基于 Linux 集群的并行环境简单架设 *

黄旭东，林 鹭
(厦门大学 数学系，福建 厦门 361005)

摘 要：并行计算在各个领域的应用越来越广泛，而基于 Linux 集群的 MPI 并行环境是一个廉价、高效的并行计算系统。介绍了两种简单的基于 Linux 集群的 MPI 并行环境的构建方法，并且提供了软件的详细配置过程。

关键词：并行计算；Linux 集群；MPI 并行环境
中图法分类号：TP393.02 **文献标识码**：A **文章编号**：1001-3695(2004)11-0254-03

Easy Construction of Parallel Environment Based on Linux Cluster

HUANG Xu-dong, LIN Lu
(Dept. of Mathematics, Xiamen University, Xiamen Fujian 361005, China)

Abstract: Nowadays parallel computing has been applied in more and more fields. Furthermore, the MPI parallel environment based on Linux Cluster is a cheap and efficient parallel computing system. Presents two kinds of easy construction methods of MPI parallel environment based on Linux Cluster, and the detailed process of software configuration is also provided.

Key words: Parallel Computing; Linux Cluster; MPI Parallel Environment

目前，集群系统已在多个领域获得应用，如传统的科学研究计算、石油勘探、天气预报、数据库、电子商务企业 IT 系统应用、生物信息处理、信号处理等。可以预见，随着对称多处理机产品的大量使用和高性能网络产品的完善，以及各种软硬件支持的增多和系统软件、应用软件的丰富，新一代高性能集群系统必将成为未来高性能计算领域的主流平台之一。简单地说，并行计算机就是用若干(几到几千)处理器并行执行一个作业，以提高计算效率；并行计算机的结构、规模和性能可以有很大的差异。以较低的投资，用若干台性能较高的 PC 机组装成集群并行计算机，采用 Linux 操作系统以及目前在各类并行机上通用的消息传递接口 MPI 并行环境(见文献[1～3])，以此为起步发展并行计算和研究，是一个合适的选择。计算机科学技术的发展在高性能计算领域为其他科学技术的发展提供了越来越广的平台；另一方面，科学技术的发展对高性能计算环境(硬、软件)不断提出更高的要求。针对特定的研究领域，在一定的财力资源下，集群并行计算机可以为数值模拟的发展提供串行计算机系统所无法比拟的高效平台。本文构建基于 Linux 的 PC 集群系统，其中 Linux 操作系统采用 RedHat 9.0；节点机由普通的 PC 机承担，通信网络使用基于 TCP/IP 协议的星型拓扑结构网络。其硬件成本比较低，比较适合在一般单位开展并行计算。

1 MPI 简介

MPI 是目前最主要的并行环境，它适用于基于分布内存的并行计算机系统的消息传递模型。它具有移植性好、功能强大、效率高等多种优点，而且有多种不同的免费、高效、实用的实现版本，几乎所有的并行计算机厂商都提供对它的支持，这是其他并行环境无法比拟的。MPI 于 1994 年产生，虽然产生的时间相对较晚，但由于它吸收了其他多种并行环境的优点，同时兼顾性能、功能、移植性等特点，在短短的几年内便迅速普及，成为消息传递并行环境的标准，其标准已由原来的 MPI-1 发展到目前的 MPI-2。

MPI-1 标准规定了如下规范：(1) MPI 是一个库，它可以被 Fortran 77 和 C 调用，从语法上说，它遵守所有对库函数/过程的调用规则，与一般的函数/过程没什么区别，从而保证遵循这些标准的 MPI 程序可以在任何平台上的可移植性；(2) 具体的 MPI 库实现由硬件供应商提供，从而开发出适合各供应商硬件的最优版本。MPI-2 规范对 MPI-1 进行了如下的扩展：动态进程；单边通信；非阻塞群集通信模式和通信之间群集通信模式；对可扩展的 I/O 的支持。MPI-1 只定义了对 Fortran 77 和 C 语言的绑定，MPI-2 将语言的绑定扩展到 Fortran 90 和 C++；对实时处理的支持；扩展了 MPI-1 的外部接口，以便使环境工具的开发更易于访问 MPI 对象。

本文将采用 MPI 的一个成熟和广泛使用的版本 MPICH-1.2.4 来构造 MPI 并行环境。软件包 MPICH-1.2.4 可以通过网上免费下载。

2 在单机上构建MPI 并行环境

由于单机 MPI 并行环境可以给 MPI 程序设计人员提供一个简单的 MPI 并行程序设计平台，所以那些有兴趣研究并行计算的老师和学生能够自行建立一个方便、廉价的操作环境；又因为 MPI 程序的可移植性，所以在单机 MPI 并行环境成功运行的 MPI 程序同样可以在大规模并行环境中成功运行。

在 PC 机上完全安装 RedHat 9.0 操作系统并设定其主机名

收稿日期：2003-11-08；**修返日期**：2004-01-05
基金项目：国家自然科学基金资助项目(10071064，10271099)；福建省自然科学基金资助项目(F0210011)

为 cluster,其中必须安装的工具包有:GCC 包,GCC-F77 包,BLAS 包,rsh 包和 rsh-server 包。注意:安装完后要将防火墙关闭。

2.1　安装和配置 MPICH-1.2.4

(1)下载 mpich-1.2.4.tar.gz 版本,并对下载的包进行解压缩

```
# tar xzvf mpich-1.2.4.tar.gz
```

(2)配置 MPICH

```
# cd mpich-1.2.4
# ./configure -prefix=/usr/local/mpi (其中-prefix 后面的路径表示安装 mpich 的路径)
```

(3)安装和编译 MPICH

```
# make
# make install
```

2.2　创建、配置文件

(1)编辑文件/etc/profile.d/mpich.sh 的内容,内容如下:

```
# !/bin/bash
export MANPATH=${MANPATH}:/usr/local/mpi/man
export PATH=${PATH}:/usr/local/mpi/bin
```

(2)编辑文件/etc/profile.d/mpich.csh 的内容,内容如下:

```
# !/bin/bash
if ($?MANPATH==0) then
setenv MANPATH:/usr/local/mpi/man
else
setenv MANPATH${MANPATH}:/usr/local/mpi/man
endif
setenv PATH${PATH}:/usr/local/mpi/bin
```

2.3　配置 rsh

(1)编辑文件/etc/hosts.equiv,内容如下:

```
cluster +
```

(2)开启 rsh 服务,命令如下:

```
# chkconfig --level 35 rsh on
# /etc/rc.d/init.d/xinetd start
```

(3)在 Root 用户下建立普通用户,如 hxd;并配置用户 hxd 的.rhosts 文件,即编辑文件/home/hxd/.rhosts,其内容如下:

```
cluster
```

(4)测试 rsh 的配置(注意:要以普通用户登录,如 hxd),命令如下:

```
# rsh cluster /bin/hostname
```

测试结果会显示本地主机名 cluster。

(5)测试 MPICH,给出 Fortran 程序测试实例如下,而 C 程序的测试类似。

```
# cp /usr/local/mpi/examples/pi3.f /home
# cd/home
# mpicc -o pi3 pi3.c
# mpirun -np 1 pi3
# mpirun -np 2 pi3
```

以上配置要注意两个要点:①为普通用户建立.rhosts 文件;②在配置单机 MPI 并行环境时要先关闭防火墙。

3　小型基于Linux 集群的MPI 并行环境的架设

假设有四台 PC 机,它们通过网线连成星型拓扑结构,其中有一台 PC 机充当主节点机且其主机名为 server,其 IP 设为 192.168.0.1;其余三台 PC 机为从节点机,它们主机名分别为 node1,node2 和 node3,且它们的 IP 分别设为 192.168.0.2,192.168.0.3 和 192.168.0.4。

在四台 PC 机上按照第 2 节中所述安装好 Linux 系统和必需的工具包(注意要将每台 PC 机的防火墙关闭)。为四台 PC 机配置相同的/etc/hosts 文件,编辑该文件且内容如下:

```
127.0.0.1       localhost.localdomain   localhost
192.168.0.1     server                  server
192.168.0.2     node1                   node1
192.168.0.3     node2                   node2
192.168.0.4     node3                   node3
```

3.1　配置 NFS

(1)四台 PC 机上都安装并启动 Portmap 服务,命令如下:

```
# /etc/rc.d/init.d/portmap start
```

(2)在 server 机中安装 NFS 相关套件并启动 NFS 服务,命令如下:

```
# /etc/rc.d/init.d/nfs start
```

(3)在 server 机中编辑文件/etc/exports,内容如下:

```
/home 192.168.0.0/24 (rw,async,no-root-squash)
/usr/local 192.168.0.0/24 (rw,async,no-root-squash)
```

其中,目录/home 和/usr/local 是分配出去的共享目录。以后每次修改了文件/etc/exports,都需要重新启动 NFS 服务。

(4)在 server 机上查看本机的共享目录名,命令如下:

```
# exportfs -rv
```

在从节点机上查看 server 上的共享目录名可用下列命令:

```
# showmount -e server
```

(5)分别在 node1,node2 和 node3 上建立相同的目录/home 和/usr/local,并分别为它们创建相同的文件/etc/fstab,使得当 server 机启动时,它们自动挂载 server 机的 NFS。文件/etc/fstab 的内容如下:

```
server:/home        /home       nfs   auto,hard,bg,intr 0 0
server:/usr/local   /usr/local  nfs   auto,hard,bg,intr 0 0
```

3.2　配置 NIS(假设以"huaixiang"作为 NIS 域名)

(1)在 server 机配置 NIS

①确认安装了 yserv 包,ybind 包和 yp-tools 包并启动 Portmap 服务。

②设定自己的域名,命令如下:

```
# domainname huaixiang
```

并在文件/etc/sysconfig/network 中加入一行

```
NISDOMAIN=huaixiang
```

③开启 NIS 服务,命令如下:

```
# /sbin/chkconfig ypserv on
# /etc/rc.d/init.d/ypserv start
```

④用 ypinit 初始化 NIS server,命令如下:

```
# /usr/lib/yp/ypinit -m
```

之后按"Ctrl+D",然后按"y"键和回车生成 NIS 数据库。

⑤更新 map,即以 Root 身份在/var/yp 目录下执行 Make 命令,以后每次在 NIS server 机上添加或删减账户时都要做此初始化动作。

⑥编辑文件/etc/nsswitch.conf,添加如下内容:

```
passwd: files nis
shadow: files nis
group: files nis
hosts: files nis dns
```

之后要重新启动 ypserv 和 yppasswdd 服务。

⑦开启 NIS 客户程序。

```
# /sbin/chkconfig ypbind on
# /etc/rc.d/init.d/ypbind start
```

⑧验证 NIS 配置。

用 ypwhich 显示主节点机的主机名;用 ypcat passwd 显示出主节点机上的用户账号(用这些账号可在任一从节点机上登录)。

(2)在 node1,node2 和 node3 上做相同的 NIS 配置

①安装 ypbind 包和 yp-tools 包。

②设定 NIS domainname,命令如下:

```
# nisdomainname huaixiang
```

并编辑文件/etc/yp.conf,添加如下一行:

```
domain huaixiang server 192.168.0.1
```

接下来,在文件/etc/sysconfig/network 中加入下面一行:

```
NISDOMAIN = huaixiang
```

③编辑文件/etc/passwd 并在文件末添加如下内容:

```
+ : : : : : :
```

④编辑文件/etc/nsswitch.conf,设置同(1)～(6)。

⑤开启 portmap 服务,如同第 3.1 节中的(1)。

⑥开启 NIS 客户程序操作,同(1)-⑦。

⑦验证 NIS 设置,同(1)-⑧。

3.3 在 node1,node2 和 node3 上设置相同的 rsh 服务,但不需要在 server 机上设置 rsh 服务

(1)安装 rsh 包和 rsh-server 包。

(2)启动 rsh 服务,命令如下:

```
# chkconfig --level 35 rsh on
# /etc/rc.d/init.d/xinetd start
```

(3)编辑文件/etc/hosts.equiv,内容如下:

```
server +
```

(4)测试 rsh,以一个 NIS 用户登录到一个节点机上并运行命令:

```
# rsh 另一台节点机名 /bin/hostname
```

结果将会显示对方的主机名。

3.4 安装和配置 MPICH

(1)只需要在主节点机 server 上安装 MPICH,且安装过程同第 2.1 节的过程一样。

(2)编辑/usr/local/mpich/share/machines.LINUX,里面定义了有哪些主机可以执行 MPI 程序。内容如下:

```
server
node1
node2
node3
```

(3)编辑文件/etc/profile.d/mpich.sh 和/etc/profile.d/mpich.csh,内容同第 2.2 节一样,并将这两个文件拷贝到所有节点机上。

(4)测试 MPICH,其操作同第 2.4 节。注意测试时要以普通用户登录并且该普通用户要能在所有节点机上的用户目录下查询到才行。

关于上述配置也要注意两个要点:①在 NFS 和 NIS 的配置时,一定要安装并开启 Portmap 服务。②对 NFS,NIS 和 rsh 进行测试时,都要关闭防火墙。

4 总结

Linux PC 集群系统是互相连接的多个独立计算机的集合,这些计算机可以是单机或多处理器系统(PC、工作站或 SMP),每个节点都有自己的存储器、I/O 设备和 Linux 操作系统。集群对用户和应用来说是一个单一的系统,它可以为用户提供低价、高效的高性能环境和快速可靠的服务。通过本文,希望能够给从事高性能计算的单位和研究人员提供一个简单、廉价和可靠的并行平台。

参考文献:

[1] 迟学斌,张林波,等.2003 年高性能计算培训班材料[M].北京:中国科学院出版社,2003.

[2] Kai Hwang. Advanced Computer Architecture Parallelism, Scalability, Programmability[M]. 北京:机械工业出版社,1999.

[3] 李贵名,俞国扬,等.基于 Linux 的 Beowulf 集群的实现[J].计算机工程,2003,29(11):49-51.

作者简介:

黄旭东(1978-),男,福建邵武人,硕士研究生,主要研究方向为并行计算;林鹭(1964-),女,福建福州人,副教授,博士,主要研究方向为矩阵计算。

(上接第 231 页)

参考文献:

[1] Duarte E P Jr, Brawerman A, Albini L C P. An Algorithm for Distributed Hierarchical Diagnosis of Dynamic Fault and Repair Events[C]. 7th International Conference, 2000. 299-306.

[2] Budhiraja N, et al. Highly-available Services Using the Primary Backup Approach[J]. Management of Replicated Data, 1992, 47-50.

[3] 张悠慧,汪东升,郑纬民. 工作站机群系统自动重构机制. 电子学报,2000,28(5):13-16.

[4] Slawomir Pilarski, Tiko Kameda. Checkpointing for Distributed Databases: Starting from the Basic[J]. IEEE Transactions on Parallel and Diatributed Systems, 1992, 3(5):602-610.

[5] Hou C J, Tsoi K S, Han C C. Effective and Concurrent Checkpointing and Recovery in Distributed Systems[J]. Computers and Digital Techniques, IEE Proceedings, 1997, 144(5):304-316.

[6] Alvisi L, Marzullo K. Message Logging: Pessimistic, Optimistic and Causal [C]. Proceedings of the 15th International Conference, 1995. 229-236.

[7] Prywes N, Rehmet P. Recovery of Software Design, State-machines and Specifications from Source Code[C]. Engineering of Complex Computer Systems, Proceedings of 2nd IEEE International Conference, 1996. 279-288.

[8] 刘心松. 具有分布式并行 I/O 节点的分布式并行服务器系统的性能研究[J]. 电子学报,2002,30(12):1801-1810.

作者简介:

陈建英(1970-),女,硕士研究生,研究方向为分布式并行数据库系统、系统容错;刘心松(1940-),男,教授,博士生导师,研究方向为宽带网络、分布并行处理、操作系统和数据库等;左朝树(1972-),男,博士研究生,研究方向为计算机网络与通信、分布式并行数据库系统和系统容错、数据库安全;陈小辉(1974-),男,硕士研究生,研究方向为分布式并行数据库系统、网络通信。