上海交通大学
SHANGHAI JIAO TONG UNIVERSITY

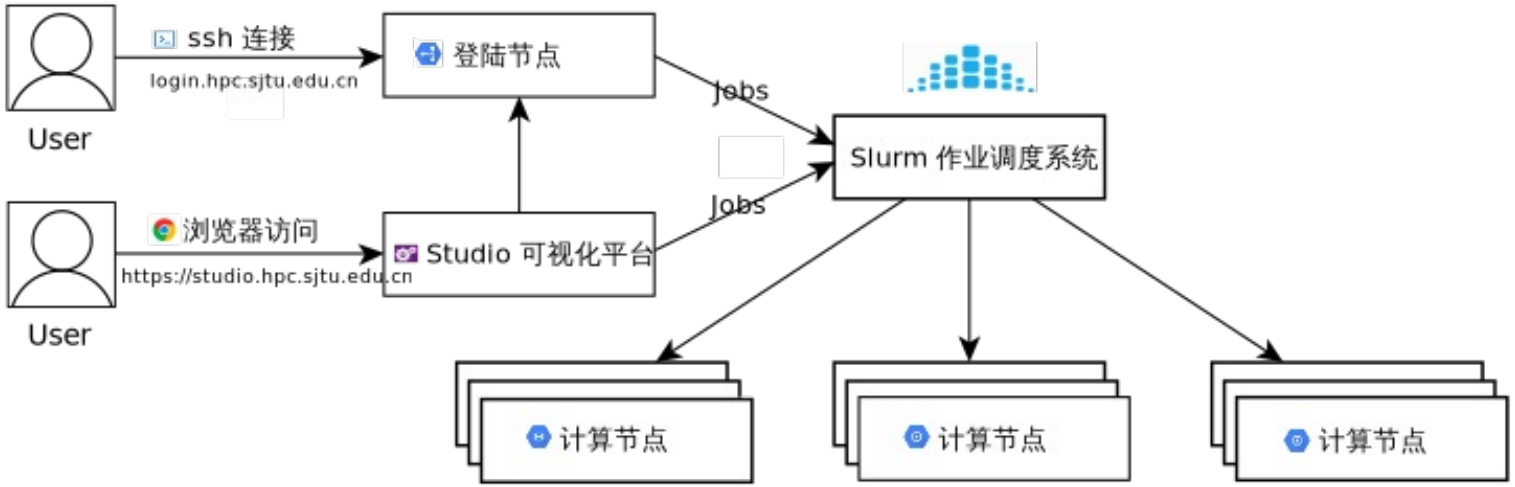
Slurm作业调度系统使用

高亦沁

2023 年 9 月 26 日

Slurm作业调度系统

- 用户将需要执行的工作编写进一个Slurm作业脚本中
- 用户从登录节点或可视化平台提交作业
- 由Slurm将作业分配到不同计算节点上运行



*相关内容参考：https://docs.hpc.sjtu.edu.cn/job/slurm.html

Slurm作业运行队列

- 用户提交作业脚本时需要指定作业运行的队列

集群	节点类型	节点数	单节点核数	单节点内存	队列	允许单作业核数	可否共享	最长运行时间
$\pi 2.0$	CPU节点（x86）	656个	40核	192G	small	1-20	可共享	7天
					cpu	40-24000	需独占	7天
					debug	测试节点	可共享	20分钟
	CPU节点（大内存）	3个	80核	3T	huge	6-80	可共享	2天
			192核	6T	192c6t	48-192	可共享	2天
AI平台	GPU节点	8个，每节点配16张V100卡	96核	1.45T	dgx2	推荐CPU配比为1:6，GPU卡数为1-128	可共享	7天
ARM平台	CPU节点（ARM）	100个	128核	256G	arm128c256g	1-12800	可共享	3天
					debugarm	测试节点	可共享	20分钟
思源一号	CPU节点	936个	64核	512G	64c512g	1-60000	可共享	7天
					debug64c512g	测试节点	可共享	1小时
	GPU节点	23个，每节点配4张A100卡	64核	160G	a100	最高CPU配比为1:16，GPU卡数为1-92	可共享	7天
					debuga100	测试节点	可共享	20分钟

*相关内容参考：https://docs.hpc.sjtu.edu.cn/job/slurm.html

Slurm作业运行队列

- 可用队列及相关信息在登录节点的欢迎界面上也会显示

思源一号集群的所有可用队列

*相关内容参考：https://docs.hpc.sjtu.edu.cn/job/slurm.html

Slurm作业调度命令

Slurm命令	功能	举例	
sbatch	提交作业	sbatch jobscript.slurm	提交名为jobscript.slurm的作业脚本
sinfo	查看集群状态	sinfo -N --state=idle	查看状态（state）为可用（idle）的节点信息
		sinfo --partition=cpu	查看cpu队列（partition）信息
		节点状态：drain(节点故障)，alloc(节点在用)，idle(节点可用)，down(节点下线)	
squeue	查看作业排队情况	squeue -l	查看作业排队细节（参数-l）信息
		squeue --state=R	查看状态为运行中（state为R）的作业
		作业状态：Running(R), Pending(PD), Completing(CG), Completed(CD), Failed, Cancelled, Node_fail	
scontrol	查看作业参数	scontrol show job [JOB_ID]	查看编号为JOB_ID的作业信息
sacct	查看作业记录	sacct --state=CD	查看状态为完成（state为CD）的作业
		sacct -S YYYY-MM-DD	查看在指定时间（YYYY-MM-DD）后的所有作业
scancel	取消作业	scancel [JOB_ID]	取消编号为JOB_ID的作业

*相关内容参考：https://docs.hpc.sjtu.edu.cn/job/slurm.html

实验1：提交作业并完成运行

实验目标

- 提交一个打印Hello World的程序并成功运行

实验步骤

1. 在Tabby中，打开思源一号的登录节点
2. 创建一个存放实验文件的新文件夹Test

   $ mkdir Test

   $ cd Test
3. 将压缩包hello.zip上传到Test文件夹中
4. 解压缩hello.zip

   $ tar -xvf hello.zip
5. 进入hello文件夹

   $ cd hello
6. 提交作业

   $ sbatch hello_world.slurm
7. 查看输出结果

   $ cat [你的作业编号].out

注：本实验中，可执行文件hello_world已提前编译，在实际使用过程中需要自行完成

*相关内容参考：https://docs.hpc.sjtu.edu.cn/job/jobsample1.html

实验2：查看作业状态和取消作业

实验目标

- 提交一个睡眠程序sleep，观察作业提交后的状态，最后将其取消

实验步骤

1. 在Tabby中，打开思源一号的登录节点
2. 将压缩包sleep.zip上传到Test文件夹中
3. 解压缩sleep.zip

   $ tar -xvf sleep.zip
4. 进入sleep文件夹

   $ cd sleep
5. 提交作业

   $ sbatch sleep.slurm
6. 使用slurm系统命令查看集群信息和作业状态

   $ sinfo

   $ squeue

   $ scontrol show job [JOB_ID]

   $ sacct
7. 取消作业

   $ scancel [JOB_ID]
8. 确认作业被成功取消

   $ sacct

*相关内容参考：https://docs.hpc.sjtu.edu.cn/job/jobsample1.html

Slurm作业脚本

脚本含义

- 在思源一号的64c512g队列上申请1核的资源，用于执行名称为hello的作业
- 作业的普通输出和错误输出分别存放在当前文件夹的[作业号.out]和[作业号.err]文件中
- 作业的内容为运行当前目录下的hello_world可执行文件

脚本样式

```
#!/bin/bash                          # 作业配置：
#SBATCH --job-name=hello             # 作业名称为hello
#SBATCH --partition=64c512g          # 申请的队列为64c512g
#SBATCH -n 1                         # 申请的核数为1核
#SBATCH --output=%j.out              # 普通输出存放在当前文件夹的[作业号.out]文件中
#SBATCH --error=%j.err               # 错误输出存放在当前文件夹的[作业号.err]文件中

./hello_world                        # 作业正文：运行当前目录下的hello_world可执行文件
```

*相关内容参考：https://docs.hpc.sjtu.edu.cn/job/jobsample1.html

Slurm交互式作业

- 交互式作业是一种特殊的队列任务
  - 普通作业：将需求的计算资源和需要执行的命令编写成作业脚本并提交，由系统指定计算节点运行
  - 交互式作业：将需求的计算资源写成命令行参数提交，系统分配计算节点后，可直接登录到申请的节点上，进行代码调试、程序编译等操作
- 举例：申请64c512g队列的1节点4核资源并启动计算节点的终端

用于提交交互式作业申请的命令

“-x abc”形式为linux命令通用的选项格式

其中x为选项，abc为选项的参数

srun命令的选项可以用于提交申请的资源

```
$ srun -p 64c512g -n 4 -N 1 --pty /bin/bash
```

-p 64c512g	-n 4	-N 1	--pty /bin/bash
队列为64c512g	核数为4	节点数为1	启动计算节点的命令行终端

*相关内容参考：https://docs.hpc.sjtu.edu.cn/job/slurm.html

module软件模块

- 超算平台提供大量预编译的编译器和软件供使用
- 用户手册中列出了所有预编译软件供参考
- 编译器和软件通过module命令加载后即可使用，用户无须再进行编译安装
- 软件可能存在多个版本，如果加载时未指定版本号，系统将加载该模块的默认版本（用D字符标识）
- 由于CPU架构不同，x86、AI和ARM的软件不通用，通过不同的平台标识区分

cpu gpu arm

编译器、MPI库、数学库、应用工具

编译器MPI库数学库应用工具

软件	描述	可用版本	平台
GNU Compiler Collection	GNU编译器套件	9.3(D)	cpu arm
Armadillo	采用c++实现的类matlab风格的矩阵运算库	10.5.0	cpu
Intel Compiler	Intel编译器套件	21.4.0	cpu
cuBLAS	NVIDIA官方提供的GPU上的高性能矩阵运算库		gpu
CUDA	NVIDIA CUDA SDK	10.2(D)	gpu
cuSOLVER	NVIDIA官方提供的GPU上的线性代数求解器		gpu
cuSPARSE	NVIDIA官方提供的GPU上的稀疏矩阵求解器		gpu
cuFFT	NVIDIA官方提供的GPU上的快速傅里叶变换运算库		gpu
NVIDIA HPC SDK	NVIDIA HPC SDK	20.11(D)	gpu
JDK	Java开发套件	12.0	cpu arm

*相关内容参考：https://docs.hpc.sjtu.edu.cn/app/module.html

module常用命令

module 命令	功能
module load [MODULE]	加载模块
module unload [MODULE]	卸载模块
module av	列出所有模块
module av keyword	列出名称中含有 keyword 的所有模块
module list	列出所有已加载的模块
module show [MODULE]	列出模块的信息，如路径、环境变量等

*相关内容参考：https://docs.hpc.sjtu.edu.cn/app/module.html

实验3：软件模块命令的使用

实验目的

- 使用软件模块，加载并使用gcc/11.2.0版

实验步骤

1. 在Tabby中，打开思源一号的登录节点
2. 查看当前加载的模组列表

```
[hpcgyq@sylogin4 ~]$ module list
No modules loaded
```

3. 查看当前的gcc版本（登录节点自带版本）

```
[hpcgyq@sylogin4 ~]$ gcc --version
gcc (GCC) 8.3.1 20191121 (Red Hat 8.3.1-5)
Copyright (C) 2018 Free Software Foundation, Inc.
This is free software; see the source for copying conditions.  There is NO
warranty; not even for MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE.
```

4. 查看可用的gcc模组

```
[hpcgyq@sylogin4 ~]$ module av gcc/

------------------------------ /dssg/share/spack/modules/icelake/linux-centos8-icelake ------------------------------
   gcc/8.3.1    gcc/8.5.0    gcc/9.3.0    gcc/9.4.0    gcc/10.3.0    gcc/11.2.0 (D)

  Where:
   D:  Default Module

Use "module spider" to find all possible modules and extensions.
Use "module keyword key1 key2 ..." to search for all possible modules matching any of the "keys".
```

5. 发现gcc/11.2.0是默认版本，直接加载gcc模块即可

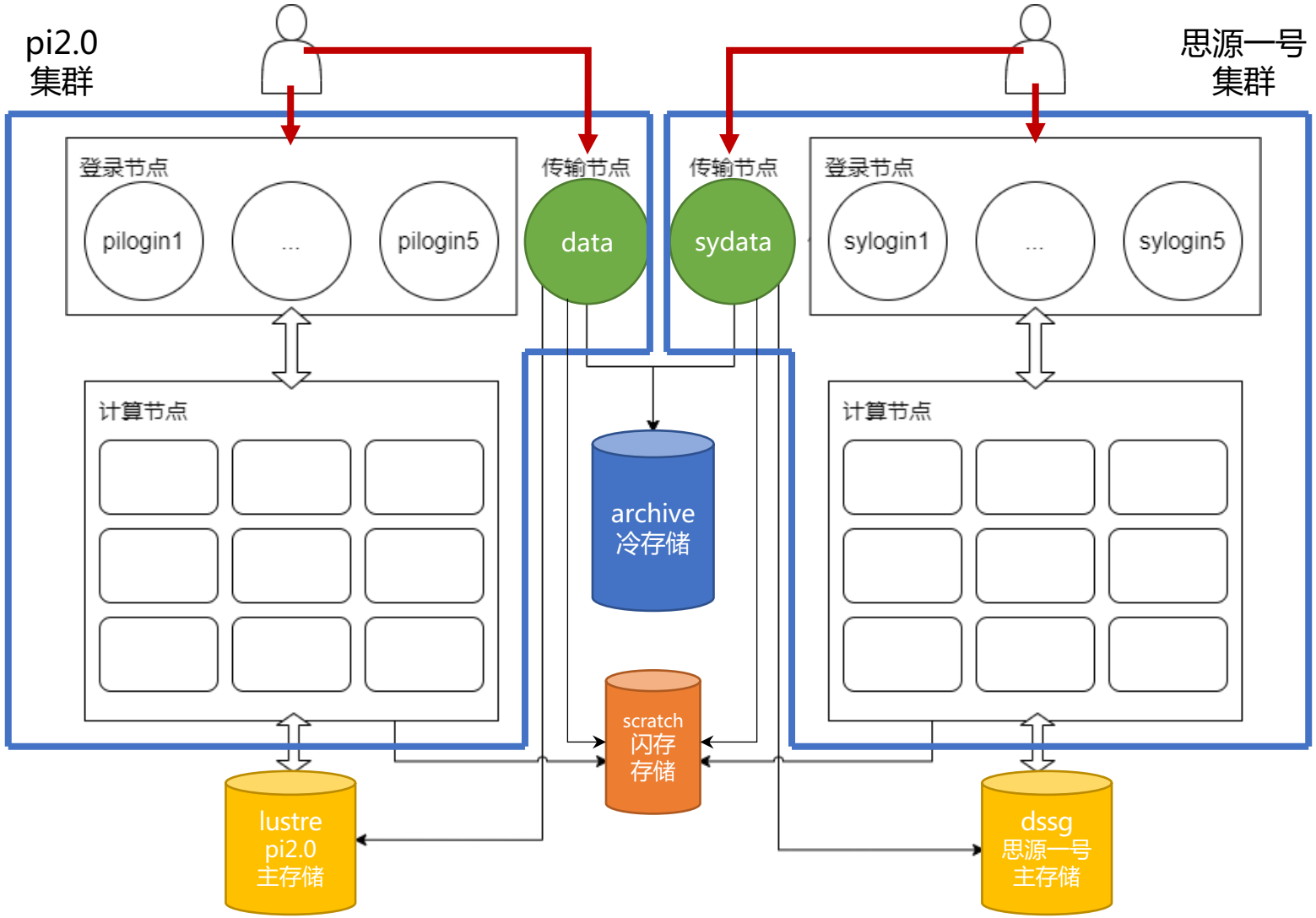
```
[hpcgyq@sylogin4 ~]$ module load gcc
[hpcgyq@sylogin4 ~]$ module list

Currently Loaded Modules:
  1) gcc/11.2.0
```

6. 查看当前gcc版本（11.2.0版加载成功）

```
[hpcgyq@sylogin4 ~]$ gcc --version
gcc (Spack GCC) 11.2.0
Copyright (C) 2021 Free Software Foundation, Inc.
This is free software; see the source for copying conditions.  There is NO
warranty; not even for MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE.
```

*相关内容参考：https://docs.hpc.sjtu.edu.cn/job/jobsample1.html

交我算集群文件系统



主文件系统：$HOME

- pi2.0和思源一号各自登录后的默认家目录
- 大容量、高可用、较高性能

全闪存文件系统：$SCRATCH

- 适用于IO密集型工作的临时目录
- 小容量、高性能
- 定期清理数据

归档文件系统：$ARCHIVE

- 存放不常用数据的目录
- 大容量、高可用、性能较低
- 提供快照功能，可恢复误删数据
- 只能通过数据传输节点访问

注意：通过cd $[文件系统]命令，可以访问不同文件系统的家目录

*相关内容参考：https://docs.hpc.sjtu.edu.cn/transport/index.html

本地和远程间的数据传输命令

- scp命令：直接拷贝文件，适用于少量大文件，且目标环境上没有数据的历史版本

  $ scp -r [源文件路径] [目标路径]

- rsync命令：进行增量传输，适用于包含大量文件的目录，或者目标文件夹已存在差异较小的历史版本

  $ rsync --archive --partial --progress [源文件路径] [目标路径]

- 绝对路径和相对路径皆可作为文件路径
  - 绝对路径：从根目录到目标文件的完整路径
  - 相对路径：目标文件基于当前所在目录的相对位置（常用，更方便）
- 如果一个路径相对于当前位置处于远程节点（如：思源一号和本地，pi2.0和思源一号等），使用以下格式：

  [用户名]@[远程节点的地址]:[文件在远程节点上的路径]

注意：在本地和远程交我算节点之间进行数据传输，需要在本地终端中执行，而不是在交我算节点终端

*相关内容参考：https://docs.hpc.sjtu.edu.cn/transport/index.html

数据传输命令练习

- 传输路径：本地->思源一号家目录

想传输的文件在哪里、叫什么　　想把这个文件放在哪里

```
$ scp -r ~/Downloads/hello_world.c hpcgyq@sydata.hpc.sjtu.edu.cn:$HOME/Test
```

数据传输命令：需要传输的是本地个人文件夹下Downloads文件夹中名叫hello_world.c的文件

以hpcgyq为用户名，登录远程sydata节点，将文件传输至默认个人目录下的Test文件夹中

- 传输路径：思源一号家目录 -> 本地

```
$ scp -r hpcgyq@sydata.hpc.sjtu.edu.cn:$HOME/Test/hello_world.c ~/Downloads
```

- 传输路径：思源一号闪存 -> 思源一号家目录

```
$ cp -r $SCRATCH/Test/hello_world.c $HOME/Test
```

注意：若需进行大规模的数据传输、或将大量数据移动至冷存储保存，请参考用户手册

*相关内容参考：https://docs.hpc.sjtu.edu.cn/transport/index.html