

基于云计算平台的虚拟集群构建技术研究

曾龙海, 张博锋, 张丽华, 何 冰, 吴耿锋, 徐炜民
(上海大学 计算机工程与科学学院, 上海 200072)

摘 要: 为了灵活构建高性能计算集群并根据用户作业需求弹性的提供计算资源以及提高物理机器的利用率, 采用了基于 Eucalyptus 开源框架构建云计算平台, 利用 OpenPBS 集群管理软件构建了虚拟集群(Virtual Cluster), 并实现了对虚拟集群的管理及监控, 提高了集群资源的利用率及实现了弹性计算的理念, 通过若干实验证实了构建虚拟集群的可行性和有效性。

关键词: 云计算; 虚拟集群; 中间件; Eucalyptus; OpenPBS

中图分类号: TP391

文献标识码: A

文章编号: 1000-7180(2010)08-0031-05

Virtual Cluster Constructing Based on Cloud Computing Platform

ZENG Long-hai, ZHANG Bo-feng, ZHANG Li-hua, HE Bing, WU Geng-feng, XU Wei-min

(College of Computer Engineering and Science, Shanghai University, Shanghai 200072, China)

Abstract: In order to build the high performance computing cluster flexibility, offer computing resources according to the need of users' jobs and improve the utilization of computers, this paper introduces the cloud computing platform which built on the open-source framework of Eucalyptus, and adopts OpenPBS which is cluster management software to create virtual cluster dynamically. In dealing with computing resources elastically offered and low utilization of computers, it develops a middleware platform to manage and monitor the virtual cluster; finally it is proved to be feasible and effective by several experiments.

Key words: cloud computing; middleware; virtual cluster; Eucalyptus; OpenPBS

1 引言

云计算平台利用虚拟化技术构建的虚拟集群, 能够动态地组织异构的计算资源, 隔离具体的硬件体系结构和多样化的软件系统平台, 灵活构建满足不同应用需求的计算环境, 提高计算资源的使用效率, 实现弹性计算系统架构, 解决高性能计算^[1]的重要途径。

本文主要论述如何构建一个基于云计算平台的虚拟集群, 虚拟集群可以根据作业的不同需求灵活地创建满足作业需求的计算资源。

2 云计算平台构建

云计算^[2]是指通过网络将分散的计算资源互

联起来构建一个巨大的计算资源池转换为三个层次的服务, 分别是软件即服务(Software as a Service)、平台即服务(Platform as a Service)和设施即服务(Infrastructure as a Service)^[3-4]。云计算平台的构建主要是基于虚拟化技术平台。

2.1 虚拟化技术

虚拟化概念^[5]用来共享昂贵的大型机系统, 对计算机资源进行重新分析, 提高计算机资源的利用率。在云计算平台中虚拟化技术应用比较广泛的有 KVM 虚拟机(Kernel-based Virtual Machine)及 Xen 虚拟机^[6]。

(1) KVM 虚拟机, 是一个开源的系统虚拟化模块, 集成在 Linux 各个主要发行的版本中, 它使用 Linux 自身的调度器进行管理, KVM 是基于硬件的

完全虚拟化,需要硬件的支持(如 Intel VT 技术或者 AMD V 技术),对硬件要求比较高;

(2) Xen 虚拟机,是一个开放源代码虚拟机监视器,由剑桥大学开发. Xen 虚拟机支持硬件系统的半虚拟化及完全虚拟化,兼容性比较好,这使得 Xen 无需特殊硬件的支持就能够达到高性能的虚拟化.

由于 Xen 虚拟机对硬件的兼容性比较好,并且可以达到高性能的虚拟化,所以本文构建的云计算平台的虚拟化技术利用 Xen 虚拟机来实现. 下面主要介绍 Xen 虚拟机在 ubuntu9.04 server 系统中的配置及使用.

① 安装 Xen 内核

在 ubuntu 操作系统下输入命令(□表示空格,下同):

```
dpkg □ -i □ linux-image-2. 6. 26-2-xen-686 _ 2.
6.26-15lenny2 _ i386. deb □ linux-modules-2. 6. 26-
2-xen-686 _ 2.6.26-15lenny2 _ i386. deb
```

其中 Xen 的 image 及 modules 需事先下载.

② 安装 xen 管理工具

在 ubuntu 操作系统下输入命令:

```
apt-get □ install □ python2. 5 □ ubuntu-xen-desk-
top
```

③ 配置虚拟机参数

建立名为 with-data 的文件,其内容如下:

```
[root] □ [data] □ [swap]
size = 4G □ size = 4G □ size = 512M
type = ext3 □ type = ext3 □ type = swap
mountpoint = / □ mountpoint = /data
options = sync □ options = nodev
errors = remount-ro
```

④ 创建虚拟机

安装完成后,重新启动系统进入加载了 Xen 内核的 ubuntu 操作系统,输入命令:

```
xen-create-image □ --hostname = vm0 □ --dhcp
□ --dist = redhat □ --arch = i386 □ --partitions = with-
data □ --boot □ --passwd □ --role udev
```

这里需要事先创建 redhat 操作系统的镜像文件.

2.2 Eucalyptus 平台

Eucalyptus 是加利福尼亚大学的 Daniel Nurni 等人实现的,是一个用于实现云计算平台的开源软件基础设施,主要依赖于 KVM 或 Xen 虚拟机来实现. Eucalyptus 是 Amazon EC2 的一个开源实现,它与 EC2 的商业服务接口兼容^[7]. Eucalyptus 体系结

构主要包括节点控制器、集群控制器和云控制器.

下面主要介绍 Eucalyptus 在 Xen 虚拟机中的配置及使用.

(1) 主节点配置

在主节点上安装云控制器、集群控制器、存储服务,在 ubuntu 系统下输入命令:

```
apt-get □ install □ eucalyptus-cloud □ eucalyptus-
cc □ eucalyptus-walrus eucalyptus-sc
```

(2) 计算节点配置

在计算节点上安装节点控制器 eucalyptus-nc,在计算节点 ubuntu 操作系统下输入命令:

```
apt-get □ install □ eucalyptus-nc
```

(3) 网络配置

Eucalyptus 平台中主节点和计算节点之间的网络通信模式为桥接模式,需要对计算节点网络链接方式更改为桥接模式.

(4) 主节点和计算节点的身份认证

产生 ssh 密钥,并将公钥拷贝到计算节点的 Eucalyptus 用户上,实现无密码连接.

3 集群系统构建

将多台同构或异构的计算机连接起来协同完成特定的任务就构成了集群系统^[8],集群系统就是一种并行处理系统. 传统的物理集群构建都是基于有限台物理机器,其拓扑结构如图 1 所示. 若物理集群需要再扩展新的计算节点则必须添加新的计算机,而且集群体系结构布局繁琐.

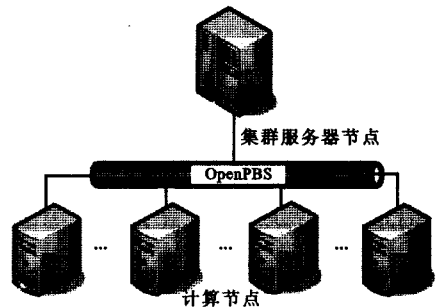


图 1 传统物理集群拓扑图

虚拟集群可以定义为,采用虚拟化技术虚拟出多台计算节点,从而构建出与物理集群相似的集群系统. 本文就采用这样的思想,基于云计算平台构建虚拟集群,实现按需计算的理念. 图 2 是本文给出的虚拟集群拓扑结构.

3.1 集群作业管理系统介绍

集群作业管理系统是集群的一个重要组成部分,目前国际上存在的集群作业管理系统种类繁多,

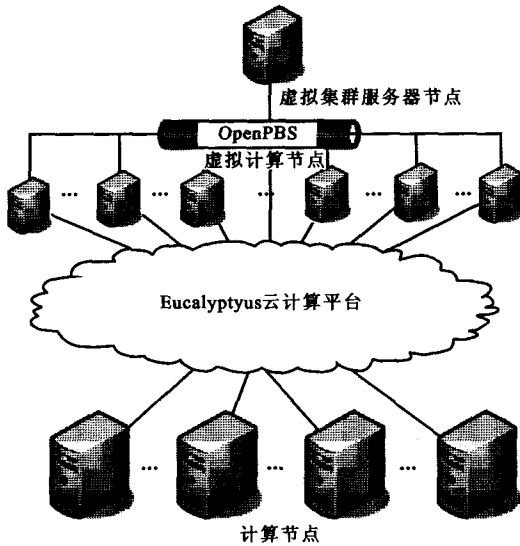


图2 虚拟集群拓扑图

其中应用比较广泛的有 PBS, CONDOR, LSF, TORQUE 等。

PBS^[9](Portable Batch System) 由于开源,易获取,并且提供完整的 API 有利于开发实现新的调度策略,支持作业的多种运行模式,成为了研究机构构建高性能计算集群应用比较广泛的作业管理系统软件。PBS 主要由三部分组成:一个资源管理器(PBS Server)、一个调度器(PBS Scheduler)以及若干个 PBS 执行器(PBS Mom),各部分功能关系如图 3 所示。

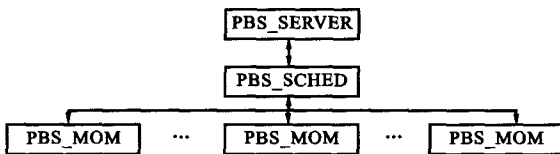


图3 PBS功能结构图

3.2 虚拟集群中间件平台开发

本文基于作业管理系统开发了一个中间件平台来对虚拟集群进行管理和监控,动态添加新的计算资源,弹性构建虚拟集群,实现用户的按需计算。中间件平台主要包括三个模块:用户作业提交模块、虚拟集群弹性构建模块、虚拟集群信息监控模块。下面分别介绍这三个模块。

3.2.1 作业提交模块

用户通过作业提交页面可以将任务递交到虚拟集群里进行运算,待作业运行完成获取运算的结果。作业提交模块流程图如图 4 所示。

主要包括:

(1) 作业文件上传、参数设置

作业文件为可执行的 C 或 C++ 程序,参数设

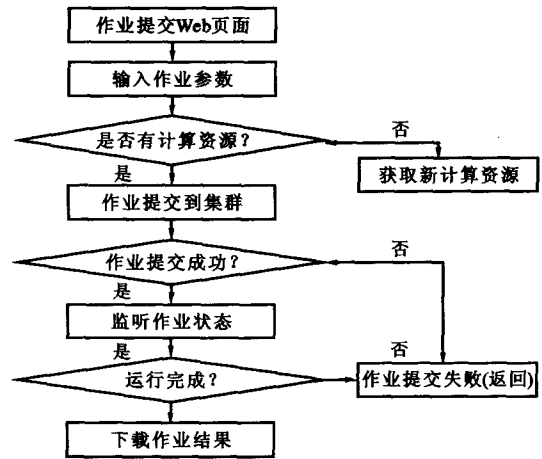


图4 作业提交流程图

置主要为所需的计算资源数目以及输出结果文件名。

(2) 作业脚本生成

主要生成 pbs 脚本信息,根据作业参数设置相应的脚本参数。

(3) 作业状态信息监控

当作业成功提交到虚拟集群,对作业信息的监控主要包括监控该作业是否等待、挂起、运行、完成。

(4) 作业运行结果获取

作业运行完成产生的作业结果可提供给用户进行下载。

作业提交模块伪代码如下:

```
Input job parameters;
If(computer_resources == true){
    submit(job);
    While(submit_job == successful){
        monitor_job_status();
        If(job_run == completed){
            If(job_result == true){
                send(jobresult, to, middleware);
            }
        } else{
            return;
        }
    }
} else{
    obtainNewComputerResources();
}
```

3.2.2 虚拟计算弹性构建模块

当虚拟集群的计算资源不能满足用户申请所需的计算资源时,中间件平台就动态的虚拟出所需的虚拟机计算资源并动态添加到虚拟集群中以满足用

户的需求.弹性构建虚拟集群流程图如图 5 所示.

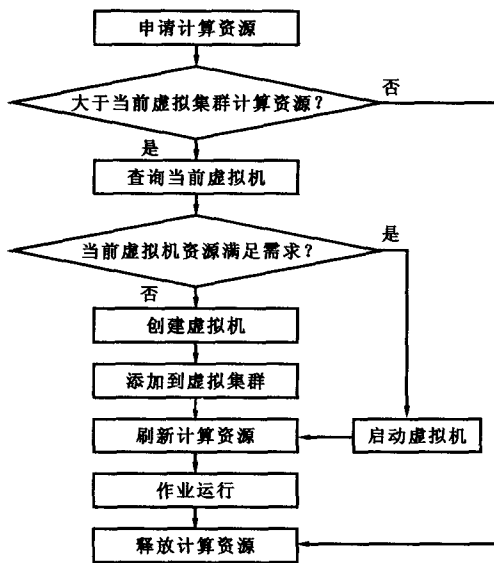


图 5 虚拟集群构建流程图

主要包括:

(1) 创建虚拟机器

基于构建的云计算平台上可以灵活的创建虚拟机器,并根据需求可以虚拟出不同的操作系统,本实验主要创建 redhat 操作系统的虚拟机.

(2) 启动虚拟机器

```
# xm console vm0
```

将新增虚拟节点添加到虚拟集群系统:

在/etc/hosts.equiv 文件下追加新增虚拟节点信息

```
cat vm0 >> /etc/hosts.equiv
```

启动新增虚拟节点执行器

```
# pbs_mom
```

(3) 刷新计算资源

查看虚拟集群空闲节点信息

```
# pbsnodes a |grep free
```

返回空闲资源数目,刷新计算资源信息.

(4) 释放计算资源

待作业运行完成,注销虚拟机器

```
# xm halt vm0
```

弹性构建虚拟集群伪代码如下:

```

If (computer_resources < job_need_computer_resources) {
    need_virtualMachine_num = job_need_computer_resources - computer_resources;
    searchCurrentVirtualMachineNum();
    While (currentVirtualMachineNum >= need_virtualMachine_num) {

```

```

        For (; currentVirtualMachineNum; currentVirtualMachineNum--) {
            startVirtualMachine();
        }
        createVirtualMachine();
        addToVirtualCluster();
        refreshComputerResource();
        jobRun();
        releaseComputerResource();
    } else {
        jobRun();
        releaseComputerResource();
    }
}

```

3.2.3 虚拟集群信息监控模块

主要完成对虚拟集群各虚拟计算节点信息的监控,当作业成功提交并运行,监控模块采用 Web 页面的方式动态的显示各虚拟节点的运行状态.

主要包括:

(1) 监听虚拟集群计算资源信息

每隔 10s 计算一次虚拟集群各节点计算资源状态信息,并以 Web 页面的形式进行计算结果显示.

(2) 监听作业申请计算资源信息

动态的监听作业提交信息,将申请的计算资源信息及时反馈到信息监控模块.

4 实验验证与结果分析

为了验证虚拟集群的弹性构建及按需计算的能力,本文在开发好的中间件平台上进行实验,实验环境如表 1 所示.

表 1 实验环境节点配置情况

机器数	CPU	核数	主频	内存
2 台	Intel Xeon 5150	四核	2.66GHz	4096Mb

4.1 集群构建灵活性实验

4.1.1 传统物理集群构建

(1) 2 台物理机器分别安装 Redhat 9.0 操作系统,计算花费时间 20 分钟(根据机器性能略有偏差).

(2) 安装 OpenPBS 2.3.6 集群管理软件,其中一台机器用作集群服务器(PBS_SERVER),另一台用作集群计算节点(PBS_MOM),构建一个具有 2 个 CPU 计算资源的集群系统平台(在/PBS 安装目录/server_priv 文件下,将 nodes 文件里计算节点信息设置为 np=2),计算花费时间 10 分钟.

4.1.2 虚拟集群构建

(1) 分别在2台物理机器安装 Ubuntu 9.04 操作系统,计算花费时间20分钟(根据机器性能略有偏差)。

(2) 安装 Xen 虚拟机(参考2.1节),构建具有两个计算节点的云计算平台(参考2.2节),其中一台机器既做云计算平台主节点又做云计算平台计算节点,计算花费时间20分钟(根据机器性能略有偏差)。

(3) 创建2台虚拟机,虚拟机配置为单核CPU,512Mb内存,Redhat9.0操作系统,同时创建两台虚拟机花费时间4分钟。

(4) 安装集群管理软件(同上),在/PBS安装目录/server_priv文件下,将nodes文件里计算节点信息设置为np=1。

可以看出,摒除云计算平台的构建时间,创建一个虚拟集群在灵活性和耗时上具有明显的优势,可以在较短的时间内同时创建多台虚拟机器。

4.2 作业运行效率对比实验

统计所有作业从递交时间到运行完成所花费的时间(如表2所示)。其中物理集群作业花费时间包括运行时间和等待时间,括号内为等待时间;虚拟集群作业花费时间包括创建虚拟机时间和运行时间,括号内为创建虚拟机时间。

表2 作业运行时间对比

作业个数	物理集群(时间/s)	虚拟集群(时间/s)
1	436.45(0.00)	440.55(0.00)
2	436.56(0.00)	441.45(0.00)
3	874.82(436.89)	681.01(232.64)
4	871.61(434.72)	808.03(348.24)
6	1 310.41(878.05)	871.54(372.16)
8	1 749.34(1 311.90)	920.40(398.28)

统计时间对比如图6所示,随着提交任务数量的增多,通过云计算平台的虚拟化技术扩展出新的虚拟机计算节点添加到虚拟集群,满足了作业对计算资源的需求,减少了作业等待的时间。所有作业从提交到完成花费的时间少于传统的虚拟集群运行所有任务所花费的时间。这说明虚拟集群技术可以有效地解决由计算资源不足而引起的作业排队等待时间过长或者根本无法运行的问题。

5 结束语

本文主要设计与实现了虚拟集群的动态构建技

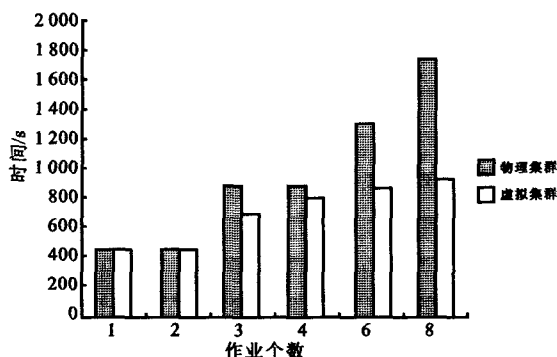


图6 作业运行花费时间

术,验证了动态构建虚拟集群的可行性,与传统的物理集群相比,虚拟集群更具灵活性。

(1) 将云计算平台和集群管理技术相结合,可以更加方便快速地构建高性能集群系统平台。

(2) 应用云计算平台的虚拟化技术可以弹性地创建虚拟的计算资源,从用户使用角度来看,减少了作业等待时间。

(3) 云计算平台可以灵活控制和分配物理机器计算资源,充分地利用计算资源来进行任务的计算,从机器利用率角度来看,提高了机器的使用效率,从而有效地解决由计算资源不足而引起的作业排队等待时间过长或者根本无法运行的问题。

下一步要做的工作是,创建多样化的应用软件平台,利用已经构建好的虚拟集群自动配置各种不同运行环境,如MPI、PVM等,以及各种科学计算软件,如CFX、Fluent、ACE、ABAQUS等,进而使虚拟集群技术实用化。

参考文献:

- [1] 刘柏,王海文,于智.高性能计算机发展现状及我国发展高性能计算机遇到的问题[J].广西轻工业,2006,22(6):94-95.
- [2] 刘鹏.云计算的定义和特点[EB/OL].[2009-02-25] <http://www.chinacloud.cn>.
- [3] Foster I, Zhao Y, Raicu I, et al. Cloud computing and grid computing 360-degree compared[C]//Grid Computing Environments Workshop. Chicago, 2008:1-10.
- [4] H L, D O. Accenture technology labs gridBatch: cloud computing for large scale data - Intensive batch[C]//CC-GRID. Shanghai:[s:n.], 2008.
- [5] Creasy R J. The origin of the VM/370 time-sharing system[J]. IBM Journal of Research and Development, 1981,25(5):483-490.

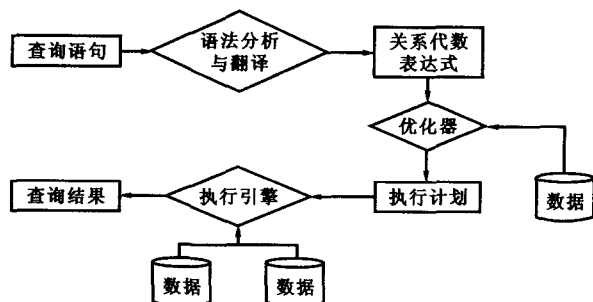


图5 查询的处理过程

(2) 查询分解. 分解器将根据元数据管理模块对查询进行分解重写, 去除查询中没有使用到的引用, 同时记录各子查询之间的组合关系。

(3) 对查询树进行优化. 选择和投影运算尽量往树叶方向下移, 把笛卡尔积运算通过附加选择和投影运算转换为连接运算。

(4) 形成查询执行策略. 说明需要访问哪些局部数据库系统, 如何组装中间结果, 按什么顺序执行连接查询及在哪里执行全局处理。

(5) 局部查询. 按照查询策略把各子查询与相应的数据源 Wrapper 进行通信, 形成各数据源的局部查询。

(6) 查询重组. 最后各 Wrapper 将子查询结果转换后传给 Mediator 进行组装, 由 Mediator 把全局查询结果返回给查询用户。

6 结束语

本文在提出了一种新型的异构数据集成系统的基础上, 分析了数据的查询处理过程. 针对现有的数据集成系统未能很好地解决底层异构数据源的“即插即用”问题, 数据查询中的全局模式同各数据源模式的转换工作繁杂, 用户视图不能很好地面向用户. 本文提出的数据集成系统以及在该系统中定义的查询有两个优点: (1) 查询过程是基于底层各异构数据

源之上的一层 Web Service, 只要可以提供 Web Service, 各异构数据源可以快速方便的查询; (2) 在数据查询中, 提出了全局可通用的数据模型, 使得各局部模式和全局模式基于同一个数据模型, 转化方便; (3) 用户的查询是基于快速生成的用户视图的, 使得用户能够方便、快速、直观地进行数据查询. 但是, 在集成数据的查询优化问题上, 本文还没有得到很好的解决, 在接下来的工作中, 将针对集成数据和本系统的特点, 提出一个更适合的查询优化方法。

参考文献:

- [1] 李军怀, 周明全, 耿国华, 等. XML 在异构数据集成中的应用研究[J]. 计算机应用, 2002, 22(9): 10-12.
- [2] 王宗杰, 侯贵法, 王成耀, 等. 基于元数据的分布异构数据集成研究[J]. 微计算机信息, 2007, 23(27): 84.
- [3] 于秀霞. 分布式数据查询优化算法的研究[J]. 长春大学学报, 2006, 16(1): 26-29.
- [4] 孙友仓, 宋彩利, 李润州. 一种基于 Web Service 的异构数据集成中间件[J]. 西安科技大学学报, 2007, 27(3): 472-474.
- [5] Sanjay Madria, Kalpdrum Passi, Sourav Bhowmick. An XML schema integration and query mechanism system[J]. Data & Knowledge Engineering, 2008(65): 266-303.
- [6] Boag S, Chamberlin D, Fernandez M F, et al. XQuery 1.0: an XML query language, W3C working draft [EB/OL]. (2002-11-15) [2009-12-19]. <http://www.w3c.org/TR/xquery/>.
- [7] Haiwei Ye, Kerherve B, Bochmann G V, et al. Pushing quality of service information and requirements into global query optimization[J]. Database Engineering and Applications Symposium, 2003(7): 170-179.

作者简介:

唐智勇 男, (1976-), 博士研究生. 研究方向为 web 服务与数据集成。

(上接第 35 页)

- [6] 杨晓伟, 廖光灯, 胡越明. 基于 Xen 的 X86 虚拟机性能调优[J]. 计算机工程, 2006, 32(24): 251-253.
- [7] Nurmi D, Wolski R, Grzegorzczak C, et al. The eucalyptus open-source cloud-computing system[C]// Cloud Computing and Applications 2008(CCA08). Germany: Hage, 2008.
- [8] Rajkumar Buyya. 高性能集群计算、编程与应用[M]. 郑伟民, 译. 北京: 电子工业出版社, 2001.
- [9] Alcaus Bayucan, Robert L Henderson. Portable batch system[M]. USA: Symposium on Cluster Computing and

the Grid, 2001.

作者简介:

曾龙海 男, (1987-), 硕士. 研究方向为人工智能与云计算。

张博锋 男, (1968-), 研究员. 研究方向为智能人机交互、云计算。

张丽华 女, (1982-), 硕士. 研究方向为人工智能与云计算。