

PBS 用户指南

1、PBS 队列介绍

目前部署在超算集群上有六个队列，队列名称分别是 batch、old、oldfat、fat、newfat、gpu。

batch : 默认队列，一般用来做作业测试；

old : 此队列共有 58 个计算节点，每个节点内存为 24G，cpu 核数为 12 核，此队列一般用于对内存需求不太大的作业；
此队列对应的节点为 c0101—c0142，c0301—c0332，除开有故障的节点，共计 58 个。

oldfat : 此队列共有 1 个胖节点，每个节点内存为 125G，cpu 核数为 32 核；
此队列对应的节点为 c0401。

fat : 此队列共有 3 个胖节点，每个节点内存为 1T，cpu 核数为 40 核，此队列一般用于对内存需求较大的作业；
此队列对应的节点为 fat01、fat02、fat03。

newfat : 此队列共有 2 个胖节点，每个节点内存为 1T，cpu 核数为 48 核，此队列一般用于对内存需求较大的作业；
此队列对应的节点为 fat04、fat05。

gpu : 此队列共有 1 个计算节点，每个节点内存为 125G，cpu 核数为 20 核，主要用于视频编码，图形处理等应用。
此队列对应的节点为 gpu01。

2、PBS 命令介绍

PBS 提供 4 条命令用于作业管理。

(1) qsub —— 用于提交作业脚本

命令格式：

```
qsub [-a date_time] [-c interval] [-C directive_prefix]
      [-e path] [-I] [-j join] [-k keep] [-l resource_list] [-m mail_options]
      [-M user_list][-N name] [-o path] [-p priority] [-q destination]
```

[-r c] [-S path_list] [-u user_list] [-v variable_list] [-V]
[-W additional_attributes] [-z]
[script]

(2) qstat —— 用于查询作业状态信息

命令格式：

qstat [-f] [-a] [-i] [-n] [-s] [-R] [-Q] [-q] [-B] [-u]

参数说明：

- f *jobid* 列出指定作业的信息
- a 列出系统所有作业
- i 列出不在运行的作业
- n 列出分配给此作业的结点
- s 列出队列管理员与 scheduler 所提供的建议
- R 列出磁盘预留信息
- Q 操作符是 destination id，指明请求的是队列状态
- q 列出队列状态，并以 alternative 形式显示
- au *userid* 列出指定用户的所有作业
- B 列出 PBS Server 信息
- r 列出所有正在运行的作业
- Qf *queue* 列出指定队列的信息
- u 若操作符为作业号，则列出其状态。若操作符为 destination id，则列出运行在其上的属于 user_list 中用户的作业状态。

常用命令示例：

a) 查看空闲节点信息（提交作业前需要查看一下各队列节点空闲状况）

```
[root@mu01 torque6]# pbsnodes -l free
fat03 free
fat04 free
fat05 free
gpu01 free
c0106 free
c0107 free
c0108 free
c0109 free
c0110 free
c0111 free
c0112 free
c0113 free
c0115 free
c0116 free
c0117 free
```

b) 查看指定作业信息 (提交作业后查看作业状态信息)

```
[root@mu01 torque6]# qstat -f 1386
Job Id: 1386.mu01
  Job_Name = work.00035.sh
  Job_Owner = wangwenlab@mu02
  resources_used.cput = 00:00:00
  resources_used.energy_used = 0
  resources_used.mem = 4608kb
  resources_used.vmem = 261920kb
  resources_used.walltime = 16:01:32
  job_state = R
  queue = old
  server = mu01
  Checkpoint = u
  ctime = Mon Apr 16 23:07:20 2018
  Error_Path = mu02:/lustre/home/wangwenlab/yangjie/hudie/fengweidie_project
              /04.correct/homolog/Papilio_machaon.protein.tblastn.shell.22556.qsub/w
              ork.00035.sh.e1386
  exec_host = c0136/0
  Hold_Types = n
  Join_Path = n
  Keep_Files = n
  Mail_Points = a
  mtime = Mon Apr 16 23:07:41 2018
  Output_Path = mu02:/lustre/home/wangwenlab/yangjie/hudie/fengweidie_projec
              t/04.correct/homolog/Papilio_machaon.protein.tblastn.shell.22556.qsub/
              work.00035.sh.o1386
  Priority = 0
  qtime = Mon Apr 16 23:07:20 2018
  Rerunnable = True
  Resource_List.feature = HPC
  Resource_List.nodename = oldPAR
  session_id = 20021
```

c) 查看所有队列作业状态

```
[root@mu01 torque6]# qstat -q

server: mu01

Queue          Memory CPU Time Walltime Node  Run Que Lm  State
-----
newfat         --    --    --    --    6  0 --  E R
gpu            --    --    --    --    0  0 --  E R
oldfat         --    --    --    --    0  0 --  E R
old            --    --    --    --   113 0 --  E R
batch          --    --    --    --    0  0 --  E R
fat            --    --    --    --    1  0 --  E R
-----
                                120  0
```

(3) qdel —— 用于删除已提交的作业

命令格式：

qdel [-W 间隔时间] 作业号

命令行参数：

qdel -p 强制清除某个作业号，一般不建议使用

例：# qdel -W 15 211 15 秒后删除作业号为 211 的作业

(4) qhold & qrls —— 作业挂起 & 作业释放

使用 qhold 命令可以挂起作业，使其不被调度执行；

使用 qrls 命令可以将挂起的作业释放，使之可以被调度执行；

命令格式：

```
qhold jobid1 jobid2 ...
```

```
qrls jobid1 jobid2 ...
```

其中 *jobidX* 代表需要操作的作业号，可以一次操作多个作业。

3、PBS 脚本文件

PBS 脚本文件由脚本选项和运行脚本两部分组成。

(1) PBS 作业脚本选项

(若无 -C 选项，则每项前面加 '#PBS')

-a *date_time* *date_time* 格式为：[[[CC]YY]MM]DD]hhmm[.SS]

表示经过 *date_time* 时间后作业才可以运行。

-c *interval* 定义作业的检查点间隔，如果机器不支持检查点，则忽略此选项。

-C *directive_prefix* 在脚本文件中以 *directive_prefix* 开头的行解释为 qsub 的命令选项。若无此选项，则默认为 '#PBS'

-e *path* 将标准错误信息重定向到 *path*

-I 以交互方式运行

-j *join* 将标准输出信息与标准错误信息合并到一个文件 *join* 中

-k *keep* 定义在执行结点上保留标准输出和标准错误信息中的哪个文件。

keep 为 o 表示保留前者，e 表示后者，oe 或 eo 表示二者都保留，n 表示皆不保留。若忽略此选项，二者都不保留。

-l *resource_list* 定义资源列表，几个常用的资源种类：

 cput=N 请求 N 秒的 CPU 时间，也可以是 hh:mm:ss 的形式。

 mem=N[K|M|G][B|W] 请求 N {k|m|g}{bytes|words} 大小的内存。

 nodes=N:ppn=M 请求 N 个结点，每个结点 M 个处理器。

-m *mail_option* *mail_option* 为 a：作业 abort 时给用户发信

 为 b：作业开始运行发信

 为 e：作业结束运行时发信

若无此选项，默认为 a。

- M *user_list* 定义有关此作业的 mail 发给哪些用户
- N *name* 作业名，限 15 个字符，首字符为字母，无空格。
- o *path* 重定向标准输出到 *path*
- p *priority* 任务优先级，整数，[-1024, 1023]，若无定义则为 0
- q *destination* *destination* 有三种形式：queue；
@server；
queue@server
- r y|n 指明作业是否可运行，y 为可运行，n 为不可运行。
- S *shell* 指明执行运行脚本所用的 shell，须包含全路径。
- u *user_list* 定义作业将在运行结点上以哪个用户名来运行。
- v *variable_list* 定义 export 到本作业的环境变量的扩展列表。
- V 表明 qsub 命令的所有环境变量都 export 到此作业。
- W *additional_attributes* 作业的其它属性
- z 指明 qsub 命令提交作业后，不在终端显示作业号。

(2) 运行脚本同 Linux 下一般的运行脚本文件

[注]：脚本文件中的 mpirun_rsh 命令行中的节点列表文件要用环境变量表示。

\$PBS_NODEFILE，这个环境变量表示由 PBS 自动分配给作业的节点列表；节点数为命令行中指定的进程数。

命令格式：

```
mpirun_rsh -np 进程数 -hostfile $PBS_NODEFILE 可执行程序名
```

4、PBS 环境下运行示例

(1) 脚本文件编辑示例

实例 1：运行 mpi 程序

命令行：#vi aaa.pbs

编辑的内容：

```
#PBS -N myjob
#PBS -o /home/jz/my.out
#PBS -e /home/jz/my.err
#PBS -l nodes=2:ppn=2
cd 目录 (脚本所在的目录)
```

```
mpirun -np 4 -hostfile $PBS_NODEFILE /home/jz/helloworld
```

解释：

原来我们都是直接在终端输入 `mpirun ;rsh.....` 这些命令执行程序，现在只要把这些提交命令放在 `.pbs` 配置文件的最后，由 PBS 来调度执行（自动分配节点和其它资源）。

`myjob` 是为要运行的程序起的任务名，可以改成你自己想要的名字。原先输出信息都是直接在屏幕上显示的，现在屏幕上的显示全部输出到文件中，上例中输出文件是 `/home/jz/my.out` 文件，可以根据自己的需要修改（目录，文件名）。程序运行时遇到的一些错误会记录在 `.err` 文件中。这样的好处是，因为对每个任务都设定了不同的输出文件，所以看结果只要打开相应文件看就可以了，不需要开多个终端，而且里面有任务的详细信息，比如实际分配的是哪些节点计算，运行时间等。

```
pbs -l nodes=2:ppn=2
```

规定使用的节点数，以及每个节点能跑多少核。

```
mpirun_rsh -np 4 -hostfile $PBS_NODEFILE /home/jz/helloworld
```

此例中 `-np` 后的 4 是并行数（ $2 \times 2 = 4$ 个 cpu），`-hostfile $PBS_NODEFILE` 不需要改变。`/home/jz/helloworld` 是你编译好的可执行文件名，需修改。

对于每个你要运行的 `mpi` 程序都需要这样一个 `.pbs` 配置文件，也就是说原来的操作是：`mpirun.....`，现在改成 2 步走：

- 1) 写个 PBS 配置文件（比如 `xxx.pbs`）；
- 2) 向 PBS 提交（`qsub xxx.pbs`）

实例 2：运行非 `mpi` 程序

有些用户并不是自己编写 `mpi` 程序，同样也可以用 PBS 提交。比如物理系运行程序时一般输入的命令是：

```
RunDMol3.sh TiFeCp2-pbe-dspp-m=1-opt
```

那么配置文件可以这样写：

```
命令行：#vi job.pbs
```

编辑的内容：

```
#PBS -N physics_job
#PBS -o /home/physics/physics_job.out
#PBS -e /home/physics/physics_job.err
#PBS -q 队列名称
#PBS -l nodes=1:ppn=2
#PBS -r y
```

cd 目录 (原来直接在节点上运行时所在的目录)

```
RunDMol3.sh TiFeCp2-pbe-dspp-m=1-opt
```

解释 :

把原来在终端直接输入的命令放到 PBS 配置文件中 , 因为只要一个节点 , 所以 nodes=1 , 至于用哪个节点系统自动分配 , 可以用 qstat 命令查询 (比如 qstat -n) 。

用户脚本文件实例 :

a)

```
#PBS -N D10
#PBS -o log/D10.out
#PBS -e log/D10.err
#PBS -q old
#PBS -l nodes=5:ppn=4
#PBS -r y
cd /newlustre/home/liaolanjie/yc/12_01_ecard
sh gatk.js D10
```

b)

```
#PBS -N test_GAPE2
#PBS -o /newlustre/home/zhangmao511/workdir/jar_second/jar/my.out
#PBS -e /newlustre/home/zhangmao511/workdir/jar_second/jar/my.err
#PBS -q fat
#PBS -l nodes=1:ppn=10
#PBS -r y
cd /newlustre/home/zhangmao511/workdir/jar_second/jar
java -jar
/newlustre/home/zhangmao511/workdir/jar_second/jar/GAPE_fat_zebrafish_Liunx_Beta_V2.jar -c
/newlustre/home/zhangmao511/workdir/jar_second/jar/zebrafish_parameters.txt
```

(2) 提交作业示例

命令行 : #qsub aaa.pbs

(3) 作业状态查询示例

qstat 后加不同参数可以查看不同的信息，查看作业的状态。

命令行：`#qstat -a`

解释：

Job id 211 是给提交的任务分配的任务号，S（常用状态：R 代表运行，Q 代表排队，E 代表正在退出，H 代表挂起，C 代表运行完毕）

命令行：`#qstat -n` 查看作业使用的节点

命令行：`#qstat jobid1 jobid2 ...` 查看指定作业号的作业（可一次查看多个作业）

命令行：`#qstat -u user1` 查看指定用户的作业

解释：该方式输出和默认略有不同，但大同小异。

命令行：`#qstat -f jobid` 查看特定作业详细信息

解释：该命令将会输出作业号为 *jobid* 的作业的详细信息。

附件：

中国科学院超级计算武汉分中心收费细则

为保障中国科学院超级计算武汉分中心用户合理及规范使用计算资源，保障用户权益，根据《中国科学院超级计算环境收费细则》，制定本细则。

一、收费方式

中国科学院超级计算武汉分中心用户使用计算资源前与分中心签订服务协议。

中国科学院超级计算武汉分中心采取预付费模式进行管理，用户使用 1 个 CPU 核运行一个小时的计算任务为 1 机时，使用后与分中心结算费用，用户机时费以自然月为周期进行结算，结算时分中心向用户提供服务对账单，系统每个月 1 日生成费用账单，用户提交计算任务前应确保计算账户有足够余额，一旦欠费将无法提交计算任务。

用户使用分中心集群软硬件资源需缴纳账户管理费，账户管理费采用包月制，内含一定量可用机时。账户管理费按自然月进行计费并收取，收取标准由系统按照用户实际使用机时自动计算。

二、收费标准

中国科学院超级计算武汉分中心环境依据运行维护成本估算，根据

《中国科学院超级计算环境收费细则》，机时费标准如下：

用户类型	CPU 收费标准 (元/CPU 小时)	GPU 收费标准 (元/GPU 小时)
中国科学院内用户	0.2	0.4
中国科学院外用户	0.4	0.8

账户管理费分为 300 元/月和 800 元/月两档，其中：

1、300 元/月档含 5000CPU 核小时，可满足用户占用 8 个 CPU 核运行一个月计算任务；

2、800 元/月档含 20000CPU 核小时，可满足用户占用 32 个 CPU 核运行一个月计算任务。

存储收费标准如下：

存储占用	收费标准 (元/T/年)
1T-50T	1000
50T-100T	2000
100T-150T	3000
150T 以上	4000

三、优惠方案

为体现对高并行度计算的支持，以及对不同服务水平的区分，提供如下建议优惠方案：

作业规模	优惠方案(收费标准的%)
------	--------------

	免费队列	正常队列	高优先级队列	独占队列
1024CPU (GPU) 及以上	0	30	65	100
512-1023CPU (GPU)	0	40	80	120
256-511CPU (GPU)	0	50	100	150
128-255CPU (GPU)	0	60	120	180
64-127CPU (GPU)	0	70	160	240
64CPU (GPU) 以下	0	100	200	400

1. 免费队列：最低优先级，有其它作业运行时将被挂起或重排队。
2. 正常队列：按正常排队策略在队列中等待。
3. 高优先级队列：更高的优先权，可抢占低级队列的资源。
4. 独占队列：相当于物理隔离的队列，用户独占这部分计算资源。用户需要使用独占队列时，应于每月 25 日之前提出机时申请，以便统筹安排计算资源。每次申请独占队列的使用期限最短为一周，最长为一个月。独占队列按照占用节点资源的起止时间进行计费。

四、重点项目资助

为实现对重点项目、重点应用的优先支持，中国科学院超级计算武汉分中心提供部分免费机时资助。

中国科学院超级计算武汉分中心为鼓励课题组等通过集群托管、软件移植等多种方式参与科研高性能计算公共平台建设，促进科研资源中心内共享。用户使用自己托管的集群设备或移植的应用软件进行的计算不收取机时费。

以下情况不计算机时费用：

1. 交任务后的排队时间；
2. 群故障或系统断电等原因造成计算任务意外中断或异常推出的；
3. 月末提交计算任务的不收取账户管理费。

五、机时奖励

凡得到高性能计算集群所形成的成果，包括论文等应标注“**The research was supported by the Wuhan Branch, Supercomputing Center, Chinese Academy of Sciences, China**”。对于利用高性能计算集群取得科研成果的课题经由用户提出申请给予机时奖励。

1. 用户可随时提交机时奖励申请，分中心每年年初集中组织一次机时奖励评定会，审核通过后的机时奖励直接充入用户账户，用于冲抵机时费；
2. 奖励机时只能用于缴纳超出机本机时部分的计算费用，不能用于冲抵账户管理费。