文章编号:1002-4026(2011)05-0056-07

基于异构模式的云计算关键技术研究

张庆科,杨波*,王琳,陈贞翔
(济南大学信息科学与工程学院,山东省网络环境智能计算技术重点实验室,山东 济南 250022)

摘要:结合云计算中 Map/Reduce 分布式编程技术引入了基于 CPU-GPU 异构混合并行编程模式,给出了该并行编程模式的原理和实现过程。该模式通过采用 CUDA 多线程并行机制提高了大规模数据处理的效率。文中对比分析了云计算中两种典型的分布式存储系统 GFS 和 HDFS,最后从宏观角度阐释了云计算虚拟化技术的三层部署架构和基本类型。

关键词:云计算;图形处理器(GPU);CUDA;并行编程模型;分布式存储;虚拟化
中图分类号:TP393 **文献标识码**:A

Research on heterogeneous model based key cloud computing technologies

ZHANG Qing-ke, YANG Bo*, WANG Lin, CHEN Zhen-xiang
(Shandong Provincial Key Laboratory of Network Based Intelligent Computing, School of Information Science and Engineering, University of Jinan, Jinan 250022, China)

Abstract: This paper presents a CPU-GPU heterogeneous parallel programming model based on the distributed programming technology of cloud computing, Map/Reduce. This paper also gives its principle and implementation process. It improves the efficiency of large-scale data processing with the multi-thread parallel mechanism, CUDA. We contrastively analyze two typical distributed storage systems, GFS and HDFS. We eventually present the three-layer architecture and basic type of virtualization technology.

Key words: cloud computing; GPU; CUDA; parallel programming model; distributed storage; virtualization

计算机、互联网和通讯技术的快速发展使得网络对海量级数据存储能力和计算能力的需求日益提升。云计算通过协同调度网络中现有的软硬件资源,实现了存储与计算服务模式的虚拟化和透明化,并以其高效、灵活、拓展性强等诸多优势而成为解决网络中海量数据存储与计算的最新方案。由于云计算的存储和计算载体主要依附于传统的基于 CPU 的数据存储处理,而受电子线路集成规模极限的限制,CPU 在性能上提升空间有限。美国 Sandia 国家实验室研究证实,受存储机制和储带宽的限制,当 CPU 处理器数目多于 16 核时,计算机性能不仅得不到提升,反而效率会大幅度下降[1]。当前图形处理器 GPU(Graphic Processing Unit)却发展迅速,目前主流 GPU 的单精度浮点处理能力已经达到了同时期 CPU 的 10 倍,存储器带宽约是 CPU 的 5 倍,在成本和功耗上,相对 CPU 而言,GPU 无需付出太大的代价即可达到同样的性能,将云计算存储和

收稿日期:2011-06-30
基金项目:国家 973 计划前期研究专项基金(2010CB635117);国家自然科学基金(60873089,60573065,60673130,90818001,F020804);山东省自然科学杰出青年基金(JQ200820)
作者简介:张庆科(1985-),男,硕士研究生,研究方向为高性能计算,数字水泥建模。Email:miczqk@hotmail.com
*通讯作者,杨波(1965-),男,博士生导师,教授,研究方向为计算机网络与智能信息处理。Email: yangbo@ujn.edu.cn

计算的载体依附于 CPU-GPU 异构协作模式不仅可以提高数据处理的效率而且可以更好的促进当前网络中海量级数据实时的存储和处理。

1　云计算概述

Google 公司于 2006 年推出“Google101 计划”时最早正式提出“云(cloud)”的概念和理论,它是继分布式运算、并行处理和网格计算后新兴的商业计算模型[2]。2007 年 IBM 将云计算概念定义为一个系统平台或者一种类型的应用程序[3]。维基百科将其定义为一种动态的、易扩展的且通常是通过互联网提供虚拟化的资源计算方式,通过这种方式,共享的软硬件资源和信息可以按需提供给计算机和其他设备[4]。中国云计算网将云定义为“云计算是分布式计算(Distributed Computing)、并行计算(Parallel Computing)和网格计算(Grid Computing)的发展,或者科学概念的商业实现”[5]。总之,云计算即通过对网络中的软硬资源进行协同调度,以冗余存储的方式确保系统的可靠性和可用性,通过虚拟化技术将海量的存储数据或计算处理程序自动拆分成多个较小的相互间耦合性较低的子数据或子程序,然后将这些子数据或子程序再交由多服务器所组成的庞大“云系统”进行并行分布计算,计算处理结果将以透明、快速、可靠的方式返回给用户的新型商业计算模式。

2　异构编程模式与云计算技术

2.1　CPU-GPU 异构编程

随着计算量需求的增加以及并行机群和多核集群的发展,高性能并行计算被广泛地应用在大量数据处理的各个领域,尤其是从 2007 年 NVIDIA 公司推出 CUDA(Compute Unified Device Architecture)计算架构以来,基于 CUP-GPU 的异构模式的计算系统和编程方式逐渐被广泛应用到高性能计算领域中,并表现出较好的性能。这种 CPU-GPU 异构模式就是一种在 CPU 和 GPU 两种不同物理体系结构下进行通信或执行某些任务的协作部署模式,在这种模式中,GPU 主要作为加速部件,CPU 主要负责任务分配和资源统一调度。该模式充分融合了 CPU 强大的的逻辑分析处理能力和 GPU 的高带宽、多线程并行计算能力,为海量数据信息处理提供了高效处理平台[6-7]。CUDA 架构采用单指令多线程(Single Instruction Multiple Thread)的执行模型 ,异构并行基本步骤为:

(1)数据初始化:变量的定义与声明;例 float ＊a_h(主机端变量),double ＊a_d(设备端变量);

(2)存储空间分配:为 CPU 端和 GPU 端变量分别分配存储空间;

主机端:a_h = (float＊) malloc (N＊sizeof (float))

设备端:CudaMalloc ((void＊＊)&a_d),N＊sizeof (float))

(3)数据传递:从主机端将数据传输到 GPU 端,通过下列函数实现;

CudaMemCpy (a_d, a_h, nBytes, cudaMemCpyHostToDevice)

(4)并行执行:调用 Kernel 函数,设定 GPU 端执行参数和函数参数,该内核函数将会被 GPU 内分配到的所有线程各执行 1 次,从而实现对数据的并行处理;

__global__ void kernel <<< GridDim,BlockDim >>> (float＊xx,float＊yy,float＊zz)

(5)结果返回:将 GPU 端计算的结果传递到主机 CPU 端口,通过下列函数调用实现;

cudaMemCpy(b_h,b_d,nBytes,cudaMemCpyDeviceToHost)

(6)释放显存:在全局存储器中回收空间,通过调用函数 cudaFree()实现。

2.2　云计算技术

云计算是分布式处理、并行计算和网格计算等概念的发展和商业实现,它在数据存储、数据管理、虚拟化、编程模式等方面具有自身独特的技术[8],表 1 总结了当前典型云厂商的云计算核心技术和相关服务。

表 1 云厂商计算关键技术与服务

Table 1 Key technologies and services of cloud manufacturer

云厂商	技术特性	核心技术	企业服务	开源
Google	存储及运算水平扩充能力	Map/Reduce 编程模型、BigTable、GFS	搜索引擎、应用托管	否
Microsoft	整合其软件及数据服务	大型应用软件开发技术	Azure 平台	否
IBM	整合其所有软件及硬件服务	网格技术、分布存储技术	虚拟资源池、云计算方案	否
Amazon	弹性虚拟平台	虚拟化技术 Xen	EC,S3,SimpleDB	是
Salesforce	弹性可定制商务软件	应用平台整合技术	Force. com	否
中国移动	坚实的网络技术丰富的带宽资源	底层集群部署技术、资源池虚拟技术、网络相关技术	BigCloud 大云平台	否

通过表 1 可以看出,当前云厂商各自采用了不同的方式和诸多技术来应用和发展云计算,在这些技术领域中,海量数据处理技术、数据分布存储技术和虚拟化技术为云计算的三大核心技术。本文从这三个技术角度进行研究分析,给出了三种技术的实现原理,同时结合目前 CPU-GPU 的异构并行编程模式重点剖析了分布式编程技术 Map/Reduce 的异构并行实现过程。

3 核心技术与原理

3.1 Map/Reduce 并行编程技术

适合云计算的编程模型必须满足大规模数据集的并行计算,当前的 HPF(High Performance Fortran)只适用于规则数据并行问题(如矩阵运算),而 MPI(Message Passing Interface)编程模型仅适用于小通讯量的计算密集型问题。支持细粒度和共享存储的基于 OpenMp(Open Multi-Processing)的编程模型适合大型数据并行,但不适合多虚拟机的任务调度。目前云计算采用的编程模型主要是 Google 的 Map/Reduce 编程技术和 Microsoft 在 2010 年底推出的 Dryad 编程技术,但被业界广泛采用的是 Map/Reduce 的分布式编程技术。

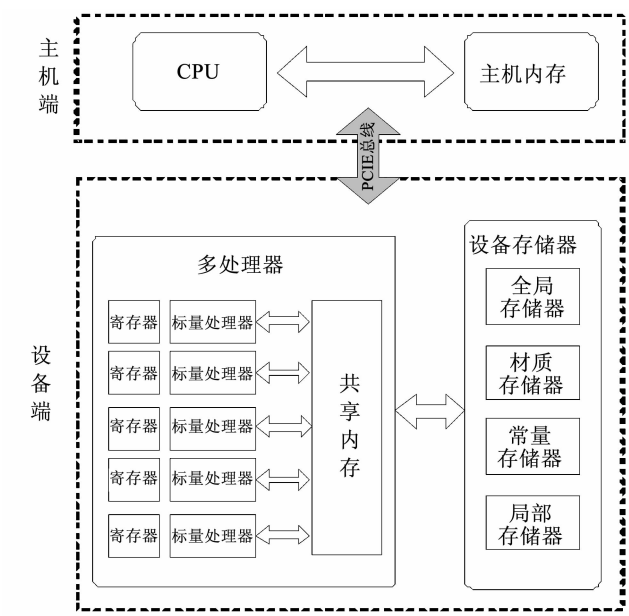

图 1 GPU-CPU 的数据通信

Fig. 1 Data communications between CPU and GPU

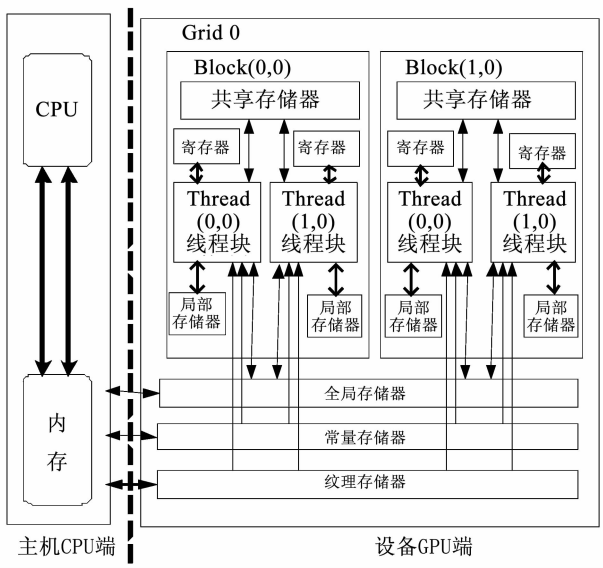

图 2 CPU-GPU 异构硬件并行架构

Fig. 2 GPU heterogeneous hardware parallel computing architecture

Map/Reduce 亦即映射/约减,它是一种简化的分布式编程模型和高效的任务调度模型,并且适用于大规模数据集的并行计算[9]。这种模式的最初功能是由 John McCarthy 于 1956 年在提出 Lisp 语言时实现。Map/Reduce 模式的思想就是将要执行的问题分解成 Map(映射)和 Reduce(约减)两个过程。首先 Map 把

一个函数应用于集合中的所有节点,将数据切割成不相关的区块,将这些区块分别调度给集群中分散的大量计算机,从而达到机体间分布式并行运算的效果。然后 Reduce 通过多个线把系统并行处理的结果集进行分类和归纳,将各节点计算的结果汇总后输出。基于 CPU-GPU 异构模式的 Map/Reduce 多线程并行机制则是在 Map 映射分配的处理机上执行这种基于 CPU-GPU 的二次机体内并行计算过程,图 1 给出了 CPU 和 GPU 通过 PCIE 总线进行数据通信的模型。在 CUDA 架构下,显示芯片执行时的最小单位是线程(thread),thread 可以组成一个线程块(block)。一个 block 中的 thread 能存取同一块共享的内存(shared memory),而且可以快速进行并行同步的计算,多个 block 构成一个计算格 Grid。其硬件并行架构如图 2 所示。

在处理分配到的数据时,设备端会以 warp 为单位(1warp = 32 thread)进行并行计算。将 Map 后的数据块分别放到各处理机 GPU 上不同的线程块(block)内的线程(thread)中执行,每个块分别赋以 GPU 中唯一的线程坐标(X,Y)和相应的标号 ID,其中 X,Y 分别表示线程 thread 在 GPU 每个 Grid 中的索引坐标。具体计算映射过程为:

$$X = \text{blockIdx.x} * \text{blockDim.x} + \text{threadIdx.x} \tag{1}$$

$$Y = \text{blockIdx.y} * \text{blockDim.y} + \text{threadIdx.y} \tag{2}$$

$$ID = \text{blockIdx.x} * \text{blockDim.x} + \text{threadIdx.x} \tag{3}$$

上式(1),(2),(3)中,blockIdx,blockDim,threadIdx,为 CUDA 编程中的内建变量,用于确定 grid 和 block 的维度,以及 block 和 thread 在其中的索引。这些内建变量只能在设备端 Device 上执行的函数中使用。每个线程块处理后的结果通过全局存储器经由 PCIE 总线返回到主机端。这种方式可以提高 Map 映射的速度,减少 Reduce 执行时间,提高了系统执行效率[10]。Map 和 Reduce 整体并行执行过程如图 3 所示,其对应的 CPU-GPU 异构并行计算过程如图 4 所示。这种编程模式不仅方便软件开发人员,而且这种技术对软件开发人员是透明的,因为它屏蔽了底层并行执行和调度的相关细节,适合于海量级数据的处理。

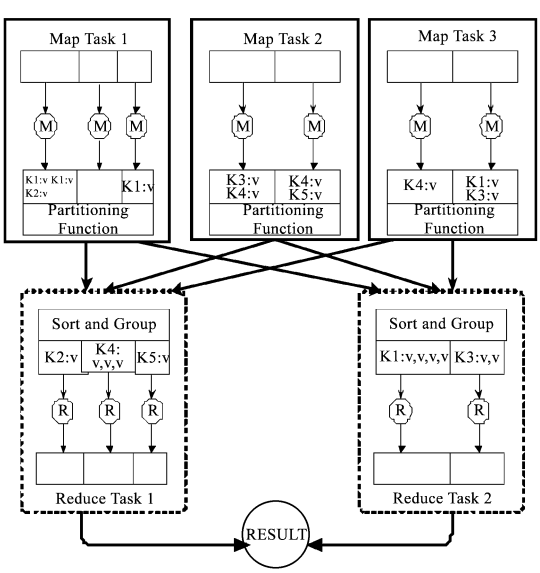

图 3　Map/Reduce 并行执行模型

Fig. 3　A Map/Reduce parallel model

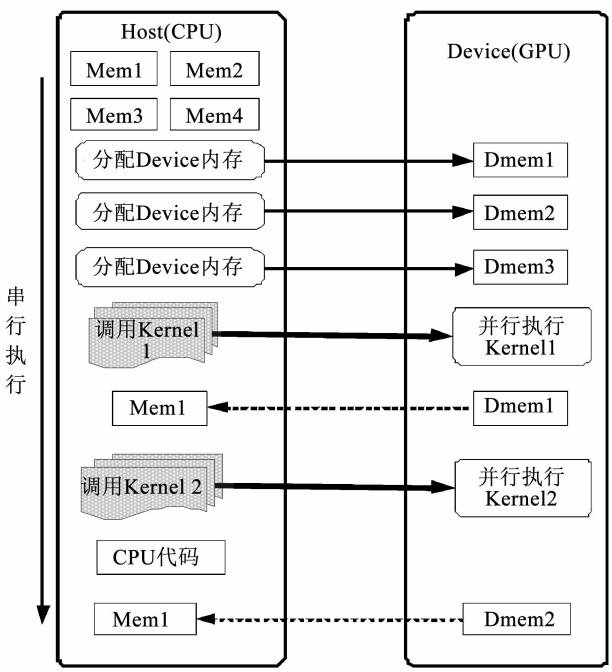

图 4　CPU-GPU 并行计算过程

Fig. 4　CPU-GPU parallel computing process

3.2　数据分布存储技术

针对云中海量级数据的存储问题,可靠性和稳定性尤为重要,因此云计算采用冗余存储的方式[11],其原理为“主从备份机制”,即将云中的文件切割后分块存储,当用户请求访问该文件时,负责文件分布的主机

Master 端会通过本地文件存储的元数据来查询“从服务器”中有关该文件的存储信息。然后将查询到的结果返回给请求客户端,再由客户端访问相应的服务器资源。其中元数据包括文件名,块的名字空间以及从文件到块的映射、副本数据。

云计算分布式数据存储中使用的技术有 Google 文件系统 GFS(Google File System)[12]和 Hadoop 的文件系统 HDFS(Hadoop File System)。这两种系统在架构上相仿,GFS 相对 HDFS 稳定性较好[13]。GFS 是一个可扩展的分布式文件系统,主要应用于大型分布式海量数据存储或访问中。一个 GFS 集群由一个主服务器(master)和大量的块服务器(chunk server)构成,并允许多客户(Client)访问(如图 5 的 GFS 结构所示)。GFS 中的文件被切分为 64 MB 的块并以冗余存储,每份数据在系统中保存 3 个以上备份。相比 GFS,HDFS 首先通过客户端联系元数据 Namenode,得到所有数据块信息以及对应的服务器位置信息(如图 6 的 HDFS 架构所示)。然后从某个数据块对应的一组数据服务器中选择一个进行访问连接。数据以包为单位从服务器返回给客户端,等数据传输完毕就断开连接。之后继续尝试连接下一个数据块对应的服务器,整个流程反复进行,直到客户端获得所有请求的数据块[14]。

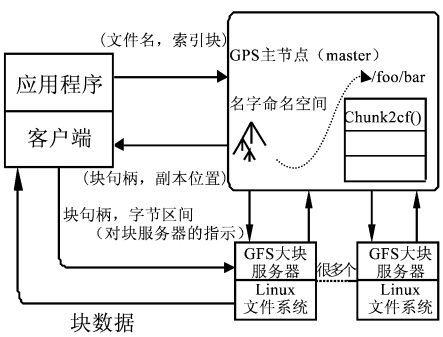

图 5 GFS 文件系统架构

Fig. 5 The architecture of GFS file system

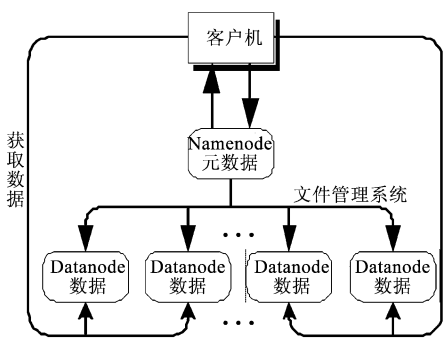

图 6 HDFS 文件系统架构

Fig. 6 The architecture of HDFS file system

GFS 和 HDFS 原理相似,首先它们都采用单一主控机 + 多台工作机的模式,由一台主控机(Master)存储系统全部元数据,并实现数据的分布、复制、备份决策。工作机存储数据,并根据主控机的指令进行数据存储、数据迁移和数据计算等。其次,GFS 和 HDFS 都通过数据分块和复制来保证系统整体稳定性。当其中一个副本失效时,系统将提供对该副本自动复制功能。但 HDFS 为了规避 GFS 的复杂度而做了简化,表 2 总结了两种分布式文件系统间的主要差异。客户端与主服务器的数据交换只限于对主服务器元数据的操作,所有数据方面的通信都直接和块服务器联系,这样来减少主服务器负载,同时,针对数据读多于写的特点,读服务被分配到多个副本所在机器,提供了系统的整体性能,提高了系统的运行效率。

表 2 GFS 和 HDFS 比较

Table 2 Comparisons between GFS and HDFS

比较项目	GFS	HDFS
中心服务器模式比较	多台物理服务器	单一中心服务器
子服务器管理模式比较	Master 通过轮询独占锁机制来快速检测块服务器生存状态	Namenode 损坏后需要等待一段时间才能获知数据 Datanode 的状态
扩展性比较	存储节点容易扩展	存储节点不容易扩展
安全模式比较	无完备的安全模式	具备安全模式

3.3 虚拟化技术

虚拟化技术源于 20 世纪 60 年代,其技术本质上是一种逻辑简化技术,它通过对底层复杂的物理结构的屏蔽来实现物理层向逻辑层的转化,即将底层的物理运动向逻辑运动的转化,最终实现软件应用与底层硬件相隔离的效果,隔离后提高了 IT 资源的利用率和灵活性。

当前典型的虚拟技术有 CPU 一级的虚拟化技术,即在底层硬件上直接运行多个操作系统;以及硬件层

上一级的虚拟化技术,如操作系统、VMware 等都是在硬件之上建立虚拟化程序;第三种是在操作系统之上的虚拟化技术,如高级语言虚拟化技术(C#,Java 等)[15]。云计算的虚拟化则是在基本虚拟化基础之上,通过整合计算机网络中计算资源、存储资源、应用等分布式资源,提供基础架构服务,平台服务以及软件服务等。图 7 给出了云计算虚拟化技术宏观结构。根据中间层虚拟化抽象的不同,又可将其划分为以下四种虚拟化类型。表 3 总结了四种云计算不同层次的虚拟化类型及相关技术原理。

表 3　云计算虚拟化类型

Table 3　A virtual type of cloud computing

类型	技术原理
存储虚拟化	将实体存储空间分割成不同的逻辑存储区域,并将信息存储到这些逻辑区域中
网络虚拟化	将底层不同网络的软硬件资源聚合成一个虚拟的整体,实现统一调度的网络实体
应用虚拟化	在操作系统和应用程序间建立虚拟环境或访问接口,实现对应用程序访问的虚拟化
桌面虚拟化	在本地计算机显示和操作远程计算机桌面,而在远程计算机执行程序或存储信息

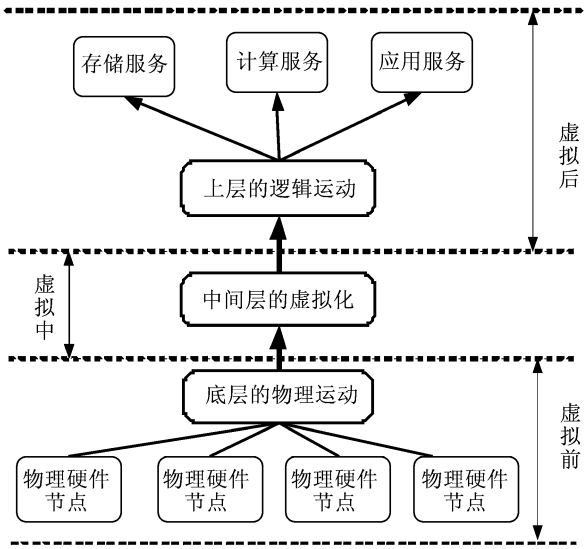

图 7　虚拟化宏观结构

Fig. 7　A macroscopic virtual structure

通过虚拟化的方式不仅有效地整合了网络中的软硬件资源,为用户提供了方便快捷的接口通道,还减少了繁琐的基础设施部署的开销。通过集群式的协同资源调度,提高了网络总体的性能。但在基础设施层次上,将虚拟机迁移到没有共享存储的其他物理主机上实现云计算系统服务的动态迁移,也是云计算系统目前面临的主要问题。

4　总结

本文将 CPU-GPU 异构编程模式与 Map/Reduce 编程技术相结合,充分发挥 GPU 多线程并行计算机制,融合 CPU 强大的逻辑事务处理能力,从整体上提高了云中海量信息处理能力,实现了云系统对 IT 资源的高效利用。分布式文件系统 GFS 和 HDFS 采用主从备份机制实现了数据在整个生命周期内有序、高效、自治,可靠的存储过程,为异构数据信息处理提供了可靠的访问载体和存储载体。云计算最终目标是实现计算和存储的虚拟化,因此虚拟化技术可视为云计算是否成熟的标志关键技术。未来云计算将会充分发挥高性能集群计算优势,通过资源的协同调度和快速高效的数据存储与处理,为云端提供更稳定、更可靠的多元化服务。

参考文献:

[1]IEEE Spectrum. Multicore is bad news for supercomputers[EB/OL]. [2011 - 05 - 12]. http://spectrum. ieee. org/computing/hardware/multicore-is-bad-news-for-supercomputers.

[2]陈康, 郑纬民. 云计算:系统实例与研究现状[J]. 软件学报,2009,20(5):1337 - 1348.

[3]JINZY Z, XING F, ZHE G. IBM cloud computing powering a smarter planet[J]. Lecture Notes in Computer Science, 2009(5931): 621 - 625.

[4]Wikipedia. Cloud computing[EB/OL]. [2011 - 05 - 12]. http://en. wikipedia. org/wiki/Cloud_computing.

[5]中国云计算网. 什么是云计算? [EB/OL]. [2011 - 05 - 12]. http://www. chinacloud. cn/show. aspx? id = 2623&cid = 17.

[6]BRDTKORB A,DYKEN C,HAGEN T,et al. State-of-the-art in heterogeneous computing[J]. Programming,2010,18(1):1 - 33.

[7]张建勋,古志民,郑超. 云计算研究进展综述[J]. 计算机应用研究,2010, 27(2): 429 -433.
[8]刘鹏. 云计算技术原理[EB/OL]. [2011 -06 -12]. http://www. Chinacloud. cn/show. aspx? id = 1929&cid = 12.
[9] ZHARIA M, KONWINSKI A, JOSEPH A D. Improving MapReduce performance in heterogeneous environments[M]// Proceedings of the 8th USENIX Symposium on Operating Systems Design and Implementation. New York: ACM Press, 2008:260 -269.
[10]张舒, 褚艳丽,赵开勇,等. GPU 高性能计算之 CUDA[M]. 北京:中国水利水电出版社,2009.
[11]陈全,邓倩妮. 云计算及其关键技术[J]. 计算机应用,2009,29(9):2562 -2567.
[12] CHEMAWAT S, GOBIOFF H, LEUNG P T. The Google file system[M]//Proceedings of the 19th ACM Symposium on Operating Systems Principles. New York: ACM Press, 2003:29 -43.
[13]SHVACHKO K, KUANG H, RADIA S,et al. The Hadoop Distributed File System[M]// Proceedings of IEEE MSST 2010,New York:IEEE Computer Society,2010:1 -10.
[14]王鹏. 云计算的关键技术与应用实例[M]. 北京:人民邮电出版社,2010.
[15]虚拟化与云计算小组. 虚拟化与云计算[M]. 北京:电子工业出版社,2009.

(上接第 41 页)

表 5 5.52% 硫酸水解样品结果

Table 5 The result of total sugar treated by pea flour residue with 5.52% sulfuric acid

水解时间/h	2	4	6	8	10
实验测定还原糖量/%	0.307	0.356	0.375	0.379	0.379
总糖/%(以还原糖计)	30.70	35.60	37.50	37.90	37.90

在实验过程中,当豌豆粉渣未进行糊化时,所测 DE 值较低,是因为淀粉颗粒中结晶区的存在会阻碍酶与底物充分接触;当淀粉充分糊化后,测其平均 DE 值为 11.35 ±0.296,在粉渣中淀粉的含量依然很高,是因为在湿法分离过程中,一些不溶蛋白和纤维素阻止了淀粉的沉淀。根据表 4,5 显示的结果,在酸水解的前 4 h,粉渣水解的速度很快,随着酸浓度的增加,粉渣水解时间缩短,最后完全水解获得还原糖浓度约为 38 mg/dL,以此推断粉渣中总糖量约为 38%。

由于豌豆蛋白具有很高的溶解度、吸水性能和乳化性能,在食品工业中用途广泛;豌豆中的膳食纤维,有助于消化,防止包括阑尾炎、心脏病和结肠癌等多种疾病。从豌豆粉渣中提取蛋白质、纤维素具有可观的经济价值。我们测得的豌豆粉渣中的蛋白质、纤维素含量较高,因此开发粉渣的综合利用及深加工具有广阔的前景,该研究为企业以后能更合理地利用粉渣奠定了基础。

参考文献:

[1]孙林,等. 粉丝生产技术[M]. 北京:中国食品出版社, 1987.
[2]阮新,耿金培,王晓洁, 等. 豌豆粉丝中硒含量的测定[J]. 食品科学,2008,29(11):527 -531.
[3]陈华,梅承英,李厚岩. 粉渣替代部分粮食喂猪试验[J]. 安徽农业技术师范学院学报, 2000, 14(3):43 -44.
[4]鲁健章,孙丽华,周彦钢. 凯氏定氮法测定鱼蛋白质含量的干扰因素分析[J]. 食品科学,2010(31):453 -456.
[5]范鹏程,田静,黄静美,等. 花生壳中纤维素和木质素含量的测定方法[J]. 重庆科技学院学报(自然科学版) ,2008,10(5): 64 -65.
[6]马耀宏. 还原糖测定技术[J]. 发酵科技通讯. 2008,32(1):20 -22.