广州超算天河二号简明使用手册

The Right Answer in Data Center PARATERA 并行®

（2019 春季版）

术语表

分区：对应文件系统上用来存储数据的位置，可以理解为机器上的一个盘符；

队列：作业管理系统把系统的计算资源划分到不同的集合，用户的作业可以提交到这些集合排队运行，称这些集合为队列。

目录

1. 系统资源简介

1.1 计算资源（节点配置）

广州超算安装的“天河二号”系统配置为：CPU 型号Intel Xeon E5-2692 12C 2.200GHz，采用TH Express-2 高速互连，有些节点配置了GPU K80 加速器。

常用队列的节点配置如下：

刀片节点：每节点 2*12 核（Xeon E5-2692V2），64G 内存；

图形界面节点：每节点 2*12 核（Xeon E5-2692V2），128G 内存；

3TB大内存节点：每节点 4*14核（Xeon E7-4850 v3），3TB内存；

GPU-K80节点：每节点2 个k80 卡 + 2*10核（Xeon E5-2660 V3），48G显存；

GPU-V100节点：每节点 4*V100 GPU（NVIDIA Tesla SXM2） + 2*14核CPU（Gold 6132），64G显存。

天河星光节点：每节点 2*12 核（Xeon E5-2692V2），128G 内存。

通过yhi 可以查看常用作业队列，各队列的节点类型对应关系如下表所示：

节点类型	队列名称
刀片节点	paratera、test（test 队列为测试队列只能使用 10 分钟）
图形界面节点	docker
3TB大内存节点	MEM_3TB（开通权限后需ssh ln41 上才能看到）
GPU -K80节点	gpu（开通权限后需ssh ln41 上才能看到）
GPU-V100节点	gpu_v100
天河星光节点	commercial、LAVA（开通天河星光平台才可以使用）

温馨提示：

登录节点仅供编译软件、拷贝数据，为避免登录节点负载过高，影响正常使用，请勿在登录节点运行程序。提交作业请使用slurm调度命令发送到计算节点，请勿在登陆节点运行作业。如有发现此类不规范操作，管理员将终止进程。并通知违规用户。两次警告无效后，每次违规操作，将会禁止账号登陆一天。

1.2 存储资源（文件系统）

“天河二号”系统采用lustre 文件系统提供大规模存储，系统包括多个存储分区，不同账号可以使用的存储分区有/WORK 分区或者/PARA 分区，登陆系统后通过 pwd 命令可以查看自己当前所在的分区。用户的家目录（通常$HOME 的位置）位于该分区下，通常用户的家目录形式为“/WORK/超算账号（或/PARA/超算账号）”，例如/WORK/pp001（或/PARA/pp001）。该分区既可用于编译和程序存储，也可提交运行作业。超算账号下默认的磁盘配额为 500G，可用命令lfs quota -uh pp001 /PARA/pp001 查看详细信息。

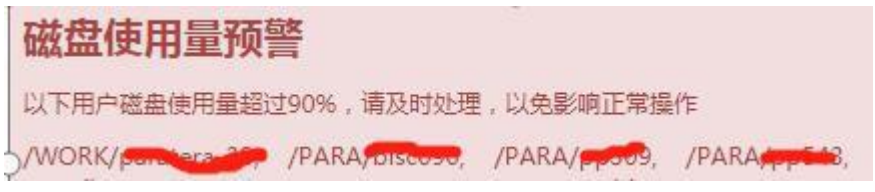
```
[ppxxx@lon6%tianhe2-B~]$ lfs quota -uh p-xxx /PARA/ppxxx
Disk quotas for user ppxxx (uid 4061):
     Filesystem      used     quota     limit     grace     files     quota     limit     grace
    /PARA/ppxxx411.3G500G      500G         -        38230         0         0         -
```

例如上面的账号ppxxx 磁盘配额为 500GB，已经使用了 411.3GB。

用户数据占用的磁盘空间超过磁盘配额后会影响数据保存或者作业运行，建议用户自己经常检查配额，及时清理不需要的数据，把有用数据下载到自己的本地。用户如果确实需要比较大的磁盘空间存储数据和执行程序，可以联系客户经理增加配额。

温馨提示：

并行@超算云服务软件上设置了存储配额报警功能，当磁盘配额不足时，登录并行@超算云服务软件界面时就会弹出存储配额不足的提醒（如下图），建议您关注此报警功能，经常查看磁盘空间使用情况，及时清理不用数据，以免作业因磁盘空间不足受到影响。



2. 使用超算应用软件

2.1 简介

“天河二号”已经部署了很多应用软件，这些软件用户都可以直接使用。

由于不同用户在“天河二号”上可能需要使用不同的软件环境，配置不同的环境变量，软件之间可能会相互影响，因而在“天河二号”上安装了 module 工具来对应用软件统一管理。module 工具主要用来帮助用户在使用软件前设置必要的环境变量。用户使用 module 加载相应版本的软件后，即可直接调用超算上已安装的软件。

2.2 基本命令

常用命令如下：

命令	功能	例子
module avail	查看可用的软件列表	
module load [modulesfile]	加载需要使用的软件	module load OpenFOAM/2.3.1
module show [modulesfile]	查看对应软件的环境（安装路径、库路径等）	module show OpenFOAM/2.3.1
module list	查看当前已加载的所有软件	
module unload [modulesfile]	移除使用module 加载的软件环境	module unload OpenFOAM/2.3.1
module purge	移除账号下module 加载的所有软件环境	module purge

module 其它用法，可使用 module --help 中查询。module 加载的软件环境只在当前登陆窗口有效，退出登陆后软件环境就会失效。用户如果需要经常使用一个软件，可以把load 命令放在~/.bashrc 或者提交脚本里面。

3. Slurm 作业管理系统

天河二号使用 Slurm 作业管理系统，采用节点独占模式，普通节点为 24 核，为避免浪费机时，使用时请尽量保证使用核数为 24 的整数倍。

作业管理系统常用命令如下：

命令	功能介绍	常用命令例子
yhi（sinfo）	显示系统资源使用情况	yhi
yhq（squeue）	显示作业状态	yhq
yhrun（srun）	用于交互式作业提交	yhrun -N 2 -n 48 -p paratera A.exe
yhbatch（sbatch）	用于批处理作业提交	yhbatch -N 2 -n 48 job.sh
yhalloc（salloc）	用于分配模式作业提交	yhalloc -p paratera
yhcancel（scancel）	用于取消已提交的作业	yhcancel JOBID
yhcontrol（scontrol）	用于查询节点信息或正在运行的作业信息	yhcontrol show job JOBID
yhacct（sacct）	用于查看历史作业信息	yhacct -u pp100 -S 03/01/17 -E 03/31/17 --field=jobid,partition,jobname,user,nnodes,start,end,elapsed,state

3.1 yhi/sinfo查看系统资源

yhi 得到的结果是当前账号可使用的队列资源信息，如下图所示：

```
[paratera_gpu1@ln1%tianhe2-C ~]$ yhi
PARTITION  AVAIL  TIMELIMIT  NODES  STATE NODELIST
MEM_128       up   infinite      2   drng cn[11803,11833]
MEM_128       up   infinite     23  alloc cn[11780-11781,11
MEM_128       up   infinite      5   idle cn[11785,11787,11
docker_128    up   infinite      2  alloc cn[11587,11627]
docker_128    up   infinite     27   idle cn[11588-11594,11
work*         up   infinite      2   comp cn[7772,12622]
work*         up   infinite    228   drng cn[4608,4614,4628
0-4882,4884,4891,4904,4909-4911,4916,4922-4923,4930-4936,49
```

其中，

第一列 PARTITION 是队列名，默认能使用的队列名为 paratera 和 work（或 bigdata），有特殊需求的账号会用到胖节点、图形界面、GPU 节点等，资源申请请联系客户经理开通权限。

第二列 AVAIL 是队列可用情况，如果显示 up 则是可用状态；如果是 inact 则是不可用状态。

第三列 TIMELIMIT 是作业运行时间限制，默认是 infinite 没有限制。

第四列 NODES 是节点数。

第五列 STATE 是节点状态，idle 是空闲节点，alloc 是已被占用节点，comp 是正在释放资源的节点，其他状态的节点都不可用。

第六列 NODELIST 是节点列表。

yhi 的常用命令选项：

命令示例	功能
yhi -n cn12345	指定显示节点 cn12345 的使用情况
yhi -p paratera	指定显示队列 paratera 情况

其他选项可以通过 yhi --help 查询

3.2 yhq/squeue查看作业状态

yhq 得到的结果是当前账号的作业运行状态，如果 yhq 没有作业信息，说明作业已退出。

```
[paratera_gz@ln1%tianhe2-C paraacct]$ yhq
             JOBID PARTITION     NAME         USER ST       TIME  NODES NODELIST(REASON)
           2061995  paratera gaussian  paratera_gz  R 8-02:38:27      1 cn10849
           2491328  paratera gaussian  paratera_gz  R   19:28:53      1 cn10713
           2495028  paratera gaussian  paratera_gz  R    4:17:36      1 cn10718
           2494543  paratera gaussian  paratera_gz  R    4:59:34      1 cn10715
           2496346      work   fluent  paratera_gz PD       0:00      6 (Resources)
```

其中，

第一列 JOBID 是作业号，作业号是唯一的。

第二列 PARTITION 是作业运行使用的队列名。

第三列 NAME 是作业名。

第四列 USER 是超算账号名。

第五列 ST 是作业状态，R 表示正常运行，PD 表示在排队，CG 表示正在退出，S 是管理员暂时挂起，只有 R 状态会计费。

第六列 TIME 是作业运行时间。

第七列 NODES 是作业使用的节点数。

第八列 NODELIST(REASON)对于运行作业（R 状态）显示作业使用的节点列表；对于排队作业（PD 状态），显示排队的原因。

yhq 的 常用命令选项：

命令示例	功能
yhq -j 123456	查看作业号为 123456 的作业信息
yhq -u pp100	查看超算账号为 pp100 的作业信息
yhq -p paratera	查看提交到 paratera 队列的作业信息
yhq -w cn123	查看使用到 cn123 节点的作业信息

其他选项可通过 yhq --help 命令查看

3.3 yhrun/srun交互式提交作业

yhrun [options] program 命令属于交互式提交作业，有屏幕输出，但容易受网络波动影响，断网或关闭窗口会导致作业中断。

yhrun 命令示例：

```
yhrun -p paratera -w cn[1100-1101] -N 2 -n 48 -t 20 A.exe
```

交互式提交 A.exe 程序。如果不关心节点和时间限制，可简写为 yhrun -p paratera -n 48 A.exe

其中，

-p paratera 指定提交作业到 paratera 队列；

-w cn[1100-1101] 指定使用节点 cn[1100-1101]；

-N 2 指定使用 2 个节点；

-n 48 指定进程数为 48，天河二号一个节点 24 核，建议使用 24 的整数倍提交作业；

-t 20 指定作业运行时间限制为 20 分钟。

yhrun 的一些常用命令选项：

参数选项	功能
-N 3	指定节点数为 3
-n 12	指定进程数为 12，天河二号普通节点 24 核，建议满核提交
-c 12	指定每个进程（任务）使用的 CPU 核为 12
-p docker_128	指定提交作业到 docker_128 队列

-w cn[100-101]	指定提交作业到 cn100、cn101 节点
-x cn[100,106]	排除 cn100、cn106 节点
-o out.log	指定标准输出到 out.log 文件
-e err.log	指定重定向错误输出到 err.log 文件
-J JOBNAME	指定作业名为 JOBNAME
-t 20	限制运行 20 分钟

yhrun 的其他选项可通过 yhrun --help 查看。

3.4 yhbatch/sbatch 后台提交作业

yhbatch 一般情况下与 yhrun 一起提交作业到后台，需要将 yhrun 写到脚本中，再用 yhbatch 提交脚本。这种方式不受本地网络波动影响，提交作业后可以关闭本地电脑。yhbatch 命令没有屏幕输出，默认输出日志为提交目录下的 slurm-xxx.out 文件，可以使用 tail -f slurm-xxx.out 实时查看日志，其中 xxx 为作业号。

yhbatch 命令示例 1（48 个进程提交 A.exe 程序）：

编写脚本 job1.sh，内容如下：

```
#!/bin/bash
yhrun -n 48 A.exe
```

然后在命令行执行 yhbatch -p paratera job1.sh 提交作业。脚本中的#!/bin/bash 是 bash 脚本的固定格式。从脚本的形式可以看出，提交脚本是一个 shell 脚本，因此常用的 shell 脚本语法都可以使用。作业开始运行后，在提交目录会生成一个 slurm-xxx.out 日志文件，其中 xxx 表示作业号。

yhbatch 命令示例 2（指定 2 个节点，4 个进程，每个进程 12 个 cpu 核提交 A.exe 程序，限制运行 60 分钟）：

编写脚本 job2.sh，内容如下：

```
#!/bin/bash
#SBATCH -N 2
#SBATCH -n 4
#SBATCH -c 12
#SBATCH -t 60
yhrun -n 4 A.exe
```

然后在命令行执行 yhbatch -p paratera job2.sh 就可以提交作业。其中#SBATCH 注释行是 slurm 定义的作业执行方式说明，一些需要通过命令行指定的设置可以通过这些说明写在脚本里，避免了每次提交作业写很长的命令行。

yhbatch 命令示例 3（单节点提交多任务）

编写脚本 job3.sh，内容如下：

```
#!/bin/bash
yhrun -n 8 A.exe &
yhrun -n 8 B.exe &
yhrun -n 8 C.exe &
wait
```

然后在命令行执行 yhbatch -N 1 -p paratera job3.sh，这里是单节点同时提交 3 个任务，每个任务使用 8 个进程。这里需要 3 个任务全部执行完毕，作业才会退出。

yhbatch 的一些常用命令选项基本与 yhrun 的相同，具体可以通过 yhbatch --help 查看。

3.5 yhalloc/salloc 分配模式作业提交

yhalloc 命令用于申请节点资源，一般用法如下：

1、执行 yhalloc -p paratera；

2、执行 yhq 查看分配到的节点资源，比如分配到 cn100；

3、执行 ssh cn100 登陆到所分配的节点；

4、登陆节点后可以执行需要的提交命令或程序；

5、作业结束后，执行 yhcancel JOBID 释放分配模式作业的节点资源。

3.6 yhcancel/scancel 取消已提交的作业

yhcancel 可以取消正在运行或排队的作业。

yhcancel 的一些常用命令示例：

命令示例	功能
yhcancel 123456	取消作业号为 123456 的作业
yhcancel -n test	取消作业名为 test 的作业
yhcancel -p paratera	取消提交到 paratera 队列的作业
yhcancel -t PENDING	取消正在排队的作业
yhcancel -w cn100	取消运行在 cn100 节点上的作业

yhcancel 的其他参数选项，可通过 yhcancel --help 查看

3.7 yhcontrol/scontrol 查看正在运行的作业信息

yhcontrol 命令可以查看正在运行的作业详情，比如提交目录、提交脚本、使用核数情况等，对已退出的作业无效。

yhcontrol 的常用示例：

```
yhcontrol show job 123456
```

查看作业号为 123456 的作业详情。

yhcontrol 的其他参数选项，可通过 yhcontrol --help 查看。

3.8 yhacct/sacct 查看历史作业信息

yhacct 命令可以查看历史作业的起止时间、结束状态、作业号、作业名、使用的节点数、节点列表、运行时间等。

yhacct 的常用命令示例：

```
yhacct -u pp001 -S 2017-09-01 -E now --field=jobid,partition,jobname,user,nnodes,nodelist,start,end,elapsed,state
```

其中，-u pp001 是指查看 pp001 账号的历史作业，-S 是开始查询时间，-E 是截止查询时间，--format 定义了输出的格式，jobid 是指作业号，partition 是指提交队列，user 是指超算账号名，nnodes 是节点数，nodelist 是节点列表，start 是开始运行时间，end 是作业退出时间，elapsed 是运行时间，state 是作业结束状态。yhacct --helpformat 可以查看支持的输出格式。

yhacct 的其他参数选项可通过 yhacct --help 查看。

4. 编译器

天河二号已配置 GNU 和 Intel 编译器，支持 C、C++、Fortran77 和 Fortran90 语言程序的开发，支持 OpenMP 和 MPI 两种并行编程模式。其中 OpenMP 为共享内存方式，只能单点并行；MPI 是分布式内存并行，支持跨节点并行。

4.1 Intel编译器

天河二号默认的 Intel 编译器为 14.0.2 版本，如需使用其他版本，可通过 module load 加载环境，例如 module load intel-compilers/15.0.1，如下图所示：

```
[    3@ln3%tianhe2-C ..    ]$ icc -v
icc version 14.0.2 (gcc version 4.4.7 compatibility)
[pp(  )@ln3%tianhe2-C pp0   ]$ module avail intel

---------------------------- /WORK/app/modulefiles ----------------------------
intel-compilers/11.1   intel-compilers/14.0.2 intel-compilers/mkl-14
intel-compilers/13.0.0 intel-compilers/15.0.1 intel-compilers/mkl-15
[pp(  )@ln3%tianhe2-C pp0..J]$ module load intel-compilers/15.0.1
[pp0..J@ln3%tianhe2-C pp0   ]$ icc -v
icc version 15.0.1 (gcc version 4.4.7 compatibility)
```

通过“which”命令可以查找命令所在路径，例如“which icc”；通过“icc -v”命令可以查询 icc 的版本。Intel 编译器的详细命令行调用则可以用“icc --help”获得。

用户经常需要使用 MKL 库，通过命令 echo $MKLROOT 可以查看 MKLROOT 环境变量确认 MKL 库的位置。

4.2 GCC编译器

天河二号默认的 GNU 编译器版本是 4.4.7，如需其他版本，可通过 module load 加载，例如 module load gcc/5.3.0，如下图所示：

```
[pp(  @ln3%tianhe2-C ~]$ gcc -v
Using built-in specs.
Target: x86_64-redhat-linux
Configured with: ../configure --prefix=/usr --mandir=/usr/share/man --infodir=/usr/share/info --with-bugurl=http://bug
zilla.redhat.com/bugzilla --enable-bootstrap --enable-shared --enable-threads=posix --enable-checking=release --with-s
ystem-zlib --enable-__cxa_atexit --disable-libunwind-exceptions --enable-gnu-unique-object --enable-languages=c,c++,ob
jc,obj-c++,java,fortran,ada --enable-java-awt=gtk --disable-dssi --with-java-home=/usr/lib/jvm/java-1.5.0-gcj-1.5.0.0/
jre --enable-libgcj-multifile --enable-java-maintainer-mode --with-ecj-jar=/usr/share/java/eclipse-ecj.jar --disable-l
ibjava-multilib --with-ppl --with-cloog --with-tune=generic --with-arch_32=i686 --build=x86_64-redhat-linux
Thread model: posix
gcc version 4.4.7 20120313 (Red Hat 4.4.7-4) (GCC)
[pp(  @ln3%tianhe2-C ~]$ module avail gcc

------------------------------------------------ /WORK/app/modulefiles -------------------------------------------------
gcc/4.7.4 gcc/4.8.4 gcc/4.9.2 gcc/5.2.0 gcc/5.3.0
```

4.3 MPI编译环境

天河二号默认使用的 mpi 版本为 mpich-3.1.3。由于天河二号系统采用了自主互连的高速网络，因此 MPI 库针对高速网络进行了优化，提供了基于 Intel 编译器和 GNU 编译器编译的 MPI 库，缺省是基于 Intel 编译器的 MPI 库。如需其他版本的 mpi，可通过 module load 加载，例如 module load MPI/Intel/MPICH/3.2-icc14-dyn

```
[pp0  @ln1%tianhe2-C local]$ module avail MPI

----------------------------------- /WORK/app/modulefiles ------------------
MPI/Gnu/MPICH/3.1                    MPI/Intel/MPICH/3.1-dbg
MPI/Gnu/MPICH/3.1-4.8.4              MPI/Intel/MPICH/3.1-dyn
MPI/Gnu/MPICH/3.1-4.9.2              MPI/Intel/MPICH/3.1-icc11
MPI/Gnu/MPICH/3.2-gcc4.4.7-dyn       MPI/Intel/MPICH/3.1-icc13
MPI/Gnu/MPICH/3.2-gcc4.9.2-dyn       MPI/Intel/MPICH/3.1-icc13-dyn
MPI/Intel/IMPI/4.1.3.048             MPI/Intel/MPICH/3.1-icc15-dyn
MPI/Intel/IMPI/5.0.2.044             MPI/Intel/MPICH/3.1-large
MPI/Intel/MPICH/2.1.5                MPI/Intel/MPICH/3.2-icc14-dyn
MPI/Intel/MPICH/3.1                  MPI/Intel/MPICH/3.2-icc14-dyn-centos
```

常见FAQ

1. 运行 2 分钟左右显示被 CANCELED。

答：不能在/HOME/ppxx 家目录提交作业，需要到/WORK/ppxx（或者/BIGDATA/ppxx）提交。

2. 作业长时间呈 CG 状态。

答：CG 是正在退出状态，不扣除机时，该状态的作业无法用 yhcancel 取消作业。

3. 作业呈 S 状态。

答：S 是挂起状态，超算维护会将作业挂起，系统恢复后作业会继续运行。

4. 超算账号能否多人使用？

答：超算账号可以多人同时使用，建议大家建立自己的文件夹，在各自的文件夹下操作。

5. yhbatch 无法提交作业。

答：查看报错信息，

1）如果提示 Invalid partition name specified，则可能是指定的队列名有误或没有权限提交到该队列；

2）如果提示 Required partition not available (inactive or drain，则执行 yhi 查看队列状态，如果队列处于 inact 状态，则可能是超算在维护；

3）如果提示 Failed to allocate resources: User’s group not permitted to submit，则可能是账号已被限制提交作业；

4）其他报错信息请联系客服。

6. 找不到某些文件。

答：先确认文件所在路径，命令行和 winscp 登录超算所在的路径不同，命令行登录所在路径为/HOME/ppxx/WORKSPACE（或者/HOME/ppxx/BIGDATA），而 winscp 登录的路径是/HOME/ppxx（或者/HOME/ppxx）。

7. 提交作业报缺库，例如 libifport.so.5: cannot open shared object file: No such file or directory。

答：按照下列步骤操作：

1）执行 locate 定位缺的库文件名，例如 locate libifport.so.5；

2）如果有结果，则系统上有这个库，将系统上的 64 位库拷到用户自己的目录；

3）设置环境变量，执行 export LD_LIBRARY_PATH=<用户自己的目录>:$LD_LIBRARY_PATH；

4）重新提交作业。

8. 在超算上怎么打开图形界面？

答：联系客服开通 vnc 远程可视化功能。

9. 操作很慢是什么原因？

答：操作很慢一般有三种情况，

1）如果是敲命令卡，比如敲 ls、cd、yhq 很卡，说明本地到超算间的网络延迟较大，建议检查网络并更换最佳链路；

2）如果是执行 ls、cd 回车后没反应，可能是超算文件系统负载较高，这时可以执行 df 确认一下，如果反应慢，则联系管理员查看文件系统问题；

3）如果是执行 yhq、yhi、yhbatch 等命令回车后没反应，可能是超算调度系统负载较高，可以联系管理员查看调度系统问题。

10. 是否支持 ipv6?

答：ipv6 需同时满足以下四个条件才能使用：

1）本地支持 ipv6 并优先使用 ipv6；

2）本地使用 windows 版云桌面客户端的 putty 配置登录超算；

3）选择云桌面的“并行专线”链路登录超算；

4）使用的是广州超算或长沙超算资源。