高性能计算平台 CPU1/GPU 队列
使用手册

南京航空航天大学高性能计算中心

2023 年 2 月

目 录

1 平台资源概述

南京航空航天大学高性能计算平台（以下简称平台）是依托工信部十三五信息化专项--南京航空航天大学智慧校园建设的校级科学仪器公共平台，平台包含了 CPU0 队列（部署于云平台）、CPU1 队列（部署于平台一期）、CPU2 队列（部署于平台二期）、GPU 队列（部署于平台一期）、GPU 人工智能队列（部署于 AI-GPU 平台）五部分，可提供 CPU 计算和 GPU 计算所需资源。

1.1 CPU1/GPU 队列

高性能计算平台 CPU1/GPU 队列硬件资源如下表所示，用户需要根据算例情况和计算需求从计算资源中选取合适的队列，确定 CPU 核数或 GPU 卡数等参数编写脚本才能提交作业。

表 1 CPU1/GPU 队列配置

队列名	CPU 类型	GPU 类型	内存	节点数
cpu	2 颗 Intel Xeon Cascade Lake Gold 6248，2.5GHz，20 核	/	192G	62
gpu4	2 颗 Intel Xeon Cascade Lake Gold 4210，2.4GHz，10 核	4 块 NVidia Tesla PCIE V100 32GB 显存	192G	29
gpu8	2 颗 Intel Xeon Cascade Lake Gold 6248，2.5GHz，20 核	8 块 NVidia Tesla PCIE V100 32GB 显存	192G	2

表 2 存储配置

型号	主要规格	套数
联想 DSS-G	硬盘 334*6T HDD 2*800G SSD 千兆网口 8*10G IB 网口 8*100GB InfiniBand 端口	1
存储容量合计：2PB 读写带宽 20GB/s		

1.2 应用软件资源

平台上已经安装常用应用软件，所在目录为/fs0/software，用户可进入该目录查看可用软件，编译或运行程序需设置正确的环境变量。

2 快速入门

本章让用户快速掌握如何使用高性能计算平台 CPU1 队列提交作业。

用户在获取上机账号后，Linux/Mac 用户可直接使用系统通过 ssh 命令：ssh 用户名@sc.nuaa.edu.cn 登录队列，Windows 用户则可以通过 ssh 客户端（例如 Xshell）登录队列（注意，队列首次登录详细步骤见第 3 章），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。

登录队列后，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；此外，队列上安装了常见的计算软件，通过 module 指令导入计算环境。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 -J test
#SBATCH -p cpu
#SBATCH -n 40
#SBATCH --error=%J.err
#SBATCH --output=%J.out

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
```

队列安装了常见的计算软件，可以通过 module 指令导入计算环境；

可以通过 module 加载平台上装有的软件环境，也可以自行安装配置需要的计算环境，下面的作业脚本加载了 intel/2017u5 的软件环境，具体可用的软件环境可使用命令 module avail 指令进行查看。

```
#!/bin/bash
#SBATCH -J test
#SBATCH -p cpu
#SBATCH -n 40
#SBATCH --error=%J.err
#SBATCH --output=%J.out

module purge
module load intel /2017u5
```

3　登录队列和文件上传下载

3.1　Windows 系统登录

（1）登录方式 1：直接用命令行 ssh 登录

以 Xshell 远程登录工具为例，输入命令：ssh user_name@sc.nuaa.edu.cn，回车。user_name 为用户平台账号，弹出用户身份验证框后输入用户密码，用户密码由管理员开户时发送至用户邮箱，注意平台账号字母均为小写字母。如图 1 所示：

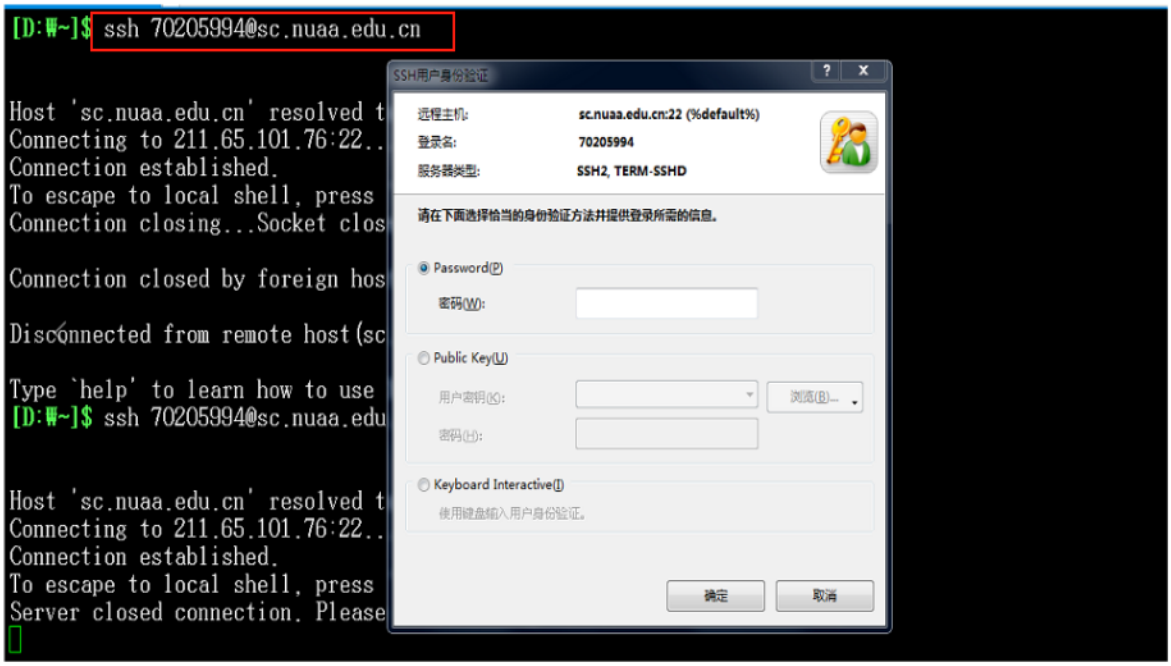

图 1 远程登录

如遇下述情况需要保留主机秘钥，选择“接受并保存”，如图 2 所示：

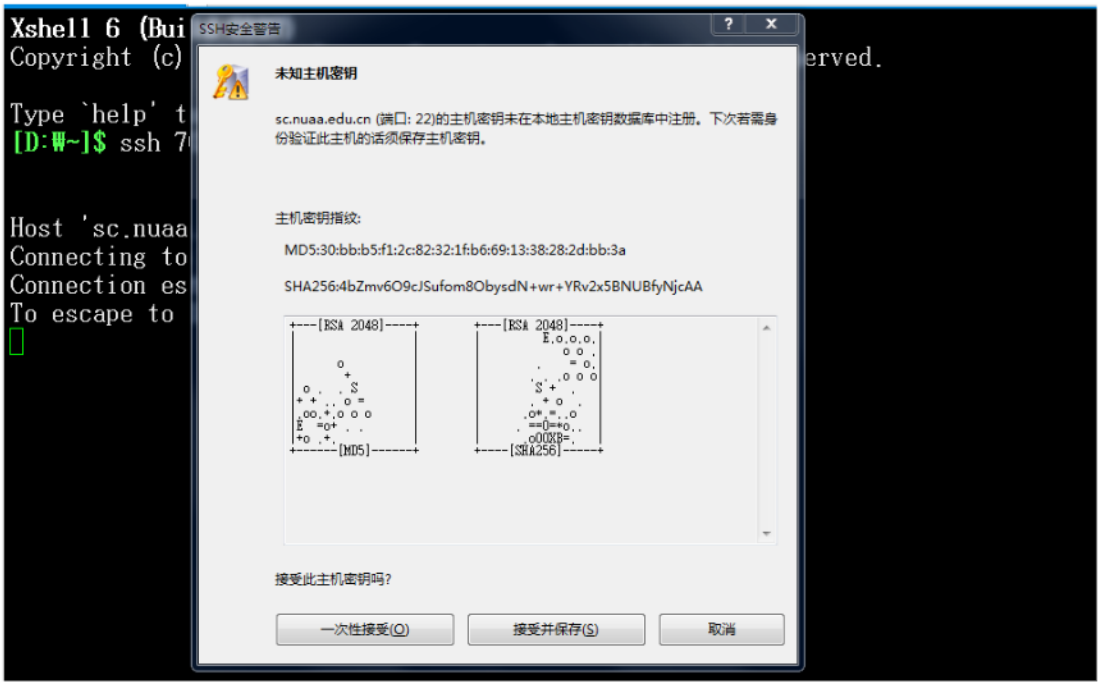

图 2 接收秘钥

如下图 3 所示表示登录成功:

图 3 登录成功

(2)登录方式 2: 新建连接，输入主机名后登录

1) 进入 Xshell，点击新建连接图标（如图 4 所示），或者点击文件 -> 新建（如图 5 所示）。

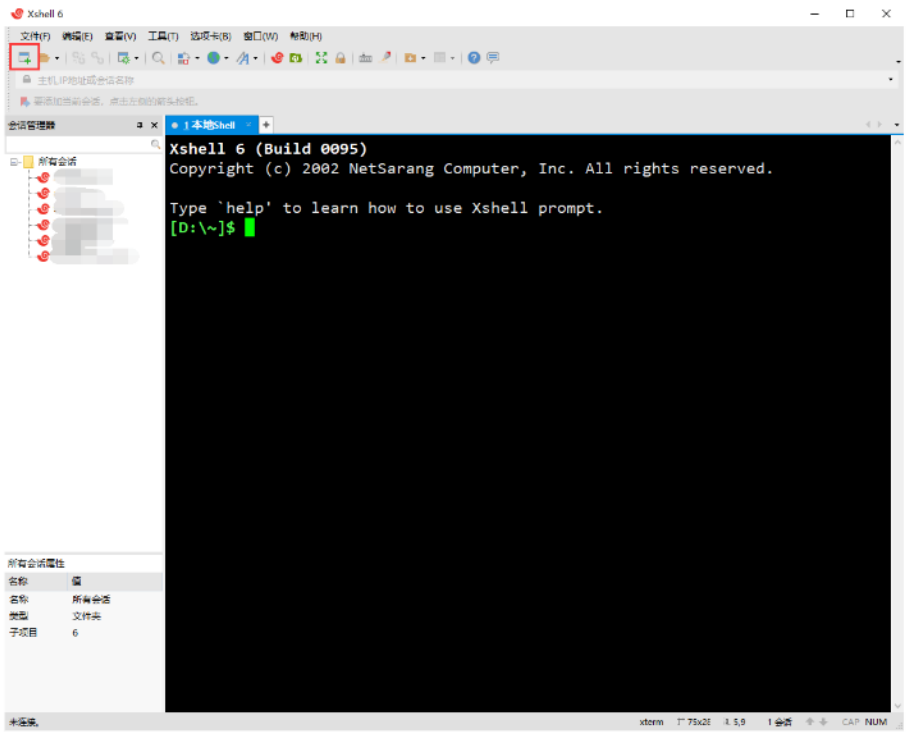

图 4 新建连接

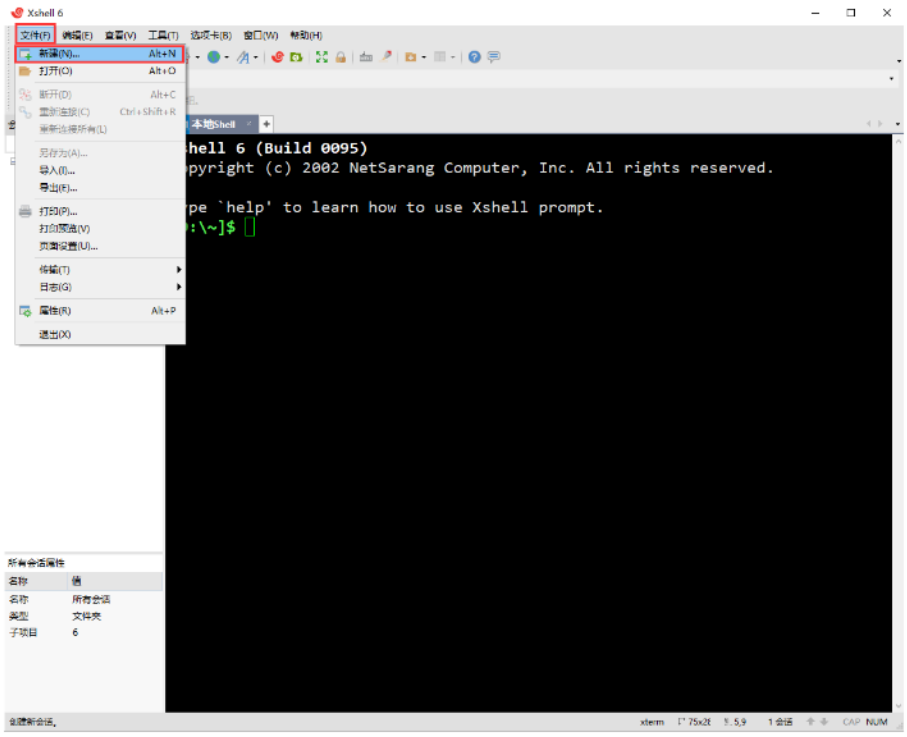

图 5 文件-连接

在弹出的新建连接对话框中输入名称，如 C1，在主机中输入 sc.nuaa.edu.cn，端口号输入 22，点击确定。

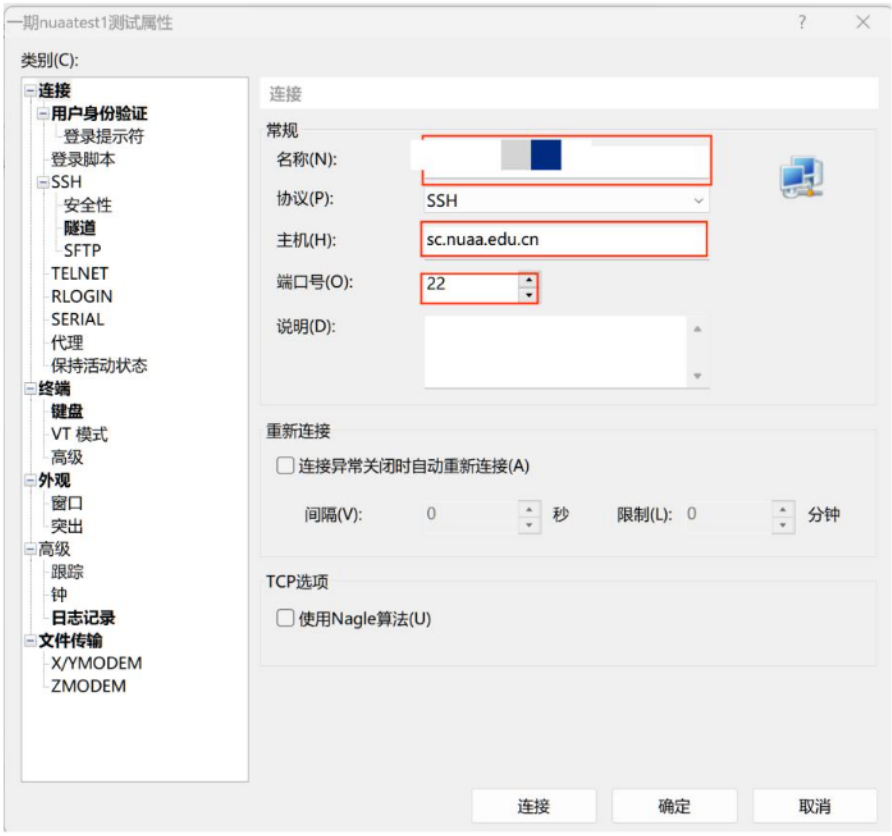

图 6 新建会话属性

在左侧会话管理器中双击 C1，如图 7。

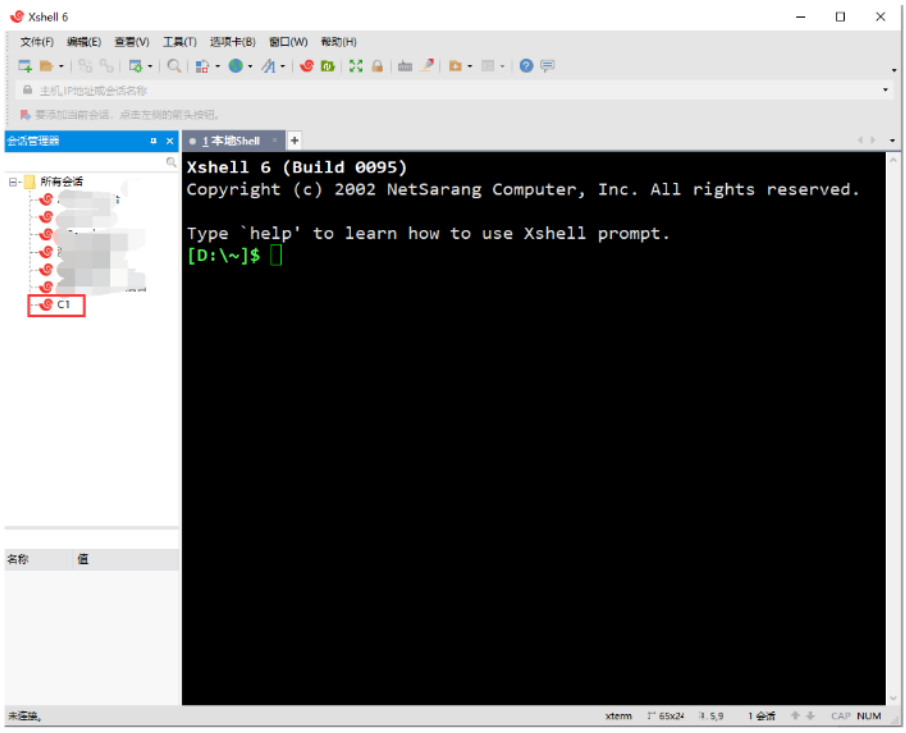

图 7 选择会话

输入用户名和密码，即为平台账户密码，如图 8、9 所示：

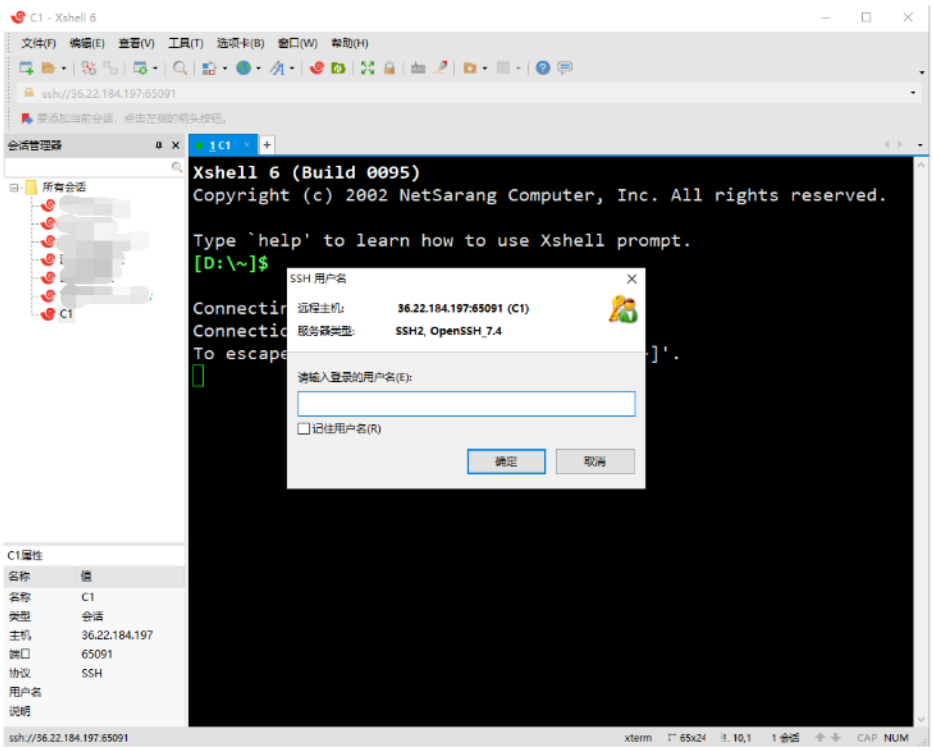

图 8 输入用户名

图 9 输入密码

3.2 Linux/Mac 系统登录

注意：因不同发行版（包括 Mac/CentOS/Ubuntu 等）命令行的差异，下述的-p 参数可能是-P，或其他可能的参数，请自行确定后再次尝试。

Mac/CentOS/Ubuntu 等系统，命令：ssh user_name@sc.nuaa.edu.cn，回车。密码由管理员开户时发送至用户邮箱。详细操作代码如图 10 所示：

```
[root@localhost ~]# ssh 70205994@sc.nuaa.edu.cn
The authenticity of host 'sc.nuaa.edu.cn (211.65.101.76)' can't be established.
RSA key fingerprint is SHA256:4bZmv6O9cJSufom8ObysdN+wr+YRv2x5BNUBfyNjcAA.
RSA key fingerprint is MD5:30:bb:b5:f1:2c:82:32:1f:b6:69:13:38:28:2d:bb:3a.
Are you sure you want to continue connecting (yes/no)? yes
Warning: Permanently added 'sc.nuaa.edu.cn,211.65.101.76' (RSA) to the list of known hosts.
Password authentication
Password
*******************************************************************************
*                          齐治交互终端 v3.3.8                                *
* 版权所有 2006-2019 浙江齐治科技股份有限公司。保留一切权利。                 *
*******************************************************************************

欢迎使用南京航空航天大学高性能计算集群

WARNING: Terminal type 'xterm-256color' is not supported, 'xterm' will be used as default.

Connecting to self@登录节点(10.12.5.12) ...
Last login: Sat Jun 27 12:31:45 2020 from 10.12.5.16
[70205994@login ~]$
```

图 10 Linux/max 系统登录命令

3.3 图形化节点登录

注：CPU1 队列无法提交图形化作业，队列提供一台图形节点仅限于数据前后处理，不要进行计算。

打开浏览器，在地址栏中输入 https://sc.nuaa.edu.cn，输入账号和密码，账户密码即为管理员开户时发送至用户邮箱的账户密码。

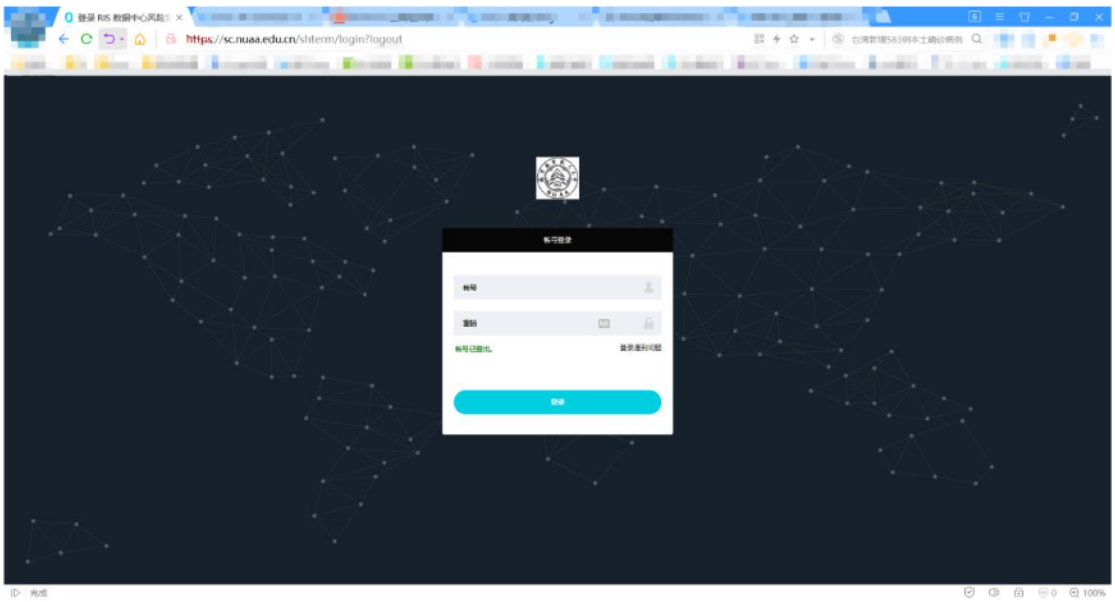

图 11 堡垒机登录界面

点击访问资产，如图 12:

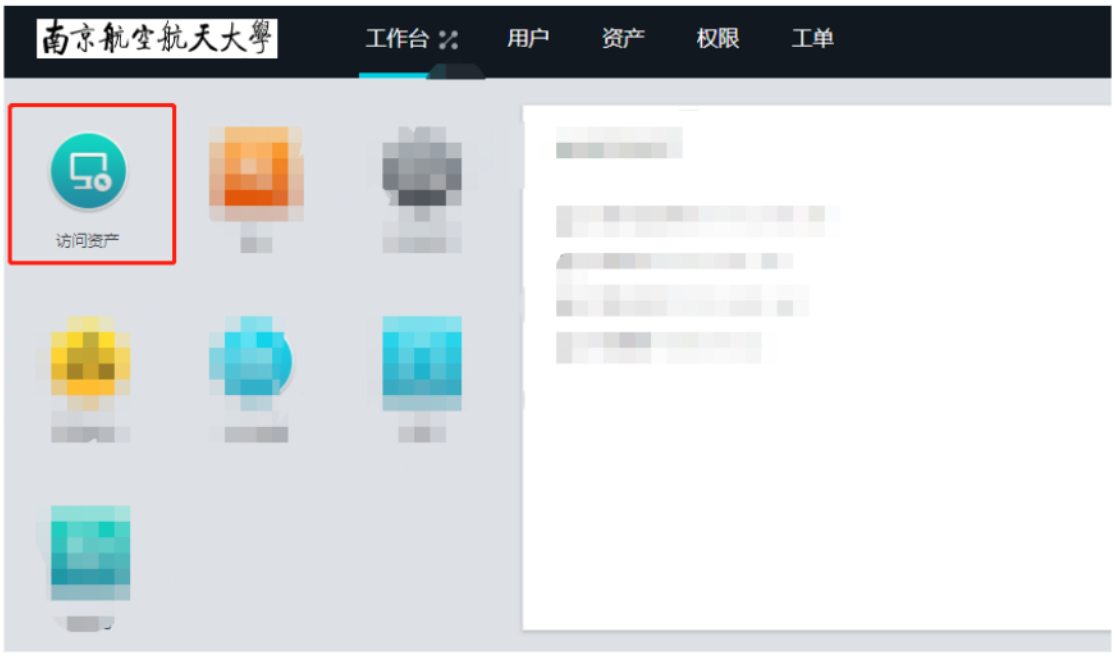

图 12 访问资产

启动图形化，如图 13:

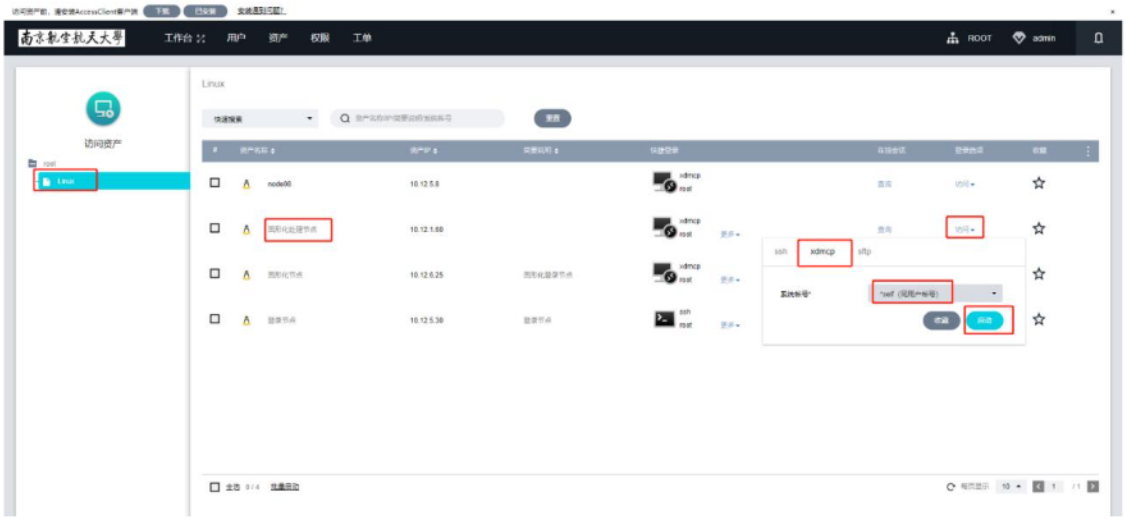

图 13 启动图形化

登陆图形化桌面，如图 14:

图 14 登录图形化桌面

启动命令行终端，如图 15:

图 15 启动命令行终端

启动应用（以 fluent20 为例），如图 16:

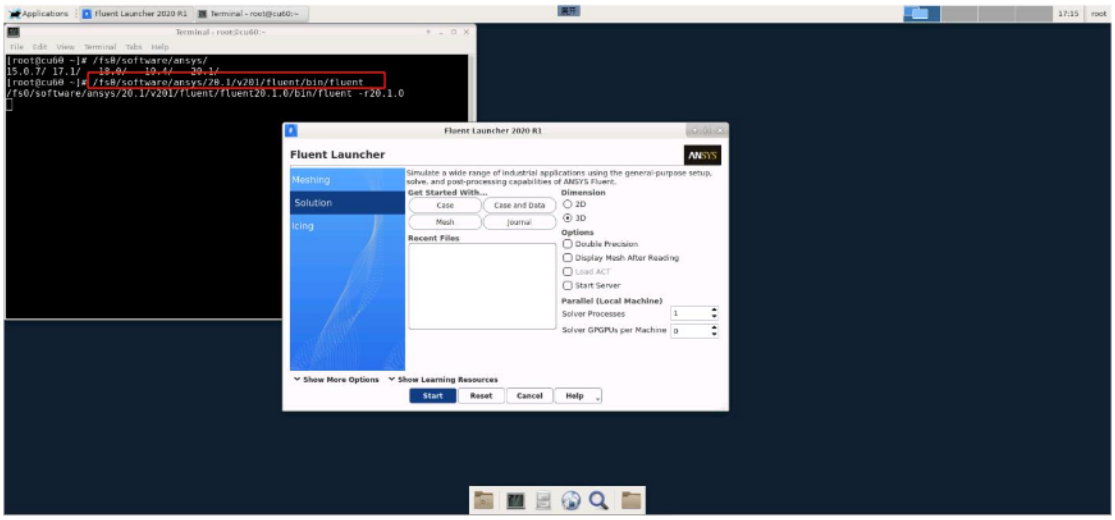

图 16 启动应用

3.4 校外用户登录 VPN

校外用户打开浏览器，在地址栏中输入 https://v.nuaa.edu.cn

图 17 校外访问 VPN

输入用户名和密码点击登录，如下图所示登录成功：

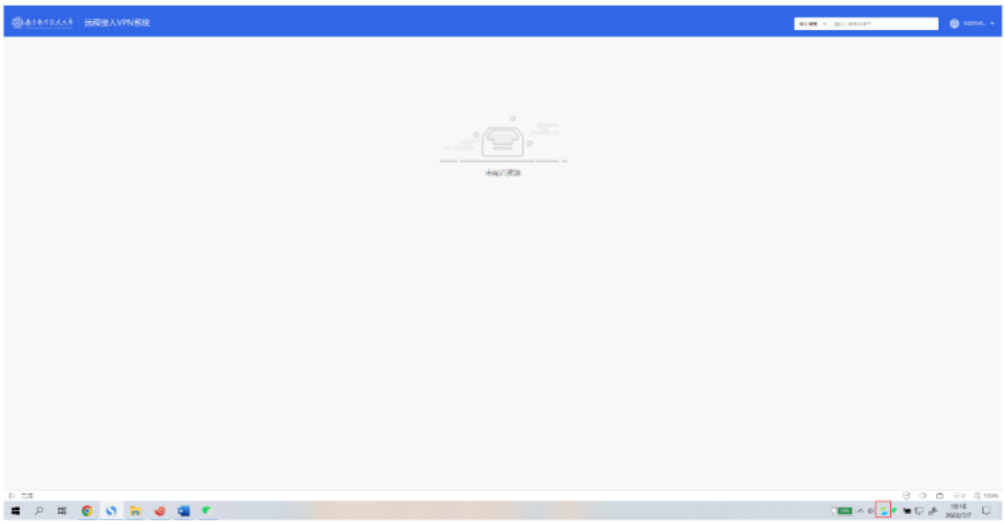

图 18 登录成功

3.5 Windows 系统文件上传下载

Windows 系统文件上传/下载需先参照 3.1 节的步骤通过 Xshell 远程登录软件使用 ssh 登录集群的登录节点，ssh 登录成功后点击 Xshell 工具栏中的“新建文件传输”图标，在提示窗口输入密码。在 Xftp 登录成功后即可用拖拽的方式从左侧本地电脑窗口将所需要上传的文件拖拽至右侧代表集群空间的窗口从而实现文件的上传。如需从集群空间下载文件，则请反向操作。如下图所示：

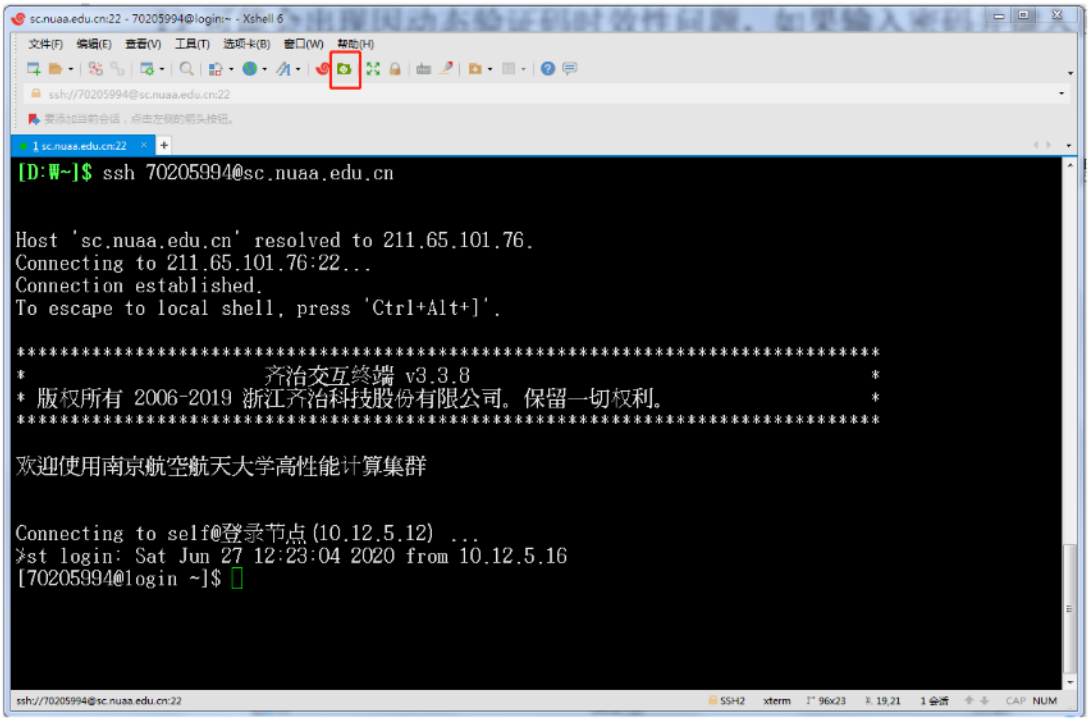

图 19 文件下载

注意:

1）集群使用 Xshell 登入可能会出现因动态验证码时效性问题，如果输入密码并键入回车键后依然没有能够登录，请再次输入密码，如此尝试 3-5 次。

2）如果出现大文件上传失败请联系工程师以协助定位并解决问题。

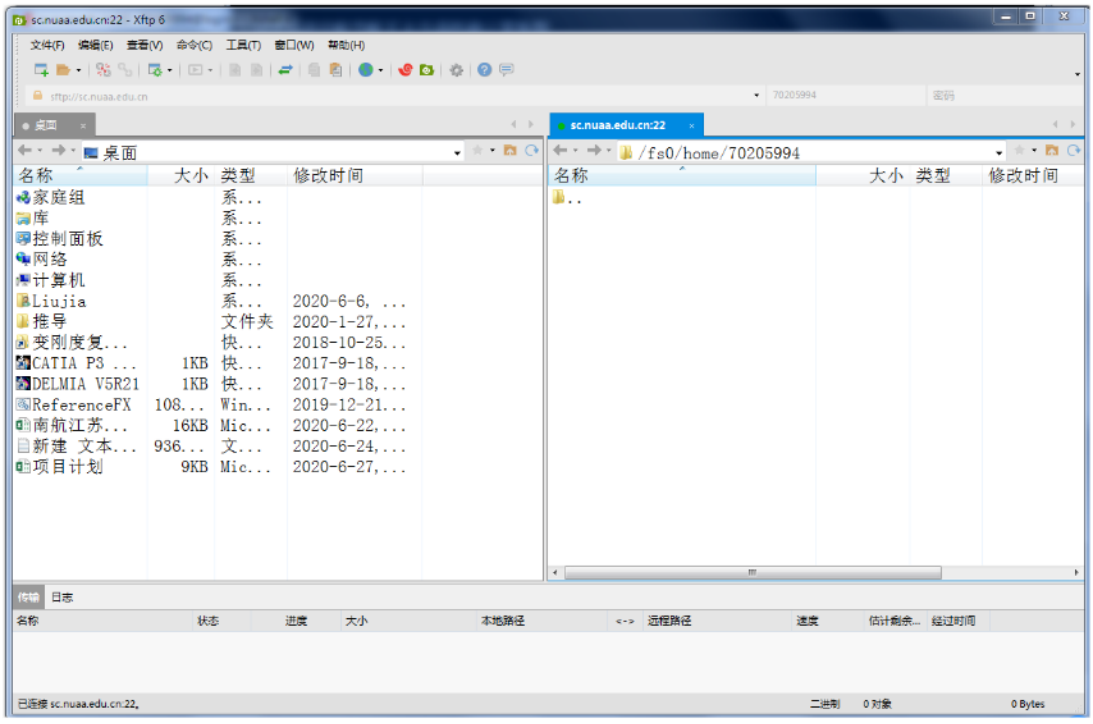

图 20 文件下载

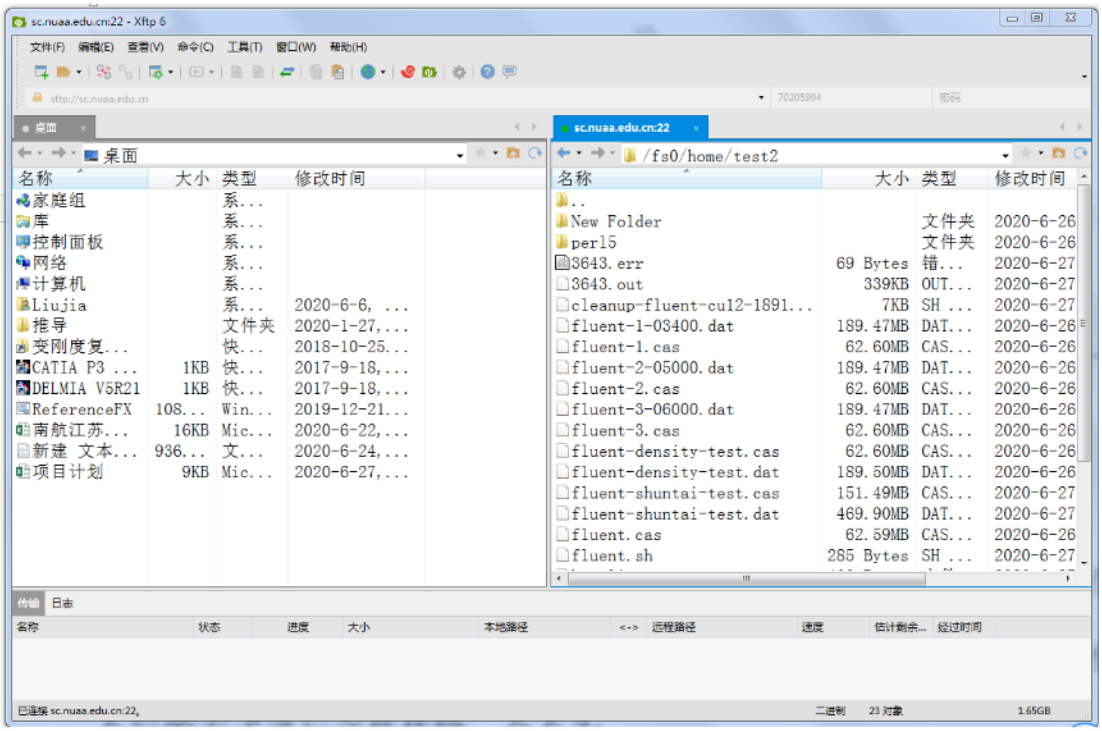

图 21 文件下载

3.6 类 Unix 系统文件上传下载

针对 Mac/CentOS/Ubuntu 等发行版的 Linux 系统，如果用户已安装了对应系统的 Xshell 软件则参照 3.2 节中的步骤。如果没有，则需要使用命令行的方式进行文件传输。命令为：sftp user_name/sc.nuaa.edu.cn/self@10.12.5.12 回车，输入密码。如图所示：

```
[70205994@login ~]$ sftp 70205994/sc.nuaa.edu.cn/self@10.12.5.12
70205994/sc.nuaa.edu.cn/self@10.12.5.12's password:
Connected to 10.12.5.12.
sftp>
sftp> cd ..
sftp> ls
70205994      frank         haidi         hyshi         jzb           pengxh        shuaizg       shuaizhigTHU
test1         test2         test4         yangguang     zh            zhouhao
sftp>
```

图 22 类 Unix 文件下载

sftp 登录后的命令行中可以使用部分当前 Linux 系统所提供的部分命令（如 ls/cd 等命令），通过输入问号(?)获取 sftp 命令行界面所支持的所有命令。如图所示：

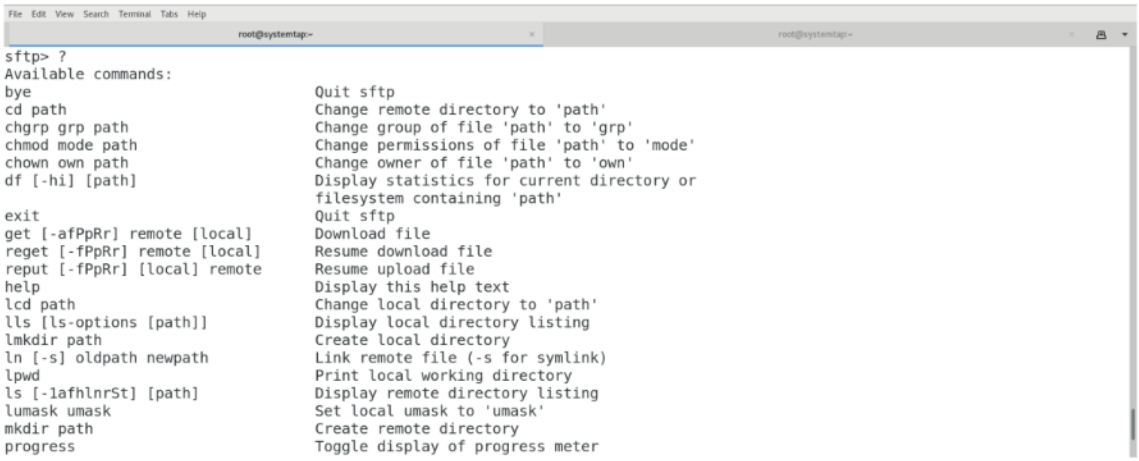

图 23 命令汇总

上传下载命令如下:

从本地电脑上传文件到集群空间:

```
#上传单个文件: sftp> put /opt/testfile
#上传文件夹: sftp> put -r /opt/testdir
```

从集群空间下载文件到本地电脑:

```
#下载单个文件: sftp> get /data/home/zhouhao/testfile
#下载文件夹: sftp> get -r /data/home/zhouhao/testdir
```

4 提交作业

Slurm 作业调度系统使用说明

CPU1 队列使用 Slurm 作业调度系统管理所有计算作业，该系统接受用户的作业请求，并将作业合理的分配到合适的节点上运行。下图为用户提交作业的示意图：

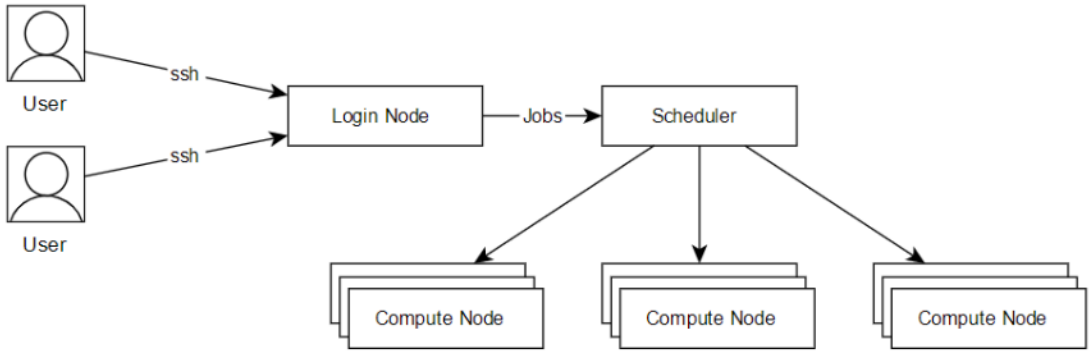

图 24 提交作业示意图

本节将介绍运行作业的两种方式，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本节还将介绍如何 sinfo、squeue、scancel 等命令具体操作。

4.1 sbatch 提交作业

注意：本文命令中所有“cpu”均代表 CPU1 队列，“gpu4”、“gpu8”分别代表 GPU 节点的 4 卡、8 卡 GPU 队列。

用户使用 sbatch 命令向作业调度系统提交作业，sbatch 可用参数十分丰富，可对作业进行非常细致的控制，这里简要介绍常用参数和方法。

运行作业的第一种方式是将整个计算过程，写到脚本中，通过 sbatch 指令提交到计算节点上执行。首先介绍一个简单的例子，假设我们的计算过程为，在计算节点上运行 hostname 指令，那么就可以如下编写作业脚本：

```
#!/bin/bash
#!/bin/bash
#SBATCH -J test
#SBATCH -p cpu
#SBATCH -n 64
#SBATCH --error=%J.err
#SBATCH --output=%J.out
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：

```
-J test          # 作业在调度系统中的作业名为 test;
-p cpu   # 作业提交的指定分区为 cpu
-n 64 # 这个作业使用 64 核运行，如果程序不支持多线程(如 openmp)，这个数不应该超过 1;
--error=%J.err   # 脚本执行的错误输出将被保存在当%j.err 文件下，%j 表示作业号;
--output=%J.ou    # 脚本执行的输出将被保存在当%j.out 文件下，%j 表示作业号;
```

除此之外，还有一些常见的参数：

```
--help      # 显示帮助信息;
-D, --chdir=<directory>          # 指定工作目录;
--get-user-env       # 获取当前的环境变量;
--gres=<list>       # 使用 gpu 这类资源，如申请两块 gpu 则--gres=gpu:2
-J, --job-name=<jobname>       # 指定该作业的作业名;
--mail-type=<type>       # 指定状态发生时，发送邮件通知，有效种类为（NONE, BEGIN, END, FAIL, REQUEUE, ALL）;
--mail-user=<user>       # 发送给指定邮箱;
-n, --ntasks=<number>       # sbatch 并不会执行任务，当需要申请相应的资源来运行脚本，默认情况下一个任务一个核心，--cpus-per-task 参数可以修改该默认值;
-c, --cpus-per-task=<ncpus>          # 每个任务所需要的核心数，默认为 1;
--ntasks-per-node=<ntasks>       # 每个节点的任务数，--ntasks 参数的优先级高于该参数，如果使用--ntasks 这个参数，那么将会变为每个节点最多运行的任务数;
-o, --output=<filename pattern>       # 输出文件，作业脚本中的输出将会输出到该文件;
-p, --partition=<partition_names>       # 将作业提交到对应分区;
-q, --qos=<qos>       # 指定 QOS;
-t, --time=<time>       # 允许作业运行的最大时间，目前未名一号和生科一号为 5 天，教学一号为两天;
-w, --nodelist=<node name list>    # 指定申请的节点;
-x, --exclude=<node name list>     # 排除指定的节点;

```

接下来是一个 GPU 作业的例子，假设我们想要申请一块 GPU 卡，并通过指令 nvidia-smi 来查看申请到 GPU 卡的信息，那么可以这么编写作业脚本：

```
#!/bin/bash
#SBATCH -J test
#SBATCH -p gpu4
#SBATCH -n 4
#SBATCH --gres=gpu:1
#SBATCH --error=%J.err
#SBATCH --output=%J.out

nvidia-smi
```

脚本中的一些参数说明如下

```
#SBATCH --gres=gpu:1              # 每个节点上申请一块 GPU 卡
```

最后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 40 个核心来运行 vasp，那么可以这么编写作业脚本:

```
#!/bin/bash
#SBATCH -J test
#SBATCH -p cpu
#SBATCH -N 2
#SBATCH --ntasks-per-node=40
#SBATCH --error=%J.err
#SBATCH --output=%J.out
# 导入 MPI 运行环境
module load intel/2017u5
# 导入 VASP 运行环境
module load vasp/5.4.4
# 执行 VASP 并行计算程序
mpirun -n 80 vasp_std
scontrol show job $SLURM_JOBID
```

4.2 salloc 交互式运行作业

运行作业的第二种方式是通过 salloc 交互式运行作业，首先需要申请计算节点，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，以下提供申请一个节点 6 个核心，并跳转到该节点上运行程序示例:

```
salloc -p cpu -N1 -n6
# 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 cu01 节点，作业 ID 为 1078858
ssh cu01          # 直接登录到刚刚申请到的节点 cu01 调式作业
scancel 1078858    # 计算资源使用完后取消作业
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
```

随后是一个 GPU 节点的使用案例;

申请一个 GPU 节点，6 个核心，1 块 GPU 卡，并跳转到节点上运行程序;

```
salloc -p gpu4 -N1   -n6 --gres=gpu:4
# 假设申请成功后返回的作业号为 1078858，申请到的节点是 gpu05
ssh gpu05 # 登录到 gpu05 上调式作业
scancel 1078858   # 计算结束后结束任务
squeue -j 1078858 # 确保作业已经退出
```

最后介绍一个跨节点使用案例;

申请两个节点，每个节点 12 个核心

```
salloc -p cpu -N2 --ntasks-per-node=80
# 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
a8u03n[05-06]节点，作业 ID 为 1078858
# 这里申请两个节点，每个节点 12 个进程，每个进程一个核心

# 根据需求导入 MPI 环境
module load intel/2017u5

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

mpirun -np 80 -machinefile slurm.hosts hostname

# 结束后退出或者结束任务
scancel 1078858
```

4.3 sinfo 查看资源空闲状态

sinfo 可查询各分区节点的空闲状态，输入 sinfo 命令，返回状态显示 idel 为空闲，mix 为节点部分核心可以使用，alloc 为已被占用，maint 为维护中，如下所示：

```
[root@cu60 ~]# sinfo
PARTITION AVAIL  TIMELIMIT  NODES  STATE NODELIST
cpu*         up   infinite      2  maint cu[02,60]
cpu*         up   infinite      3    mix cu[01,03,62]
cpu*         up   infinite      7  alloc cu[04,17-21,61]
cpu*         up   infinite     50   idle cu[05-16,22-59]
gpu4         up   infinite      1   mix$ gpu07
gpu4         up   infinite      1 alloc$ gpu04
gpu4         up   infinite      1  maint gpu08
```

sinfo 的一些常用参数：

```
--help      # 显示 sinfo 命令的使用帮助信息;
-d          # 查看集群中没有响应的节点;
-i <seconds>      # 每隔相应的秒数，对输出的分区节点信息进行刷新
-n <name_list>      # 显示指定节点的信息，如果指定多个节点的话用逗号隔开;
-N      # 按每个节点一行的格式来显示信息;
-p  # <partition> 显示指定分区的信息，如果指定多个分区的话用逗号隔开;
-r      # 只显示响应的节点;
-R      # 显示节点不正常工作的原因;
```

4.4 squeue/sq 查看作业队列

用户可以通过 squeue 或 sq 命令查看提交作业的排队情况，如下所示输入 sq 命令：

```
sq
```

默认情况下 squeue 或 sq 输出的内容如下，分别是作业号，分区，作业名，用户，作业状态，运行时间，节点数量，运行节点(如果还在排队则显示排队原因)。

```
JOBID PARTITION    NAME      USER ST     TIME   NODES NODELIST(REASON)
```

squeue 的常见参数：

```
    --help      # 显示 squeue 命令的使用帮助信息;
    -A <account_list>      # 显示指定账户下所有用户的作业，如果是多个账户的话用逗号隔开;
    -i <seconds>       # 每隔相应的秒数，对输出的作业信息进行刷新
    -j <job_id_list> #显示指定作业号作业信息，如果是多个作业号的话用逗号隔开;
    -n <name_list> #显示指定节点上作业信息，如果指定多个节点的话用逗号隔开;
    -t <state_list>   # 显示指定状态的作业信息，如果指定多个状态的话用逗号隔开;
    -u <user_list>   # 显示指定用户的作业信息，如果是多个用户的话用逗号隔开;
    -w <hostlist>    # 显示指定节点上运行的作业，如果是多个节点的话用逗号隔开;
```

4.5 scancel 取消作业

用户可以通过 scancel 命令取消账号中已提交的作业，如下所示：

```
    # 取消作业 ID 为 123 的作业
    scancel 123
```

也可通过 scancel 命令取消自己账号上所有作业：

```
    # 注意 whoami 前后不是单引号
    scancel -u user_name
```

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;
    -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;
    -q <qos>                  # 取消指定 qos 的作业;
    -t <job_state_name> # 取消指定作态的作业，"PENDING", "RUNNING" 或 "SUSPENDED";
    -u <user_name>            # 取消指定用户下的作业;
```

5 常见问题及注意事项

5.1 Xshell 工具在哪下载？

答：可以通过 Xshell 官方网站申请免费版 Xshell，具体申请地址如下：https://www.xshell.com/zh/free-for-home-school/

5.2 我想用的计算软件没有怎么办？

答：计算软件可自行安装到宿主目录中，如不会安装请联系集群管理员协助安装。

5.3 我要用的计算软件作业脚本不会写怎么办？

答：用户所需具体脚本可通过询问集群管理员寻找目标软件的作业模板。如 ansys、comsol、vasp、lammps 等。

5.4 提交作业报错如下错误

sbatch:error:Batch job submission failed:Invalid account or account/partition combination specified

答：导致如上报错的原因较多，大概原因有：脚本中队列设置错误，slurm 账号没有绑定等。具体问题需联系管理员逐个排查。

5.5 作业没有运行，并且显示 QOSGrpCpuLimit

答：出现此英文字符是因为账号核数有所限制，正在运行的作业达到核数限制后，其他作业状态则变为正在排队。

5.6 怎么修改账号密码

答：登录高性能计算平台管理系统修改密码，具体步骤可参考《高性能计算平台二期平台管理系统用户端操作手册》。

（使用流程及使用规范、常用计算软件使用及常用命令和脚本模板请参考高性能计算平台用户手册）