LSF at SLAC

Using the SIMES Batch Cluster

Neal Adams
Stanford Linear Accelerator Center
neal@slac.stanford.edu

Useful LSF Commands

bsub	submit a batch job to LSF
bjobs	display batch job information
bkill	kill batch job
bmod	modify job submission options
bqueues	display batch queue information
busers	displays information about batch users
lshosts	display LSF host information

For more details use: *man <command_name>*.

Useful LSF Commands

- bqueues

```
87 iris01 neal/bin> bqueues
QUEUE_NAME      PRIO STATUS          MAX JL/U JL/P JL/H NJOBS  PEND   RUN  SUSP
…
simesq          192  Open:Active       -    -    -    -   728     0   728     0
simesgpuq       191  Open:Active       -    -    -    -    64     0    64     0
…
short           185  Open:Active       -    -    -    -     0     0     0     0
medium          180  Open:Active       -    -    -    -   153   102    51     0
long            175  Open:Active       -    -    -    -   897   757   140     0
xlong           170  Open:Active       -    -    1    2  1636  1359   277     0
xxl             165  Open:Active     160   64    -    1    56     1    55     0
…
```

- busers

```
85 iris01 neal/bin> busers
USER/GROUP          JL/P    MAX  NJOBS   PEND    RUN  SSUSP  USUSP    RSV
neal                   -      -      0      0      0      0      0      0

79 sprocket sf/neal> busers kemper
USER/GROUP          JL/P    MAX  NJOBS   PEND    RUN  SSUSP  USUSP    RSV
kemper                 -      -    384      0    384      0      0      0
```

Useful LSF Commands

- lshosts

```
50 sprocket sf/neal> lshosts
HOST_NAME      type    model  cpuf ncpus maxmem maxswp server RESOURCES
farmboss1     LINUX AMD_2400   6.7     4 15976M 16386M    Yes (linux linux64 rhel40 master)
farmboss2     LINUX AMD_2400   6.7     4 15976M 16386M    Yes (linux linux64 rhel40  master)
farmnfs        SUN5   UF_900   2.8     2  4096M  7209M    Yes (solaris sol9 master)
farmhand       SUN5  UT1_440   1.0     1   256M  2220M    Yes (bcs solaris sol9)
sunlics1       SUN5  UT1_440   1.0     2  2048M  5731M    Yes (lics solaris sol10)
sunlics2       SUN5  UT1_440   1.0     2  2048M  3673M    Yes (lics solaris sol10)
sunlics3       SUN5  UT1_440   1.0     2  2048M  5726M    Yes (lics solaris sol10)
sprocket      LINUX   PC_200   0.5     1  2009M  4094M    Yes (linux linux64 rhel40 dungheap)
adam         MACOSX  G5_2000   4.8     -      -      -    Yes (macosx ppc_darwin)
[…]

51 sprocket sf/neal> lshosts simes0001
HOST_NAME      type    model  cpuf ncpus maxmem maxswp server RESOURCES
simes0001      LINUX INTEL_26  11.0     8 15918M 32767M    Yes (bs linux linux64 rhel50 simes)
```

Using bsub

- To submit batch jobs to the SLAC LSF cluster use the *bsub* command.

  bsub [*bsub options*] *command* [*arguments*]

  For example:

  bsub -o *outputfilename* date -u

Using bsub

Example of a simple bsub:

```
iris01 sf/neal> bsub hostname
Job <235254> is submitted to default queue <short>.

iris01 sf/neal> bjobs
JOBID   USER    STAT  QUEUE      FROM_HOST   EXEC_HOST   JOB_NAME   SUBMIT_TIME
235254  neal    PEND  short      iris01                  hostname   Mar  4 19:17

iris01 sf/neal> bjobs
JOBID   USER    STAT  QUEUE      FROM_HOST   EXEC_HOST   JOB_NAME   SUBMIT_TIME
235254  neal    RUN   short      iris01      yili0146    hostname   Mar  4 19:17

iris01 sf/neal> bjobs 235254
JOBID   USER    STAT  QUEUE      FROM_HOST   EXEC_HOST   JOB_NAME   SUBMIT_TIME
235254  neal    DONE  short      iris01      yili0146    hostname   Mar  4 19:17
```

Using bsub

Output from my simple batch job:

```
Job <hostname> was submitted from host <iris01> by user <neal>.
Job was executed on host(s) <yili0146>, in queue <short>, as user <neal>.
</u/sf/neal> was used as the home directory.
</u/sf/neal> was used as the working directory.
Started at Sun Mar  4 19:21:19 2007
Results reported at Sun Mar  4 19:21:57 2007
Your job looked like:

------------------------------------------------------------
# LSBATCH: User input
hostname
------------------------------------------------------------

Successfully completed.
Resource usage summary:
CPU time   :      0.22 sec.
Max Memory :         3 MB
Max Swap   :        11 MB
Max Processes  :         3
Max Threads    :         3

The output (if any) follows:
yili0146
```

Using bsub

Default behavior using bsub at SLAC.

- Job will be submitted to the default *short* job queue.
- Output will be returned via email.
- Job will be scheduled on a host of the same OS type.

  SUN5
  LINUX
  MACOSX
  WINDOWS

A few useful bsub options.

- Submit with a CPU limit (normalized): bsub -c

  Example: bsub -c 24:00 date

- Submit with a RUN limit (wallclock): bsub -W

  Example: bsub -W 24:00 date

- Submit with a jobname: bsub -J "job_name"

  Example: bsub -J "Date_job" date

Batch Job Scheduling Policy

- By default LSF is configured for FCFS scheduling.
- SLAC uses fairshare scheduling in the general queues.
- Fairshare controls how resources are shared between competing users or user groups.
- Job priorities are dynamic and change based upon your usage in the queues over the last few days. (Usage values decay over a period of hours.)

The SIMES Cluster

- SIMES Servers (simesfarm)
  - 64 Dell 1950 Dual-CPU Quad-Core Intel(R) Xeon(R) CPU X5355 @ 2.66GHz
  - 20 Dell C6100 Dual-CPU Hex-Core Intel(R) Xeon(R) CPU X5675 @ 3.07GHz
  - 752 cores (job slots)
  - ~380GB local /scratch space (simes0001-64); ~64GB (simes0065-84)
  - Infiniband
- Dedicated LSF MPI queue (simesq)
  - Access controlled via LSF user group (simes)
- Two login nodes accessible from SLAC interactive servers.
  - simes0001
  - simes0002

The SIMES GPU Cluster

- SIMES GPU Servers (simesgpufarm)
  - 4 Colfax Dual-CPU Quad-Core Intel(R) Xeon(R) CPU X5520 @ 2.27GHz
  - 72 cores (job slots)
  - 8 nVidia Tesla (Fermi) GPUs
- Dedicated LSF MPI queue (simesgpuq)
  - Access controlled via LSF user group (simes)
- One login node accessible from SLAC interactive servers.
  - simes-gpu
  - SuperMicro Intel(R) Xeon(R) CPU E5520 @ 2.27GHz

The SIMES Cluster

```
iris01 sf/neal> bqueues -l simesq

QUEUE: simesq
  -- SIMES MPI queue.

PARAMETERS/STATISTICS
PRIO NICE STATUS          MAX JL/U JL/P JL/H NJOBS  PEND   RUN SSUSP USUSP  RSV
192    0  Open:Active       -    -    -    -   288     0   288     0     0    0

[…]

USERS: simes/
HOSTS:  simesfarm/
ADMINISTRATORS:  moritzb
RES_REQ:  select[type==LINUX] span[]

JOB_CONTROLS:
    TERMINATE    [kill -CONT -$LSB_JOBRES_PID -$LSB_PAMPID; kill -TERM -$LSB_JOBRES_PID -
     $LSB_PAMPID]
```

Submitting jobs to the SIMES Cluster

Using bsub to submit OpenMPI jobs to simesq.

```
71 iris01 sf/neal> bsub  -q simesq –a mympi -n 10 -o ~neal/tmp/simestest.out ~neal/MPI/openmpi/
    rhel40/hello

Job <381464> is submitted to queue <simesq>.

78 sprocket sf/neal> bjobs -l 381464

Job <381464>, User <neal>, Project <none>, Status <DONE>, Queue <simesq>, Job Priority <50>,
          Command <pam -g 1 mympirun_wrapper /u/sf/neal/MPI/openmpi/rhel40/hello>
Tue Jul  8 16:04:22: Submitted from host <iris01>, CWD <$HOME>, Output File <
                     /u/sf/neal/tmp/simestest.out>, 10 Processors Requested;
Tue Jul  8 16:04:31: Started on 10 Hosts/Processors <8*simes0005> <2*simes0004>
                     , Execution Home </u/sf/neal>, Execution CWD </u/sf/neal>;

Tue Jul  8 16:04:35: Done successfully. The CPU time used is 1.3 seconds.
```

Submitting jobs to the SIMES Cluster

Using bsub to submit OpenMPI job to simesq overriding default queue ptile option.

```
iris01 sf/neal> bsub -q simesq –a mympi -n 4 -R "span[ptile=1]" -o ~neal/tmp/simes4.out ~neal/MPI/
    openmpi/rhel40/hello
Job <354171> is submitted to queue <simesq>.
```

The span[ptile=1] option will tell LSF to schedule the job using 1 processor on each host.

```
iris01 sf/neal> bjobs -l 354171

Job <354171>, User <neal>, Project <none>, Status <DONE>, Queue <simesq>, Job Priority <50>,
        Command <pam -g 1 mympirun_wrapper /u/sf/neal/MPI/openmpi/rhel40/hello>
Tue Jul  8 13:00:04: Submitted from host <iris01>, CWD <$HOME>, Output File
       < /u/sf/neal/tmp/simes4.out>, 4 Processors Requested, Requested Resources <span[ptile=1]>;
Tue Jul  8 13:00:13: Started on 4 Hosts/Processors <1*simes0049> <1*simes0035>
                     <1*simes0021> <1*simes0028>, Execution Home </u/sf/neal>,
                     Execution CWD </u/sf/neal>;
Tue Jul  8 13:00:16: Done successfully. The CPU time used is 0.7 seconds.
```

Submitting jobs to the SIMES Cluster

Using bsub to submit OpenMPI job to simesq requesting scratch > 100GB.

```
iris01 sf/neal> bsub -q simesq -a openmpi -n 10 -R "scratch > 100" -o ~neal/tmp/simesfun.out ~neal/MPI/
     openmpi/rhel40/hello
Job <387815> is submitted to queue <simesq>.

iris01 sf/neal> bjobs -l 387815

Job <387815>, User <neal>, Project <none>, Status <DONE>, Queue <simesq>, Job P
                     riority <50>, Command <pam -g 1 openmpirun_wrapper /u/sf/n
                     eal/MPI/openmpi/rhel40/hello>
Tue Jul  8 17:36:44: Submitted from host <iris01>, CWD <$HOME/bin>, Output Fi
                     le </u/sf/neal/tmp/simesfun.out>, 10 Processors Requested,
                     Requested Resources <scratch > 100>;
Tue Jul  8 17:36:54: Started on 10 Hosts/Processors <8*simes0033> <2*simes0058>
                     , Execution Home </u/sf/neal>, Execution CWD </u/sf/neal/b
                     in>;
Tue Jul  8 17:36:58: Done successfully. The CPU time used is 1.2 seconds.

108 sprocket neal/bin> lsload -I scratch simes0033 simes0058
HOST_NAME       status scratch
simes0058           ok   377.0
simes0033           ok   377.0
```

Submitting jobs to the SIMES Cluster

Using bsub to submit OpenMPI job to simesq requesting two processors on each host and scratch > 300GB.

```
41 iris02 sf/neal> bsub -q simesq –a mympi -n 4 -R "span[ptile=2] && scratch>300" -o ~neal/tmp/
    scratchit.out ~neal/MPI/openmpi/rhel40/hello
Job <390941> is submitted to queue <simesq>.

42 iris02 sf/neal> bjobs -l 390941

Job <390941>, User <neal>, Project <none>, Status <DONE>, Queue <simesq>, Job P
                     riority <50>, Command <pam -g 1 mympirun_wrapper /u/sf/n
                     eal/MPI/openmpi/rhel40/hello>
Tue Jul  8 18:20:50: Submitted from host <iris02>, CWD <$HOME>, Output File </u
                     /sf/neal/tmp/scratchit.out>, 4 Processors Requested, Reque
                     sted Resources <span[ptile=2] && scratch>300>;
Tue Jul  8 18:21:02: Started on 4 Hosts/Processors <2*simes0020> <2*simes0006>,
                     Execution Home </u/sf/neal>, Execution CWD </u/sf/neal>;
Tue Jul  8 18:21:06: Done successfully. The CPU time used is 0.3 seconds.

108 sprocket neal/bin> lsload -I scratch simes0030 simes0006
HOST_NAME       status scratch
simes0006           ok   377.0
simes0030           ok   377.0
```

LSF Documentation

- SLAC specific LSF documentation.

  http://www.slac.stanford.edu/comp/unix
  Click on "High Performance"

- Platform LSF documentation.

  http://www.slac.stanford.edu/comp/unix/package/lsf/currdoc/html/index.html
  http://www.slac.stanford.edu/comp/unix/package/lsf/currdoc/pdf/manuals/

Problem Reporting

Send email to:

unix-admin@slac.stanford.edu