ISSN 1009-3044
Computer Knowledge and Technology 电脑知识与技术
Vol.18, No.32, November 2022
E-mail:wltx@dnzs.net.cn
http://www.dnzs.net.cn
Tel:+86-551-65690963 65690964

# 基于Rocks cluster的集群高性能计算系统的搭建

杨寅

(上海工程技术大学,上海 201600)

**摘要**:高性能计算系统在国民经济的许多领域有非常重要的应用。全世界高性能计算机500TOP甚至是一个国家综合国力的体现。基于Linux的Rocks cluster系统提供了一种可行的高性能集群计算机的搭建方案。介绍了Rocks cluster的特点,提供了Rocks cluster服务器的详细部署方法,并给出了集群管理和监控的方案。旨在提高集群系统的易用率,为高性能计算系统集群应用系统提供了一个易部署、高效且稳定的构建方法。

**关键词**:高性能计算系统;集群;HPC;Linux;Rocks Cluster

**中图分类号**:TP315 **文献标识码**:A

**文章编号**:1009-3044(2022)32-0068-03

DOI:10.14004/j.cnki.ckt.2022.1985

开放科学(资源服务)标识码(OSID):

## 1 概述

### 1.1 高性能计算机集群

超级计算机于20世纪60年代问世,几十年后,矢量计算机成为20世纪90年代的主导设计。拥有数万个现成处理器的大规模并行超级计算机现在已经成为常态。中国的超级计算发展也十分迅速,截至目前已部署天河、神威·太湖之光等高性能计算机。

高性能计算集群(High Performance Computing Cluster),通常是指使用多个处理器或以几台计算机组成的计算系统,用于各种领域的大型计算任务,包括量子力学、气候研究、分子建模和密码分析。可以由多台超级计算机组成,也可以通过由网络连接的PC机组成的集群。由于在某些廉价而通用的计算平台上运行并行计算集群可以提供极佳的性能价格比,所以近年来这种解决方案越来越受到用户的青睐[1]。

### 1.2 高性能计算系统

现如今,大多数现代超级计算机使用Linux内核作为操作系统,但每台超算的设计人员都会对他们使用的系统做了特定更改,并且没有行业标准,主要原因是不同计算机硬件架构的差异。例如IBM Blue Gene超级计算机在计算节点上使用CNK操作系统、天河超级计算机使用Slurm,以及OpenHPC。然而。对于非专业人员来说,集群的建立,安装、管理和监控一直是一致和困扰的问题。

Rocks集群是一个基于SMP的分布式架构开发的Linux系统。作为业内一流的免费集群操作系统,Rocks集群具有低成本、易扩展、结构精简、多软件可继承等特点,向广泛的科学应用提供集群计算能力,提高研究者们的工作效率。

本文选用Rocks6.2版本以最便捷的方式搭建高性能计算集群平台,低成本建立计算机集群,为普通用户大规模计算任务提供了快捷有效的解决方案。

## 2 Rocks集群结构

Rocks Cluster在软件结构上由FRONTEND(前端节点)和NODE(计算节点)组成。一般来说,前端节点可通过局域网交换机连接到多个计算节点,前端节点可配置一个网络适配器连接到公共网络(外网),通过前端节点上的外网端口,计算节点可以同时实现对公共网络(外网)的访问。Rocks Cluster的具体结构如图1所示:

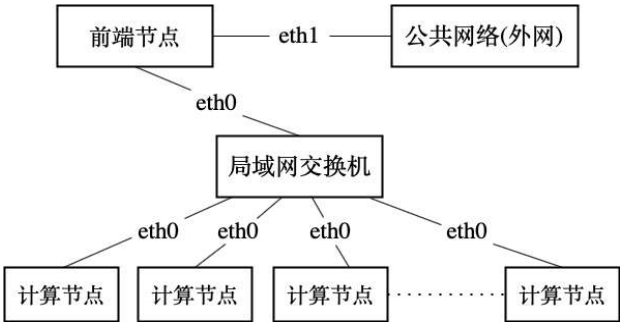


**图1 Rocks Cluster的具体结构**

另外,Rocks系统中的模块化Roll机制也是其关键组成部分之一,它可以为特定的专业领域定制Roll包。在使用GEON Roll构建集群前端节点时,所有计算节点都将安装Geo专用软件。通过这种机制,我们可以轻松地将某些专业软件安装到Rocks集群系统中。

Rocks是一份完整的光盘机群解决方案,它面向

收稿日期:2022-03-06
基金项目:上海市大学生创新创业计划(S202110856037)
作者简介:杨寅(2000—),男,贵州人,主要研究方向为大数据处理。



(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

x86及IA64的Red Hat Linux COTS机群。组建一套Rocks cluster并不需要任何机群方面的经历，实际上，机群架构师将能找到一条灵活的并且标题化的方式来重新设计整个软件栈，而这对大多数用户而言则适当地隐藏了[2]。尽管Rocks包括了在任何机群软件结构中都应指望的工具（PBS、Maui、GM支持、Ganglia等），它的安装简易性则是独一无二的[3]。

## 3 安装Rocks集群

在硬件方面，主节点为一台IBM System x3650 M4服务器，两颗CPU：2.6GHz Intel Xeon E5-2670，64GB内存，2.4TB硬盘。

系统软件方面，在Rocks cluster的官方网站http://www.rocksclusters.org下载rocks 6.2 ISO安装包，安装包内提供的工具详情如下表：

**表1　Rocks ISO安装包工具**

| 卷名 | 描述 | 功能 |
|---|---|---|
| Area51 | 系统安全相关的服务和应用程序 | 分析集群上文件和内核的完整性 |
| Base | 基础卷 | Rocks基本安装包 |
| SGE | 分布式资源管理软件 | 保证集群内的资源得到有效利用 |
| ZFS-linux | ZFS设备驱动程序 | 用于构建可靠的存储系统 |
| Fingerprint | 应用程序依赖性软件 | 分析、管理与跟踪应用程序依赖关系 |
| Ganglia | 集群监控软件 | 监控系统性能 |
| HPC | 性能峰值测试 | 并行运行的已配置软件工具的应用 |
| HTCondor | 高吞吐量计算 | 计算系统吞吐量 |
| Kernel | 核心卷 | Rocks启动引导包 |
| KVM | 系统虚拟化模块 | 让进程有专属的内核和用户模式 |
| OS | CentOS安装卷 | 安装Linux系统 |
| Perl | 实际抽取与回报语言 | 从CPAN模块创建基于Rocks的RPM |
| Java、Python | 编程语言 | 编写计算机程序 |

以上准备工作完成后，安装内容如下：

### 3.1 RAID磁盘阵列设置

Redundant Array of Independence Disks(RAID)磁盘阵列是一种数据存储虚拟化技术，它将多个物理磁盘驱动器组件按不同的方式组合起来形成一个或多个逻辑单元，以实现数据冗余、提升性能。下面是常见RAID方案的对比：

**表2　RAID方案对比**

| RAID等级 | RAID0 | RAID1 | RAID3 | RAID5 | RAID6 | RAID10 |
|---|---|---|---|---|---|---|
| 别名 | 条带 | 镜像 | 专用奇偶校验条带 | 分布奇偶校验条带 | 双重奇偶校验条带 | 镜像加条带 |
| 容错性 | 无 | 有 | 有 | 有 | 有 | 有 |
| 冗余类型 | 无 | 有 | 有 | 有 | 有 | 有 |
| 热备份选择 | 无 | 有 | 有 | 有 | 有 | 有 |
| 读性能 | 高 | 低 | 高 | 高 | 高 | 高 |
| 随机写性能 | 高 | 低 | 低 | 一般 | 低 | 一般 |
| 连续写性能 | 高 | 低 | 低 | 低 | 低 | 一般 |
| 需要磁盘数 | n≥2 | 2n(n≥1) | n≥3 | n≥3 | n≥4 | 2n≥n≥4 |
| 可用容量 | 100% | 50% | (n-1)/n | (n-1)/n | (n-2)/n | 50% |

磁盘阵列卡进行RAID1+0模式后，磁盘容量减半获得容错能力[4]。综合可靠性与性能，选用RAID 10作为系统磁盘阵列。设置步骤如下：

(1) 服务器启动后，等待进入阵列卡提示出现，按“Ctrl + H”进入RAID卡选项，点击“start”进入配置界面。

(2) 在配置界面中，进入“Configuration Wizard”选项，选择“new configuration”创建新配置，确认后系统将清除磁盘内所有数据。

(3) 在下一个界面中选择“manual configuration”进行手动分配磁盘空间，根据RAID 10的配置方式，将四个硬盘为一组添加到“Drives groups”，按此方式添加两组。

(4) 将制作好的两组硬盘放入“Span”，在“virtual disk definition”配置界面选择RAID Level为RAID 10，其他参数保持默认。更新数据“Update Size”后开始磁盘阵列创建。

(5) 保存设置并返回主配置界面，可以看到阵列设置为RAID 10，VD状态为最佳，两组共8块硬盘状态在线。

### 3.2 Rocks主节点安装

(1) 插入Kernel/Boot Roll光盘，从光盘启动服务器，在安装选项界面输入Build。

(2) 选择安装的必要功能，包括area51、kernel、base、OS、SGE、HPC、Ganglia、Java、Python等。

(3) 输入集群信息，集群名称保持默认，输入接口eth0与eth1的外网地址、私网IP、外网网关、外网DNS等。

(4) 配置系统root的密码，设置时区为Asia/Shanghai，硬盘自动分区。

(5) 系统主节点开始自动安装，安装完成后会自动重启，至此主节点安装完成。



(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

### 3.3 Rocks计算节点安装

Rocks最大的优点在于把你所需要的软件在系统安装时都一次安装好，不需要安装者更多地配置[5]。计算节点可以从主节点获取所需文件，自行安装。

(1) 以root用户登录主管理节点，打开终端，执行# insert-ethers命令，选择计算节点属性Compute。

(2) 通过交换机连接主节点与计算节点，将子节点服务器启动，BIOS设置引导顺序为：CD、PXE（网络引导）、Hard Disk。主节点识别成功后，子节点将自行安装。安装完成后，子节点自动重新启动并加入计算机群。

(3) 更改计算节点的根密码以提高Rocks系统的安全性。执行命令如下：# rocks set host sec_attr compute attr=root_pw，系统确认用户自定义密码后即设置完成。

(4) 至此Rocks集群搭建完毕，可通过命令# rocks list host查看集群主节点及计算节点状态。

### 3.4 Rocks系统网络配置

节点设置完成后，此时系统仍然无法连接互联网，不能及时进行远程调试和提交运行任务。经调试，系统还需要修改以下参数：

(1) 使用命令# ifconfig查看网络配置，再通过vim修改ONBOOT=yes，将Rocks系统的双网卡eth0和eth1网络连接设置为开机自动启用。

(2) 修改完成后使用命令# service network restart重启网络。

(3) 至此，网络配置完成。可以使用# ping IP查看网络状态。

## 4 Rocks集群管理和监控

### 4.1 集群用户管理

Rocks的用户管理方式是使用411安全信息服务来代替传统NIS(Network Information Service)服务。这个系统设计的目的是提高拓展性、安全性与意外处理能力，并且411服务与NIS服务的使用和配置方式与NIS服务几乎没有区别[6]。

### 4.2 集群任务管理

Rocks安装包内包括Sun Grid Engine(SGE)分布式资源管理软件，保证集群内的资源得到有效利用并且不需要NFS进行操作。使用它可以极大提升用户提交计算任务的效率，并更加合理分配和提高整个集群的计算资源[7]。

### 4.3 集群监控系统

Ganglia是一个开源的分布式监控系统，由美国加利福尼亚大学发起，用于集群和网格这样的高性能计算系统。Ganglia偏向于操作系统低层一点的监控，主要是收集集群中的各个机器的CPU使用情况、内存使用情况、磁盘IO、网络IO、磁盘容量等[8]。它基于针对集群整体的层次结构设计，利用诸如用于数据表示的XML、用于紧凑、可移植的数据传输的XDR以及用于数据存储和可视化的RRDtool工具。Ganglia使用精心设计的数据结构和算法来实现高效的节点调度与高并发性。

Rocks系统安装包中已包含ganglia卷，我们仅需要使用命令# rocks enable roll ganglia激活roll中的ganglia安装程序，并执行语句# cd /export/rocks/install切换到对应安装包目录进行安装。

安装完成后，可以在浏览器中进入http://localhost/ganglia生成系统实时状态显示，向用户提供直观的Web界面。

## 5 总结

本文基于Rocks cluster系统搭建一组高性能计算集群。介绍了当今高性能计算机集群与系统的现状，针对Rocks集群结构展开说明，详细描述基于Rocks构建高性能计算集群系统的过程，并且讨论服务器RAID磁盘阵列的设置过程，最后对Rocks计算节点进行配置与部署，为用户提供一个快速易用的集群搭建方案。

## 参考文献：

[1] 张澜宇,邹溪.PC实验室的高性能运算平台的实施[J].实验室科学,2016,19(5):53-56,62.

[2] Rocks Cluster Distribution: Users Guider[M]. UC Regents, 2011.

[3] 林先津.桌面虚拟化技术在分布式设备管理中的研究与应用[J].实验技术与管理,2013,30(4):68-70.

[4] 赖策,祝元仲.虚拟机基于不同存储池模式下的磁盘性能测试分析[J].轻工科技,2020,36(7):90-91.

[5] 张洋,陈文波,李廉.基于Rocks的高性能集群平台搭建与应用[J].计算机工程与科学,2008,30(11):137-139,143.

[6] 徐明,罗明宇,蔡文举,等.Rocks集群应用系统发布技术研究[J].自动化技术与应用,2015,34(3):38-41.

[7] 侯建军.数字电子技术基础[M].北京:高等教育出版社,2003.

[8] 狄晓娇.企业级Hadoop平台实现的相关技术[J].中国新通信,2016,18(4):89-90.

【通联编辑：王力】



(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net