# cngrid2 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）

cngrid2 超算有两套计算资源系统，两套系统对应不同的存储，通过 sinfo 可以查看 cngrid2 系统的常用作业队列。第一套系统各队列的节点和网络配置如下：

v3_64_single 队列： CPU 型号 E5-2640 v3 @ 2.60GHz，16 核，64GB 内存，节点间采用万兆以太网互连，单节点作业使用本队列，性能更高，但每个作业节点使用上限不超过 4 个节点。

v2_all 队列：CPU 型号 E5-2670 v2 @ 2.50GHz，20 核，64GB 或者 128GB 内存，节点采用 IB 高速互连，通信速度快。

pg2_128_pool 队列：CPU 型号 E5-2670 v2 @ 2.50GHz，20 核，128GB 内存，节点间采用 IB 高速互连，建议需要大内存的作业使用本队列。

v2_test 队列：CPU 型号 E5-2670 v2 @ 2.50GHz，20 核，64GB 内存，节点采用 IB 高速互连，通信速度快。

第二套系统队列的节点和网络配置如下：

v3_ib 队列：CPU 型号 E5-2678 v3 @ 2.50GHz，24 核，64GB 内存，节点间采用 IB 高速互联，通信速度快。

**温馨提示：**

v2_test 队列是测试队列，只能运行 10 分钟，仅用来调试使用，如需正常计算请提交到其他队列。

v2_all 队列和 pg2_128_pool 队列不允许提交单节点作业，单节点作业请提交到 v3_64_single 队列。

v3_ib 队列是新资源，需要单独申请开通使用，需要在 ln11 登录节点上提交作业。

## 1.2 存储资源（文件系统）

“cngrid2”系统采用并行文件系统提供大规模存储。第一套系统的存储位于并行文件系统/public1 下面，第二套系统的存储位于并行文件系统/public4 下面，登陆系统后通过 pwd 命令可以查看自己当前所在的分区。

第一套系统的用户如果有图形使用需求，我们会协助将家目录（HOME）迁移到/public2 下面，在 public2 安装图形软件，但是仍然建议用户的工作目录和数据放在/public1 下面，速度更快。

# 2. 使用超算应用软件

## 2.1 简介

由于不同用户在“cngrid2”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“cngrid2”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 intel/15.0.2 |
| module show [modulesfile] | 查看对应软件的环境（安装路径、库路径等） | module show intel/15.0.2 |
| 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 intel/15.0.2 |

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

cngrid2 使用 Slurm 作业管理系统，采用节点独占模式，v2_all 队列和 pg2_128_pool 队列每个节点 20 核，v3_64_single 队列每个节点 16 核，v3_ib 队列一个节点 20 核，为避免浪费机时，使用时请尽量保证满核提交（即为单节点核数的整数倍）。

作业管理系统常用命令如下：

| 命令 | 功能介绍 | 常用命令例子 |
|---|---|---|
| sinfo | 显示系统资源使用情况 | sinfo |
| squeue | 显示作业状态 | squeue |
| srun | 用于交互式作业提交 | srun -N 2 -n 48 -p v2_all A.exe |
| sbatch | 用于批处理作业提交 | sbatch -N 2 -n 48 job.sh |
| salloc | 用于分配模式作业提交 | salloc -p v2_all |
| scancel | 用于取消已提交的作业 | scancel JOBID |
| scontrol | 用于查询节点信息或正在运行的作业信息 | scontrol show job JOBID |
| sacct | 用于查看历史作业信息 | sacct -u pp100 -S 03/01/17 -E 03/31/17 --field=jobid,partition,jobname,user,nnodes,start,end,elapsed,state |

## 3.1 sinfo 查看系统资源

sinfo 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：

```
[deploy@ln4%cngrid2 ~]$ sinfo
PARTITION    AVAIL  TIMELIMIT  NODES  STATE NODELIST
v2_all*         up   infinite      1 drain* c9002
v2_all*         up   infinite      7  down* c[0802,1906-1909,2406,9708]
v2_all*         up   infinite      7   drng c[0508-0510,1304,2306,9406,9502]
v2_all*         up   infinite     11  drain c[0807,0909,1303,1307,1503,1706,2302,2308,2508,
2609,9501]
v2_all*         up   infinite    296  alloc c[0101-0110,0201-0210,0407-0410,0503,0506-0507,
0602-0610,0701-0710,0801,0803-0804,0806,0808-0809,0901-0908,0910,1001-1010,1102-1110,1201-1
210,1301-1302,1305-1306,1308-1310,1401-1410,1501-1502,1504-1510,1602-1610,1701-1704,1707-17
10,1801-1808,1810,1901,1904-1905,1910,2001-2010,2101-2110,2201-2206,2208-2210,2301,2303-230
5,2307,2309-2310,2401-2405,2407-2410,2501-2507,2603-2608,2610,2701-2710,8707,9007,9102-9110
,9201-9210,9301-9310,9401-9405,9407-9408,9410,9503-9510,9601-9610,9701-9707,9709-9710,9801-
9810,9901-9905]
v2_all*         up   infinite     64   idle c[0301-0310,0401-0406,0501-0502,0504-0505,1705,
2207,2509-2510,2601-2602,8701-8706,8708-8710,8801-8810,8901-8910,9001,9003-9006,9008-9010,9
```

其中，

第一列 PARTITION 是队列名。

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。

sinfo 的常用命令选项：

| 命令示例 | 功能 |
|---|---|
| sinfo -n c12345 | 指定显示节点 c12345 的使用情况 |
| sinfo -p v2_all | 指定显示队列 v2_all 情况 |

其他选项可以通过 sinfo --help 查询

## 3.2 squeue 查看作业状态

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

具体示例见下图：

```
@ln1%cngrid2 ~]$ squeue
JOBID     PARTITION             NAME     USER ST       TIME  NODES NODELIST(REASON)
   487622         v3_ib Ret540_mgz_full.   deploy  R 2-17:59:00      2 e[3102-3103]
   487790  pg2_128_pool 01-mpiexec-4-thr   pg2034  R 2-06:49:05      1 c0104
   487795  pg2_128_pool 01-mpiexec-4-thr   pg2034  R 2-06:25:10      1 c0607
   487796  pg2_128_pool 01-mpiexec-1-thr   pg2034  R 2-06:18:36      1 c0608
   488051       vip_17       c02_wlh.sh   pg3125  R 2-04:46:37      1 e1804
   488055  pg2_128_pool       neb-z-u.sh   pg2087  R 2-04:24:38      2 c[9108,9208]
   488064       v2_all     ych-re400Pr1   pg2057  R 2-04:37:45      4 c[4508-4510,4901]
   488164  pg2_128_pool       neb-y-u.sh   pg2087  R 2-04:16:28      2 c[9104-9105]
   488167       v2_all         cp2k.sub   pg2213  R 2-04:13:58      1 c0708
   488170  v3_64_single        molpro.sh   pg3027  R 2-03:58:04      2 d[1605-1606]
   488310       v2_all         Ni-clm.sh   pg3089  R 2-01:38:25      2 c[4401-4402]
   488373       v2_all         Ru-clm.sh   pg3089  R 2-01:05:07      2 c[1906,2603]
```

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。

squeue 的 常用命令选项：

| 命令示例 | 功能 |
|---|---|
| squeue -j 123456 | 查看作业号为 123456 的作业信息 |
| squeue -u pg2011 | 查看超算账号为 pg2011 的作业信息 |
| squeue -p v2_all | 查看提交到 v2_all 队列的作业信息 |
| squeue -w c123 | 查看使用到 c123 节点的作业信息 |

其他选项可通过 squeue --help 命令查看

## 3.3 srun 交互式提交作业

srun [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。

srun 命令示例：

```
srun -p v2_all -w c[1100-1101] -N 2 -n 40 -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 srun -p v2_all -n 40 A.exe

其中，

-p v2_all 指定提交作业到 v2_all 队列；

-w c[1100-1101] 指定使用节点 c[1100-1101]；

-N 2 指定使用 2 个节点；

-n 40 指定进程数为 40，cngrid2 超算 v2_all 队列和 pg2_128_pool 队列每一个节点 20 核，v3_64_single 队列每个节点 16 核，建议使用单节点核数的整数倍提交作业；

-t 20 指定作业运行时间限制为 20 分钟。

srun 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 20 | 指定进程数为 20，cngrid2 普通队列（v2_all 队列和 pg2_128_pool 队列）每个节点 20 核，建议满核提交 |
| -c 20 | 指定每个进程（任务）使用的 CPU 核为 20 |
| -p v3_64_single | 指定提交作业到 v3_64_single 队列 |
| -w c[100-101] | 指定提交作业到 c100、c101 节点 |
| -x c[100,106] | 排除使用 c100、c106 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 |

srun 的其他选项可通过 srun --help 查看。

## 3.4 sbatch 后台提交作业

sbatch 一般情况下与 srun 一起提交作业到后台，需要将 srun 写到脚本中，再用 sbatch 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。sbatch 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。

sbatch 命令示例 1（40 个进程提交 A.exe 程序）：

编写脚本 job1.sh，内容如下：

```
#!/bin/bash
srun -n 40 A.exe
```

然后在命令行执行 sbatch -p v2_all 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。

sbatch 命令示例 2（指定 2 个节点，4 个进程，每个进程 10 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 10

#SBATCH -t 60

srun -n 4 A.exe
```

然后在命令行执行 sbatch -p v2_all 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。

sbatch 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
srun -n 5 A.exe &
srun -n 5 B.exe &
srun -n 5 C.exe &
srun -n 5 C.exe &
wait
```

然后在命令行执行 sbatch -N 1 -p v2_all job3.sh，这里是单节点同时提交 4 个任务，每个任务使用 5 个进程。这里需要 4 个任务同时运行，并且全部执行完毕，作业才会退出。

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

## 3.5 salloc 分配模式作业提交

salloc 命令用于申请节点资源，一般用法如下：

1、执行 salloc -p v2_all；
2、执行 squeue 查看分配到的节点资源，比如分配到 c100；
3、执行 ssh c100 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 scancel JOBID 释放分配模式作业的节点资源。

## 3.6 scancel 取消已提交的作业

scancel 可以取消正在运行或排队的作业。

scancel 的一些常用命令示例：

| 命令示例 | 功能 |
| --- | --- |
| scancel 123456 | 取消作业号为 123456 的作业 |
| scancel -n test | 取消作业名为 test 的作业 |
| scancel -p v2_all | 取消提交到 v2_all 队列的作业 |
| scancel -t PENDING | 取消正在排队的作业 |
| scancel -w c100 | 取消运行在 c100 节点上的作业 |

scancel 的其他参数选项，可通过 scancel --help 查看

## 3.7 scontrol 查看正在运行的作业信息

scontrol 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。

scontrol 的常用示例：

```
scontrol show job 123456
```

查看作业号为 123456 的作业详情。

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

## 3.8 sacct 查看历史作业信息

sacct 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。

sacct 的常用命令示例：

```
sacct -u pg2011 -S 2017-09-01 -E now --field=jobid,partition,jobname,user,nnodes,nodelist,start,end,elapsed,state
```

其中，-u pg2011 是指查看 ppg2011 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。sacct --helpformat 可以查看支持的输出格式。

sacct 的其他参数选项可通过 sacct --help 查看。

# 4 编译器

cngrid2 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

cngrid2 可通过 module load 加载 intel 编译环境，例如 module load intel/18.0.2-fast，如下图 1 所示：

```
[deploy@ln1%cngrid2 ~]$ module avail mpi
------------------------- /public1/soft/modulefiles -------------------------
mpi/intel/5.0.3.049-intel15.0.6       mpi/mvapich2/2.2-gcc
mpi/intel/5.0.3.049-intel15.0.6-fast  mpi/openmpi/2.0.0
mpi/intel/18.0.2                      mpi/openmpi/3.0.1-pmi-gcc
mpi/intel/18.0.2-fast                 mpi/openmpi/3.0.1-pmi-icc15
mpi/mpich/3.2.1-icc18                 mpi/openmpi/3.0.1-pmi-icc18
[deploy@ln1%cngrid2 ~]$ module avail intel
------------------------- /public1/soft/modulefiles -------------------------
intel/15.0.6  intel/15.0.6-fast  intel/18.0.2  intel/18.0.2-fast
[deploy@ln1%cngrid2 ~]$ module load intel/18.0.2-fast
```

图 1 module 使用示意图

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。

## 4.2GCC 编译器

cngrid2 默认的 GNU 编译器版本是 4.8.5，如需其他版本，可通过 module load 加载，例如 module load gcc/5.3.0。

## 4.3 MPI 编译环境

cngrid2 使用的 MPI 编译环境必须通过 module load 加载。例如 module load mpi/intel/18.0.2-fast

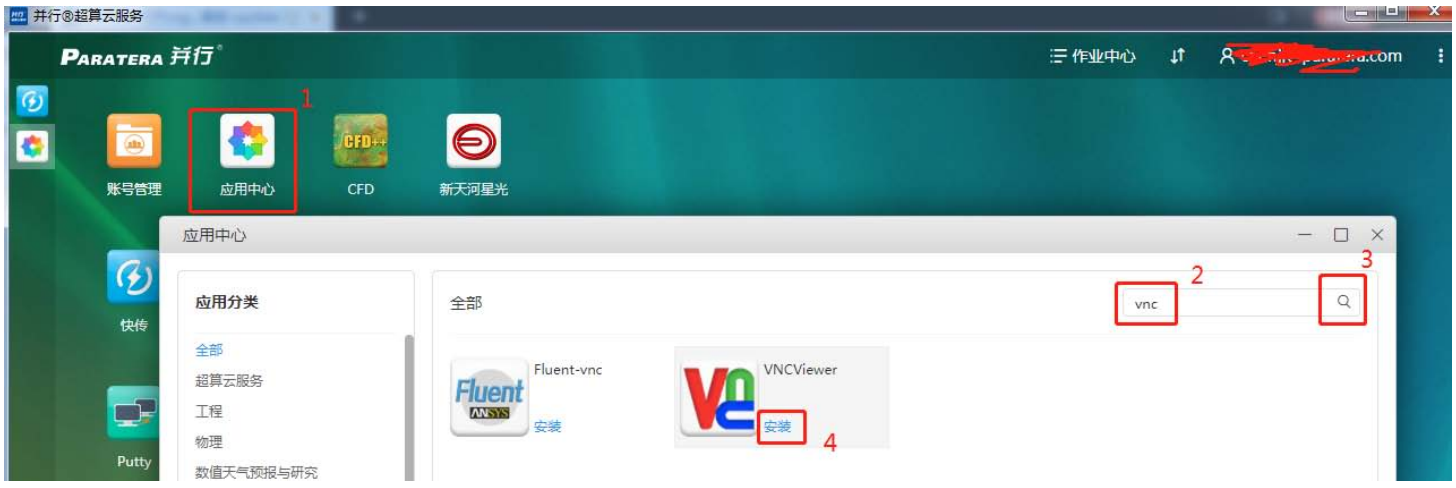
**温馨提示：**

- 编译软件建议采用 intel 编译器（intel/18.0.2-fast）和 intel mpi（mpi/intel/18.0.2-fast）；后台提交作业建议加载相应 intel 库（intel/18.0.2）和 intel mpi 库（mpi/intel/18.0.2）。
- cngrid2 上部署了一些开源软件，比如 python，lammps，namd 等软件，可执行 module avail 查看，module load 直接调用。
- cngrid2 可以连外网，您可以连网安装一些软件。

# 5 图形化界面

cngrid2 可以打开图形化界面，具体方法如下：

1. 联系我们开通可视化功能（可以在微信群里提出）
2. 开通后，在“并行@超算云服务软件”的“应用中心”安装 VNC 软件，操作方法见下图




3. 点击并行@超算云服务软件界面的 VNC 软件，选择自己账号，点击连接
4. 即可使用图形界面
5. 结束使时，一定要点击停止 vnc 服务，结束作业，以免浪费机时