours dedicated
- AvgXF: Average expansion factor (runtime + queuetime)/walltime
- MaxXF: Max expansion factor
- Effic: Job efficiency
- WCAcc: Walltime accuracy

SciNet tools

Available from "module load extras"

diskUsage: filesystem usage summary

```
-------------------------|----------------|-----------------------
.                        |   Space Limits |        File Limits
fs         name      type |     size quota |    files quota avg_size
-------------------------|----------------|-----------------------
home       scinet    GRP  |  871.24G    __ |  8036671     __  113.67K
scratch    scinet    GRP  |   26.40T   80T |  1760236   10M   15.73M
scratch2   scinet    GRP  |    0.00K   80T |        1   10M    0.00K
project    scinet    FSET |  408.51G    5T |   305454    1M    1.37M
archive    scinet    GRP  |    4.66T  100T |  4815233   10M    1.01M
reserved1  scinet    GRP  |   12.78M    __ |      179    __   73.12K
-------------------------|----------------|-----------------------
home       nolta     USR  |    4.15G  200G |   113037    __    38.46K
scratch    nolta     USR  |   72.05G   50T |    53623    2M    1.38M
scratch2   nolta     USR  |   23.00M   20T |     4061    1M    5.80K
project    nolta     USR  |    0.00K    __ |        0    __    0.00K
project2   nolta     USR  |   52.78T    __ |  1658958    __   33.36M
-------------------------|----------------|-----------------------
```

Important to note how many files you have, not just total space.

Parallel filesystems hate having lots of small files; prefer a few big files.

quota: shorter diskUsage

```
$ quota
Retrieving user quotas for user nolta:

-----------------------------------------------------------
Filesystem         size       quota     #files      limit
-----------------------------------------------------------
home             4.15GB       200GB     113040         __
scratch         72.05GB        50TB      53623         2M
scratch2        23.00MB        20TB       4061    1000000
project          0.00KB          __          0         __
project2        52.78TB          __    1658958         __
-----------------------------------------------------------
```

qsum: summary of job queue by user

`showq` can be a bit verbose.

```
$ qsum
---------------------------------------------
qsum: queue summarized by user
      collected at Wed Mar 11 10:08:09 EDT 2015
-----------+--------------+--------------------
   RUNNING   |   REQUESTED  | USER      GROUP
 Cores Jobs  |  Cores Jobs  | NAME      NAME
-----------+--------------+--------------------
  2672   334 |   3992   499 | jpa14     dgrant
  2400     7 |    200     1 | vbooth    vbooth
  1888    15 |      0     0 | jtitanta  mkarttu
  1728     9 |      0     0 | zhangt68  mthomson
  1664     3 |      0     0 | levineza  wadsley
  1600     1 |      0     0 | ltobal    groth
  1344     7 |      0     0 | kholghy   mthomson
  1056     2 |      0     0 | ken       groth
  1024     1 |      0     0 | upierre   siemens
  1024     1 |      0     0 | mbehzad   ashgriz
   968     5 |      0     0 | treist    zingg
   899   892 |    717   507 | {atlas}
   832     2 |      0     0 | neaves    mthomson
   768    12 |  62144   971 | trangdo   jchoy4
   736     1 |   8192     2 | malvarez  pen
   720     3 |      0     0 | skhosrav  zingg
   672    10 |      0     0 | cwang     jpolanyi
   672     3 |      0     0 | armin     mthomson
   640    10 |      0     0 | haojia    csingh
```

jobperf: what's the per-node performance of a job?

```
$ jobperf 12345678
-------------------------------------------------------------------------------------
                   RUNNING         IDLE      USER       MEMORY(MB)   PROCESS NAMES
    HOSTNAME     #  %CPU  %MEM   DISK SLEEP   NAME       USED AVAIL   (excl:bash,ssh,ssh
-------------------------------------------------------------------------------------
gpc-f109n001     1    2%  0.0%      0    4    user123    2196  13872
gpc-f109n002     1    2%  0.0%      0    2    user123    1771  14296
gpc-f109n003     1    2%  0.0%      0    2    user123    1594  14474
gpc-f109n004     1    0%  0.0%      0    2    user123    1592  14476
gpc-f109n005    11  223%  0.0%     16    4    user123    1621  14447   26*python
gpc-f109n006     8  148%  2.4%     17    4    user123    1683  14385   24*python
gpc-f109n007     1    2%  2.4%      0   28    user123    1863  14204   24*python
gpc-f109n008     1    2%  2.4%      0   28    user123    1838  14229   24*python
-------------------------------------------------------------------------------------
```

more job scripts

```
jobError <jobID | jobNAME>
    displays on realtime the error output of a given job

jobOutput <jobID | jobNAME>
    displays on realtime the standard output of a given job

jobcd <jobID | jobNAME>
    allows users to quickly move into the working directory of a given job

jobscript <jobID | jobNAME>
    displays the submission script used when submitting a given job

jobssh <jobID | jobNAME>
    allows users to connect to the head-node of a given job

jobtop <jobID | jobNAME>
    allows users to see "top" on the head-node of a given job
```

Experimental tools

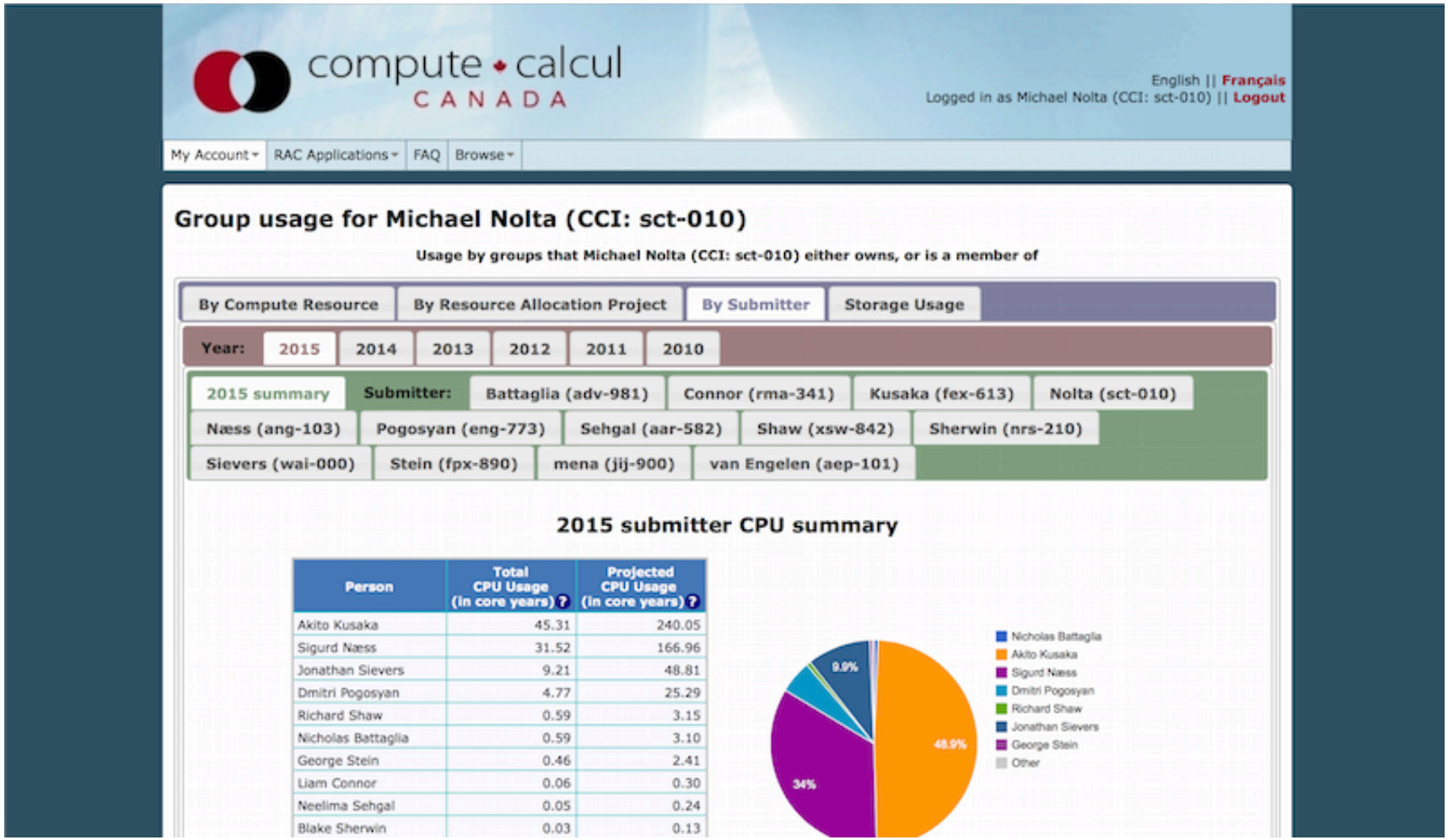
Things we're working on

Experimental: why aren't my jobs running?

- Job priority is determined by:
  - your allocation;
  - cycles used in the last 14 days.
- An allocation is a percentage of the machine.
- Allocations are handed out by the RAC every fall.
- Currently the default allocation is 0.013% of the GPC (~4 cores).
- Use more/less than your allocation in a 14 day window, and your priority increases/decreases.

How much of your total allocation have you used?

Go to `ccdb.computecanada.ca > My Account > View Group Usage`



Experimental: why aren't my jobs running?

Currently no way for users to easily understand why their priority is what it is.

```
$ /scinet/gpc/scinetcli/scinet usage

Account ifk-120-af has used 118.9% of its rolling allocation.
The trailing daily usage graph looks like this: ▂▃▅▂▁▁▁▁▁▁▁▁▁▁
As a result, jobs are unlikely to run for the next 2 days.

+----------+-------------------------------------+
| User     |   Trailing 14 day usage [node-days] |
|----------+-------------------------------------|
| akito    |                          1476.41    |
| sievers  |                           308.068   |
| sigurdkn |                            78.6308  |
| nsehgal  |                             1.479   |
| bsherwin |                             1.14469 |
| gstein   |                             0.0322917 |
| nolta    |                             0.00871528 |
+----------+-------------------------------------+
```

Thanks!