中科曙光 Sugon

调度系统概述-原理和部署

HPC产品事业部

2022-05-23

携手成就梦想

目录

调度系统概述-基本概念

■ 资源（Resource）

- ✓ 作业运行过程中使用的可量化实体都是资源；
- ✓ **包括硬件资源（节点、内存、CPU 、GPU等）和软件资源（ License ）；**

■ 集群（Cluster）

- ✓ 包含**计算、存储、网络**等各种资源实体且彼此联系的资源集合；
- ✓ 在物理上，一般由计算处理、互联通信、I/O 存储、操作系统、编译器、运行环境、开发工具等多个软硬件子系统组成；
- ✓ 节点是集群的基本组成单位，从角色上一般可以划分为**管理节点、登陆节点、计算节点、存储节点等。**

■ 作业（Job）

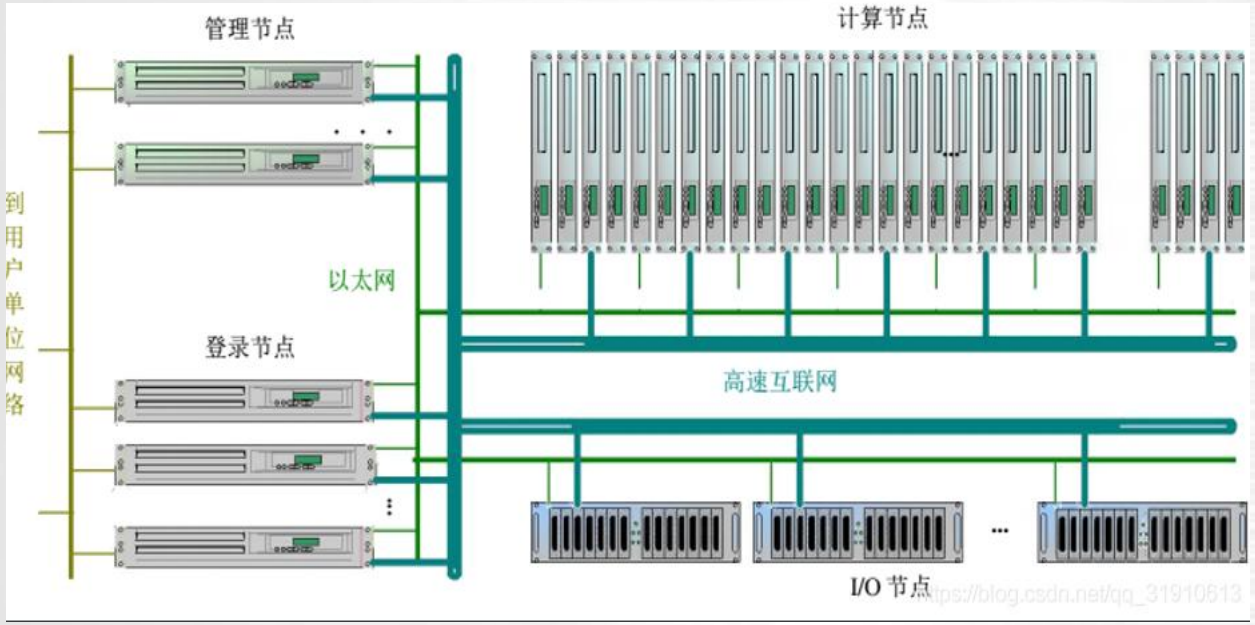
- ✓ **物理构成，一组关联的资源分配请求，以及一组关联的处理过程**；
- ✓ 交互方式，可以分为交互式作业和非交互式作业；
- ✓ 资源使用，可以分为串行作业和并行作业；

调度系统概述-基本概念（续）

- **分区（Partition）**
  - ✓ 带名称的作业容器；
  - ✓ 用户访问控制；
  - ✓ 资源使用限制；

- **作业调度系统（Job Schedule System）**
  - ✓ 负责监控和管理集群中资源和作业的软件系统；
  - ✓ 通常由资源管理器、调度器、任务执行器，以及用户命令和API组成；

调度系统概述-节点角色划分



登陆节点

和普通用户交互的主要节点。主要用来接受用户的 ssh 连接和文件传送，同时可以用来编译程序，修改代码和通过任务管理系统提交任务到计算节点。登陆节点还可细分为最外侧的负责 VPN 对接外网和入流量分流负载均衡的节点，以及普通的用来浏览文件编译程序提交任务的节点，和负责高带宽大文件传输的专用节点等。

控制节点

普通用户无法登陆，一般具有单独的管理网络，作业管理，资源分配等功能。还可细分为提供资源管理软件，提供账户管理，提供数据库后端，提供监控软件后端等不同功能的节点。分类越细，高可用就越好。每一种功能，还可以进一步包括主节点和备用从节点，从而防止单点故障。

存储节点（IO）

通过网络文件系统，如 Parastor300S，共享给登陆节点和计算节点使用。通常与这些节点通过高速互联网络，比如 `InfiniBand` 相连接，带给用户调用本地文件的速度。还可细分为存储元数据的节点，存储文件内容的节点，备份数据的节点等。

计算节点

用来进行计算任务的节点，占据了集群中的绝大多数节点。还可细分为不同硬件特性的计算节点。比如大内存节点用来解决内存瓶颈的问题，现在最大内存可达 `3T`。又比如多 `GPU` 节点，用来进行机器学习等任务。还有具有本地固态硬盘的节点，用来满足需要高速 IO 的计算任务的需求等等。

调度系统概述-面临的问题和主要用途

集群架构面临的问题

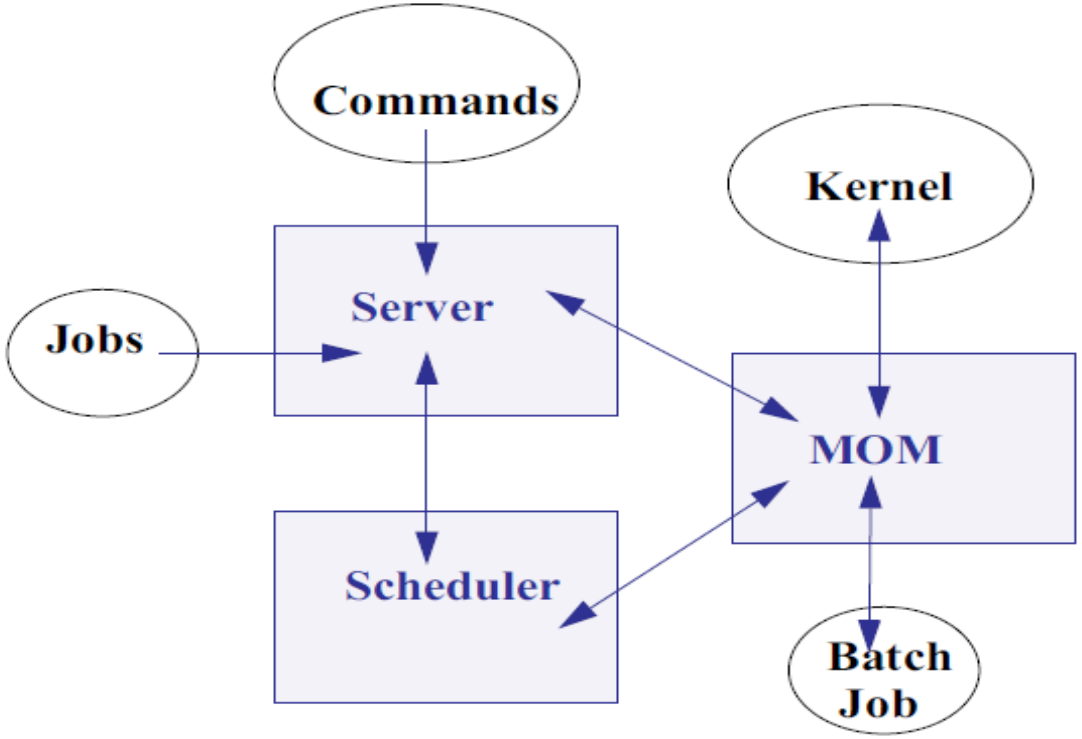
- **机群结构的松散性**
  - ✓ 异构资源/异构系统;
- **作业运行容易冲突**
  - ✓ 多用户作业并发提交;
- **资源使用不受控制**
  - ✓ OS无法实现全局管理;

调度系统的主要用途

- **系统资源整合**
  - ✓ 管理异构资源和异构系统;
- **多任务管理**
  - ✓ 统一管理任务，避免冲突;
- **资源访问控制**
  - ✓ 基于策略的资源访问控制;

调度系统是面向集群的操作系统。

调度系统概述-典型架构



主控服务Server：
主要负责资源和作业的监控和管理功能。

调度服务Scheduler：
主要负责定义和执行调度策略，包括配额管理。

执行代理MOM：
主要负责监控资源状态，以及作业的启停和监控。

访问接口：
用户访问系统的统一入口。

调度系统概述-SLURM基本概念

Slurm简介

Slurm：Slurm是一个开源，高度可扩展的集群管理工具和作业调度系统，可以简单理解为一个多机的资源和任务管理系统。主要以下提供三种关键功能：

资源分配：

在特定时间段内为用户分配计算资源，进行独占或非独占访问权限，以便他们可以执行作业。简单的说就是为用户作业提供对计算资源的授权和分配。

作业管理：

它提供了对节点上的作业节进行启动、执行和监控作业的框架。

作业调度：

通过管理待处理作业的队列来仲裁资源的争用。例如根据优先级或不同当调度策略调整资源的分配顺序。

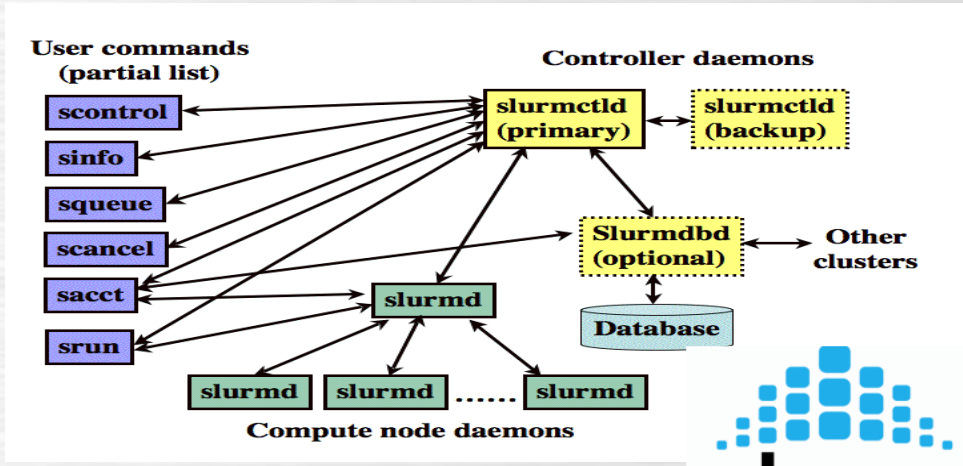
调度系统概述-功能特性

IBM Blue Gene/Q　Cray XT

天河超级计算机　硅立方

IBM_Sequoia　……

slurm workload manager Version 17.02



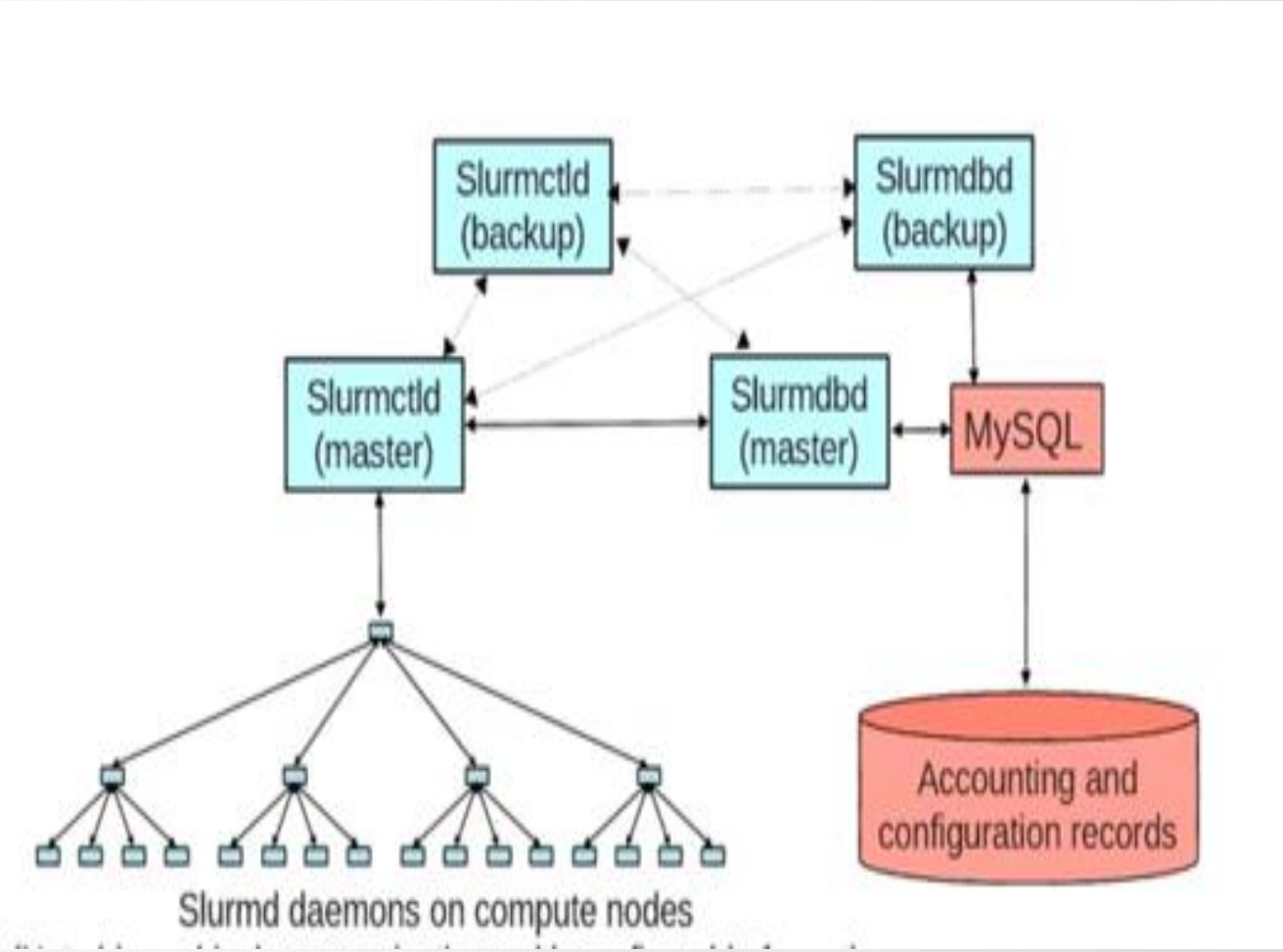
slurm workload manager Version 17.02

SLURM调度核心 -2016年6月Top500前十名有5台

- 高性能;
- 灵活性（众多插件）;
- 扩展性很高，支持数百万处理器核心的调度;
- 节点容错，高稳定性;
- 安全性高、易移植性（不修改内核）好;
- 支持各种常用的HPC操作系统(AIX、Linux、Solaris);
- MPI支持较好，作业抢占、进程/线程绑定、作业依赖等轻松支持;
- 网络拓扑调度，内置支持Tree、3D-Tours等多种算法。

调度系统概述-运行架构

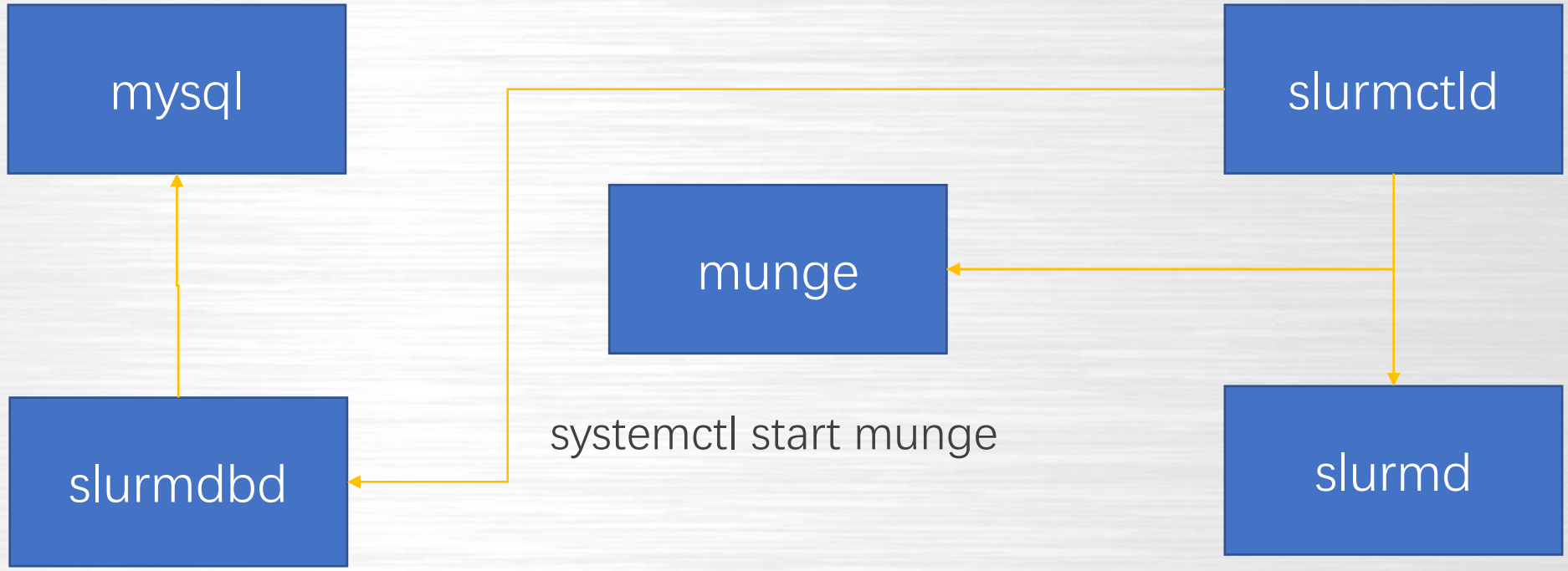
- **主控服务slurmctld**
  故障切换 资源监控
  队列管理 作业调度
- **记账存储服务slurmdbd**
  记账数据 配置信息
  故障切换
- **数据库MySQL**
  记账和配置信息存储
- **计算代理slurmd**
  启动任务 监控任务
- **认证服务munge**
  内部通信认证



调度系统概述-服务启动

/etc/init.d/my_mysqld start

systemctl start slurmctld



systemctl start munge

systemctl start slurmdbd

systemctl start slurmd

目录

安装部署-角色与服务

Centos7.6	192.168.99.125	CPU：2GHz*40 内存：4GB 磁盘：20GB	管理节点与计算节点

Slurm安装-环境准备

中科曙光 Sugon

配置SSH无密码登录

配置主机A免密登录到主机B

（方法一）

1.进入到我的home目录 cd ~/.ssh

2. ssh-keygen（四个回车），会生成两个文件id_rsa（私钥）、id_rsa.pub（公钥）

3.将公钥拷贝到要免登陆的机器上:

ssh-copy-id -i ~/.ssh/id_rsa.pub root@192.168.99.125

（方法二）

1.在主机A生产密钥对: ssh-keygen -t rsa， 会在.ssh目录下产生密钥文件

2.拷贝主机A的公钥到主机B: scp /root/.ssh/id_rsa.pub B:/root/.ssh/

3.将主机A的公钥加到主机B的授权列表.ssh/authorized_keys（若不存在，手动创建）: cat id_rsa.pub >> authorized_keys

4.授权列表authorized_keys的权限必须是600，chmod 600 authorized_keys

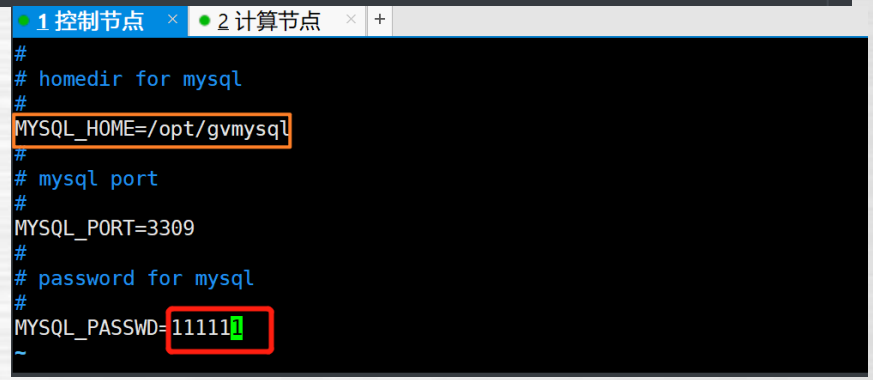
Slurm安装-环境准备

- 1:查看防火状态
- systemctl status firewalld
- service iptables status
- 2:暂时关闭防火墙
- systemctl stop firewalld
- **service iptables stop**
- 3:永久关闭防火墙
- systemctl disable firewalld
- **chkconfig iptables off**

Slurm安装-环境准备

- 这里数据库有以下要求

参数名称	参数取值	备注
Mysql安装路径	/opt/gvmysql	保持默认值
Mysql使用端口	3309	保持默认值
Mysql管理员密码	xxx	用户手工设定（111111）



- 修改mysql文件夹中的config文件
- vim config令MYSQL_PASSWD=111111
- yum -y install autoconf #安装缺少的库
- 执行 sh install_mysql_linux.sh
- ln -s /opt/gvmysql/my_mysql.sock /tmp/mysql.sock
- 添加环境变量vim /etc/profile

export PATH=$PATH:/opt/munge/bin:/opt/gvmysql/bin:/opt/gridview/slurm/bin:/opt/gridview/slurm/sbin

Slurm安装-环境准备

中科曙光 Sugon

安装munge

MUNGE (MUNGE Uid 'N' Gid Emporium)是一种用于创建和验证凭证的身份验证服务。它允许进程在一组具有公共用户和组的主机中验证另一个本地或远程进程的UID和GID。

1) 先确认环境是否已经安装Libgpg-error和libgcrypt，如果已经安装则不需要再安装这2个包。

```
rpm -qa | grep -E 'libgpg|libgcrypt'
```

2) 若未安装，则下载Libgpg-error和libgcrypt最新版。分别解压缩两个包，进入相应目录执行

解压：tar -jxf libgpg-error-1.27.tar.bz2 &&&& tar -xvf libgcrypt-1.7.5.tar.gz

编译安装：

```
./configure&&make&&make install
```

3) 在Linux系统中添加munge组和用户

```
groupadd -g 1100 munge创建一个新的组，并添加组 ID
useradd -g munge -u 1100 munge
```

Slurm安装-环境准备

安装munge

4) 下载munge最新版，解压缩，进入目录执行

编译安装：

./configure --prefix=/opt/munge --with-crypto-lib=libgcrypt

make&&make install

5) 生成秘钥（计算节点拷贝管理节点秘钥）

管理节点：echo -n "foo" | sha1sum | cut -d ' ' -f1 > /opt/munge/etc/munge/munge.key

计算节点： scp /opt/munge/etc/munge/munge.key gv244:/opt/munge/etc/munge/

chmod 0400 /opt/munge/etc/munge/munge.key

6) **设置使用root用户启动munge，修改munge安装目录/opt/munge/etc/sysconfig/munge中USER="root"**

7) **安装包目录中的src/etc 下修改munge.service的User和Group为root**，并放到/usr/lib/systemd/system下，添加执行权限

cp munge.service /usr/lib/systemd/system

systemctl daemon-reload

systemctl start munge.service #启动munge服务

8）验证

munge -n | ssh ip /opt/munge/bin/munge

Slurm安装-环境准备

安装slurm

1) 下载slurm最新版，解压缩并进入目录，执行

管理节点：./configure --prefix=/opt/gridview/slurm --with-munge=/opt/munge --with-mysql_config=/opt/gvmysql/bin

计算节点：./configure --prefix =/opt/gridview/slurm --with-munge=/opt/munge

make&& make install

PS: ./configure --help命令可以列出所有编译参数，常用编译参数含义为：

--prefix=PREFIX 指定程序存放路径，默认/usr/local/bin

--with-munge=path 指定munge安装的路径

--with-mysql_config=path 指定存在mysql_config二进制文件的目录的路径

--disable-debug 禁用调试符号并优化编译

2）安装pmi，pmi2(安装其他附带工具包如openlava,方法类似)

进入slurm安装包目录下的contribs/pmi

make&& make install

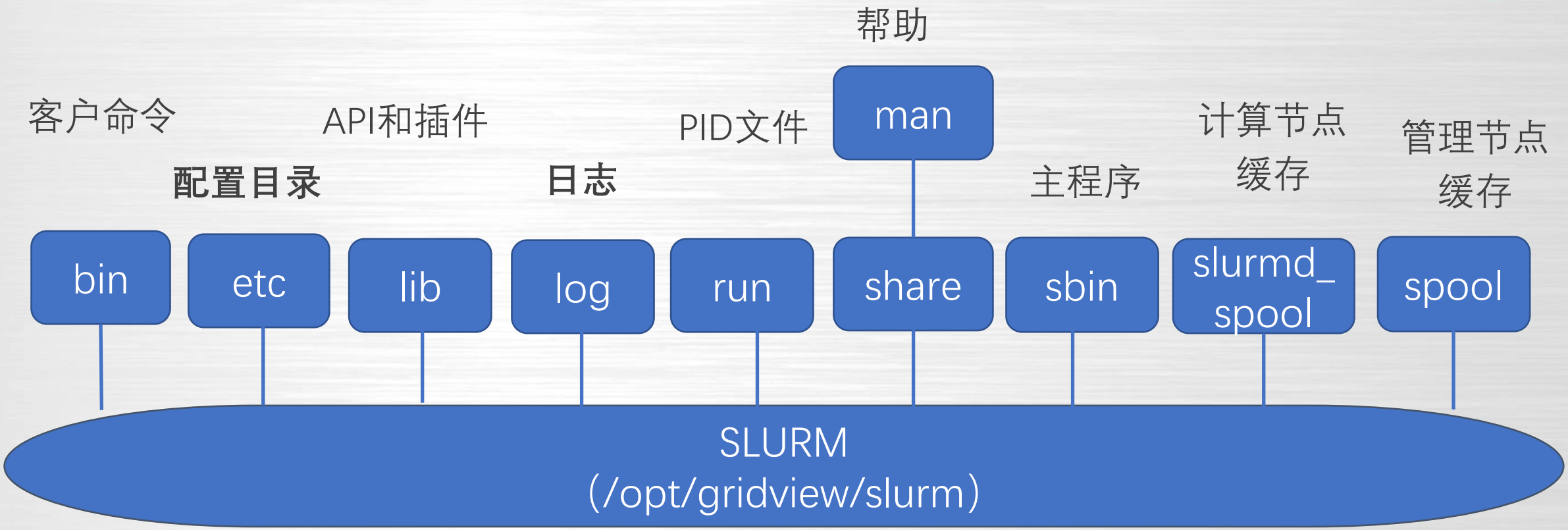
进入slurm安装包目录下的contribs/pmi2

make&& make install

3) 准备工作目录

mkdir /opt/slurm18/etc 配置文件存放目录，默认为安装目录下的etc文件夹

安装部署-软件目录结构



说明：

1. **所有节点的配置目录etc需要借助共享存储全局共享（/g1/gv_share/pia_etc）**
2. **主备节点之间需要同步spool目录实现高可用（/opt/gridview/slurm/spool）**

安装部署-配置文件概述

配置类型	文件名称	主要内容	其它说明
主配置文件	slurm.conf	包含主要的调度配置参数，包括调度策略、运行配置、日志配置、记账采集、权限控制、容错配置、认证方式、作业前后处理等等。	必选
记账存储服务配置文件	slurmdbd.conf	包含slurmdbd使用的配置参数，包括认证方式、权限控制、运行配置、日志配置和数据库访问配置。	必选
节点配置文件	slurm_node.conf	包含所有计算节点的配置参数，如节点名、CPU核数、内存、默认状态等。	可选，可合并到slurm.conf
分区配置文件	slurm_partition.conf	包含所有分区的配置参数，如分区名、节点列表、优先级、合法账号、默认时间、默认内存、默认状态等。	可选，可合并到slurm.conf
拓扑配置文件	topology.conf	包含所有节点与交换机以及交换机之间的层次网络链接关系。	可选，由参数TopologyPlugin=topology/tree控制
通用资源配置文件	gres.conf	定义计算节点包含的GRES资源，如名称、类型、设备等。	可选，由参数GresTypes和Gres控制

安装部署-主配置文件slurm.conf

中科曙光 Sugon

```
ClusterName=cluster_gvm03  #集群名称
ControlMachine=gvm03       #主用节点
BackupController=gvm04     #备用节点
AuthType=auth/munge          #内部认证
CryptoType=crypto/munge      #加密方式
MaxJobCount=200000           #最大作业数20万
JobSubmitPlugins=lua         #提交参数过滤
KillOnBadExit=1              #task异常作业清理
ProctrackType=proctrack/cgroup  #进程跟踪插件
ReturnToService=1              #禁用自动恢复
SlurmctldPort=6817           # 主控服务端口
SlurmdPort=6818              #计算代理端口
SlurmUser=slurmadm           #slurmctld运行用户
SlurmdSpoolDir=/opt/gridview/slurm17/slurmd_spool/
# 计算代理缓存
StateSaveLocation=/opt/gridview/slurm17/spool   #
slurmctld本地文件缓存
TopologyPlugin=topology/tree     #拓扑调度
```

```
TaskPlugin=task/affinity          # 任务启动cpuset
MinJobAge=300                      #作业内存保留时
间
SlurmctldTimeout=30               #主备切换时间
SlurmdTimeout=300                  #计算代理响应时
间
FastSchedule=1                      #快速调度作业
SchedulerType=sched/backfill        #启用回填
SchedulerPort=7321                  #调度器端口
SelectType=select/cons_res          #资源选择算法
SelectTypeParameters=CR_Core_Memory  #基于
Core和内存调度
SchedulerParameters=batch_sched_delay=3,defer,sche
d_min_interval=10,sched_interval=30,default_queue_d
epth=100,bf_max_job_test=100,bf_interval=30   # 调
度参数
AccountingStorageTRES=cpu,mem  #TRES指标配置
```

安装部署-主配置文件slurm.conf

```
PriorityType=priority/multifactor        #优先级策略
PriorityDecayHalfLife=30          #半衰期时长
PriorityCalcPeriod=5              #FS统计间隔
PriorityWeightFairshare=100       #FS权重
PriorityWeightPartition=1000      #分区权重
ClusterName=pia                   #集群名
PreemptMode=requeue,gang                #抢占策略
PreemptType=preempt/partition_prio    #队列优先级
DebugFlags=NO_CONF_HASH          # 调试标识
PrivateData=accounts,events,jobs,reservations,usage,users     # 权限控制
HealthCheckInterval=60                  #检查间隔
HealthCheckProgram=/usr/sbin/nhc      #检查工具
AccountingStorageEnforce=associations,limits  #组织关联和资源限制
AccountingStorageHost=gvm04    #主用记账服务
AccountingStorageBackupHost=gvm03 #备用记账服务
AccountingStoragePort=7031          #记账服务端口
```

```
AccountingStorageType=accounting_storage/slurmdbd
#启用slurmdbd
AccountingStorageUser=root          #记账服务
AccountingStoreJobComment=YES   #记录作业注释
JobCompType=jobcomp/none     #禁止生成comp日志
JobAcctGatherFrequency=300    #作业采集间隔
JobAcctGatherType=jobacct_gather/linux    #启用Linux插件
SlurmctldDebug=3                        #日志级别
SlurmctldLogFile=/opt/gridview/slurm17/log/slurmctld.log
JobRequeue=1                  # 允许重新排队(默认为1)
SlurmdDebug=3                           #日志级别
SlurmdLogFile=/opt/gridview/slurm17/log/slurmd_%h.log
SuspendTime=1800                        #
include slurm_node.conf                 #引入节点配置
include slurm_partition.conf            #引入分区配置
```

安装部署-记账存储服务配置slurmdbd.conf

```
AuthType=auth/munge                                          # 内部认证类型
DbdHost=gvm04                                                # slurmdbd服务节点
DbdBackupHost =gvm03                                         # 备用服务节点
DbdPort=7031                                                 # 记账存储服务监控端口
SlurmUser=slurmadm                                           # 运行用户
DebugLevel=3                                                 # 日志级别
PrivateData=accounts,events,jobs,reservations,usage,users    # 权限控制
LogFile=/opt/gridview/slurm17/log/slurmdbd.log               # 日志路径
StorageType=accounting_storage/mysql                         # 启用mysql
StorageHost=gvm05                                            # 数据库主机
StorageBackupHost=gvm06                                      # 数据库备机
StoragePort=3308                                             # 数据库端口
StoragePass=root                                             # 密码
StorageUser=root                                             # 用户名
StorageLoc=gv_slurm_db                                       # 数据库示例
```

安装部署-节点配置slurm_node.conf

```
NodeName=cmac[0011-0260] NodeAddr=cmac[0011-0260] CPUs=32 Boards=1 SocketsPerBoard=2 CoresPerSocket=16 ThreadsPerCore=1 RealMemory=385437 State=UNKNOWN
NodeName=cmac[0261-1538] NodeAddr=cmac[0261-1538] CPUs=32 Boards=1 SocketsPerBoard=2 CoresPerSocket=16 ThreadsPerCore=1 RealMemory=191913 State=UNKNOWN
```

可选参数:
MemSpecLimit： 保留内存的大小
Weight：节点权重，用于节点选择
Feature： 节点特征，用于节点选择
Gres： 通用资源(如GPU)，如 GRES=gpus:2
Reason： 节点状态异常（down、drain、fail等）时的原因。
State： 可选的状态包括DOWN 、FAIL、FAILING 、UNKNOWN 、BUSY、IDLE、CLOUD、FUTURE。不要直接配置成BUSY（报错）和IDLE，而应该配置为UNKNOWN（默认）。

注意事项:
1. 节点配置的变更需要同时重启slurmctld和slurmd服务

安装部署-分区配置slurm_partition.conf

```
PartitionName=serial    Nodes=cmac[0011-0034] Priority=1000  OverSubscribe=FORCE:1 Default=NO  AllowAccounts=ALL DefaultTime=15-00:00:00 MaxTime=INFINITE DefMemPerCPU=10240 LLN=YES State=UP
PartitionName=serial_op    Nodes=cmac[0011-0034] Priority=1000  OverSubscribe=FORCE:1 Default=NO  AllowAccounts=nwp,nwp_op,nwp_sp,lijuan,nwp_pd,nwp_qu DefaultTime=15-00:00:00 MaxTime=INFINITE
 DefMemPerCPU=10240 LLN=YES State=UP
PartitionName=largemem    Nodes=cmac[0035-0260] Priority=1000  OverSubscribe=FORCE:1 Default=NO  AllowAccounts=ALL QOS=normal_qos DefaultTime=15-00:00:00 MaxTime=INFINITE DefMemPerCPU=10240 S
tate=UP
PartitionName=normal    Nodes=cmac[0035-1538] Priority=1000  OverSubscribe=FORCE:1 Default=NO  AllowAccounts=ALL QOS=normal_qos DefaultTime=15-00:00:00 MaxTime=INFINITE DefMemPerCPU=5120 Stat
e=UP
PartitionName=operation Nodes=cmac[0035-1538] Priority=2000  OverSubscribe=FORCE:1 Default=NO  AllowAccounts=nwp,nwp_op,nwp_sp,lijuan,nwp_pd,nwp_qu,nwpbj_ex DefaultTime=15-00:00:00 MaxTime=IN
FINITE State=UP
```

参数简介:

OverSubscribe:

EXCLUSIVE:独占节点，启用了Select/cons_res

FORCE[:X]:强制节点（在X作业间）共享，忽略用户自身请求

YES:允许作业共享，考虑用户--oversubscribe 请求。

PreemptMode:

队列级的抢占模式，覆盖全局配置

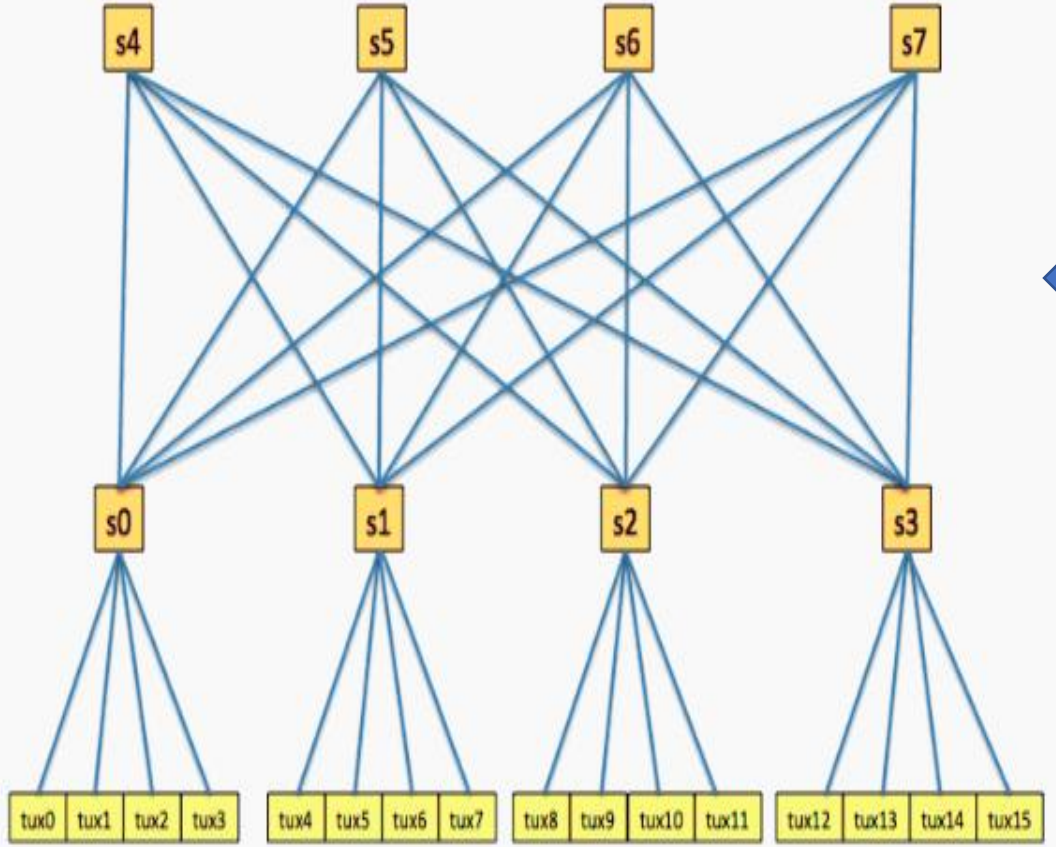
PriorityJobFactor:

用于multifactor Priority。

State:

UP（正常） DOWN(接收不调度) DRAIN（调度不接收） INACTIVE（DOWN+DRAIN）

安装部署-拓扑配置示例




```
# topology.conf
# Switch Configuration
SwitchName=s0 Nodes=tux[0-3]
SwitchName=s1 Nodes=tux[4-7]
SwitchName=s2 Nodes=tux[8-11]
SwitchName=s3 Nodes=tux[12-15]
SwitchName=s4 Switches=s[0-3]
```

配置格式:

（1）交换机-节点

SwitchName=X Nodes=<node name>

（2）交换机-交换机

SwitchName=X Switches=<switch name>

安装部署-拓扑配置topology.conf

```
……
SwitchName=ibsw97
Nodes=gv0942,gv0934,gv0941,gv0933,gv0940,gv0932,gv0939,gv0931,gv0946,gv0938,gv0945,gv0937,gv0944,gv0936,gv0943,gv0935,gv0891,gv0892
SwitchName=ibsw98
Nodes=gv0926,gv0918,gv0925,gv0917,gv0924,gv0916,gv0923,gv0915,gv0930,gv0922,gv0929,gv0921,gv0928,gv0920,gv0927,gv0919,gv0893,gv0894
SwitchName=ibsw99
Nodes=gv0910,gv0902,gv0909,gv0901,gv0908,gv0900,gv0907,gv0899,gv0914,gv0906,gv0913,gv0905,gv0912,gv0904,gv0911,gv0903,gv0883,gv0884
SwitchName=ibsw102 Switches=ibsw3,ibsw2,ibsw1
SwitchName=ibsw103 Switches=ibsw4,ibsw6,ibsw5
SwitchName=ibsw104 Switches=ibsw9,ibsw7,ibsw8
SwitchName=ibsw105 Switches=ibsw12,ibsw11,ibsw10
.....
```

安装部署-通用资源配置gres.conf

- 支持通用资源，必须在slurm.conf配置文件中明确指定要管理哪些资源。

参数	解释
GresTypes	e.g. *GresTypes=gpu,mic*
Gres	e.g. Gres=gpu:tesla:2,gpu:kepler:2

安装部署-通用资源配置gres.conf

参数	解释
Name	通用资源的名称（必须与slurm.conf中的GresTypes值匹配）
Count	此节点上可用的此类型资源数。默认值为1
CPUs	指定可以使用该资源的CPU index number
File	e.g. *File=/dev/nvidia[0-3]*
Type	指定设备类型。

```
NodeName=e01r1n[01-09,11-19],e01r2n[00-19],e01r3n[00-19],e01r4n[00-19],e02r1n[00-19],e02r2n[00-19],e02r3n[00-19],e02r4n[00-19],e03r1n[00-19],e03r2n[00-19],e03r3n[00-19],e03r4n[00-19],e04r1n[00-19],e04r2n[00-19],e04r3n[00-19],e04r4n[00-19],e05r1n[00-19],e05r2n[00-19]
5r3n[00-19],e05r4n[00-19],e06r1n[00-19],e06r2n[00-19],e06r3n[00-19],e06r4n[00-19],e07r1n[00-19],e07r2n[00-19],e07r3n[00-19],e07r4n[00-19],e08r1n[00-19],e08r2n[00-19],e08r3n[00-19],e08r4n[00-19],e09r1n[00-19],e09r2n[00-19],e09r3n[00-19],e09r4n[00-19],e10r1n[00-19],e1
n[00-19],e10r3n[00-19],e10r4n[00-19],e11r1n[00-19],e11r2n[00-19],e11r3n[00-19],e11r4n[00-19],e12r1n[00-19],e12r2n[00-19],e12r3n[00-19],e12r4n[00-19],e13r1n[00-19],e13r2n[00-19],e13r3n[00-19],e13r4n[00-19],e14r1n[00-19],e14r2n[00-19],e14r3n[00-19],e14r4n[00-19],e15r1
0-19],e15r2n[00-19],e15r3n[00-19],e15r4n[00-19],e16r1n[00-19],e16r2n[00-19],e16r3n[00-19],e16r4n[00-19],e17r1n[00-19],e17r2n[00-19],e17r3n[00-19],e17r4n[00-19],e18r1n[00-19],e18r2n[00-19],e18r3n[00-19],e18r4n[00-19],e19r1n[00-19],e19r2n[00-19],e19r3n[00-19],e19r4n[0
9],e20r1n[00-19],e20r2n[00-19],e20r3n[00-19],e20r4n[00-19],i01r1n[01-09,11-19],i01r2n[00-19],i01r3n[00-19],i01r4n[00-19],i02r1n[00-19],i02r2n[00-19],i02r3n[00-19],i02r4n[00-19],j07r1n[00-19],j07r2n[00-19],j07r3n[00-19],j07r4n[00-19],j08r1n[00-19],j08r2n[00-19],j08r3
0-19],j08r4n[00-19],j09r1n[00-19],j09r2n[00-19],j09r3n[00-19],j09r4n[00-19],j10r1n[00-19],j10r2n[00-19],j10r3n[00-19],j10r4n[00-19],j15r1n[00-19],j15r2n[00-19],j15r3n[00-19],j15r4n[00-19],j16r1n[00-19],j16r2n[00-19],j16r3n[00-19],j16r4n[00-19],j17r1n[00-19],j17r2n[0
9],j17r3n[00-19],j17r4n[00-19],j18r1n[00-19],j18r2n[00-19],j18r3n[00-19],j18r4n[00-19],j19r1n[00-19],j19r2n[00-19],j19r3n[00-19],j19r4n[00-19],j20r1n[00-19],j20r2n[00-19],j20r3n[00-19],j20r4n[00-19],j12r1n[00-19],j12r2n[00-19],j12r3n[00-19],j12r4n[00-19],j13r1n[00-1
j13r2n[00-19],j13r3n[00-19],j13r4n[00-19],j14r1n[00-19],j14r2n[00-19],j14r3n[00-19],j14r4n[00-19],j11r1n[00-19],j11r2n[00-19],j11r3n[00-19],j11r4n[00-19],b01r1n[01-09,11-19],b01r2n[00-19],b01r3n[00-19],b01r4n[00-19],b02r1n[00-19],b02r2n[00-19],b02r3n[00-19],b02r4n[0
9],b03r1n[00-19],b03r2n[00-19],b03r3n[00-19],b03r4n[00-19],b04r1n[00-19],b04r2n[00-19],b04r3n[00-19],b04r4n[00-19],b05r1n[00-19],b05r2n[00-19],b05r3n[00-19],b05r4n[00-19],b06r1n[00-19],b06r2n[00-19],b06r3n[00-19],b06r4n[00-19],b07r1n[00-19],b07r2n[00-19],b07r3n[00-1
b07r4n[00-19],b08r1n[00-19],b08r2n[00-19],b08r3n[00-19],b08r4n[00-19],b09r1n[00-19],b09r2n[00-19],b09r3n[00-19],b09r4n[00-19],b10r1n[00-19],b10r2n[00-19],b10r3n[00-19],b10r4n[00-19],b11r1n[00-19],b11r2n[00-19],b11r3n[00-19],b11r4n[00-19],b12r1n[00-19],b12r2n[00-19],
r3n[00-19],b12r4n[00-19],b13r1n[00-19],b13r2n[00-19],b13r3n[00-19],b13r4n[00-19],b14r1n[00-19],b14r2n[00-19],b14r3n[00-19],b14r4n[00-19],b15r1n[00-19],b15r2n[00-19],b15r3n[00-19],b15r4n[00-19],b16r1n[00-19],b16r2n[00-19],b16r3n[00-19],b16r4n[00-19],b17r1n[00-19],b17
[00-19],b17r3n[00-19],b17r4n[00-19],b18r1n[00-19],b18r2n[00-19],b18r3n[00-19],b18r4n[00-19],b19r1n[00-19],b19r2n[00-19],b19r3n[00-19],b19r4n[00-19],b20r1n[00-19],b20r2n[00-19],b20r3n[00-19],b20r4n[00-19],c01r1n[01-09,11-19],c01r2n[00-19],c01r3n[00-19],c01r4n[00-19],
r1n[00-19],c02r2n[00-19],c02r3n[00-19],c02r4n[00-19],c03r1n[00-19],c03r2n[00-19],c03r3n[00-19],c03r4n[00-19],c04r1n[00-19],c04r2n[00-19],c04r3n[00-19],c04r4n[00-19],c05r1n[00-19],c05r2n[00-19],c05r3n[00-19],c05r4n[00-19],c06r1n[00-19],c06r2n[00-19],c06r3n[00-19],c06
[00-19],c07r1n[00-19],c07r2n[00-19],c07r3n[00-19],c07r4n[00-19],c08r1n[00-19],c08r2n[00-19],c08r3n[00-19],c08r4n[00-19],c09r1n[00-19],c09r2n[00-19],c09r3n[00-19],c09r4n[00-19],c10r1n[00-19],c10r2n[00-19],c10r3n[00-19],c10r4n[00-19],c11r1n[00-19],c11r2n[00-19],c11r3n
-19],c11r4n[00-19],c12r1n[00-19],c12r2n[00-19],c12r3n[00-19],c12r4n[00-19],c13r1n[00-19],c13r2n[00-19],c13r3n[00-19],c13r4n[00-19],c14r1n[00-19],c14r2n[00-19],c14r3n[00-19],c14r4n[00-19],c15r1n[00-19],c15r2n[00-19],c15r3n[00-19],c15r4n[00-19],c16r1n[00-19],c16r2n[00
],c16r3n[00-19],c16r4n[00-19],f11r1n[01-09,11-19],f11r2n[00-19],f11r3n[00-19],f11r4n[00-19],f12r1n[00-19],f12r2n[00-19],f12r3n[00-19],f12r4n[00-19],f13r1n[00-19],f13r2n[00-19],f13r3n[00-19],f13r4n[00-19],f14r1n[00-19],f14r2n[00-19],f14r3n[00-19],f14r4n[00-19],f15r1n
-19],f15r2n[00-19],f15r3n[00-19],f15r4n[00-19],f16r1n[00-19],f16r2n[00-19],f16r3n[00-19],f16r4n[00-19],f17r1n[00-19],f17r2n[00-19],f17r3n[00-19],f17r4n[00-19],f18r1n[00-19],f18r2n[00-19],f18r3n[00-19],f18r4n[00-19],f19r1n[00-19],f19r2n[00-19],f19r3n[00-19],f19r4n[00
],f20r1n[00-19],f20r2n[00-19],f20r3n[00-19],f20r4n[00-19],h03r1n[00-19] Name=dcu Type=Hygon File=/dev/dri/renderD131 Cores=24-31
[root@admin04 ~]#
```

安装部署-日志管理概述

- **日志文件范围**

  主控服务slurmctld　　记账存储服务slurmdbd
  计算代理服务slurmd　　认证服务munge

- **日志级别参数**

  配置参数:

  Slurm.conf:　　SlurmctldDebug　SlurmdbdDebug
  slurmdbd.conf:　　DebugLevel

  日志级别:

  调度系统支持的日志级别包括:

quiet	fatal	error	info	verbose	debug	debug 2	debug 3	debug 4	debug 5
0	1	2	**3（默认级别）**	4	5	6	7	8	9

- **日志转储设置**

  通过操作系统的logrotate工具管理实现自动的日志滚动保存。

安装部署-日志文件说明

日志文件:

- **主进程日志slurmctld.log**

  存在于调度系统管理节点的/opt/gridview/slurm17/log目录，记录主进程运行日志，涉及作业提交、作业调度、作业控制、状态监控等各个方面的正常和异常信息。

- **记账存储服务日志slurmdbd.log**

  存在于调度系统管理节点的/opt/gridview/slurm17/log目录，主要记录跟数据库相关的各种操作日志。

- **计算代理日志slurmd_{hostname}.log**

  存在于调度系统计算节点的/opt/gridview/slurm17/log目录，记录计算节点服务的运行日志。

- **认证服务日志munged.log**

  存在于每一个调度相关节点的/opt/gridview/munge/log/munge/munged.log目录，主要记录调度系统各组件通信过程中产生/销毁各种凭证（credential）的日志。

安装部署-日志转储设置

转储配置：

通过三个logrotate配置文件分别实现slurmctld/slurmdbd、slurmd、munge服务的日志文件转储。

日志转储配置	节点分布	对应服务
/etc/logrotate.d/slurm	管理节点	主控服务slurmctld 记账存储服务slurmdbd
/etc/logrotate.d/slurmd	计算节点	计算代理slurmd
/etc/logrotate.d/munge	所有节点	认证服务munged

手工测试：

```
logrotate  -f   /etc/logrotate.d/slurmd
```

安装部署-日志转储设置-主服务转储配置

```
/opt/gridview/slurm17/log/slurmdbd.log
/opt/gridview/slurm17/log/slurmctld.log
{
        compress                # 启用gzip压缩
        missingok           # 日志不存在不报错退出
        nocopytruncate      # 转储不清空
        nodelaycompress     # 转储时立即压缩
        nomail              # 禁用邮件通知
        notifempty          # 为空时不转储
        noolddir            # 原目录保存
        rotate 3            # 保存3个转储文件
        sharedscripts       # 多个文件同时处理
        daily               # 转储周期
        dateext             # 转储后缀为年月日
        size 200M           # 条件大于200M
```

```
        # 文件转储后的操作
        postrotate
                for daemon in $(scontrol show daemons)
                do
                        killall -SIGUSR2 $daemon
                done
                ps -fe|grep slurmdbd |grep -v grep
                if [ $? -ne 0 ]
                then
                echo "no slurmdbd process"
                else
                killall -SIGUSR2 slurmdbd
                fi
        endscript
}
```

安装部署-日志转储设置-计算代理转储配置

```
/opt/gridview/slurm17/log/slurmd_*.log
{
        compress                 # 启用gzip压缩
        missingok          # 日志不存在不报错退出
        nocopytruncate         # 转储不清空
        nodelaycompress      # 转储时立即压缩
        nomail                  # 禁用邮件通知
        notifempty               # 为空时不转储
        noolddir                 # 原目录保存
        rotate 3                 # 保存3个转储文件
        sharedscripts            # 多个文件同时处理
        daily                    # 转储周期
        dateext                  # 转储后缀为年月日
        size 50M                # 条件大于50M


        # 文件转储后的操作
        postrotate
                for daemon in $(scontrol show daemons)
                do
                        killall -SIGUSR2 $daemon
                done
          endscript


}
```

安装部署-日志转储设置-认证服务转储配置

```
/opt/gridview/munge/log/munge/*.log
{
        compress                # 启用gzip压缩
        missingok           # 日志不存在不报错退出
        nocopytruncate      # 转储不清空
        nodelaycompress     # 转储时立即压缩
        nomail              # 禁用邮件通知
        notifempty          # 为空时不转储
        noolddir            # 原目录保存
        rotate 3            # 保存3个转储文件
        sharedscripts       # 多个文件同时处理
        daily               # 转储周期
        dateext             # 转储后缀为年月日
        size 50M            # 条件大于50M
```

```
        # 文件转储后的操作
        postrotate
                ps -fe|grep munged |grep -v grep
                if [ $? -ne 0 ]
                then
                        echo "no munged process"
                else
                        killall -SIGUSR2 munged
                fi
        endscript

}
```