北航星空高性能计算集群使用手册

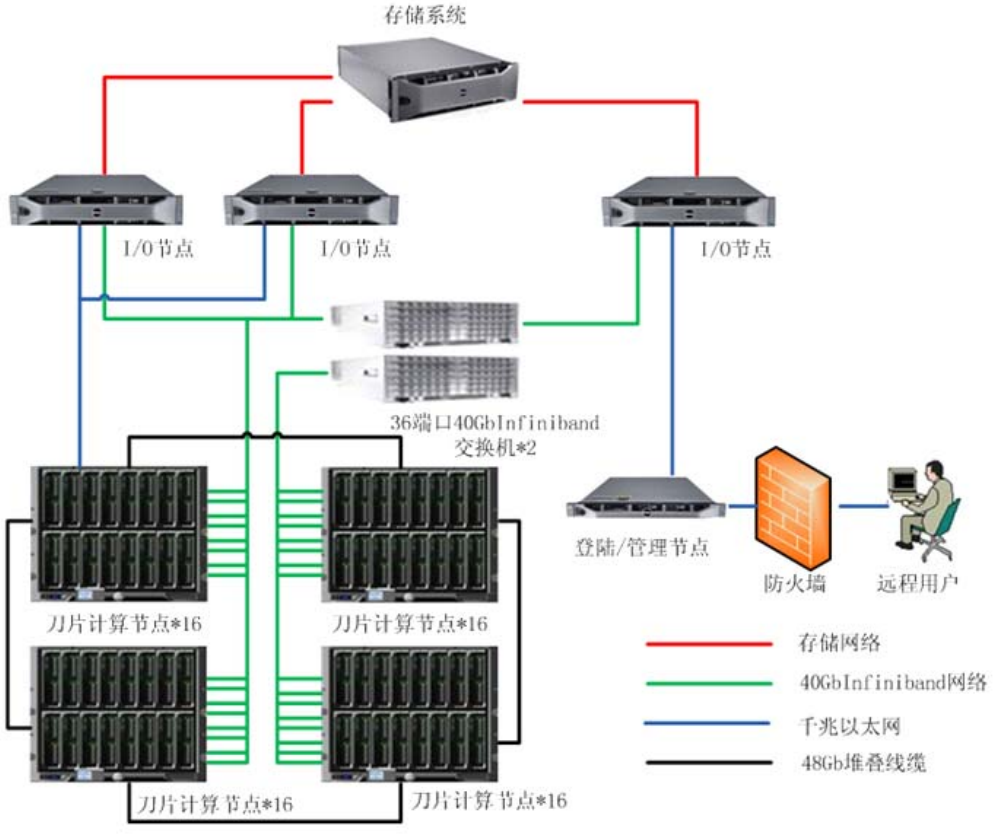
本文包括如下内容：高性能计算硬件集群介绍、高性能软件环境介绍、高性能集群HOWTO、高性能计算集群账号申请，高性能集群性能（包括理论值和 linpack 值）。

1、 星空高性能计算集群简介

星空高性能计算集群分为 star1 和 star2 两个集群。

Star2 高性能计算集群由 16 个刀片服务器、1 台机架式服务器组成，共 256 核处理器，访问 IP 地址是 202.112.136.138

Star1 高性能计算集群由 64 个刀片服务器、4 台机架式服务器组成，其中刀片服务器作为计算节点，机架服务器作为管理节点和 IO 节点。每台刀片服务器有 2 颗 Intel Xeon E5530(2.4GHz)处理器和 24GB(12×2GB) DDR3 1066GHz 内存，总计算能力一共有 512 核处理器，再打开超线程后，一共有 1024 核处理器。计算网络采用 40Gb 的高带宽、低延迟的 Infiniband 网络，计算节点与 IO 节点之间也通过 Infiniband 网络连接，保证读写密集型程序的高带宽和低延迟需求。管理网络通过机箱背板上的交换机模块和千兆以太网交换机互联，所有计算节点、管理节点、IO 节点都采用千兆以太网互联。操作系统为 Red Hat Enterprise Linux Server release 5.4。星空高性能计算集群只能通过 ssh 登录，访问 IP 地址为：202.112.136.140。集群拓扑结构参见下图：



星空高性能计算集群安装 Rocks 集群管理软件，装有 Intel C++、Fortran 编译器和支持以太网及 Infiniband 高速网络的 MPI 并行环境（包括 Open MPI 1.4.1 和 MVAPICH 1.2.0），还装有 ANSYS 流体动力学软件。IO 节点采用 Lustre（版本 1.8.3）并行文件系统，总存储空间 5.1T。

星空高性能计算集群由北航网络信息中心维护管理，面向全校师生提供科研或者学习方面的高性能计算服务。如果通过使用该集群，有论文或者研究成果发表，请答谢北航网络信息中心。

申请账号请发邮件到，每人都有一个账号，不能多人共享一个账号。集群计算资源为共享资源，请所有用户相互尊重，遵循使用规章制度，不要独占资源或者进行其它一些与高性能计算不相关的活动。如果违反使用规章制度，账号将被封锁，并且该用户将可能遭受网络信息中心其它服务黑名单处罚。如果有任何问题和建议请发邮件到，或者拨打电话 82317005－836。

2、 星空高性能计算集群环境和使用步骤

2.1. 高性能计算集群环境

2.1.1. 用户程序和数据存储

目前，提供用户主目录“/home/*”和“/mnt/luster/public”两个地方进行存储，其中：

1. 用户主目录“/home/*”目录用于存放用户编写的程序。
2. “/mnt/luster/public”作为公共的存储空间，为所有用户提供模式运算和应用所需的存储。用户可在/mnt/lustr/public 目录下建立子目录，用于存放大规模的计算数据。请注意该空间为公共使用空间，管理员将定期删除回收该存储空间，用户的所有计算数据以及结果应及时下载备份到本地并删除。

2.1.2. 编译器

集群的所有节点都安装了 GNU 64 位和 Intel 64 位编译器，GNU 为操作系统自带，Intel 编译器安装在/opt/intel 下，都支持 C、C++、Fortran 程序编译。

编译器	安装路径	版本
GNU 编译器	/usr/bin	4.1.2
Intel 编译器	/opt/intel	10.1.022

2.1.3. 软件库（待建设和完善）

fftw2.1.5
Intel mkl7.2.1
Intel ipp4.1
Intel Vtune
Intel Trace Analyzer
Intel Trace Collector

2.1.4. MPI 环境

系统默认编译环境：GNU 编译器，路径为/usr/mpi/gcc/mvapich-1.2.0/bin。

MPI 产品	安装路径	网络类型	版本	MPI 标准
MPICH2				

GNU 编译器	/opt/mpich2/gnu/	以太网	1.1.1p1	支持 MPI-1 标准
Intel 编译器	opt/mpich2-intel-10/	以太网	1.2.1	支持 MPI-2 标准
OpenMPI				
GNU 编译器	/opt/openmpi/	以太网	1.3.3	支持 MPI-2 标准
MVAPICH				
GNU 编译器	/usr/mpi/gcc/mvapich-1.2.0/	Infiniband	1.2.7	支持 MPI-1 标准
Intel 编译器	/usr/mpi/intel/mvapich-1.2.0/	Infiniband	1.2.7	支持 MPI-1 标准
OpenMPI				
GNU 编译器	/usr/mpi/gcc/openmpi-1.4.1/	Infiniband	1.4.1	支持 MPI-1 标准
Intel 编译器	/usr/mpi/intel/openmpi-1.4.1/	Infiniband	1.4.1	支持 MPI-1 标准

2.1.5. 通信网络

- InfiniBand Native
- InfiniBand ipoib
- 千兆以太网

2.2. 使用步骤

2.2.1. 登录

使用本集群必须先向网络信息中心申请账号，一人一个账号。登录地址、账号和初始密码获知后，就可以登录集群，初次登录后请用 passwd 命令及时修改密码。当您登录到本集群时，您所处的位置是登录节点。在登录节点上您只能编译、提交作业，切记不能将登录节点作为计算节点。

如果您不希望每次连接的时候都输入密码，可以在您的机器上生成一个密钥，并把相应的公钥拷贝的您的用户主目录下的.ssh 子目录下的 authorized_keys 中。

登录地址：

2.2.2. 上传并行程序代码

上传传递源代码和数据文件，可以通过 scp 或者 sftp。Linux 用户可以通过 scp 传递文件，命令为“scp programname YourUsername@Server IP:~/”；Window 用户可以通过 SSH Secure Shell、CuteFTP Pro 等支持 sftp 的客户端软件或者 Secure CRT、Putty 等支持 SSH 协议登录的客户端软件进行登录和数据传递。

请及时将计算所产生的数据和结果下载到本地并及时删除。

2.2.3. 并行程序变编译

本系统提供并行的 MPI 程序编译环境：

1、对于 fortran 程序，可以采用“mpif77 example.f”或“mpif90 example.f”编译。

2、对于 c 程序，可以采用“mpicc example.c”编译。

3、的、对于 c++程序，可以采用“mpicxx example.cpp”。

注意：如果不使用默认路径“/usr/mpi/gcc/mvapich-1.2.0/bin”的 GNU MPI 环境，请指明全路径。

2.2.4. 编写作业脚本

作业脚本编写参考 2.3 作业脚本样例及相关文档。

特别注意：

1、资源申请时请注意单个节点内存不能超过 20G。

2、使用 Infiniband 的 mpi 时使用“mpirun_rsh -hostfile $PBS_NODEFILE -np *** *****”。

3、使用以太网的 mpi 时使用“mpirun -machinefile $PBS_NODEFILE -np *** ****”。

2.2.5. 提交作业

集群采用 Open PBS 作业管理系统管理作业，使用 qsub 命令提交作业。假设上一步编写的脚本名为 YourJob.sh，提交命令如下：

$qsub YourJob.sh

2.2.6. 查看运行结果和数据清理

运行“qstat”或者“showq”查看运行结果，并及时下载运算结果和数据同时将其从服务器上删除。

2.3. 作业脚本样例

2.3.1. 以太网作业脚本示例

以太网作业脚本样例

```
#!/bin/sh -f

#PBS -N cpitest
#PBS -l nodes=2:ppn=8
#PBS -l walltime=03:00:00
#PBS -l mem=12mb
#PBS -q default

nprocs=`wc -l < $PBS_NODEFILE`
cd $PBS_O_WORKDIR

/usr/mpi/intel/openmpi-1.4.1/bin/mpirun -np $nprocs -machinefile $PBS_NODEFILE $PBS_O_WORKDIR/cpi
```

说明：

1. 脚本中的粗红体部分可根据用户需要替换
2. 如果使用上述脚本，源代码必须支持以太网的 MPI 编译环境编译连接，如 mpicc、mpif77、mpif90 等等。

3. #PBS -N 后为当前作业的名字
4. #PBS -l nodes= 指定了当前作业申请的资源，2:ppn=8 表示使用 2 个计算节点，每个节点使用 8 个处理器。
5. #PBS -l walltime= 指定了当前作业申请的计算时间，03:00:00 表示该作业需要 3 小时运行时间，一旦超过该事件，作业将被终止。
6. #PBS -l mem= 指定了当前作业估计申请的内存量。
7. #PBS -q 指定了作业将被提交到的队列名。
8. $PBS_O_WORKDIR 为 OpenPBS 的环境变量，表示当前作业的工作目录，即运行 qsub 命令提交作业脚本时用户所在的目录
9. cpi 为例子中并行程序的名称，可替换其他程序。

2.3.2. Infiniband 网作业脚本示例

Infiniband 网作业脚本样例，使用 gnu 编译器，MPI 环境为 mvapich-1.2.0

```
#!/bin/sh -f

#PBS -N linpacktest
#PBS -l nodes=20:ppn=8
#PBS -l walltime=03:00:00
#PBS -l mem=1024mb
#PBS -q default

nprocs=`wc -l < $PBS_NODEFILE`
cd $PBS_O_WORKDIR

/usr/mpi/intel/mvapich-1.2.0/bin/mpirun_rsh -np $nprocs -hostfile $PBS_NODEFILE $PBS_O_WORKDIR/xhpl
```

说明：
1. 脚本中的粗体部分可根据用户需要替换。
2. 脚本中以#PBS 开头的各行的含义与上一个脚本相同。

xhpl 为并行程序，可以按需替换其他程序。

2.4. PBS 相关命令

1. qsub 作业脚本 ：提交一个作业，如果成功返回一个作业编号，格式为[序号.管理节点名]，序号为数字
2. qdel 作业序号 ：删除一个作业，如果加-p 则为强制删除，作业号只需给出数字序号
3. qstat ：查看当前系统的作业运行状况，系统显示目前队列中所有作业的最主要信息
4. qstat -a：查看当前系统的作业运行状况，系统显示目前队列中所有作业的详细信息
5. qstat -f 作业序号：查看对应的正在运行作业的详细信息
6. qhold 作业序号 ： 阻塞指定作业
7. qrls 作业序号 ： 释放指定作业