Fastone Slurm使用指南

目录

1. Fastone Compute Cloud通过Slurm集群运行任务介绍

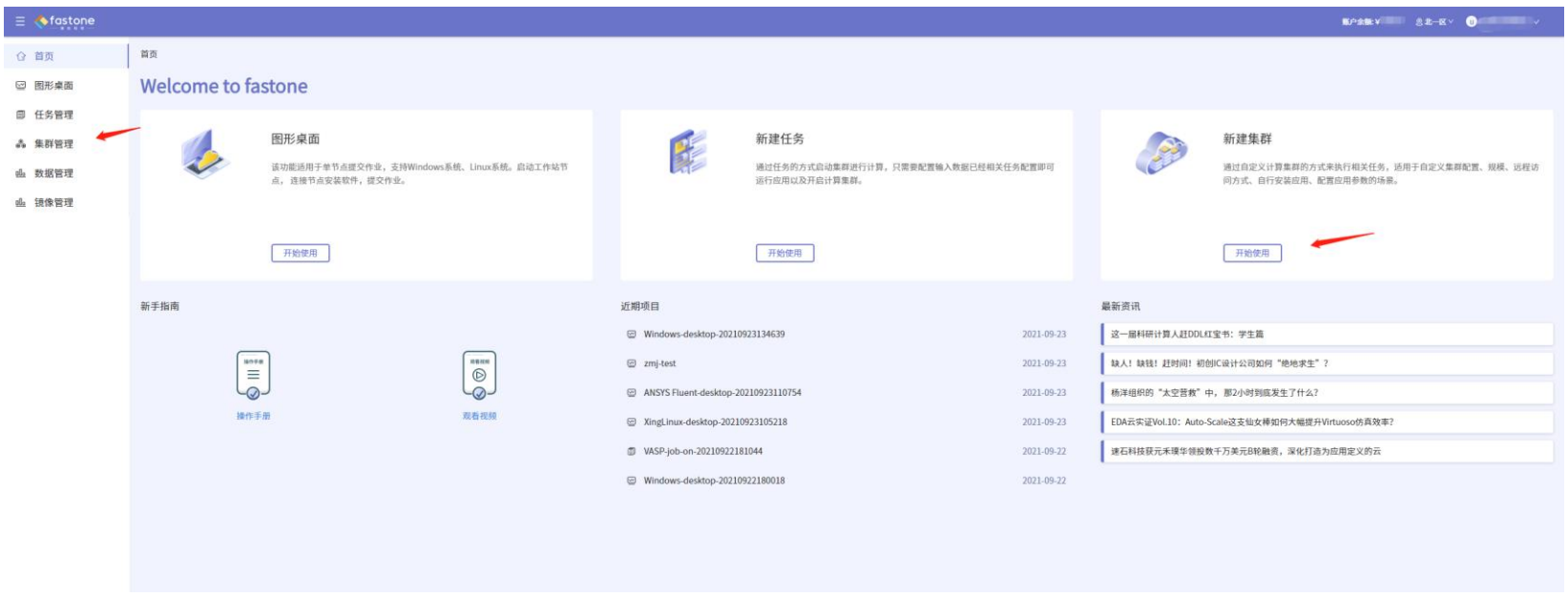
1.1 Fastone Compute Cloud集群管理功能

Fastone Compute Cloud（速石计算云以下简称FCC）集群管理功能实现自动构建slurm分区以让用户直接在开启的管理节点中执行任务。开启的机器只有一台管理节点，系统在以此节点为slurm头节点，自动构建可自动伸缩的slurm集群且分区计算节点数量为0。用户登录管理节点使用slurm命令提交任务后，系统将根据slurm命令自动开启满足其要求的机器运行任务。开启的计算节点，在连续10分钟内无任务运行时，机器将会自动关闭。每个slurm分区最多可开100台机器。

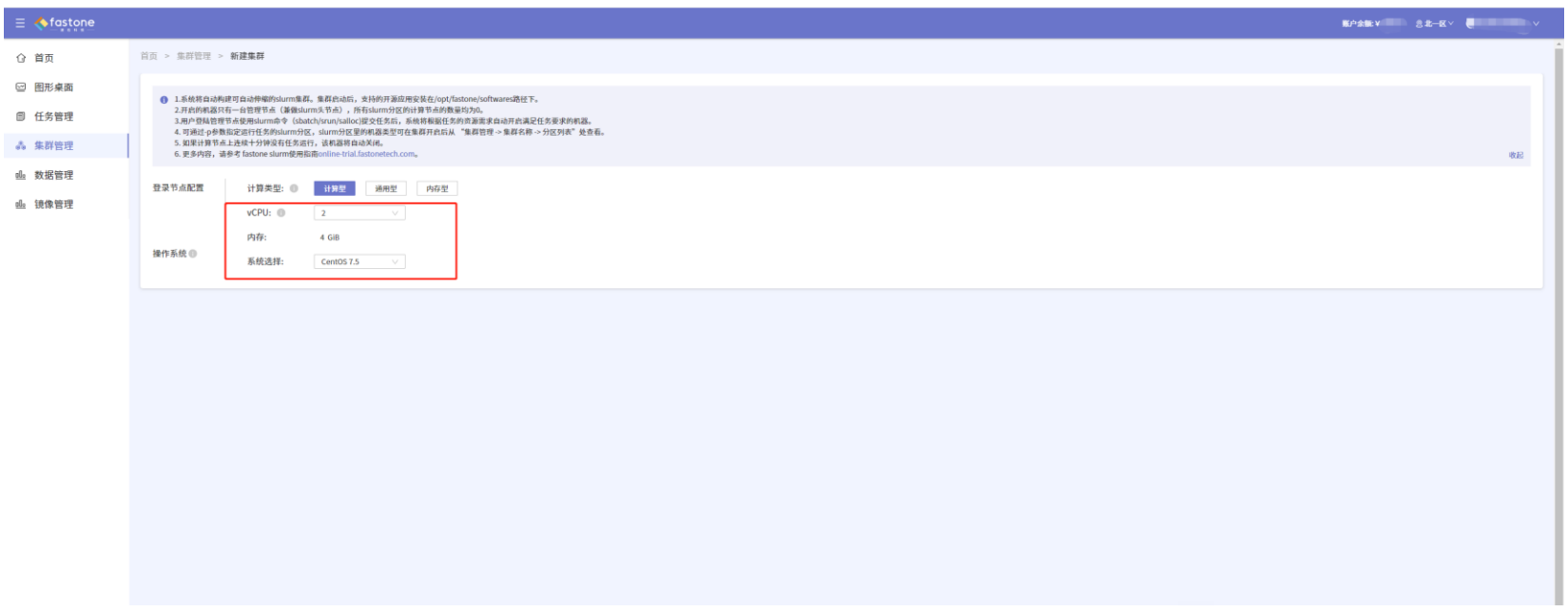
此教程针对Fastone Compute Cloud使用手册中的运行任务方式三：通过Slurm集群运行任务做详细介绍

1.2 使用Slurm集群运行任务

1. 点击集群管理->创建集群，或者点击首页->新建集群开始使用



2. 根据需要管理的节点数量选择合适的配置，开启一台slurm管理节点



3. 点击ssh，打开ssh远程桌面

以下针对于slurm命令的介绍，均在此台管理节点机器上执行

2. Slurm 任务管理系统

2.1 slurm介绍

Slurm（ Simple Linux Utility for Resource Management, Slurm官方网站http://slurm.schedmd.com/ ）是适用于大型和小型Linux集群的开源，容错且高度可扩展的集群管理和作业调度系统。作为集群工作负载管理器，Slurm具有三个关键功能。首先，它在一段时间内为用户分配对资源（计算节点）的独占和/或非独占访问权限，以便他们可以执行工作。其次，它提供了一个框架，用于在分配的节点集上启动，执行和监视工作。最后，它通过管理待处理的工作队列来仲裁资源争用。

Slurm 利用分区(partition)对 CPU、内存、网络等资源进行分类，以便将不同需求的任务运行到不同计算节点上。用户需利用 slurm 命令将该任务所需要的 CPU 核等资源等提交到特定的分区中，等任务申请的资源得到满足后，任务才开始运行。任务运行受分区、账户、服务质量(QOS)等限制。

更多详细命令请查看在此网页查看 https://slurm.schedmd.com/man_index.html

2.2 slurm常用命令

任务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	显示任务状态	squeue
srun	用于交互式任务提交	srun -n 2 -p p1-c1-2 hostname
sbatch	用于批处理任务提交	sbatch -n 2 job.sh
salloc	用于分配模式任务提交	salloc -p p1-c1-2
scancel	用于取消已提交的任务	scancel JOBID
scontrol	用于查询节点信息或正在运行的任务信息	scontrol show job JOBID
sacct	用于查看历史任务信息	sacct -u user1 -S 03/01/17 -E 03/31/17 --field=jobid,partition,jobname,user,nnodes,start,end,elapsed,state

2.3 slurm查看系统资源及任务状态

2.3.1 sinfo 查看系统资源

sinfo 查看当前账号可使用的分区信息，如下图所示。每个可使用27个分区，即27种机器类型。最多可开启100个节点。

其中：

第一列 PARTITION 是分区名称，每个分区开启的机器类型相同。

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 是任务运行时间限制，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，allocated、alloc 是节点已被占用，drain 是已失去活力的节点，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 p1-c1-2-1	指定显示节点 p1-c1-2-1 的使用情况
sinfo -p p1-c1-2	指定显示分区 p1-c1-2 情况

其他选项可以通过 sinfo --help 查询

```
-bash-4.2$ sinfo
PARTITION AVAIL  TIMELIMIT  NODES  STATE NODELIST
p1-c1-2*     up   infinite      1  drain p1-c1-2-2
p1-c1-2*     up   infinite      1    mix p1-c1-2-1
p1-r1-2      up   infinite      0    n/a
p1-m1-2      up   infinite      0    n/a
p1-c1-4      up   infinite      0    n/a
p1-r1-4      up   infinite      0    n/a
p1-m1-4      up   infinite      0    n/a
p1-c1-8      up   infinite      0    n/a
p1-r1-8      up   infinite      0    n/a
p1-m1-8      up   infinite      0    n/a
p1-g1-1      up   infinite      0    n/a
p1-c1-16     up   infinite      0    n/a
p1-r1-16     up   infinite      0    n/a
p1-m1-16     up   infinite      0    n/a
p1-r1-32     up   infinite      0    n/a
p1-m1-32     up   infinite      0    n/a
p1-g1-4      up   infinite      0    n/a
p1-c1-36     up   infinite      0    n/a
p1-c1-48     up   infinite      0    n/a
p1-r1-48     up   infinite      0    n/a
p1-m1-48     up   infinite      0    n/a
p1-r1-64     up   infinite      0    n/a
p1-m1-64     up   infinite      0    n/a
p1-g1-8      up   infinite      0    n/a
p1-c1-72     up   infinite      0    n/a
p1-c1-96     up   infinite      0    n/a
p1-r1-96     up   infinite      0    n/a
p1-m1-96     up   infinite      0    n/a
```

2.3.2 squeue 查看任务状态

squeue 得到的结果是当前账号的任务运行状态，如果 squeue 没有任务信息，说明任务已退出。

具体示例见下图：

```
-bash-4.2$ squeue
             JOBID PARTITION     NAME     USER ST       TIME  NODES NODELIST(REASON)
                11   p1-r1-2 hostname [redacted] PD       0:00      1 (Nodes required for job are DOWN, DRAINED or reserved for jobs in higher priority partitions)
                12   p1-c1-4 run_flue [redacted] PD       0:00      1 (Nodes required for job are DOWN, DRAINED or reserved for jobs in higher priority partitions)
```

其中：

第一列 JOBID 是任务号，任务号是唯一的。

第二列 PARTITION 是任务运行使用的分区名称。

第三列 NAME 是任务名。

第四列 USER 是FCC平台账号。

第五列 ST 是任务状态，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 是任务运行时间。

第七列 NODES 是任务使用的节点数。

第八列 NODELIST(REASON)对于运行任务（R 状态）显示任务使用的节点列表；对于排队任务（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 11	查看任务号为 11 的任务信息
squeue -u user1	查看FCC计算云账号user1的任务信息
squeue -p p1-c1-4	查看提交到 p1-c1-4 队列的任务信息
squeue -w p1-c1-4-1	查看使用到 p1-c1-4 节点的任务信息

其他选项可通过 squeue —help 命令查看

2.4 slurm三种提交任务命令

2.4.1 srun 交互式提交任务

srun [options] program 命令属于交互式提交任务，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任务中断。

srun 命令示例：

```
srun -p p1-c1-4 -n 4 --pty /run_fluent.sh -i /test-data/run_fluent.jou -s 3ddp
```

交互式运行fluent任务。

其中：

-p p1-c1-4 指定提交任务到 p1-c1-4 分区

-n 4 指定进程数为 4，p1-c1-4 分区每一个节点4 核

--pty /run_fluent.sh 指定运行脚本run_fluent.sh，一般是客户自己写的脚本

-i /test-data/run_fluent.jou 指定运行fluent任务的输入文件

-s 3ddp 设置运行fluent任务的求解器

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 4	指定进程数为 4
-p p1-c1-4	指定p1-c1-4分区，开启此分区内的机器
-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	指定任务名为 JOBNAME
-t 20	限制运行 20 分钟

srun 的其他选项可通过 srun --help 查看。

2.4.2 sbatch 后台提交任务

sbatch 一般情况下与 srun 一起提交任务到后台，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，提交任务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为任务号。

sbatch 命令示例 1（4 个进程提交hostname命令）：编写脚本 run_fluent.sh，内容如下：

```
#!/bin/bash
srun -n 4 hostname
```

然后在命令行执行sbatch -p p1-c1-2 job1.sh 提交任务。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。任务开始运行后，在提交目录会生成一个 slurm-xxx.out 日志文件，其中 xxx 表示任务号。

sbatch 命令示例 2（指定 2个进程，每个进程 2个 cpu 核提交 hostname 程序，限制运行 10 分钟）：编写脚本 job2.sh，内容如下：

```
!/bin/bash

#SBATCH -

n 2

#SBATCH -c 2

#SBATCH -t 10

srun -n 4 hostname
```

然后在命令行执行sbatch -p p1-c1-2 job2.sh 就可以提交任务。其中#SBATCH 注释行是 slurm 定义的任务执行方式说明，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，避免了每次提交任务写很长的命令行。

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。

2.4.3 salloc 分配模式任务提交

salloc 命令用于申请节点资源，一般用法如下：

1、执行 salloc -p p1-c1-2；
2、执行 squeue 查看分配到的节点资源，比如分配到 p1-c1-2-1；
3、执行 ssh p1-c1-2-1 登录到所分配的节点；
4、登陆节点后可以执行需要的提交命令或程序；
5、任务结束后，执行 scancel JOBID 释放分配模式任务的节点资源。

2.5 slurm管理任务命令

2.5.1 scancel 取消已提交的任务

scancel 可以取消正在运行或排队的任务。

scancel 的一些常用命令示例：

命令示例	功能
scancel 11	取消任务号为 11 的任务
scancel -n test-001	取消任务名为 test-001 的任务
scancel -p p1-c1-2	取消提交到 p1-c1-2 队列的任务
scancel -t PENDING	取消正在排队的任务
scancel -w p1-c1-2-1	取消运行在 p1-c1-2-1节点上的任务

scancel 的其他参数选项，可通过 scancel --help 查看

2.5.2 scontrol 查看正在运行的任务信息

scontrol 命令可以查看正在运行的任务详情，比如提交目录、提交脚本、使用核数情况等，对已退出的任务无效。

scontrol 的常用示例：

```
scontrol show job 7
```

```
-bash-4.2$ scontrol show job 7
JobId=7 JobName=run_fluent-2020.sh
   UserId=u18017580612(2004) GroupId=u18017580612(2005) MCS_label=N/A
   Priority=4294901754 Nice=0 Account=(null) QOS=(null)
   JobState=PENDING Reason=Nodes_required_for_job_are_DOWN,_DRAINED_or_reserved_for_jobs_in_higher_priority_partitions Dependency=(null)
   Requeue=1 Restarts=0 BatchFlag=0 Reboot=0 ExitCode=0:0
   RunTime=00:00:00 TimeLimit=UNLIMITED TimeMin=N/A
   SubmitTime=2021-09-24T02:00:39 EligibleTime=2021-09-24T02:00:39
   AccrueTime=2021-09-24T02:00:39
   StartTime=Unknown EndTime=Unknown Deadline=N/A
   SuspendTime=None SecsPreSuspend=0 LastSchedEval=2021-09-24T02:05:18
   Partition=p1-c1-2 AllocNode:Sid=head-1:31528
   ReqNodeList=(null) ExcNodeList=(null)
   NodeList=(null)
   NumNodes=1-1 NumCPUs=2 NumTasks=2 CPUs/Task=1 ReqB:S:C:T=0:0:*:*
   TRES=cpu=2,mem=1M,node=1,billing=2
   Socks/Node=* NtasksPerN:B:S:C=0:0:*:* CoreSpec=*
   MinCPUsNode=1 MinMemoryNode=1M MinTmpDiskNode=0
   Features=(null) DelayBoot=00:00:00
   OverSubscribe=OK Contiguous=0 Licenses=(null) Network=(null)
   Command=/opt/fastone/softwares/ansys-2020/run_fluent-2020.sh
   WorkDir=/fastone/users/u18017580612
   Power=
```

查看任务号为 7 的任务详情。

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

2.5.3 sacct 查看历史任务信息

sacct 命令可以查看历史任务的起止时间、结束状态、任务号、任务名、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 user1 -S 2020-11-01 -E now --field=jobid,partition,jobname,user,nnodes,nodelist,start,end,elapsed,state
```

其中：

-u user1 是指查看 user1 账号的历史任务

-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是指任务号

partition 是指提交队列

user 是指FCC账号

nnodes 是节点数

nodelist 是节点列表

start 是开始运行时间

end是任务退出时间

elapsed 是运行时间

state 是任务结束状态

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。