



Production environment on FERMI

[silvia.giuliani@Cineca.it](mailto:silvia.giuliani@ Cineca.it)

[a.marani@Cineca.it](mailto:a.marani@ Cineca.it)



www.cineca.it



PROGRAMMING SPACE

- HOME

> cd \$HOME

/fermi/home/userexternal/....

- 50 GB **quota**
 >cindata (check your space usage)
- **Backup:** yes



PRODUCTION SPACE

- **SCRATCH**

```
>cd $CINECA_SCRATCH
```

```
/gpfs/scratch/userexternal/....
```

- No **quota**
 >cindata (check your space usage)
- No **backup**
- **Cleaning** procedure (everyday the clean procedure deletes all files older than 30 days)
 → NOT YET IMPLEMENTED



ARCHIVING SPACE

- **CINECA_DATA**

>cd \$CINECA_DATA

/shared/data/userexternal/....

- **100 GB quota**

- >cindata (check your space usage)

- you can ask for more when applying for projects (ISCRA)

- **No Backup**

- **\$CINECA_PROJECT** → NOT YET IMPLEMENTED

(from user space to project space)

- PROFILES

>module av <profile_name>

- **profile/base (default):** contains the basics and well tested modules
- **profile/front-end:** contains the modules compiled for front-end nodes
- **profile/advanced:** contains the application modules to be tested

> module load <profile_name>

List of applications to run on back-end nodes

>module av

----- /cineca/prod/modulefiles/base/applications -----

abinit/6.12.3	crystal09/1.01	qe/5.0bqq
amber/12(default)	dl_poly/4.03(default)	siesta/3.1
bigdft/1.6.0	gromacs/4.5.5(default)	vasp/5.2.12
cp2k/2.3(default)	lammps/20120816	vasp/5.3.2
namd/2.9	cpmd/v3.15.3(default)	

- Load a specific module
 > module load <module_name>
- Show the variables set by a specific module
 > module show <module_name>
- Retrieve informations on a specific module
 > module help <module_name>



- command line
 - > `./myexe`
 - On **Front-end** nodes
- batch mode
 - On **Back-end** or **Front-end** nodes
 - > `lsubsubmit job.cmd`



EXECUTION

Front End nodes

- **Pre and Post** processing
- **Data transfer**
- **Serial** execution (1 core)
- Executables compiled with **Front-End compilers**
 - >front-end-gnu/4.4.6
 - >front-end-xl/1.0
- **Command line** execution (up to 10 min)
- **Batch execution** (up to 6 h)

- **USER EXECUTABLE**

>edit job.cmd

- **Shell** interpreter path
#!/bin/bash
- **Load Leveler Scheduler** Keywords
@
@
@
.....
- **Variables** initialization
- **Execution** line
./myexe <options>

- **MODULE EXECUTABLE**

- **Shell** interpreter path
#!/bin/bash
- **Load Leveler Scheduler** Keywords
@
@
@
.....
- **Variables** initialization
module load <module_name>
- **Execution** line
exe <options>

[illegible]



EXECUTION

Back End nodes

- **Parallel** execution
- Executable compiled with serial or parallel
BE compilers
 - >bgq-gnu/4.4.6
 - >bgq-xl/1.0
- NO **command line** execution
- **Batch** execution (from 64 compute nodes up to 2048 compute nodes, wall clock time up to 24 h)
- **Runjob** command
 - >runjob <options>
 - >man runjob

- **USER EXECUTABLE**

- **Shell** interpreter path
#!/bin/bash
- **Load Leveler Scheduler** Keywords
@
@
@
.....
- **Variables** initialization
- **Execution** line
>runjob <runjob_options> : ./myexe
 <myexe_options>

- **MODULE EXECUTABLE**

- **Shell** interpreter path
#!/bin/bash
- **Load Leveler Scheduler** Keywords
@
@
@
.....
- **Variables** initialization
module load <module_name>
- **Execution** line
>runjob <runjob_options> :
\$MODULE_HOME/bin/exe <exe_options>

General LL KEYWORDS

Back end nodes

```
# @ job_name = myname
# @ output = $(job_name).$(jobid).out
# @ error = $(job_name).$(jobid).err
# @ environment = COPY_ALL # export all variables
                             from your submission shell

# @ job_type = bluegene
# @ wall_clock_limit = 00:10:00 # execution time h:m:s
# @ bg_size = 64 # compute nodes number
# @ notification = always|never|start|complete|error
# @ notify_user = <email_address>
# @ account_no = <budget_name> # saldo -b
# @ queue # indicates keyword section is completed
```



RUNJOB OPTIONS

man runjob

--exe path name for the executable to run
`runjob --exe <exe_name>`

--args Arguments for the executable specified by --exe
`runjob --exe <exe_name> --args <option1> --args <option2>`

It's easier to use ':' syntax instead of --exe and --args

--ranks-per-node number of ranks (MPI tasks) per compute node. Valid values are 1 (default), 2, 4, 8, 16, 32 and 64

bg_size = 64

runjob --ranks-per-node 1 : ./exe <options>

--np total number of ranks (MPI tasks)

bg_size = 64

runjob --np 64 --ranks-per-node 1: ./exe <options>

runjob --np 256 --ranks-per-node 4: ./exe <options>

#serial job:

runjob --np 1 --ranks-per-node 1: ./exe <options>

--envs Sets the environment variable to export on the compute nodes

```
bg_size = 64
```

```
#MPI/OpenMP job (foreach MPI task 16 threads)
```

```
runjob -n 64 --ranks-per-node 1 --envs
```

```
OMP_NUM_THREADS=16 : ./exe <options>
```

--exp-env Exports an environment variable from the current environment to the job

```
bg_size = 64
```

```
export OMP_NUM_THREADS = 16
```

```
runjob -n 64 --ranks-per-node 1 --exp-env
```

```
OMP_NUM_THREADS : ./exe <options>
```

@ bg_shape =

MD(A)xMD(B)xMD(C)xMD(D) #midplanes
number in the A,B,C,D dimensions

@ bg_rotate = **true**|false

@ bg_connectivity = torus|**mesh**|either|

Xa Xb Xc Xd #type of connectivity



BLUEGENE LL KEYWORDS

@ bg_connectivity = Mesh # default

- **for requests <= 1 midplane** (512 compute nodes)
bg_size = 64 | 128 | 256 | 512
- **for requests > 1 midplane**
bg_size = (512)x2 | (512)x3 | (512)x4



BLUEGENE LL KEYWORDS

@ bg_connectivity = Torus

- **for requests \geq 1MD**
bg_size = 512 | (512)x2 | (512)x4



BLUEGENE LL KEYWORDS

@ bg_connectivity = Mesh #default

@ bg_shape = distribution of midplanes on A, B,
C, D directions

MD(A)XMD(B)XMD(C)XMD(D)

Fermi machine (20 midplanes)

1X5X2X2

@ bg_connectivity = Torus

@ bg_shape = distribution of midplanes on A,
B, C, D directions

The midplanes must have Torus connectivity in
all dimensions

comp. nodes **512**

512*2

512*4

512*5

512*10

1x1x1x1

1x1x1x2 and 1x1x2x1

1x1x2x2

1x5x1x1

1X5X1X2

EXAMPLE

4 midplanes

#@bg_size = 2048

#@connectivity = Mesh



1X2X2X1

1X2X1X2

1X1X2X2

1X4X1X1

4 midplanes # @ bg_shape = 1X1X2X2
 # @ bg_connectivity = Mesh

By default # @ bg_rotate = true. The scheduler
can consider all possible rotations of the given
shape



1X1X2X2
1X2X1X2
1X2X2X1
~~1X4X1X1~~



BLUEGENE LL KEYWORDS

4 midplanes

#@bg_size = 2048

#@connectivity = Torus



1X1X2X2



LL COMMANDS

llsubmit

llsubmit job.cmd

llq

llq -u \$USER

[sgiulian@fen07 ~]\$ llq -u amarani0

Id	Owner	Submitted	ST	PRI	Class	Running	On

fen04.7334.0	amarani0	9/21 15:11	I	50	parallel		

1 job step(s) in query, 1 waiting, 0 pending, 0 running, 0 held, 0 preempted

llq -s <job_id>

Provides information on why a selected list of jobs remain in the NotQueued, Idle, or Deferred state.



"llq -s" output

- **[sgiulian@fen07 ~]\$ llq -s fen04.7334.0**
- ===== EVALUATIONS FOR JOB STEP fen04.fermi.cineca.it.7334.0 =====
- Step state : Idle
- Considered for scheduling at : Mon 24 Sep 2012 10:31:45 AM CEST
- Top dog estimated start time : Tue 25 Sep 2012 08:48:07 AM CEST
- Minimum initiators needed: 1 per machine, 1 total.
- 8 machines can run at least 1 tasks per machine, 128 tasks total.
- Not enough resources to start now.
- Shape 1x1x1x4 does not fit machine 1x5x2x2.
- Shape 1x1x4x1 does not fit machine 1x5x2x2.
- Shape 4x1x1x1 does not fit machine 1x5x2x2.
- Shape 2x1x1x2 does not fit machine 1x5x2x2.
- Shape 2x1x2x1 does not fit machine 1x5x2x2.
- Shape 2x2x1x1 does not fit machine 1x5x2x2.
- MP "R00-M0" is busy.
- MP "R00-M1" is busy.
- MP "R01-M0" is busy.
- MP "R01-M1" is busy.
- MP "R20-M0" is busy.
- MP "R20-M1" is busy.
- MP "R21-M0" is busy.
- MP "R21-M1" is busy.
- MP "R40-M0" is busy.
- MP "R30-M0" is busy.
- MP "R10-M0" is busy.
- MP "R41-M0" is busy.
- MP "R31-M0" cannot be used by job class.
- MP "R40-M1" is busy.
- MP "R30-M1" is busy.
- This step is a top-dog.

BG_SIZE = 2048 # 4 MD
BG_CONNECTIVITY = MESH

The job is a top dog.



"llq -s" output

```
[sgiuilian@fen07 proveMPI]$ llq -s fen03.7942.0
```

```
===== EVALUATIONS FOR JOB STEP fen03.fermi.cineca.it.7942.0 =====
```

```
Step state           : Idle
Considered for scheduling at   : Tue 25 Sep 2012 09:52:23 AM CEST
```

```
Minimum initiators needed: 1 per machine, 1 total.
8 machines can run at least 1 tasks per machine, 128 tasks total.
Not enough resources to start now.
Shape 2x1x1x1 does not fit machine 1x5x2x2.
MP "R00-M0" is busy.
MP "R01-M0" is busy.
MP "R20-M0" is busy.
MP "R21-M0" is on drain list.
MP "R40-M0" is not AVAILABLE (state="LoadLeveler Drained").
MP "R41-M0" is busy.
MP "R30-M0" is not AVAILABLE (state="LoadLeveler Drained").
MP "R31-M0" cannot be used by job class.
MP "R10-M0" is busy.
MP "R11-M0" cannot be used by job class.
MP "R00-M1" is busy.
MP "R21-M1" is on drain list.
MP "R40-M1" is not AVAILABLE (state="LoadLeveler Drained").
MP "R30-M1" is not AVAILABLE (state="LoadLeveler Drained").
MP "R10-M1" is busy.
MP "R01-M1" is busy.
MP "R41-M1" is busy.
MP "R31-M1" cannot be used by job class.
```

```
Not enough resources for this step to be backfilled.
```

```
This step can not become a top-dog. Global MAX_TOP_DOGS limit of 1 reached.
```

```
BG_SIZE =1024 # 2 MD
BG_CONNECTIVITY = MESH
```

```
The job is not a top dog and it can
not be backfilled.
```



"llq -s" output

- [sgiulian@fen07 proveMPI]\$ **llq -s fen04.7546.0**
- ===== EVALUATIONS FOR JOB STEP fen04.fermi.cineca.it.7546.0 =====
- Step state : Idle
- Considered for scheduling at : Mon 24 Sep 2012 01:56:00 PM CEST
- Minimum initiators needed: 1 per machine, 1 total.
- 8 machines can run at least 1 tasks per machine, 128 tasks total.
- Not enough resources to start now.
- Shape 1x1x1x3 does not fit machine 1x5x2x2.
- Shape 1x1x3x1 does not fit machine 1x5x2x2.
- Shape 3x1x1x1 does not fit machine 1x5x2x2.
- MP "R00-M0" is busy.
- MP "R00-M1" is busy.
- MP "R01-M0" is busy.
- MP "R01-M1" is busy.
- MP "R20-M0" is busy.
- MP "R20-M1" is busy.
- MP "R21-M0" is busy.
- MP "R21-M1" is busy.
- MP "R40-M0" is busy.
- MP "R41-M0" is busy.
- Not enough resources for this step as top-dog.
- Shape 1x1x1x3 does not fit machine 1x5x2x2.
- Shape 1x1x3x1 does not fit machine 1x5x2x2.
- Shape 3x1x1x1 does not fit machine 1x5x2x2.
- MP "R00-M0" is busy.
- MP "R00-M1" is busy.
- MP "R01-M0" is busy.
- MP "R01-M1" is busy.
- MP "R20-M0" is busy.
- MP "R20-M1" is busy.
- MP "R21-M0" is busy.

BG_SIZE = 1536 # 3 MD

BG_CONNECTIVITY = TORUS

The job will not start. It's not possible to have the TORUS connection for all directions.



LL COMMANDS

llq -l <job_id>

- Specifies that a more verbose output will be generated for job_id
- In particular you'll be notified about the bgsizes you requested and the real bgsizes allocated:

```
.....  
.....  
  
BG Size Requested: 1024  
BG Size Allocated: 1024  
BG Shape Requested:  
BG Shape Allocated: 1x1x1x2  
BG Connectivity Requested: Mesh  
BG Connectivity Allocated: Torus Torus Torus Torus  
.....  
.....
```

llcancel

>llcancel <job_id>



LL CLASSES

BE nodes

- **Debug**

2 racks with 16 I/O nodes

- TEST - Short time (64 compute nodes, 30 min)

@ wall_clock_limit = up to 24h

@ bg_size = 64

- **Longdebug**

2 racks with 16 I/O nodes

- Long time (64 compute nodes, > 30 min)

- **Parallel**

8 racks with 8 I/O nodes

- PRODUCTION (from 128 to 2048 compute nodes)

@ wall_clock_limit = up to 24h

@ bg_size = from 64 to 2048

- **Special**

2 racks with 16 I/O nodes

- I/O intensive jobs (from 64 to 512 compute nodes)

@ class = special

- **Keyproject**

8 racks with 8 I/O nodes

- Very parallel jobs (ask to user support superc@cineca.it)

@ class = keyproject



SUPERC MODULE

>module load superc

jobtyp (provides useful information about job in the LL queues - user, tasks, times, ...)

- For using

> jobtyp <job_id>

sstat (provides useful information about the system status - jobs in the LL queues, allocated nodes, ...)

- For using

> sstat

sstat2 (provides a more complete information about the system status - Midplane avail/down/drained, jobs in the LL queues, allocated nodes, ...)

- For using

> sstat2

bgtop (draws a full-terminal display of nodeboards and jobs)

>bgtop

loadHPC (calculates aggregate statistics of LL jobs)

>loadHPC



saldo -b

Prints budget info of your username:

- validity range
- consumed resources both on the local cluster and on all clusters
- percentage for accounts enabled for given usernames

saldo -r

Prints daily resources usage report on the local cluster for

- selected username (-u)
- selected account (-a)

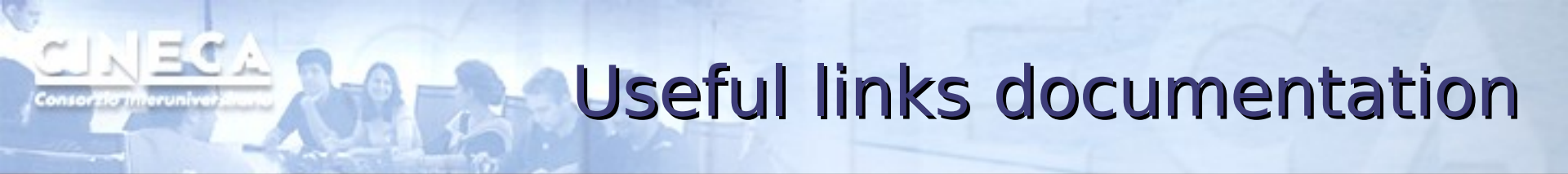
-----Resources used from 201101 to 201212-----				
date	username	account	localCluster Consumed/h	num.jobs



CONSUMED RESOURCES

- Remember that you are consuming the **ALLOCATED** resources and not necessarily the **REQUESTED** resources

$(\text{allocated compute nodes}) * (16\text{cores}) * (\text{execution time})$



Useful links documentation

Job command file keyword descriptions IBM

- http://publib.boulder.ibm.com/infocenter/clresctr/vxrx/index.jsp?topic=/com.ibm.cluster.loadl.v5r1.load100.doc/am2ug_sbmbgjbs.htm

FERMI's User guides

- <http://www.hpc.cineca.it/content/ibm-fermi-user-guide>
- <http://www.hpc.cineca.it/content/batch-scheduler-loadleveler-0>