闽都创新实验室高性能计算集群使用说明

v. 2022-1

2022 年 10 月 11 日

目录

1 高性能计算平台简介

闽都创新实验室公共测试平台高性能计算集群共有 102 个节点，配置如下：

节点类别	型号	主要规格	节点数
登录节点	PR2710P	32 核 (2*Intel Xeon Gold 6226R) 64 GB 内存 (4*16GB) 200 TB 硬盘 (外接磁盘阵列)	1
普通节点	PR2740TP	32 核 (2*Intel Xeon Gold 6226R) 192 GB 内存 (12*16GB) 600 GB SAS 硬盘 InfiniBand	84
高 I/O 节点	PR2710P	32 核 (2*Intel Xeon Gold 6226R) 192 GB 内存 (6*32GB) 4 TB 硬盘 (7*600GB SAS RAID0)	12
大内存节点	PR2710P	32 核 (2*Intel Xeon Gold 6226R) 512 GB 内存 (16*32GB) 2.4 TB SAS 硬盘	4
编译调试节点	同高 I/O 节点		1

集群配置了常用的电子结构、固体物理、分子动力学软件，如 Gaussian，Molpro，Q-Chem，NWChem，VASP，Gromacs 等，采用 slurm 队列管理系统。集群于 2022 年 5 月 10 日投入试运行，2022 年 10 月 15 日正式上线运行。

集群管理员联系方式：

陈 俊
海西院工作区 2 号楼 1325 室
chenjun@fjirsm.ac.cn

张 豪
海西院工作区 2 号楼 1011 室
zhanghao1206@fjirsm.ac.cn

对本文档的任何建议和意见，或发现错误，请联系集群管理员（chenjun@fjirsm.ac.cn）。

2 快速入门

在获取上机账号后，Linux/Mac 用户可通过命令行登录集群，Windows 用户则可以通过 ssh 客户端登录集群；登录集群后，编写作业脚本，并通过 sbatch 指令将作业提交到计算节点上执行。集群上安装了常见的计算软件，安装位置为 /share/soft 。

先登录集群，Linux/Mac 用户直接使用系统自带的终端通过 ssh 命令登录：

```
ssh user_name@ip_address
```

Windows 用户可使用 Xshell 客户端进行登陆。下载 Xshell，选择免费许可版，为了方便文件传输，可同时下载并安装 Xftp，安装完后点击软件左上角新建连接，输入 IP 和用户名密码即可登录。更多信息请参考 3 连接集群 章节。

登录后，编写作业脚本，并通过 sbatch 命令将作业提交到计算节点上执行。用 squeue 命令查看作业状态，scancel 命令中止作业，sinfo 命令查看集群节点使用状态。

假设我们的计算过程为：在计算节点上运行 hostname 指令，那么就可以这么编写作业脚本；

```
#!/bin/bash
#SBATCH -o job.%j.out
#SBATCH -p debug
#SBATCH -J myFirstJob
#SBATCH -N 1
#SBATCH -n 1

hostname
```

假设上面作业脚本的文件名为 job.sh，通过以下命令提交：

```
sbatch job.sh
```

集群提供了几种常见计算软件的任务提交命令，如 Gaussian，VASP 等。

Gaussian 的任务提交命令为 qg16，假设用户的输入文件为 h2o.gjf，输入文件内指定使用 32 核进行计算，则可以通过以下命令将任务提交到 small 队列的 1 台节点：

```
qg16 h2o.gjf small
```

VASP 的任务提交命令为 qvasp，假设用户在当前目录下已准备完整相关输入文件（如 INCAR，POSCAR，POTCAR，KPOINTS），则可以通过以下命令将任务提交到 small 队列的 1 台节点并使用 32 核进行 MPI 并行计算（用户需指定 VASP 版本，如 vtst，gam 等）：

```
qvasp vtst 32 small
```

3 连接集群

Linux/Mac 用户直接使用电脑自带的命令行终端通过 ssh 登录，通过 scp 或者 rsync 传输文件；

Windows 用户可以通过 Xshell+Xftp 进行登录和传输数据；

Mac 用户可以通过 Cyberduck 传输数据（需在创建书签时设置“传输文件”类型为“使用浏览器连接”）；

需要用到图形界面的 Windows 用户可以使用 MobaXterm 等客户端。

集群 IP 地址若忘记，请询问集群管理员。

3.1 Linux/Mac 登录和传输数据

Linux/Mac 用户通过自带的命令行终端，以 ssh 的方式登录；

打开终端后通过以下命令连接：

```
# user_name 为上机账号，ip_address 为所要连接集群的 IP
ssh user_name@ip_address
```

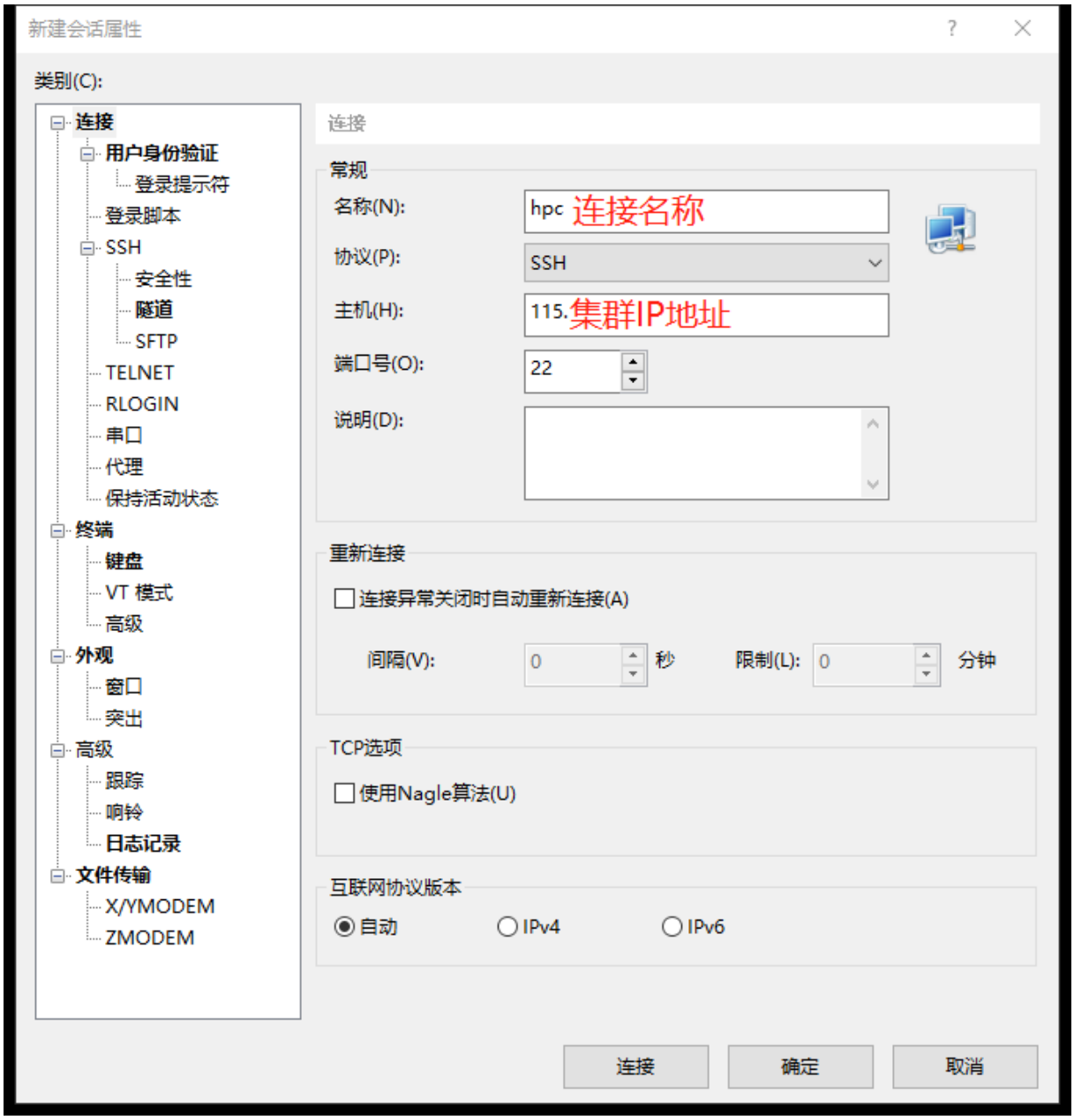
通过 scp 或者 rsync 传输数据：

```
scp [-r] file username@ip_address:~/
# 把本地的名为 file 的文件上传到集群的 home 目录下
# 如果 file 是一个目录的话还需要加上参数 r
scp [-r] username@ip_address:~/file ./
# 把服务器上 home 下名为 file 的文件下载到本地当前位置
rsync [-Pr] file username@ip_address:~/
# 把本地的名为 file 的文件同步到集群的 home 目录下
# 如果 file 是一个目录的话还需要加上参数 r
rsync [-Pr] username@ip_address:~/file ./
# 把服务器上 home 下名为 file 的文件同步到本地当前位置
```

3.2 Windows 用户使用客户端

Windows 用户可以使用客户端登录集群、上传数据，常见的客户端有 Xshell+Xftp，下载Xshell+Xftp时选择免费许可版，安装后通过以下方式登录集群；

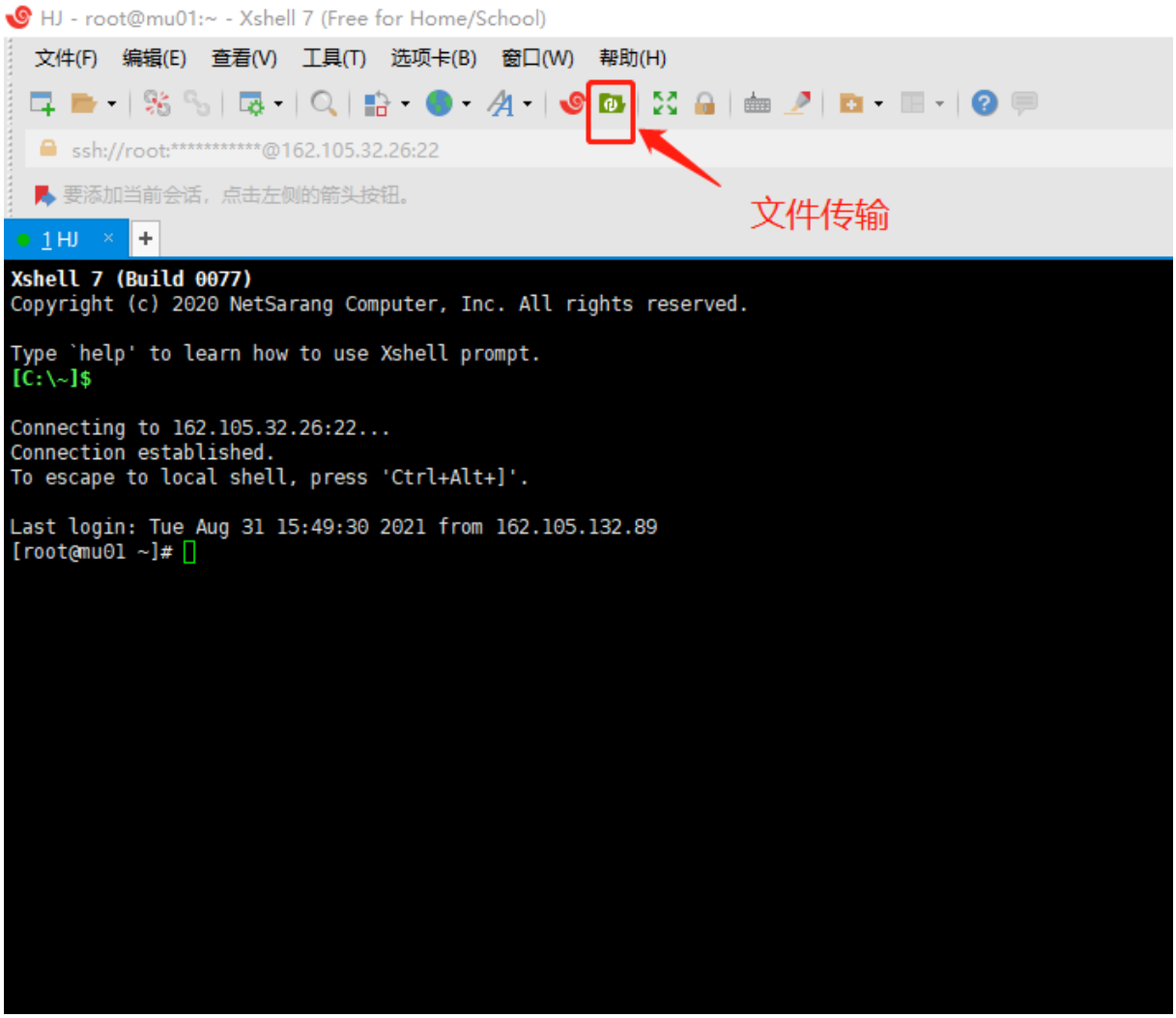
首先，打开 Xshell 客户端后，点击左上角"文件"»"新建"，输入连接的名称和集群的 IP 地址；



然后，点击连接按钮，输入用户名和密码，此处验证方式一定要选择 Keyboard Interactive 而不是 Password；

下次登录集群时，点击" 文件"»" 打开"，找打对应的连接名称，选择连接即可。

最后，登录后，在安装 Xftp 的前提下，打开文件传输按键，如下图所示；



3.3 图形界面

通常情况下，只需通过命令行即可提交并运行任务，有些情况下，用户会使用到图形界面来进行调试等。

3.3.1 Linux

Linux 用户使用图形界面时，只需在登录时加入-X 参数即可使用图形界面：

```
1 ssh -X user_name@ip_address
```

3.3.2 Mac

MacOS 用户使用图形界面时需要安装第三方支持 X11 的软件 Xquartz。安装完后，在登录时加入-X 参数即可使用图形界面：

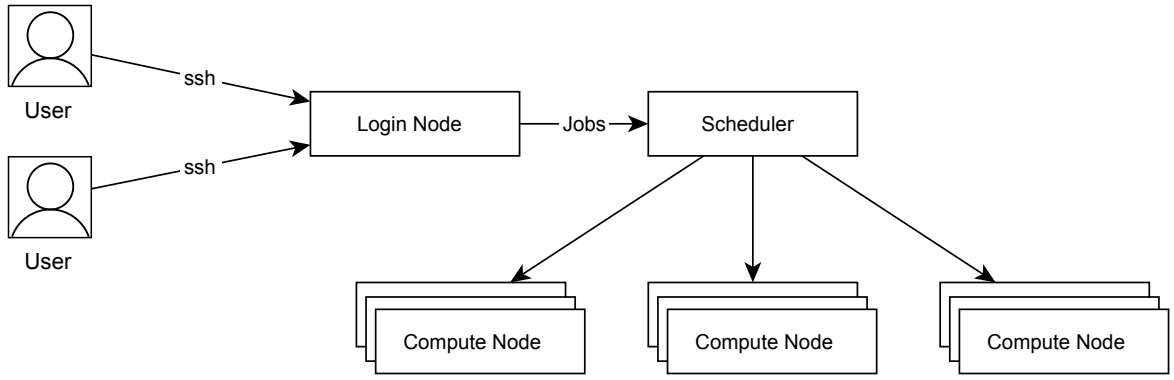
```
1 ssh -X user_name@ip_address
```

3.3.3 Windows

Windows 用户推荐使用 MobaXterm 客户端，在默认情况下 MobaXterm 会自动打开 X11 的转发。也可使用 Xshell+Xmanager 来使用图形界面，但 Xmanager 为收费软件。也可使用 Putty+Xming 来使用图形界面。

4 提交作业

登录后进入的是登录节点，并不能直接运行程序，需要通过作业调度系统，将作业提交到计算节点上执行；



运行作业的方式有两种：

一种是将计算过程写成脚本，通过 sbatch 指令提交到计算节点执行；

另一种是通过 salloc 申请到计算节点，再 ssh 连接到计算节点进行计算；

通过 sinfo 查看计算节点空闲状态；

可以通过 squeue 查看已经提交作业的排队情况；

通过 scancel 取消已经提交的作业。

4.1 sbatch 提交作业

将整个计算过程，写到脚本中，通过 sbatch 指令提交到计算节点上执行；

先介绍一个简单的例子，随后介绍例子中涉及的参数，接着介绍 sbatch 其他一些常见参数，最后再介绍 GPU 和 MPI 跨节点作业案例。

首先是一个简单的例子；

假设我们的计算过程为：在计算节点上运行 hostname 指令，那么就可以这么编写作业脚本；

```
1 #!/bin/bash
2 #SBATCH -o job.%j.out
3 #SBATCH -p debug
4 #SBATCH -J myFirstJob
5 #SBATCH -N 1
6 #SBATCH -n 1
7
8 hostname
```

假设上面作业脚本的文件名为 job.sh，通过以下命令提交：

```
sbatch job.sh
```

随后我们介绍脚本中涉及的参数；

```
-o job.%j.out   # 脚本执行的输出将被保存在当job.%j.out文件下，%j表示作业号；
-p debug        # 作业提交的指定分区（队列）为debug；
-J myFirstJob   # 作业在调度系统中的作业名为myFirstJob；
-N 1            # 申请节点数为1，如果作业不能跨节点（MPI）运行，申请的节点数应不超过1；
-n 1            # 每个节点上运行一个任务，默认情况下也可理解为每个节点使用一个核心，如果程序不支持多线程（如openmp），这个数不应该超过1；
```

其中：

-p 指定作业的运行分区 (队列)。提交作业时必须指定分区，每个分区有不同的属性，如 theory/community 分区，每个节点核心数为 32，内存为 192G。可以通过 sinfo 查看分区的空闲状态；

下面是一个跨节点多核心的例子；

假设我们想用两个节点，每个节点 32 个核心来运行 VASP，那么可以这么编写作业脚本；

```
#!/bin/bash
#SBATCH -o job.%j.out
#SBATCH -p community
#SBATCH -J myFirstMPIJob
#SBATCH -N 2
#SBATCH --ntasks-per-node=32

# 写入环境变量
 ulimit -s unlimited
 export OMP_NUM_THREADS=1
 export PATH=/share/soft/vasp/vasp_6.3.0_libxc_5.2.2_wannier90_3.1.0/bin:/slurm/bin:/slurm/sbin:/usr/lib64/qt-3.3/bin:/usr/local/sbin:/usr/local/bin:/usr/sbin:/usr/bin:/root/bin:/share/xena/bin:.
```

```
12  export LD_LIBRARY_PATH=/slurm/lib:/lib64:/usr/
      lib64:/usr/local/lib64:/usr/lib64/atlas:/usr/
      lib64/atlas-sse3
13
14  source /share/soft/intel/2018.4/bin/compilervars.
      sh intel64
15  source /share/soft/intel/2018.4/mkl/bin/mklvars.sh
       intel64
16  source /share/soft/intel/2018.4/impi/2018.4.274/
      bin64/mpivars.sh
17
18 # 生成 machinefile
19 srun hostname -s | sort -n >slurm.hosts
20
21 # 执行MPI并行计算程序
22 mpirun -n 64 -machinefile slurm.hosts vasp_std >
     log
```

4.2 salloc 交互运行作业

salloc 申请计算节点，然后登录到申请到的计算节点上运行指令；

salloc 的参数与 sbatch 相同，该部分先介绍一个简单的使用案例；随后介绍一个 GPU 的使用案例；最后介绍一个跨节点使用案例；

首先是一个简单的例子；

申请一个节点 6 个核心，并跳转到该节点上运行程序；

```
1 salloc -p debug -N1 -n6 -t 2:00:00
2 # salloc 申请成功后会返回申请到的节点和作业ID等信
    息，假设申请到的是gibbs-io13节点，作业ID为
    1078858
3 ssh gibbs-io13     # 直接登录到刚刚申请到的节点gibbs
    -io13调式/运行作业
4 scancel 1078858    # 计算资源使用完后取消作业
5 squeue -j 1078858 # 查看作业是否还在运行，确保作业
    已经退出，避免产生不必要的费用
```

下面介绍一个跨节点使用案例；

申请两个节点，每个节点 32 个核心

```
salloc -p community -N2 -ntasks-per-node=32
# salloc 申请成功后会返回申请到的节点和作业ID等信息，假设申请到的是gibbs[55-56]节点，作业ID为1078858
# 这里申请两个节点，每个节点32个进程，每个进程一个核心

# 根据需求导入MPI环境
source /share/soft/intel/2018.4/impi/2018.4.274/bin64/mpivars.sh

# 根据以下命令生成MPI需要的machine file
srun hostname -s | sort -n > slurm.hosts

mpirun -np 64 -machinefile slurm.hosts vasp_std

# 结束后退出或者结束任务
scancel 1078858
```

4.3 sinfo 查看资源空闲状态

通过 sinfo 可查询各分区节点的空闲状态；

sinfo 命令显示集群的所有分区节点的空闲状态，idel 为空闲，mix 为节点部分核心可以使用，alloc 为已被占用；

```
sinfo
```

也可只显示指定分区的空闲状态，如查看 community 分区的状态：

```
sinfo -p community
```

4.4 squeue 查看作业队列

squeue 可查看提交作业的排队情况及运行情况，PD 表示排队状态，R 表示运行状态；

运行状态的作业会显示已分配的节点名，这时也可以通过 ssh 直接登录到相应计算节点，再用 top 命令查看系统状态，或用 df 命令查看计算节点本地硬盘使用情况。

4.5 scancel 取消作业

介绍几个使用使用案例，分别是，取消指定作业、取消自己上机账号上所有作业、取消自己上机账号上所有状态为 PENDING 的作业，最后介绍 scancel 常见的参数。

取消指定作业；

```
# 取消作业ID为123的作业
scancel 123
```

取消自己上机上号上所有作业；

```
# 注意whoami前后不是单引号，而是键盘左上的Shift+~
scancel -u `whoami`
```

取消自己上机账号上所有状态为 PENDING 的作业；

```
# 注意whoami前后不是单引号，而是键盘左上的Shift+~
scancel -t PENDING -u `whoami`
```

scancel 常见参数；

```
--help          # 显示scancel命令的使用帮助信息；
-A <account>    # 取消指定账户的作业，
                #如果没有指定job_id，将取消所有作业；
-n <job_name> # 取消指定作业名的作业；
-p <partition_name> # 取消指定分区的作业；
-t <job_state_name> # 取消指定作态的作业，如
        # "PENDING", "RUNNING" 或 "SUSPENDED";
-u <user_name> # 取消指定用户下的作业；
```

5 使用软件

平台上预安装了部分常用软件，用户可进入 /share/soft 目录查看，可进入相应的子目录查看软件版本（若有多版本的软件同时存在）。

后期也将可用 module 指令进行查看和加载使用已有软件。

用户还可以自行将软件编译安装到自己的 HOME 目录下。

5.1 Gaussian

Gaussian 是一个功能强大的量子化学综合软件包，平台并未购买该软件，用户需要确保版权。福建物构所结构化学国家重点实验室具有 Gaussian 的版权。

Gaussian 软件的使用需要用户属于 gaussian 组，有使用需求的用户需在申请账号时在申请表中勾选。

平台预装了多个版本的 Gaussian 软件，如 G09D01，G16A03SSE4，G16C01AVX，用户可按需使用。

Gaussian 软件的使用需在 shell 环境变量中作如下设置：

```
# 指定特定版本Gaussian的安装目录
export g16root='/share/soft/gaussian/G16C01AVX'
# 加载软件相关环境变量
source ${g16root}/g16/bsd/g16.profile
# 指定scratch目录。用户若发现计算节点的/tmp目录被写满，请务必联系管理员（chenjun@fjirsm.ac.cn）
export GAUSS_SCRDIR=/tmp
# 使用g16命令对输入文件 h2o.gjf 进行计算
g16 h2o.gjf
# 计算完后会生成输出文件 h2o.log ，以及用户在输入文件中指定的 .chk 文件等
```

平台也提供了 G16C01AVX 版本的简要提交方式，用户需将 /share/xena/sample-script-theory 目录（理论组）或 /share/xena/sample-script-community 目录（实验组）添加到用户的 shell 环境变量中，使用 qg16 命令提交计算任务。

```
# 用vim编辑 ~/.bashrc ，将以下内容加到文件的最后
export PATH=/share/xena/sample-script-theory:${PATH}
# 或
export PATH=/share/xena/sample-script-community:${PATH}
# 保存文件并退出vim，source 使环境变量生效
source ~/.bashrc
# 提交输入文件 h2o.gjf 进行计算
```

```
qg16 h2o.gjf
sinfo
```

5.2 VASP

VASP 是维也纳大学 Hafner 小组开发的进行电子结构计算和量子力学-分子动力学模拟软件包。它是目前材料模拟和计算物质科学研究中最流行的商用软件之一。平台并未购买该软件，用户需要确保版权。

平台预装了多个版本的 VASP 软件，如 5.4.4，5.4.4_mbd，5.4.4_sol，5.4.4_vtstcode，5.4.4_vtstcode_sol，6.2.0，6.2.0_libxc_4.3.4，6.3.0_libxc_5.2.2_wannier90_3.1.0 等。

其中，

_sol 版本可用于电化学溶剂化模型计算；

_vtstcode 版本可用于 CI-NEB 方法的过渡态优化；

_mbd 版本可用于计算 DFT-MBD@rsSCS/FI 方法的色散校正；

_libxc 版本支持相应的 libxc 泛函库；

_wannier90 版本支持调用 wannier90 程序。

以上所有版本均使用 intel 2018.4 版本编译器套装进行编译，因此软件使用需预先加载 intel 相关环境变量。然后，用户需将想要使用的版本下 bin 目录添加到环境变量 PATH 中，方可使用。每个版本下又分别提供了 vasp_std，vasp_gam，vasp_ncl 等三个可执行文件。

例如，用户想要使用 vasp_5.4.4_vtstcode_sol 对电化学体系进行 CI-NEB 过渡态结构搜索，需设置以下环境变量：

```
# 加载环境变量
export OMP_NUM_THREADS=1
source /share/soft/intel/2018.4/bin/compilervars.sh
    intel64
source /share/soft/intel/2018.4/mkl/bin/mklvars.sh
   intel64
source /share/soft/intel/2018.4/impi/2018.4.274/
   bin64/mpivars.sh
export PATH=/share/soft/vasp/vasp_5.4.4
   _vtstcode_sol/bin:${PATH}
# 使用MPI进行并行计算
mpirun -n 32 vasp_std
# 计算完成后会生成OUTCAR等文件
```

平台也提供了 VASP 的简要提交方式，与前述 Gaussian 的简要提交方式一样，用户需将 /share/xena/sample-script-theory 目录（理论组）或 /share/xena/sample-script-community 目录（实验组）添加到用户的 shell 环境变量中，使用 qvasp 命令提交计算任务。

6 常见问题

1. 为什么 sinfo 查看对应的分区有空闲节点，但是我的作业却还在排队

闽都创新实验室高性能计算平台采用 slurm 作业调度系统，整个队列中可能有需要占用多节点的高优先级任务正在等待资源，调度器会一定程度上为这些作业保留资源，以确保它们能够运行。

2. error: Job submit/allocate failed: Invalid partition name specified

错误原因：未指定正确的 partition，可通过 sinfo 指令获取可用的分区。

3. (QOSMaxWallDurationPerJobLimit)

错误原因：指定-t, –time= 参数时，时间超过 qos 允许的时长，通过以下命令可以查看相应 qos 允许运行的最大时长。可通过 sinfo 指令，在 TIMELIMIT 列获取可用分区的相应时长限制。

4. batch job submission failed: Requested node configuration is not available

错误原因：申请资源的节点配置不匹配，如 theory/community 的每个节点只有 32 个核心，但你如果单节点申请的核心数超过 32，就会报错。