浪潮天梭TS10K集群培训



inspur 浪潮



第一章 集群介绍

inspur 浪潮

一、什么是集群？

“联结起来的计算机的集合，整个系统使用起来像一个单一的一体的计算资源”

- Greg Pfister, *In Search of clusters*



清华地球模拟器

二、为什么用集群？

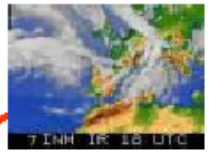
通俗的说，就是单一计算节点的资源已经无法满足我们的计算需求，比如内存的容量，cpu的计算能力。也许一个算例我们要计算数周甚至数月，这时候我们就需要用到集群来处理这些计算问题。



Government-Classified Work

Government - Research

OAK RIDGE NATIONAL LABORATORY



Instituto Nacional de Meteorología España

(Severe) Weather Prediction & Climate Modeling



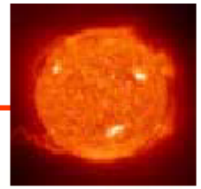
Automotive Design & Safety





Drug Discovery & Genomic Research





Aircraft/Spacecraft Design & Fuel-Efficiency



Oil Exploration & Energy Research



Basic Scientific Research

inspur 浪潮

三、常见的几种并行计算体系结构

- SMP
- DSM （CC-NUMA）
- MPP
- Cluster

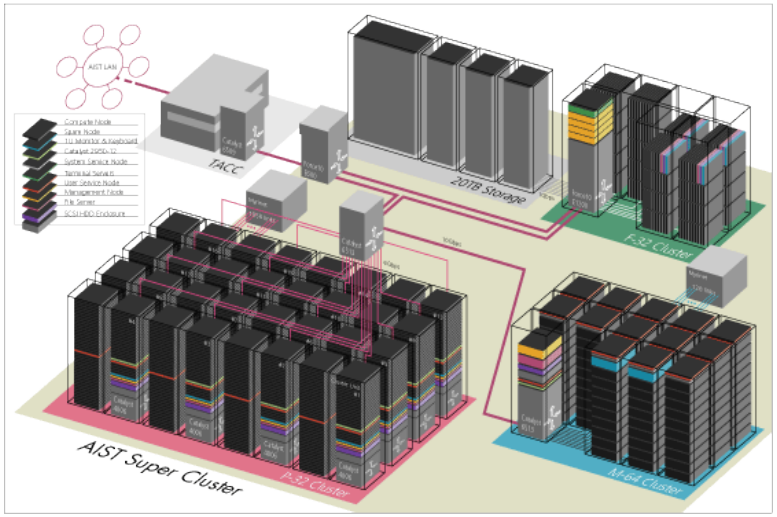
我们通常说的集群为Cluster集群。

并行计算机的发展

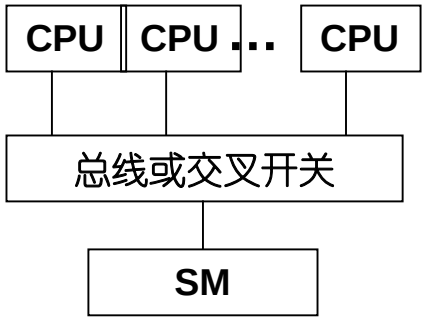
- (1)并行机的萌芽阶段（1964-1975）
- (2)向量机/ SMP的发展和鼎盛阶段(1976-1990)
- (3)MPP出现和蓬勃发展阶段（1990-1995）
- (4)Cluster出现，并成为并行计算机的主流（1995年后）



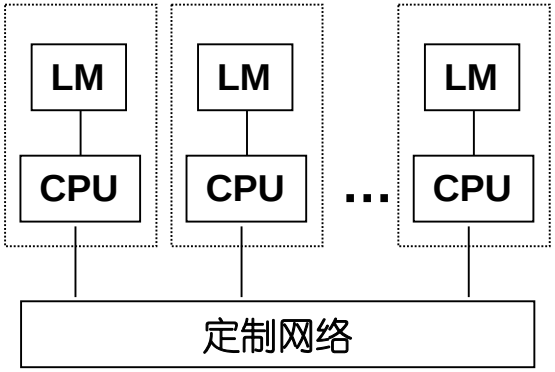




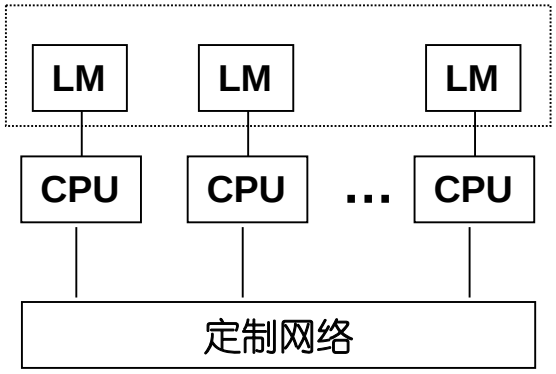
并行计算体系结构对比



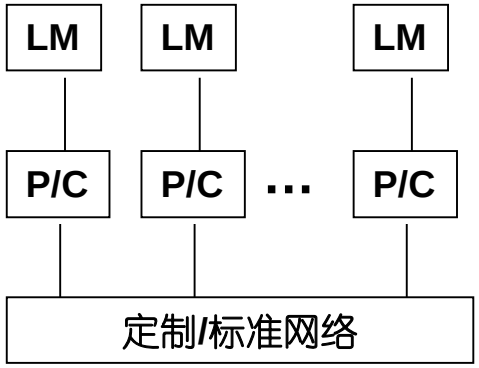
(a) SMP, 物理上单一地址空间



(b) MPP, 物理/逻辑上多地址空间



(c) DSM, 逻辑上单一地址空间

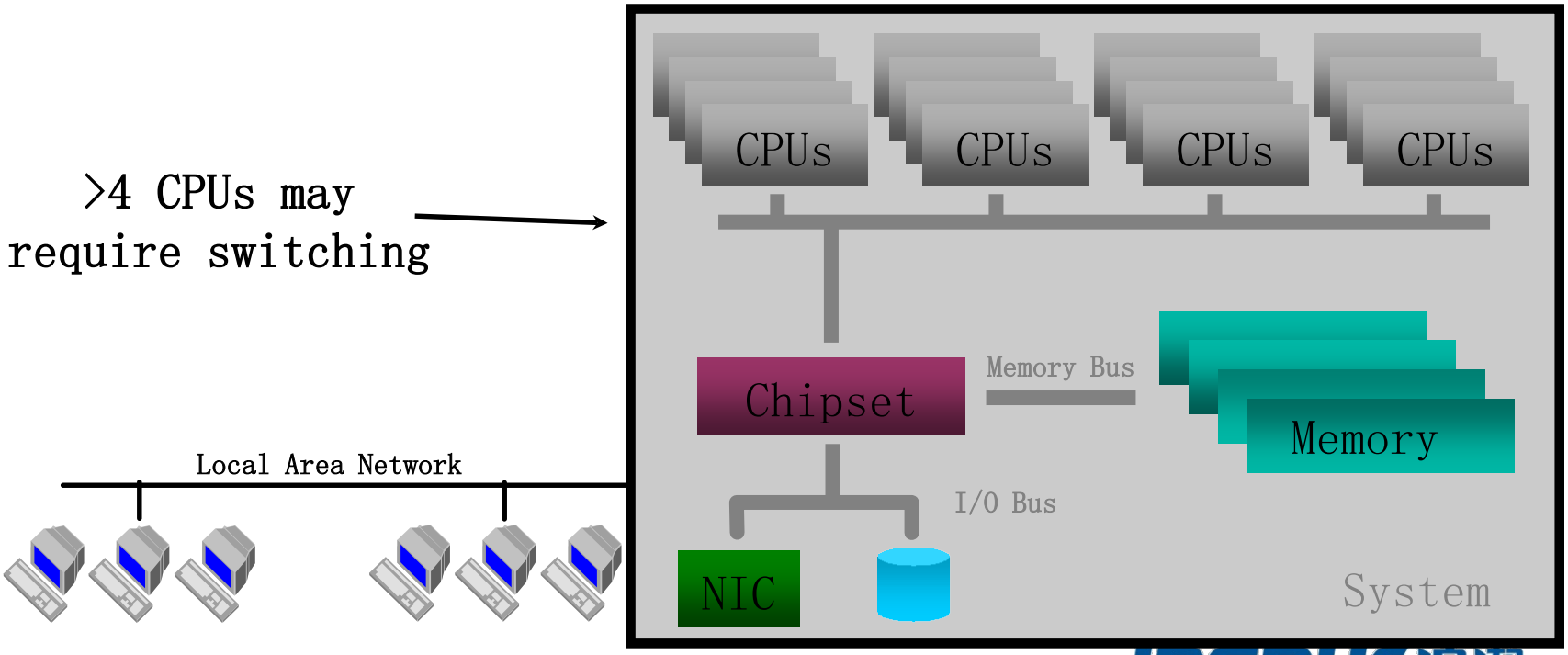


(d) Cluster, 物理/逻辑上多地址空间

inspur 浪潮

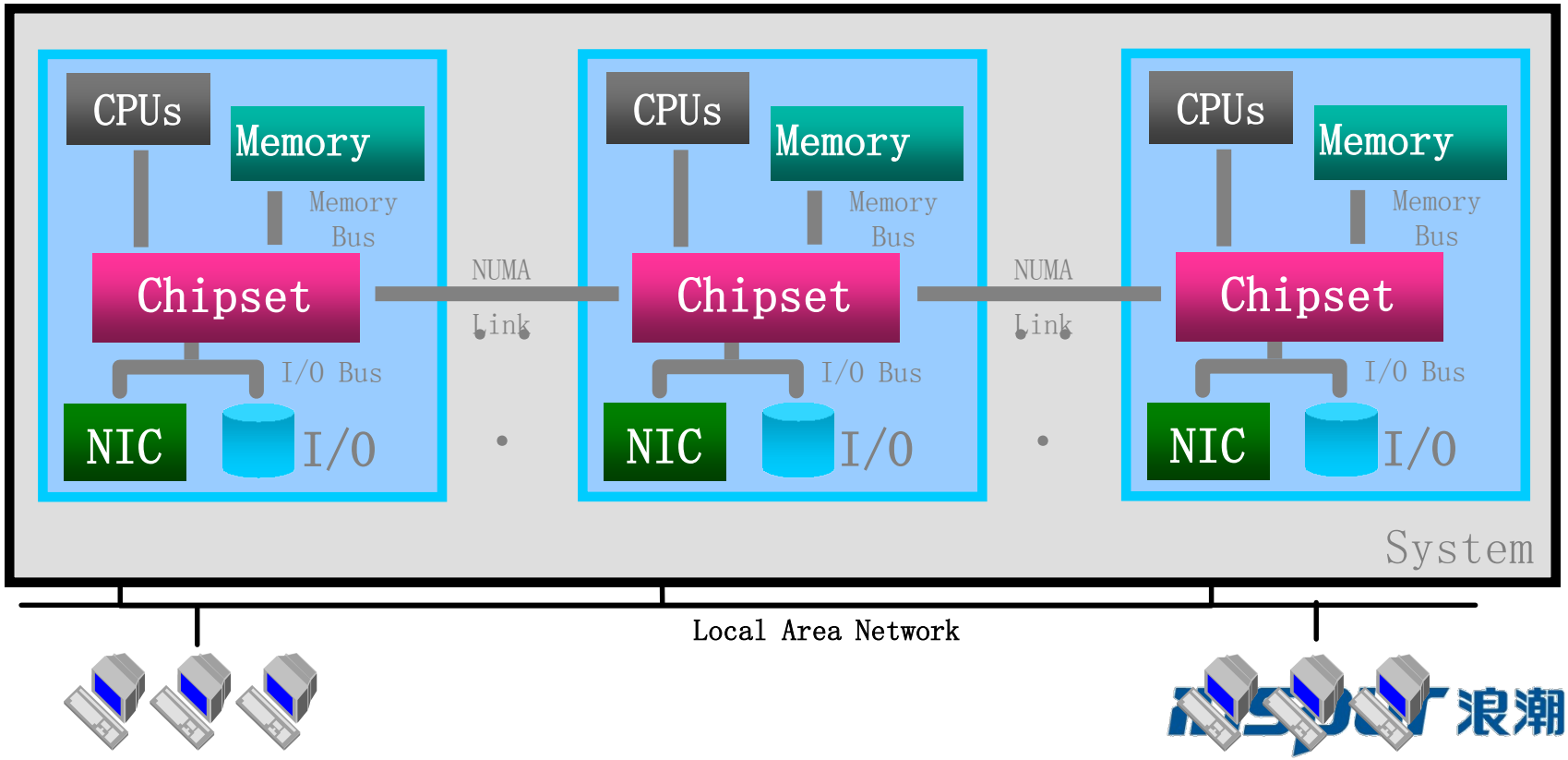
SMP- Symmetric MultiProcessing

- 多个CPU连接于统一的内存总线
- 内存地址统一编址，单一操作系统映像
- 可扩展性较差，一般CPU个数少于32个
- 目前商用服务器多采用这种架构



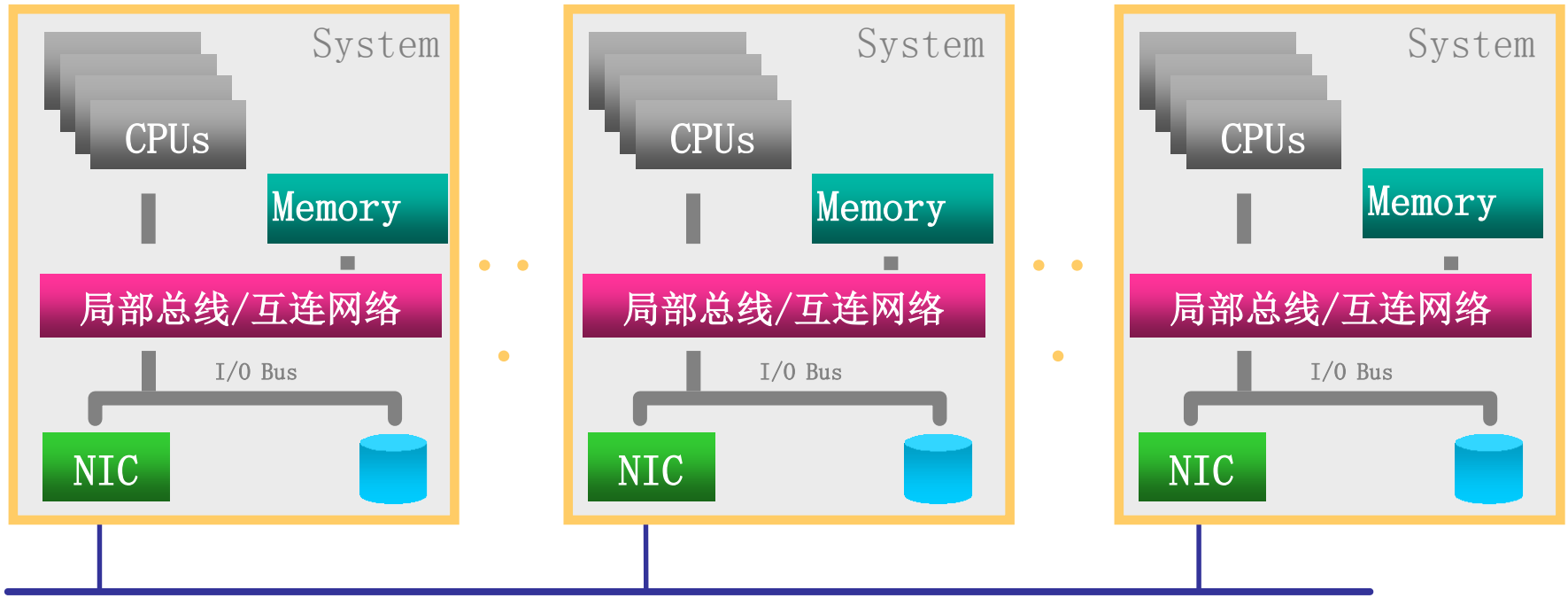
DSM- Distributed Shared Memory

- 物理上分布存储、所有内存模块统一编址
- 非一致内存访问（NUMA）模式
- 基于Cache的数据一致性，又称CC-NUMA
- 节点数可扩展到几百个，小型机多为此类架构



MPP- Massively Parallel Processors

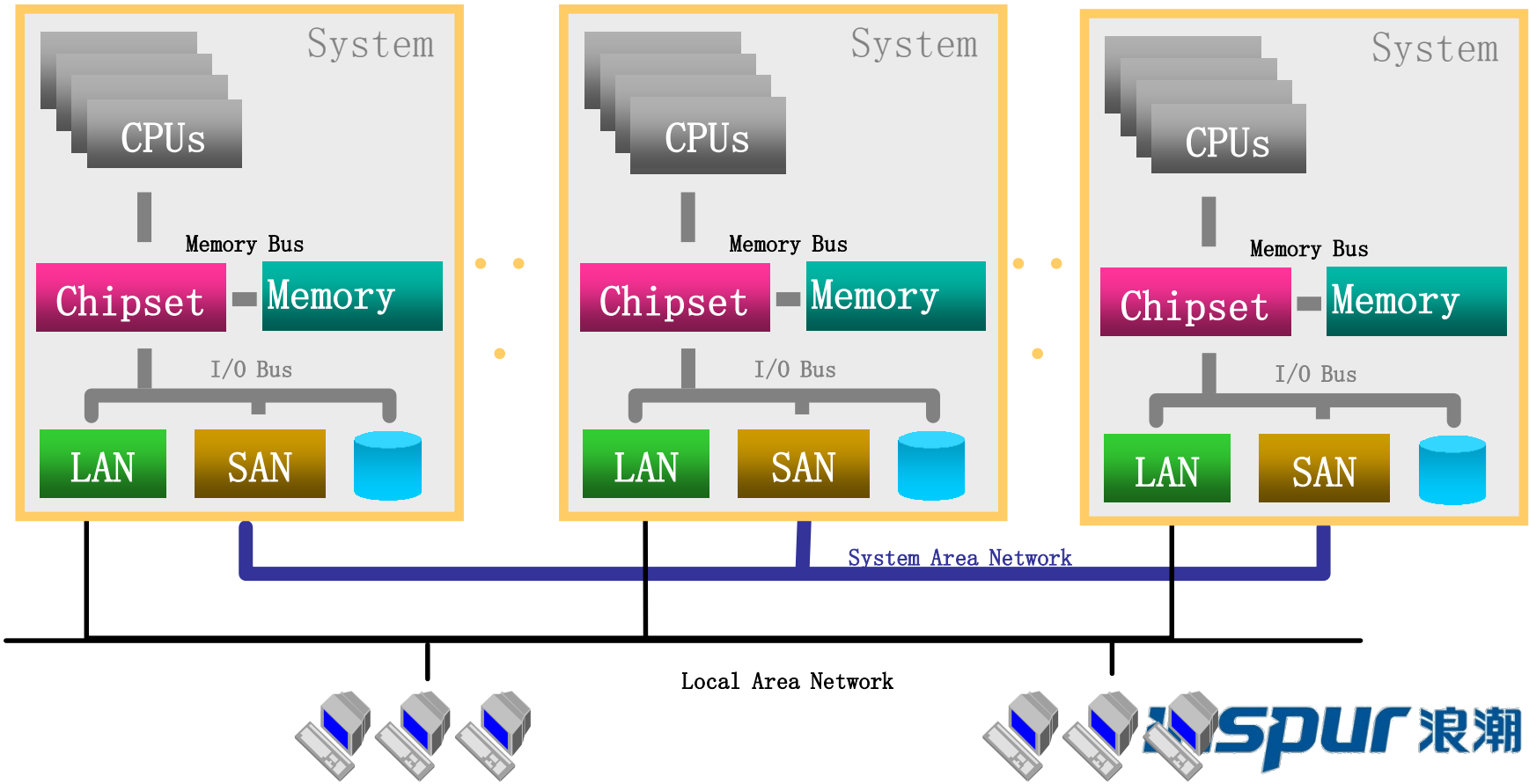
- 节点个数可达成百上千个
- 节点类型可以是DM、SMP、DSM
- 节点之间采用专用高速互连设备
- 排列在Top500前面的多数系统属于这种类型



inspur 浪潮

Cluster

- 独立的节点、商品化网络连接
- 开放的硬件设备、操作系统、应用编程接口
- 节点数可达几百个，性能已接近超级计算机系统
- 近年来发展很快，已广泛应用到高性能科学计算领域



并行计算机

- 并行计算机就是由多个处理单元组成的计算机系统，这些处理单元相互通信和协作以快速、高效求解大型复杂问题。



对于集群，我通常关注：

处理单元有多少
处理单元的功能有多强
处理单元之间怎样连接
处理单元的数据如何传递
各处理单元如何相互协作
并行程序如何编写

常见的HPC应用软件

- **CAE(力学/工程/设计)**
  - ABAQUS Standard, Explicit
  - Ansys (ANSYS)
  - CFD++ (Metacomp)
  - CFD-ACE+ (ESI)
  - Feko* (EMSS)
  - Fluent* (Fluent)
  - FIRE* (AVL)
  - HyperMesh* (Altair)
  - LS-Dyna* (LSTC)
  - Mafia* (CST)
  - MSC.Nastran* (MSC.Software)
  - NX Nastran* (UGS)
  - PAM-Crash* (ESI)
  - PowerFLOW* (EXA)
  - Star-CD* (CD/Adapco)
- **CAD**
  - Pro/ENGINEER* (PTC)
  - NX* (UGS)
  - Parasolid* (UGS)
  - Tc Vis* (UGS)
- **能源**
  - Eclipse* (Schlumberger)
  - Petrel / GigaViz (Schlumberger)
  - Omega2 (Western Geco)
  - VIP / Nexus (Landmark Graphics)
  - Geoprobe (Landmark Graphics)
  - Promax / SeisSpace (Landmark Graphics)
  - Geodepth* (Paradigm Geo)
  - Focus* (Paradigm Geo)
  - Geocluster (CGG)
  - GeoEast (BGP)
  - iCluster (SinoPec)
- **金融分析**
  - RiskWatch (Algorithmic)
  - RMDS (Reuters)
  - Thomson Financial
  - Credient, Intellitracs (Sungard)
  - CPS (Morgan-Stanley)
- **动漫渲染**
  - Alias – Maya
  - Discreet – 3ds max
  - mental images – mental ray
  - Pixar RenderMan

生命/化学

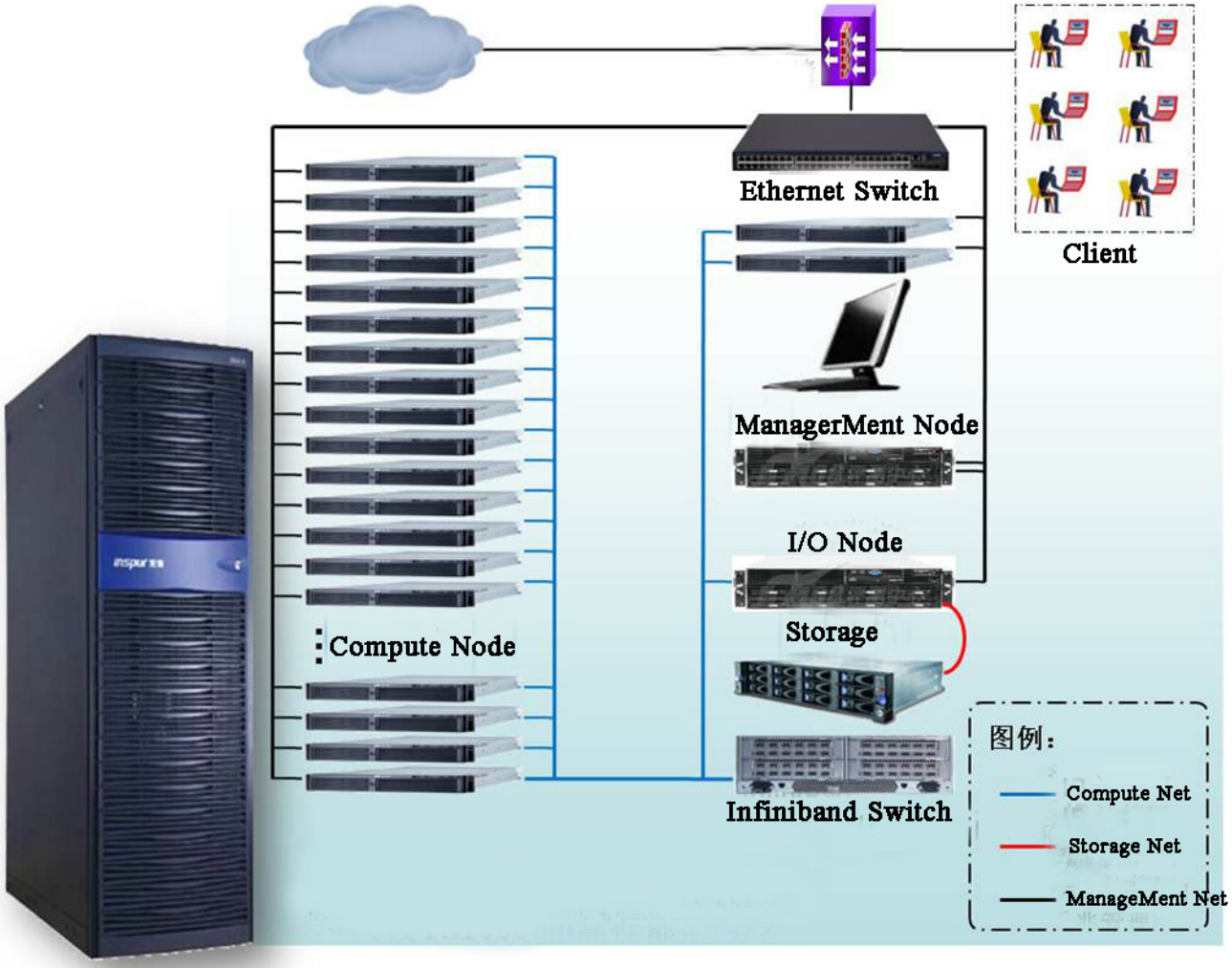
BLAST
Gaussian
AMBER
GAMESS
Gromacs
HMMER
NAMD

气象预报

MM5
WRF
CCSM3
POP
CAM
Aladin
Hirlam
UM
GRAPES (CMA)

inspur 浪潮

三、集群有哪些部分组成？



- **计算节点：**承载集群的计算任务
- **管理节点：**用户登录、管理调度整个集群、任务提交等，提供对整个系统的监控管理
- **存储节点：**存储大量的计算数据，高可用，高安全

inspur 浪潮

天梭TS10000硬件系统组成



- 计算节点
  - TS860
  - NX5440M4
  - GPU,MIC
- 登录/管理节点
  - NF5270M4
- 存储节点
  - NF5270M4
- 计算网络
  - Infiniband网络
  - 1000M以太网
- 管理网络
  - 1000M以太网

浪潮TSCE3.1实现远程统一监控管理

inspur 浪潮

天梭TS10000软件系统组成

- **Redhat linux操作系统**
- 浪潮集群监控管理软件
- 集群作业调度系统
  - TSCE作业调度系统
- 集群并行环境
  - MPICH MPI消息传递接口软件
  - INTEL MPI
  - InfiniBand专用MPI
- 集群开发环境
  - GNU C,C++ 、GNU F77/90
  - Intel编译器C/C++/Fortran、Intel高性能数学库MKL



补充说明

1、所有计算计算节点均挂载了mu01节点的/opt目录，且ypbind的服务器为mu01节点，/opt目录为intel编译器安装目录。

2,所有计算计算节点均挂载了存储节点的/lustre存储，存储空间为150T，所有用户数据放置在/lustre/home目录下

2、管理节点、计算节点安装Redhat as6.5企业版64位。

3、其余更详细的情况请参考资产配置总表和日常维护手册

2

第二章 集群使用

一、集群并行环境内容

- How can we begin?





一、集群并行环境内容

需要的条件：

1、客户端与集群连通

2、客户端如果为linux可以直接ssh登录到管理节点；如果为windows系统，需要安装相应软件









3、一个合法账号！

4、相应操作知识积累

一、集群并行环境内容

- 账号有哪些内容？

✓ 1、属性、密码

环境变量.bashrc、默认bash、密码复杂度

✓ 2、家目录

nfs网络文件系统使所有节点均有相同的家目录

✓ 3、Mail

✓ 4、用户为集群用那么所有节点之间必须互通

一、集群并行环境内容

那么集群如何添加用户？

如果每台服务器依次添加集群用户，并且每个节点内均有相同的内容，那么如果集群规模达到成千上万呢？

解决方案：

1、 nis服务
2、 ntp服务
3、 nfs网络文件分区
4、 ssh、 rsh
5、 mpi程序、相应的编译器、应用软件

二、集群开发环境

Intel面向高性能计算环境的工具库

- **Intel® C++ Compiler Linux*** 版
- **Intel® Fortran Compiler Linux*** 版
- 英特尔® 数学核心函数库**Linux** 版
- 英特尔® **MPI** 库 **Linux** 版
- 英特尔® 跟踪分析器和跟踪采集器
- 英特尔® **VTune** 分析器 **Linux** 版





inspur 浪潮

Intel® C++ Compiler Linux* 版

提供广泛的高级优化、多线程和处理器支持

其中包括处理器自动调度、矢量化功能、自动并行处理、OpenMP*、数据预取和循环展开，以及用于并行处理、数学处理和多媒体库的高度优化的 C++ 模板

支持三种平台架构：IA-32 、 IA-64 和Intel®64体系结构

支持最新多核处理器

在命令行上使用编译器

在调用编译器之前，必须给编译器环境脚本 iccvars.sh（或 iccvars.csh）使用 "source" 命令，以设置环境变量：

```
source /opt/intel/composer_xe_2013.5.192/bin/iccvars.sh intel64
```

要编译 C 语言源文件（例如，my_source_file.c），请使用以下命令：

```
$ icc my_source_file.c
```

要编译 C++ 源文件，请使用此命令：

```
$ icpc my_source_file.cpp
```

成功编译之后，会在当前目录下创建一个可执行文件 a.out。

通过使用 -help 选项调用编译器，可以从命令行查看编译器选项的摘要：

```
$ icc -help
```


inspur 浪潮

英特尔® 数学核心函数库Linux 版

英特尔® MKL可提供经过高度优化和广泛线程化的数学例程，专为需要极致性能的科学、工程及金融等领域的应用而设计

“英特尔 MKL 集群版”通过 LAPACK（解算器与特征值解算器）提供线性代数功能；另外还通过 1、2、3 级 BLAS 为复杂的数学软件提供所需的矢量、矢量-矩阵及矩阵-矩阵运算。对于喜欢使用 FORTRAN 90/95 编程语言的用户，则有机会通过专门设计的接口（参数个数已减少）来调用 LAPACK 驱动程序与计算子例程。

添加环境变量

source /opt/intel/composer_xe_2013.5.192/mkl/bin/mklvars.sh intel64

- 线性代数 - BLAS 和 LAPACK
- 线性代数 – ScaLAPACK
- 线性代数 - 稀疏矩阵解算器
- 快速傅立叶变换 (FFT)
- 向量数学库
- 矢量随机数生成器
- LINPACK 基准

inspur 浪潮

英特尔® MPI 库 Linux 版

英特尔® MPI 库 3.1 可在多种结构上实施高性能的 MPI-2 规范，旨在提高应用在基于英特尔架构的集群上的性能

支持以下互连结构的任意组合：

共享内存

通过 DAPL* 实现 RDMA 功能的网络结构，如 InfiniBand* 与 Myrinet*套接字，例如, TCP/IP over Ethernet*、Gigabit Ethernet* 以及其它互连结构。

使用“英特尔 MPI 库”涉及以下步骤：

添加环境变量

source /opt/intel/impi/4.1.1.036/bin64/mpivars.sh

编译与链接应用程序

使用适当的 mpi 命令编译 MPI 程序,例如，使用 GNU* C 编译器的 mpicc 命令编译 C 代码，如下所示：

$ mpicc <installdir>/test/test.c

选择网络结构或设备

要选择特定的结构组合，请将 I_MPI_DEVICE 环境变量设置为某个支持的结构

运行 MPI 程序

要启动与“英特尔 MPI 库”链接的程序，请使用 mpiexec 命令：

$ mpirun –genv I_MPI_DEVICE rdma –n ./test

inspur 浪潮

MPI应用

- MPICH的使用
  - 最简单MPI程序

```
[test@mu01 ~]$ cat test.c
#include<stdio.h>
#include<mpi.h>
int main( int argc,char *argv[] )
{
    MPI_Init( &argc, &argv );
    printf( "Hello, world!\n" );
    MPI_Finalize();
    return 0;
}
```

MPI程序的编译

- mpicc -o hello hello.c

MPI程序的运行

- mpirun -np 4 ./test

```
[test@mu01 ~]$ mpirun -np 4 ./test
Hello, world!
Hello, world!
Hello, world!
Hello, world!
```

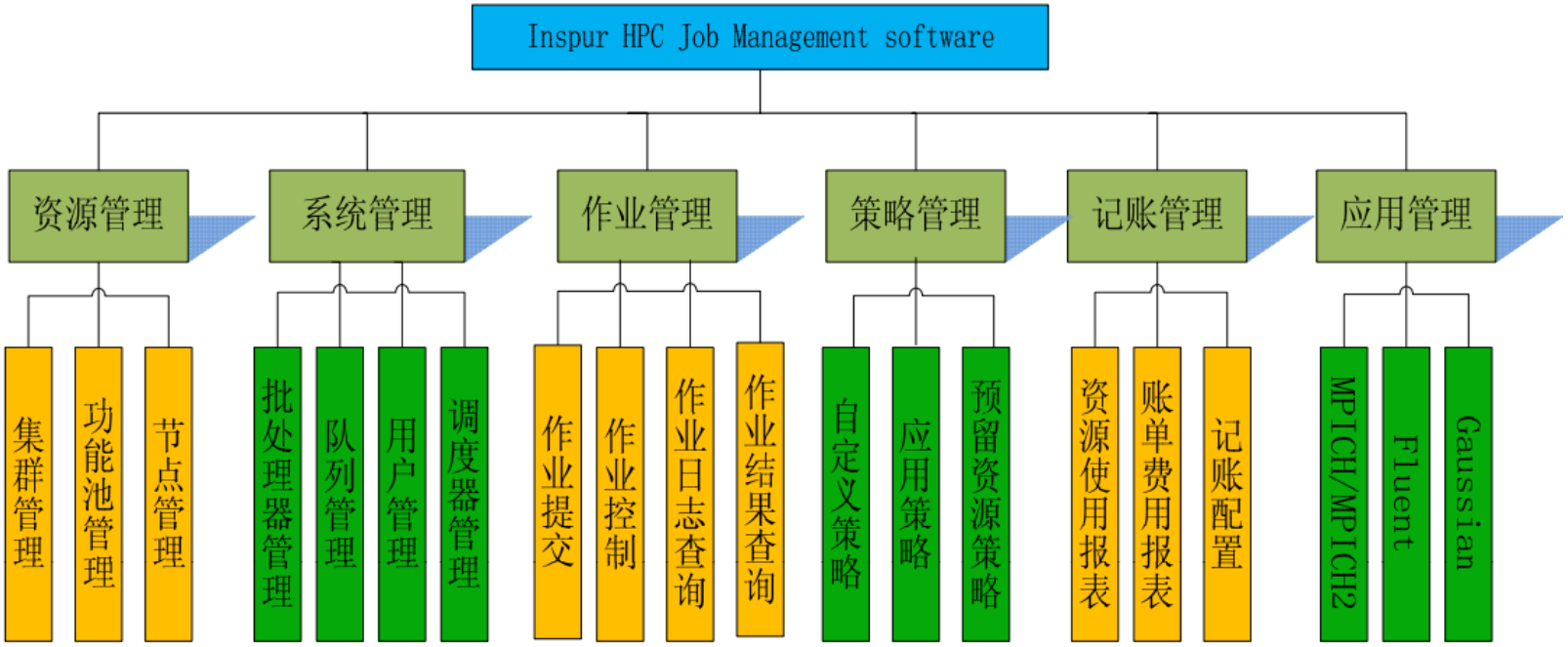
应用介绍（集群与胖节点）

目前集群底层采用redhat 企业版 64位as6.4操作系统，安装应用软件：

intel C、 fortran、 mkl、 mpi

ABINIT g09 wien2K yambo vasp espresso siesta

HPC作业管理软件



- 可以运行在单机、集群、多集群之上，接受批处理作业、shell脚本；作业运行前可以设置作业的调度策略，进行作业控制，然后运行作业，并且把输出结果给提交者。
- 包括资源管理、系统管理、作业提交、作业控制、作业日志管理、策略管理、记账和报表管理、以及应用管理



第三章 集群维护管理

inspur 浪潮

一、集群系统管理

A：SSH无密码访问的实现
B：查看硬件的物理状况
C：系统备份方法

SSH无密码访问的实现

在用户建立之后，ssh无密码访问一般都是已经实现了的，如果需要自己做一遍无密码访问，需要按照如下步骤：
（1）在用户建立之后，在用户所在的目录下，有adduser的一个目录，进入到adduser目录按照步骤执行以下程序:
以inspur用户为例：

```
[inspur@ln01 adduser]$ ./adduser2 inspur
spawn ssh-keygen -t dsa
Generating public/private dsa key pair.
Enter file in which to save the key (/home/inspur/.ssh/id_dsa):
Created directory '/home/inspur/.ssh'.
Enter passphrase (empty for no passphrase):
Enter same passphrase again:
Your identification has been saved in /home/inspur/.ssh/id_dsa.
Your public key has been saved in /home/inspur/.ssh/id_dsa.pub.
The key fingerprint is:
c5:e2:3b:be:ed:e9:40:72:46:98:b6:9e:b5:aa:c6:16 inspur@ln01
The key's randomart image is:
+--[ DSA 1024]----+
|            |
|     o .    |
|    + o o   |
|   . + o    |
|    o S     |
|   E. B o   |
|  . .o =    |
|   +  o + . |
|  o... o+=  |
+-----------------+
[inspur@ln01 adduser]$ ./adduser3 inspur 123456a?
```

等待结束后，无密码访问就基本通过了

SSH无密码访问的实现

测试账户ssh无密码访问是否可用

[inspur@ln01 adduser]$ tentakel -g allm 'date‘

使用以上命令如果可以正常输出的话，说明无密码访问是没有问题的

二、系统应用

A：集群管理软件tentakel的使用
B：机器开关机说明

A. 集群管理软件tentakl的使用

Tentakl是集群常用软件，十分方便

[root@mu01 test]# vi /etc/tentakel.conf

添加需要管理的机器名和群组

详细使用试验

例如关闭所有机器
1 先关闭计算结点
[root@mu01 ~]# tentakel -g allm “/sbin/poweroff” 使用千兆网络，一次关闭所有的刀片结点

2 关闭mu01
[root@mu01 ~]# /sbin/poweroff

inspur 浪潮

新建一个用户并提交作业流程解析

- step1
- ssh无密码访问
- 1 将所有将要使用ssh登录的机器主机明写入hostfile中，每一行一个主机名
- 2 按照循序执行下面的脚本
- [test@mu01 ssh-configure]$ pwd
- /opt/ssh/ssh-configure
- [test@mu01 ssh-configure]$ ls
- hostfile step1.exp step2.exp
- [test@mu01 ssh-configure]$ ./step1.exp test
- [test@mu01 ssh-configure]$ ./step2.exp test TEST_PASSWORD

- 说明：
- 1所有主机是可解析的（/etc/hosts文件）
- 2 两个可执行脚本实际上都是autoexpect脚本
- 3 在执行step2.exp时需要将TEST_PASSWORD替换为test密码
- 4 如两个脚本中的任何一个执行出现错误，必须按照循序，重新再执行一次这两个脚本。

新建一个用户并提交作业流程解析

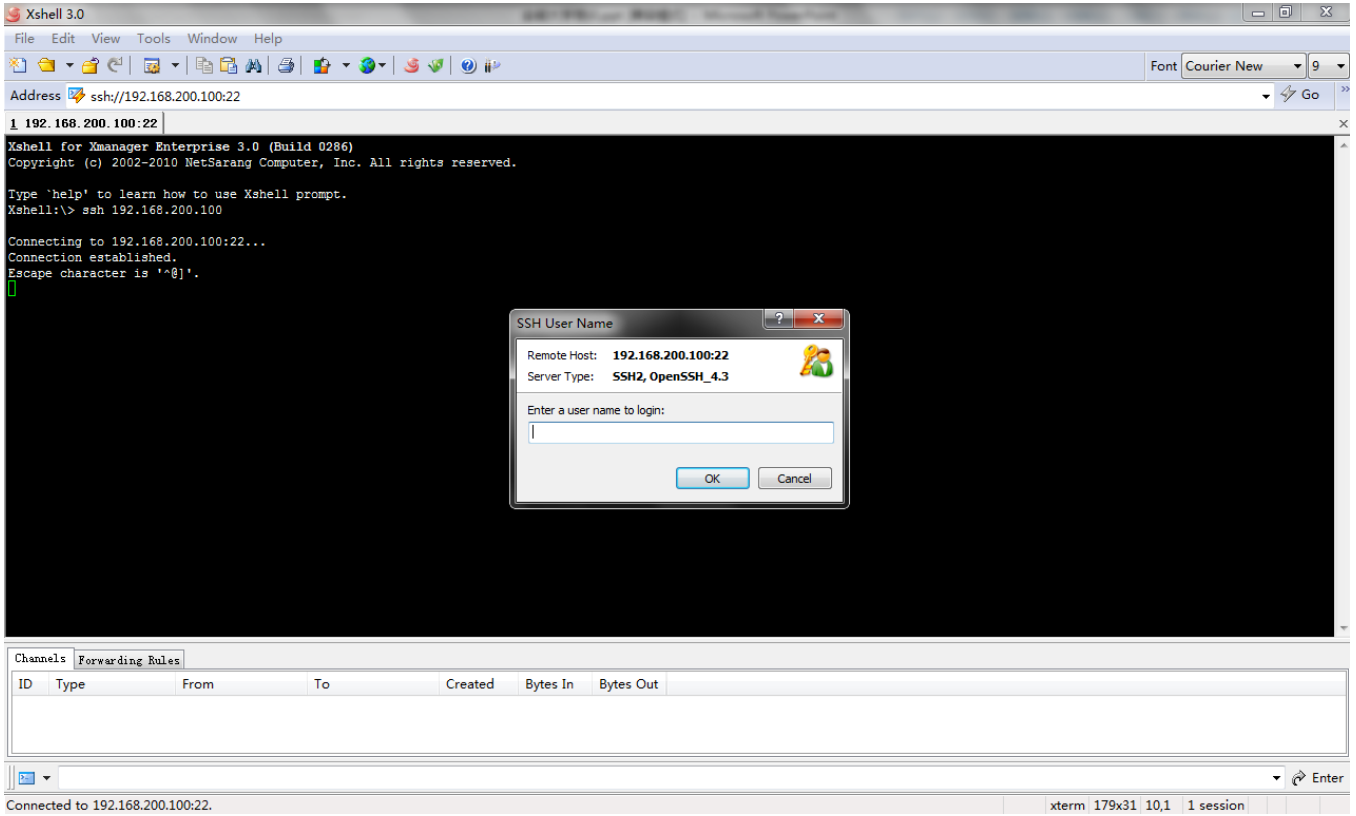
- **Step2　增加环境变量**
- 修改.bashrc
- 在NIS基础上，节点间实现ssh无密码访问后，还需要针对环境变量进行相关设定。
- 只需要在mu01节点上修改（/home/test目录是共享的），使用tentakel工具在所有节点上执行source命令即可。
- 在bashrc添加对应的环境变量,环境变量中可以指定Intel c、Intel fortran、MKL和MPI的安装路径。这儿修改后就不需要在pbs脚本中指定了。
- 以下仅是一个实例，以实际应用中安装路径为准。如果需要用intel软件包的话
- [root@mu01 ～]#vi　/home/inspur/.bashrc
- source /opt/intel/composer_xe_2011_sp1/bin/compilervars.sh intel64
- source /opt/intel/mkl/bin/intel64/mklvars_intel64.sh
- source /opt/intel/impi/4.0.3/bin64/mpivars.sh

新建一个用户并提交作业流程解析

- step3
- 测试用户
- 新用户添加后，尝试使用新用户test登录
- [root@mu01 ~]# su test
- [test@mu01 ~]# ssh cu01
- 修改密码
- [test@cu01 ~]# yppasswd
- Ctrl+d
- [test@mu01 ~]#matlab

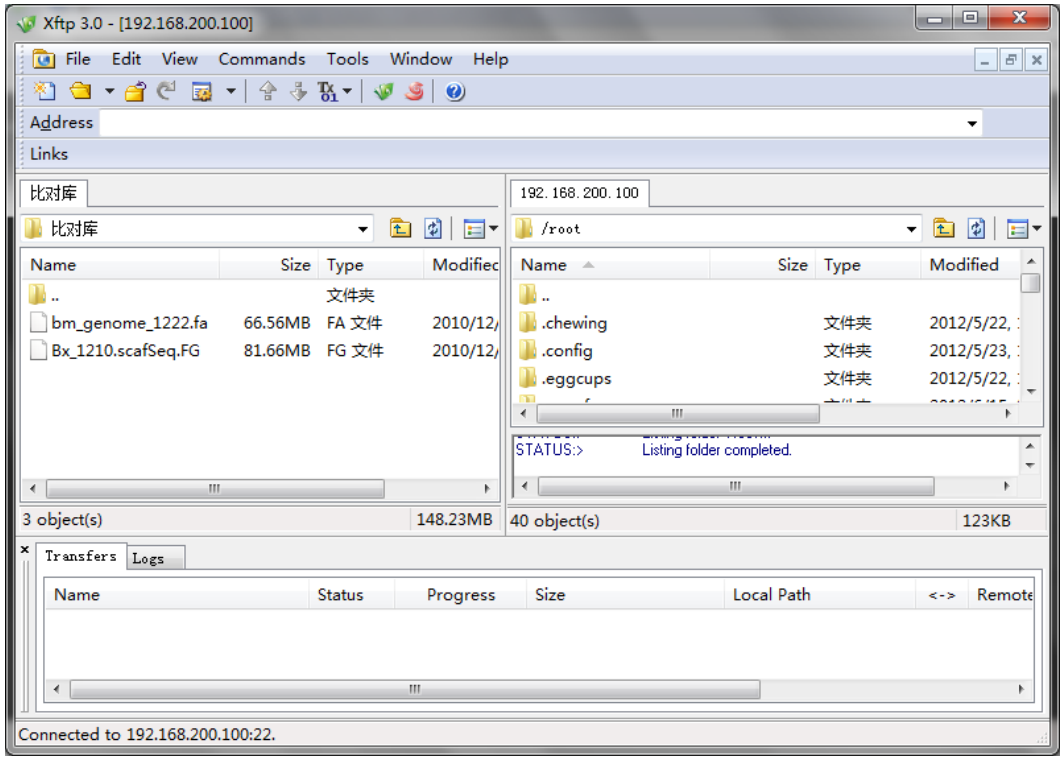
新建一个用户并提交作业流程解析

- step4
- 提交作业
- 打开ssh工具



新建一个用户并提交作业流程解析

- step5
- 打开sftp工具上传需要计算的文件



inspur 浪潮

新建一个用户并提交作业流程解析

- step6
- pbs脚本
- 需要根据软件的不同编写不同的PBS递交脚本关于怎么书写脚本，在下面的课程中会将，下面我们以已经写好的脚本为例

```
[inspur@ln01 cpi]$ cat cpi.sh
#!/bin/sh -f

#PBS -N cpi
#PBS -l nodes=2:ppn=20
#PBS -q mic

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

/opt/intel/impi/5.0.2.044/intel64/bin/mpirun -genv I_MPI_DEVICE rdma  -np $nprocs  $PBS_O_WORKDIR/cpi
```

新建一个用户并提交作业流程解析

- step7
- 提交作业
- 使用普通用户，登陆到mu01节点
- [inspur@mu01 ~]$ qsub cpi.sh
- 每个作业提交后必有一个作业号码，请牢记

新建一个用户并提交作业流程解析

- step8
- 如何查看作业号及作业状态
- 使用qstat -n
- 计算完毕后查看输出结果，以hello为例，这里的名字取决于你在脚本中设定的作业名字,后面的数字15即是你的作业号
- hello.e15 hello.o15
- 其中e代表错误输出，o代表正确输出

linux远程VNC桌面连接

redhat为我们提供远程图形桌面访问功能，需要开启vnc服务

开启vnc服务，:2为通讯的端口，可以自行设定，但不能重复

输入你需要设定的登录该图形界面的账号密码，注意输入时候屏幕无提示

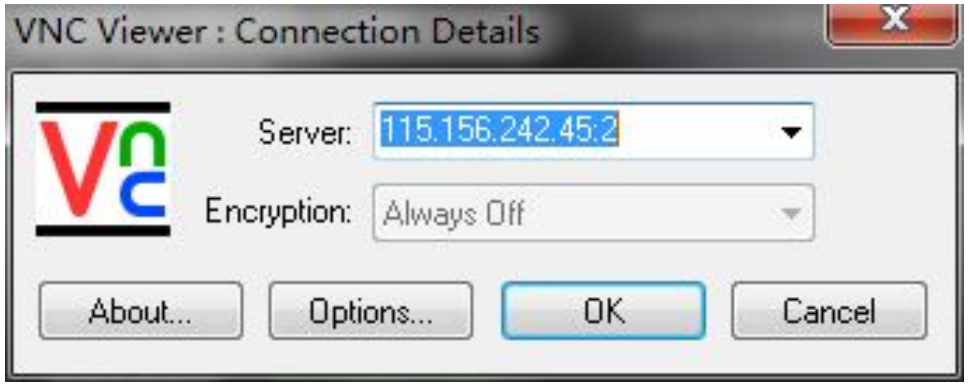
```
[inspur@mu01 ~]$ vncserver :2

You will require a password to access your desktops.

Password:
Verify:
A VNC server is already running as :2
```

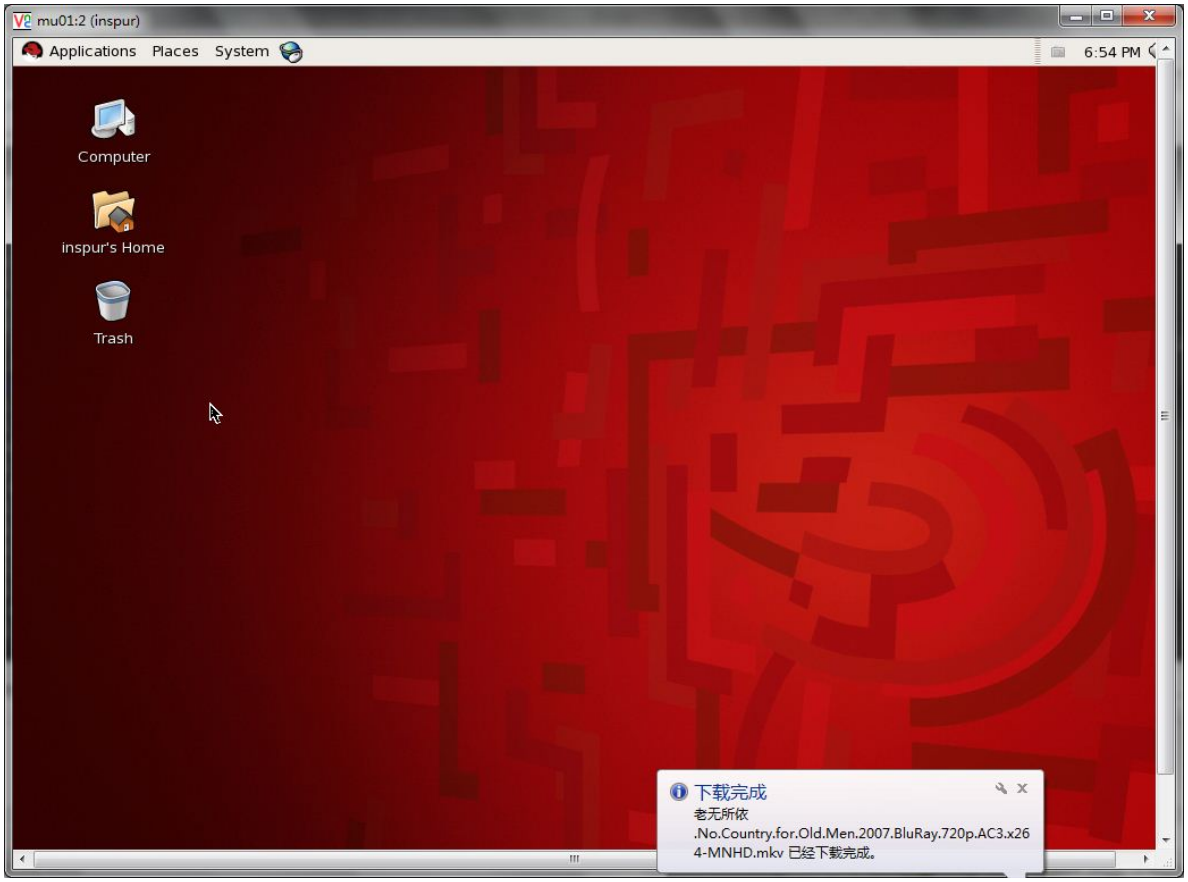
linux远程VNC桌面连接

通过vnc客户端进行连接，输入ip:端口号 端口号为前面vncserver启动时候设定的port号，输入账号，敲回车输入密码



linux远程VNC桌面连接

连接后显示，注意前期你用什么用户启动vncserver，那么你登录的就是该用户的图形环境



inspur 浪潮

linux远程VNC桌面连接

vnc修改密码：

```
[inspur@mu01 .vnc]$ ls
cu01:3.log  cu01:3.pid  mu01:2.log  mu01:2.pid  passwd  xstartup
[inspur@mu01 .vnc]$ rm -rf passwd
```

删除 /home/用户名/.vnc/passwd文件，然后重新登录时候就需要自己设定密码了。

linux远程VNC桌面连接

vnc修改密码：

```
[inspur@mu01 .vnc]$ ls
cu01:3.log  cu01:3.pid  mu01:2.log  mu01:2.pid  passwd  xstartup
[inspur@mu01 .vnc]$ rm -rf passwd
```

删除 /home/用户名/.vnc/passwd文件，然后重新登录时候就需要自己设定密码了。

vncserver退出：这里的:2 仍然是你前期指定的port号

```
[inspur@mu01 abaqus]$ vncserver -kill :2
Killing Xvnc process ID 15007
```

注意vncserver重启后需要自行手动启动。

开关机按钮

管理节点和计算节点：



机器硬件识别常用命令

- 1、CPU的识别：[test@mu01 /]$ cat /proc/cpuinfo

```
[inspur@cu01 ~]$ cat /proc/cpuinfo |grep GHz
model name      : Intel(R) Xeon(R) CPU E5-2620 0 @ 2.00GHz
model name      : Intel(R) Xeon(R) CPU E5-2620 0 @ 2.00GHz
model name      : Intel(R) Xeon(R) CPU E5-2620 0 @ 2.00GHz
model name      : Intel(R) Xeon(R) CPU E5-2620 0 @ 2.00GHz
model name      : Intel(R) Xeon(R) CPU E5-2620 0 @ 2.00GHz
model name      : Intel(R) Xeon(R) CPU E5-2620 0 @ 2.00GHz
model name      : Intel(R) Xeon(R) CPU E5-2620 0 @ 2.00GHz
model name      : Intel(R) Xeon(R) CPU E5-2620 0 @ 2.00GHz
model name      : Intel(R) Xeon(R) CPU E5-2620 0 @ 2.00GHz
model name      : Intel(R) Xeon(R) CPU E5-2620 0 @ 2.00GHz
model name      : Intel(R) Xeon(R) CPU E5-2620 0 @ 2.00GHz
model name      : Intel(R) Xeon(R) CPU E5-2620 0 @ 2.00GHz
```

- 2、内存的识别：[test@cu01 root]$ free
- 3、硬盘的识别：

  [root@mu01 /]# fdisk -l或者[root@mu01 /]# df -h

- 4、基于千兆以太网的查看

  [root@mu01 /]# ethtool eth1

集群常见故障的解决

- PBS node状态方面

  cu01　　offline

  此故障首先在mu01下用root用户执行pbsnodes -c

  cu01，执行完毕后查看状态时否为free

  cu02　　down

  此故障较多为硬件故障引起，请在mu01下ping这台机器看是否通，如不通，请重启此台机器，如果无法启动此机器，请报修

集群常见故障的解决

- 文件系统以及NFS方面

cu01-14计算节点：

```
[inspur@cu01 ~]$ cat /etc/rc.local
#!/bin/sh
#
# This script will be executed *after* all the other init scripts.
# You can put your own initialization stuff in here if you don't
# want to do the full Sys V style init stuff.

touch /var/lock/subsys/local
mount -t nfs -o vers=3 mu01:/opt /opt
mount -t nfs -o vers=3 mu01:/home /home
mkdir /dev/cpuset
mount -t cpuset cpuset /dev/cpuset
```

谢谢大家!



inspur 浪潮