基于 PC 集群的 MPI 并行环境的搭建

希润高娃
内蒙古师范大学计算机与信息工程学院　内蒙古　010022

摘要：集群是充分利用计算资源的一个重要概念，PC 集群是最易构建的分布式并行计算环境。MPI 是应用最广的并行程序设计平台。本文通过实例阐述 PC 集群及 PC 集群上的 MPI 并行计算环境的搭建。

关键词：集群；PC 集群；MPI；并行

0　引言

高性能仍是计算机领域的重要目标。并行计算的不断发展推动着高性能计算的发展，利用集群系统的并行计算技术已被认为是未来解决大数据量问题或商业计算的导向。集群能够充分利用现有的闲置计算机资源，而且还能够使用较低配置的软硬件，来实现较高性能的计算。随着处理器技术和高性能网络技术的飞速发展，以及软件工具的改善，PC 的性能不断的提高，价格越来越廉价，性价比已远远优于工作站，而 PC 集群也成为了高性价比的并行计算资源。所以基于 PC 的并行计算环境及其上的并行算法的研究是很有价值和意义的。

MPI(Message Passing Interface)是基于消息传递编写并行程序的一种用户界面，是一个并行函数库的标准，是可移植的标准平台。MPI 具有内定和用户自定义的数据汇集操作功能，可以把大量数据进行打包后传输；能有效地管理消息、缓存区，具有结构化缓存，扩充数据类型及异构性，能实现完全异步的点对点通信，发送与接收操作可以完全与计算操作交叉进行。MP 移植性好、功能强大、效率高、硬件平台支持度高等多种优点，使它成为了研究和应用的重要并行环境。

1　集群

1.1　集群系统结构

集群是一组协同工作的服务实体，用以提供比单一服务实体更具扩展性和可用性的服务平台，属于分布式存储的并行计算机结构，典型的集群系统结构如图 1 所示。

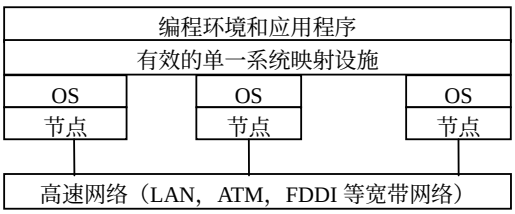


图 1　集群系统结构图

集群通过计算机高速网络将一群工作节点连接起来，充分利用各个节点的各种资源，统一调度，协调处理，以实现高效率的并行计算。其高速网络可以是普通的局域网也可以是高速网络，而通信协议一般是 TCP/IP 协议。集群的节点可以是高端微机也可以是低端微机，而且可以是单机也可以是多处理器系统。每个节点都有独立的存储器、I/O 设备和操作系统，各节点除了可以作为单一的计算资源供交互式用户使用外，还可以协同工作表现为单一、集中的计算资源供并行计算任务使用。

1.2　集群系统分类

根据不同的标准，集群系统可有多种分类方式。其中针对集群系统的使用目的可分为三类。

(l) 高性能计算集群(High Performance Compute clustering)：以高性价比替代用于并行计算的传统超级计算机。使用高速商用互联网络；可以包含数十个至上万个 PC 机或工作站；可以提供接近甚至超过传统超级计算机的计算能力。但是价格只是具有同等计算能力的超级计算机的几十分之一。

作者简介：希润高娃(1981-)，女，硕士研究生，研究方向：计算机系统结构。

(2) 高可用性集群(High-availability clustering)：掩盖硬件以及软件的易错性，对外提供不间断的服务。高可用性集群致力于使服务器系统的运行速度和响应速度尽可能快。

(3) 负载均衡集群(Load-balance clustering)：在多个节点间分配网络或计算负载。

这三种类型的集群通常会发生混合与重叠。本文主要围绕高性能计算集群进行介绍。

2 PC 集群的建立与 MPI 环境的安装

2.1 硬件选择与安装

硬件是集群系统的物理基础，硬件的安装比软件安装较为标准化。如何选择适合于搭建集群系统的硬件，主要取决于对集群的功能、性能以及价格等因素的要求。

单个节点构建，包括 CPU、内存、总线和 I/O 系统等；节点连接包括网卡、交换机等。

PC 节点跟普通 PC 并没有很大的区别，只是性能上可能有所差别，如 CPU 的主频较高，内存容量较大，I/O 系统速度较快等。

节点连接时，连接各节点的通信子系统是整个集群系统的性能瓶颈。集群系统中使用的网络技术有很多，包括以太网、光纤分布式数据接口 FDDI(Fiber Distributed Data Interface)、异步传输模式 ATM(Asynchronous Transfer Model)、Myrinet、高性能并行接口 HiPPI(High Performance Parallel Interface)以及可扩展一致性接口 SCI(Scalable Coherent Interface)等，但由于高速以太网的普及，采用以太网会带来较高的性价比。以太网系统包括网线、网卡及交换机等。选择网卡主要考虑带宽、接口类型、总线类型等。交换机的技术参数较多，主要考虑机架插槽数和扩展槽数、最大可堆叠数、背板吞吐量、缓冲区大小和 MAC 地址表大小等。

本文中的实例是包含四个 PC 节点的小型集群，每个节点的配置有，CPU：Pentium Ⅳ 2.0G、内存：2G 、硬盘：160G IDE、主板：ASUS P4T Socket 423 网卡：100Mbps Ethernet，在一个网段的局域网中，四台 PC 机可以互相 Ping 的通，通过 DHCP 获得 IP 地址，互相可以 ssh；节点连接没有使用交换设备。

2.2 软件的选择与安装

2.2.1 软件选择

目前，流行的操作系统大致分为如下几类：Windows 系列、各种 UNIX、Linux 以及 Mac OS。对构建高性能计算集群，Linux 是最佳的选择。本文中所创建的集群中使用的是 Red Hat Enterprise Linux-5U4。

Linux 是目前兼容硬件平台及各种文件系统最为完备的操作系统，良好的软硬件兼容性保证了 Linux 系统能够使用于异构集群，同时开放源码的发展模式也保证了现有集群软件方便地迁移至 Linux 下运行。Linux 系统对 UNIX 系统的兼容使得目前基于 UNIX 的集群系统可以在很小的修改下就能够成功运行于 Linux。

MPICH 是最重要的一种 MPI 实现，是一个与 MPI 规范同步发展的版本。本文将采用 MPI 的一个成熟和广泛使用的版本 mpich2-1.3.2p1 来构造 MPI 并行环境。

2.2.2 软件安装

四个节点的操作相同，下面以节点 1 的操作为例进行说明。

(1) 配置 IP 地址及网络环境

如表 1 所示，设置 IP 地址与主机名。

表 1 节点主机名与 IP 地址分配表

Pc	主机名	IP
节点 1	Node1	192.168.0.164
节点 2	Node2	192.168.0.145
节点 3	Node3	192.168.0.151
节点 4	Node4	192.168.0.118

① 通过编辑/etc/sysconfig/network 文件永久修改主机名。编辑内容为：

NETWORKING=yes

NETWORKING_IPV6=no

HOSTNAME=node1

② 在其中一台机器上修改/etc/hosts 文件，将每个机器的 IP 地址与主机名关联起来，然后把该 hosts 文件拷贝到其他三台机器的相应目录下。hosts 文件更改如下：

127.0.0.1 localhost.localdomain localhost

192.168.0.164 node1

192.168.0.145 node2

192.168.0.151 node3

192.168.0.118 node4。

(2) 创建 SSH 信任连接

① 在一台机器上生成 SSH 密钥对，命令为：#ssh-keygen -t dsa。

② 进入 ssh 目录，生成 authorized_keys 文件，命令为：

#mv id_rsa.pub authorized_keys ；authorized_keys 文件用于保存所有允许以当前用户身份登录的 ssh 客户端用户的公钥内容。

③ 建立本身的信任连接，命令为：#ssh node1，按提示输入 YES。

④ 把生成的 authorized_keys 文件拷贝到其它机器的 ssh 目录下。

⑤ 对每个机器执行 ssh 命令来确认每台机器的信任连接已建立。

(3) 安装 MPICH2

① 从 MPICH 官方网站下载源代码包，解压缩，创建安装目录，进入 mpich2 解压目录，编译并且设置安装目录 /usr/loca/mpi，进行安装。

② 进入 mpd 目录，安装进程管理器，最新版的 MPICH2 的默认进程管理器为 hydra，本文中使用的是 mpd。

```
#cd src/pm/mpd
#./configure –prefix=/usr/local/mpi
#make && make install
```

③ 通过编辑.bashrc 文件修改环境变量，修改后的.bashrc 文件如下：

```
# User specific aliases and functions
   alias rm='rm -i'
   alias cp='cp -i'
   alias mv='mv -i'
   PATH="$PATH:/usr/local/mpi/bin"
#Source global definitions
   if [ -f /etc/bashrc ]; then
    . /etc/bashrc
   fi
```

④ 修改/usr/local/mpi/etc/mpd.conf 文件，设置密码，内容为：secretword=myword；设置文件读取权限和修改时间。

```
#touch /usr/local/mpi/etc/mpd.conf
#chmod 600 /usr/local/mpi/etc/mpd.conf
```

⑤ 最后创建主机名称集合文件 etc/mpd.hosts，文件内容为四个主机名。

(4) 查错

启动单节点和集群，上面运行测试程序，运行正常的话，利用 mpdcheck，对每个节点和每两个节点进行查错。

3 测试

在建立的集群上运行了 MPICH 的圆周率程序和分块矩阵乘法程序，运行结果如表 2 所示。

表 2 测试结果对照表

程序	问题规模	单机单进程运行时间(S)	集群四进程并发运行时间(S)
圆周率	10000	0.010941	0.022534
圆周率	100000000	73.898167	18.460739
矩阵乘	10*10	0.00813	0.02123
矩阵乘	300*300	0.25022	0.20228

4 结论

利用集群系统的研究与应用已成为热点，它具有高性价比、高可扩展性、灵活的系统结构等其它并行系统不可代替的优势。而在 PC 普及的今天，PC 集群具有很多的优点及很大的发展空间。本文的实例表明，基于 PC 集群的 MPI 并行环境是易构建的，对于大量数据计算，性能更佳。由于参与运算的机器需要通过网络传递一些消息，如果计算量不大的话，通信开销影响会比较明显，反不如单机来得快。

参考文献

[1]徐巍,李玉榕.MPICH 在集群系统中的运用.福建工程学院学报.

[2]镇方雄.PC 集群的建立和 MPI 并行环境的实现与编程.咸宁学院学报.

[3]邹凯.Linux Cluster and Storage.

[4]都志辉,李三立,陈渝等.高性能计算之并行编程技术-MPI 并行程序设计.清华大学出版社.

The Building of MPI parallel environment on PC Cluster

Xi Rungaowa
Computer Information and Engineering College,Inner Mongolia Normal University,Inner Mongolia Autonomous Region,010022,China

Abstract: Cluster is an important concept to make full use of computing resources.PC cluster is the most easy to construct distributed parallel computing environment.MPI is one of the most widely used parallel programming platform.This paper presents building of PC cluster and the MPI parallel computing environment on PC cluster, through the examples.

Keywords:cluster; PC cluster; MPI; parallel