



集群系统的构建

— 用户培训

寇大治

上海超级计算中心

<http://www.ssc.net.cn>





- 集群系统概述
- 硬件的选择与安装
- 软件的选择与安装
- 集群系统性能评测



集群系统概述

- 集群：非单一系统镜像的松耦合体系
 - 系统由若干节点构成，
 - 所有节点通过高速网络互联，
 - 作业通过消息传递通信方式分布到各个计算节点上。



集群的优点

- 计算机硬件的发展（CPU、MEM、主板）；
- 计算机网络的发展（互联技术）；
- 计算机体系结构的发展（NUMA）；
- 并行计算的发展（MPI）；
- 价格便宜并且易于构建；
- 易于扩展和升级。



重要特征

- 集群的各节点都是一个完整的系统（工作站、PC机、SMP机器、PS3等等）；
 - 硬件
 - 软件
- 互联网络通常使用商品化网络（以太网、FDDI、Myrinet、InfiniBand）；
- 网络接口与节点的I/O总线松耦合相连；



集群的分类

- 根据不同的标准，可有多种分类方式
- 针对集群系统的使用目的可将其分为三类：
 - 高性能计算集群
 - 负载均衡集群
 - 高可用性集群



典型集群系统

- Berkeley NOW
- Beowulf
 - COTS (Commodity off the shelf)
- LAMP
 - SMP—cluster



- 集群系统概述
- **硬件的选择与安装**
- 软件的选择与安装
- 集群系统性能评测



硬件的选择与安装

- 节点构建
 - 处理器、内存、总线、磁盘与I/O
- 系统构建
 - 网络适配器、交换机
- 集群系统示例
 - 曙光4000A
 - 曙光5000A（魔方）



集群系统示例

● 曙光4000A:

- CPU: Opteron 850 2.4GHz (单核心CPU)
- 内存: 16GB (4X4GB)
- 硬盘: 37GB SCSI
- 主板: 泰安, 四路
- 网络: 百兆、千兆、MYRINET



集群系统示例

- 曙光5000A（魔方）：
 - CPU: Opteron 8347 1.9GHz（四核心CPU）
 - 内存: 64GB（4X4X4GB）
 - 硬盘: 137GB SAS
 - 主板: 泰安，四路
 - 网络: 千兆、InfiniBand



- 集群系统概述
- 硬件的选择与安装
- **软件的选择与安装**
- 集群系统性能评测



软件的选择与安装

- 操作系统
- 调度系统
- 并行环境



操作系统

- Windows
- UNIX
- Linux



单一系统映像

- 单一系统映像SSI (Single System Image)
 - **单一系统:** 尽管系统中有多处理器, 用户仍然把整个集群视为一个单一的系统来使用
 - **单一控制:** 逻辑上, 最终用户或系统用户使用的服务都来自只有唯一接口的同一个地方
 - **对称性:** 用户可以从任一个节点上获得集群服务, 也就是说, 对于所有节点和所有用户, 除了那些对一般访问权限作保护的服务和功能外, 所有集群服务和功能都是对称的
 - **位置透明:** 用户不用了解真正执行服务的物理设备的位置



单一系统映像

- 单一登录 (Single Sign On)
- 单一文件管理 (Single File Hierarchy)
- 单一作业管理系统 (Single Job Management System)



单一登录

● 两种方法

- 网络信息系统NIS (Network Information System)
- Shell脚本 (Shell Script)



网络信息系统NIS

● 服务器端:

① 设置NIS域名:

在文件/etc/sysconfig/network中添加如下一行:

```
NISDOMAIN=Cluster
```

② 初始化数据库:

运行/usr/lib/yp目录下的ypinit命令初始化系统信息数据库:

```
[node0] # /usr/lib/ypinit -m
```

③ 启动守护进程:

在/etc/rc.d/rc3.d目录下增加两个链接,以便系统在启动时自动载入守护进程:

```
[node0] # cd /etc/rc.d/rc3.d
```

```
[node0] # ln -s ../init.d/ypserv S61ypserv
```

```
[node0] # ln -s ../init.d/yppasswdd S61yppasswd
```



网络信息系统NIS

● 客户端:

在客户节点上配置NIS，填入域名Cluster以及服务器名node0；并在文件/etc/passwd中加入以下一行：

```
+ : 0 : 0 : : :
```



Shell脚本

- **NIS**: 设置方便, 但是采用客户 / 服务器模式, 在应用启动的时有可能产生瓶颈。
- 为了提高性能, 我们可以采用**Shell脚本**的方式来完成同样的工作。
- 例如, 对用户信息的管理, 可以创建一个名为 **cluster_user** 的脚本, 负责在其所在结点上创建新的用户, 而后再将相关的配置文件 (**/etc/passwd**、**/etc/groups**) 复制到集群系统中所有其它节点, 这里要注意的一点是用户的主目录应该建立在全局共享的分区中 (**NFS**共享)。对其它的一些信息, 比如 **/etc/hosts**和**/etc/host.equiv**等文件可同样处理。



单一文件管理

- **NFS**是一种**Unix/Linux**之间通过网络共享文件的标准方式。使用**NFS**，就能够透明的安装和访问网络上远程主机的文件系统，将其安装（**mount**）在本地的文件系统中，类似于**Windows**下的映射驱动器。
- 示例



单一文件管理

服务器端：

(1)启动服务进程：**NFS**服务器需要使用守护进程，通过在目录 `/etc/rc.d/rc3.d` 之下增加链接可以使系统在启动时自动载入这两个进程：

```
[node0] # cd /etc/rc.d/rc3.d
```

```
[node0] # ln -s ../init.d/nfs S60nfs
```



单一文件管理

(2) 设置共享目录：首先，在根目录下建立目录/home。

```
[node0] # mkdir home
```

然后，在文件/etc/exports当中增加以下几行。

```
/home    node1 (rw)
```

```
.....
```

```
/home    node.. (rw)
```

这几行的意思是将服务器上的/home目录进行共享，设置节点可以访问，**rw**表示允许读和写（缺省为只读）。



单一文件管理

客户端:

在文件/etc/fstab当中加入:

```
node0:/home          /home          nfs
```

这样就完成了NFS在集群系统中的设置，以后所有用户的主目录都可以设置在/home中。



单一作业管理系统

- 用户可以透明地从任一节点提交一项作业，作业可以调度为以批处理、交互或并行的模式运行
- 用户服务器、任务调度器、资源管理器
- 典型的作业管理系统
 - SGE (Sun Grid Engine)
 - Torque/PBS (Portable Batch System)
 - LSF (Load Sharing Facility)



并行环境

- MPI
- OpenMP



MPICH的安装

①MPICH是一个开放源码的软件，所以可以从网上免费获取它的源代码。用户可以直接从MPICH的主页下载最新的软件包mpich.tar.gz。

②使用如下命令解压缩源代码：

```
[node0] # tar -zxvf mpich.tar.gz
```

解压缩后会得到mpich的目录。



MPICH的安装

③进入该目录，并执行位于该目录下**configure**脚本，为下一步编译源代码进行准备。该配置脚本可以接受很多的参数（Options），通过运行命令如下：

```
[node0] # ./configure -help
```

可以获取更详细的参数信息。这里只列举几个最常用的参数：

-prefix: 指定mpich的安装目录。

--with-device: 指明所使用的通信系统类型。例如使用**ch_p4**表示通常的TCP/IP通信系统。

--with-arch: 指明所使用的操作系统的类型。

运行如下命令完成前期配置：

```
[node0] # ./configure --prefix=/home/MPICH \  
--with-device=ch_p4 \  
--with-arch=LINUX
```



MPICH的安装

④最后运行如下命令完成MPICH的编译和安装:

```
[node0] # make
```

```
[node0] # make install
```



MPICH的配置

- ①第一步要进行rsh的配置，使系统中不同节点之间的rsh操作不需要密码的输入。这首先需要在每个节点的/etc/hosts文件中写入计算节点(主机名,IP地址)；然后在/etc/host.equiv中写入所有的集群节点主机名，下面是node0上的这两个文件内容：

```
[node0] # cat /etc/hosts
192.168.0.10          node0
192.168.0.11          node1
192.168.0.12          node2
.....
[node0] # cat /etc/host.equiv
node0
node1
node2
.....
```



MPICH的配置

②第二步要更改MPICH的节点列表文件，该文件位于 `/PATH/TO/MPICH/share` 目录下。这里我们使用的 `arch` 参数是 `LINUX`，相应的列表文件为 `machines.LINUX`。在这个文件中要写明集群系统中所有的节点的主机名。该文件以一定的方式指明了实际执行 `MPI` 程序时进程是如何分配到各个节点上的。

以下是示例集群中的 `machines` 文件：

```
[node0] # cat /home/MPICH/share/machines.LINUX
node0
node1
node2
.....
```



MPICH的运行

- MPICH中最常用的两个命令就是mpicc和mpirun。
- mpicc是一个MPI编译器，它负责将源程序编译为可执行文件，它最常用的参数是-o用来指明输出文件。

```
[node0] # cd /home/MPICH/examples
```

```
[node0] # ../bin/mpicc cpi.c -o cpinew
```

- mpirun则是用来执行一个编译好的MPI程序。下面是它最常用的一些参数：

-np <np>：用来指明所要生成的进程数。

-machinefile <machinefile name>：缺省时使用的machines文件是前面介绍过的位于share目录下的machines.LINUX；但通过这个参数可以指定一个临时的machines文件，从而使用不同的进程指派方式。

一个标准的mpirun命令如下：

```
[node0] # ../bin/mpirun -np 10 cpinew -machinefile  
./new_machine_file
```



MPICH的运行

```
[node0] # ./mpirun -np 10 cpi
Process 1 on node1
Process 4 on node4
Process 3 on node3
Process 7 on node7
Process 8 on node8
Process 9 on node9
Process 2 on node2
Process 6 on node6
Process 5 on node5
Process 0 on node0
pi is approximately 3.1416009869231249, Error is
0.0000083333333318
wall clock time = 0.015806
```



- 集群系统概述
- 硬件的选择与安装
- 软件的选择与安装
- **集群系统性能评测**



集群系统性能评测

- 基准测试程序（Benchmark）
 - LINPACK、LAPACK、BLAS、BLACS、Livermore Loops、Dhrystone、Whetstone、NAS、SPEC、Sim
 - LinPACK: Top500的标准测试程序



附：曙光5000A（魔方）

- TOP500.org组织发布的第32次全球超级计算机五百强榜单中排名第10，WINDOWS HPC 2008
- 曙光公司联合微软公司在曙光5000A上安装部署了Windows HPC Server 2008进行Linpack测试



附：曙光5000A（魔方）

- SUSE Linux Enterprise Server (x86_64) VERSION 10 PATCHLEVEL 2 kernel: 2.6.16.60-0.31-smp



附：曙光5000A（魔方）

- 曙光5000A所有节点均采用Quad-Core AMD Opteron™-2347代号Barcelona的处理器，每一个节点都是四路四核心的平台，单核心频率1.9GHz，L2 2MB，L3 2MB，内存采用DDRII667 4GB单根，每节点共16根，总计64GB内存，即平均每个处理器核心4GB内存。
- 10台 Voltaire 288口 Infiniband 交换机，交换机内最小延迟1.8us，跨交换机最小延迟2.5us





附：曙光5000A（魔方）

- 每**10**个节点以刀片的形式组成一个机箱（目测箱高约**7U**），每**5**个机箱组成一个机柜（目测柜高约**45U**），一个机柜内**5**个机箱的下部有约**7U**高的水冷模块。
- 每个核心的理论峰值计算能力为 **$1.9G*4=7.6G$** ，即一箱的峰值计算能力为**1216G**，即每个机柜的峰值计算能力约为**6T（6080G）**
- 不带水冷系统功耗**700**千瓦，含水冷系统的功耗为**1兆瓦**



Thank You

寇大治

上海超级计算中心

<http://www.ssc.net.cn>