HZ BOOKS
华章教育

高等学校计算机专业规划教材

分布式计算、云计算与大数据

林伟伟 刘波 编著

Distributed Computing, Cloud Computing and Big Data

机械工业出版社
China Machine Press

高等学校计算机专业规划教材

分布式计算、云计算与大数据

林伟伟 刘波 编著

Distributed Computing, Cloud Computing and Big Data

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

分布式计算、云计算与大数据 / 林伟伟，刘波编著 . —北京：机械工业出版社，2015.10
（高等学校计算机专业规划教材）

ISBN 978-7-111-51777-1

I. 分… II. ①林… ②刘… III. 计算机网络－数据处理－高等学校－教材 IV. TP393

中国版本图书馆 CIP 数据核字（2015）第 240762 号

本书将传统的分布式计算与新兴的云计算、大数据等技术综合起来，以应用需求为背景讲解技术原理和应用方法，主要内容包括：传统分布式计算的基本原理和核心技术，云计算的原理、架构、实现技术及安全问题，大数据的分析模型、存储平台、编程技术及电商大数据分析技术等。

本书适合作为高等学校计算机专业高年级本科生和研究生教材，也适合作为相关技术人员的参考读物。

出版发行：机械工业出版社（北京市西城区百万庄大街 22 号　邮政编码：100037）
责任编辑：曲　熠　　责任校对：殷　虹
印　　刷：北京市荣盛彩色印刷有限公司　　版　　次：2015 年 11 月第 1 版第 1 次印刷
开　　本：185mm × 260mm　1/16　　印　　张：30
书　　号：ISBN 978-7-111-51777-1　　定　　价：59.00 元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换
客服热线：（010）88378991　88361066　　投稿热线：（010）88379604
购书热线：（010）68326294　88379649　68995259　　读者信箱：hzjsj@hzbook.com

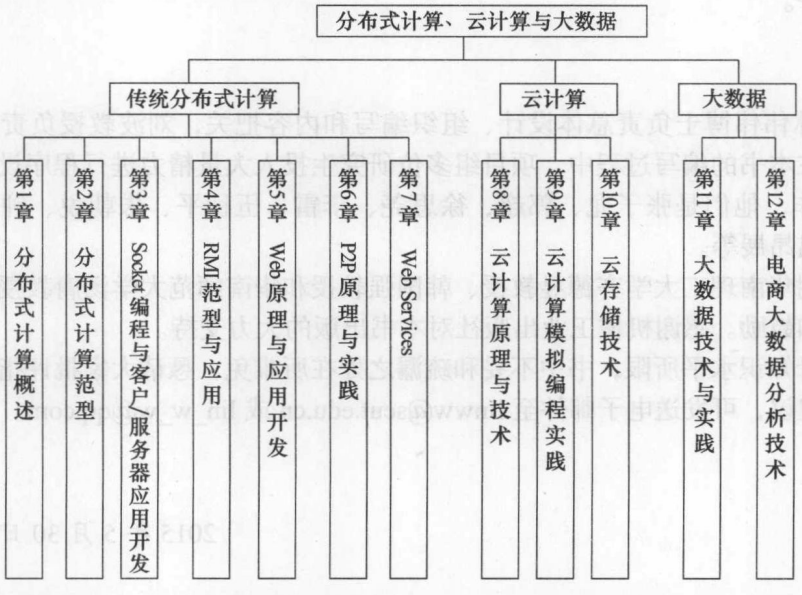
前　言

背景

分布式计算从20世纪六七十年代发展到现在，一直是计算机科学技术的理论与应用的热点问题，特别是最近几年，随着互联网、移动互联网、社交网络应用的发展，急需分布式计算的新技术——云计算、大数据，以满足和实现新时代计算机的应用需求。云计算、大数据等新技术本质上是分布式计算的发展和延伸，现有的书籍一般很少把经典的分布式计算与新兴的云计算、大数据等技术综合起来，并以应用需求为背景来剖析这些技术的原理和应用方法，本书正是为了适应这一新的发展趋势和需求而编写的，希望对云计算、大数据等新技术的研究与应用起到一定的作用。

内容规划

本书包含传统分布式计算、云计算和大数据三方面的内容，具体内容包括：传统分布式计算的基本原理、核心技术、相关开发技术与方法（Socket、RMI、P2P、Web Services）；云计算概述与原理、云计算架构与实现技术（Google、Amazon的云计算技术）、云计算研究现状与发展方向、云计算模拟编程实践、云存储技术、云计算安全问题与技术；大数据的分析计算模型（PRAM、BSP、LogP、MapReduce、Spark内存计算等）、大数据存储平台（Hadoop[HDFS/HBase]、Cassandra、Redis、MongoDB等）、大数据分析处理技术（Impala、Hadoopdb、Spark等）、大数据编程技术及研究现状、电商大数据分析技术等。全书共12章，各章之间的层次关系如下：



教学资源与使用方法

本书提供配套的 PPT 课件和课后习题参考答案，使用本书进行教学的教师可以从华章网站（www.hzbook.com）下载或发送电子邮件至 linww@scut.edu.cn 或 lin_w_w@qq.com 向编者索取。

本书可以作为计算机及相关专业高年级本科生和研究生的教材，建议在学习过操作系统、计算机网络、面向对象编程语言之后学习本课程。本书内容可根据不同的教学目的和对象进行选择，例如，对于本科类的分布式计算相关课程，可以选择分布式计算相关章节（第 1 ~ 7 章）重点讲解；对于本科类的云计算相关课程，可以选择分布式计算和云计算相关章节（第 1 ~ 10 章）重点讲解；对于本科类的大数据相关课程，可以选择分布式计算和大数据相关章节（第 1 ~ 7 和 11、12 章）重点讲解；对于研究生的课程，可以选择云计算和大数据相关章节重点讲解。根据本书的定位，建议每章讲授的最低学时分配如下：

章号	建议重点讲授章节	建议学时
第 1 章	1.1，1.2，1.3.1，1.3.2，1.3.3	2
第 2 章	所有内容	2
第 3 章	3.2，3.5，3.6，3.7	4
第 4 章	4.2，4.3，4.4	4
第 5 章	5.1.3，5.3，5.4，5.7.5	6
第 6 章	6.2，6.4	2
第 7 章	7.1.4，7.1.5，7.2.2，7.3	3
第 8 章	8.1，8.2，8.3，8.4	3
第 9 章	9.1，9.3，9.4	4
第 10 章	10.2，10.4	4
第 11 章	11.2，11.3.4，11.3.5，11.4，11.5，11.6	8
第 12 章	12.2，12.3，12.4	8

此外，本书的教学应该有相应的实验课程，建议实验课程学时数不少于理论课程学时数的三分之一。

致谢

本书由林伟伟博士负责总体设计、组织编写和内容把关，刘波教授负责全书审校和整体润色。在本书的编写过程中，项目组多位研究生投入大量精力进行程序设计与资料收集、整理工作，他们是张子龙、郭超、徐思尧、李雷、伍秋平、朱朝悦、钟坯平、吴文泰、杨超、温昂展等。

衷心感谢华南理工大学齐德昱教授、韩国强教授和华南师范大学汤庸教授对本书编写工作的指导和鼓励。感谢机械工业出版社对本书出版的大力支持。

由于编者知识水平所限，书中不妥和疏漏之处在所难免，恳请大家批评指正。如果有任何问题和建议，可发送电子邮件至 linww@scut.edu.cn 或 lin_w_w@qq.com。

林伟伟

2015 年 5 月 30 日于广州

目　录

第1章　分布式计算概述

本章首先介绍分布式计算的定义、优缺点等基本知识，然后讨论分布式系统的定义、特征和经典案例，最后重点讨论分布式计算的进程间通信等相关技术，这些内容将为后续章节打下基础。

1.1　分布式计算的概念

1.1.1　定义

分布式计算是计算机科学的重要研究内容，主要研究对象是分布式系统。简单地说，一个分布式系统是由若干通过网络互联的计算机组成的软硬件系统，且这些计算机互相配合以完成一个共同的目标（往往这个共同的目标称为“项目”）。

分布式计算的一种简单定义是在分布式系统上执行的计算。更为正式的定义是，分布式计算研究如何把一个需要非常巨大的计算能力才能解决的问题分成许多小的部分，然后把这些部分分配给许多计算机进行处理，最后把各部分的计算结果合并起来得到最终的结果。本质上，分布式计算是一种基于网络的分而治之的计算方式。

1.1.2　分布式计算的优缺点

在计算机网络出现之前，单机计算是计算的主要形式。自20世纪80年代以来，由于Web的促进，分布式计算得到飞速发展。分布式计算可以有效利用全世界联网机器的闲置处理能力，帮助一些缺乏研究资金的、公益性质的科学研究，加速人类的科学进程。分布式计算的优点如下：

1）低廉的计算机价格和网络访问的可用性。今天的个人计算机（Personal Computer，PC）比早期的大型计算机具有更出众的计算能力，但体积和价格不断下降。再加上Internet连接越来越普及且价格低廉，大量互连计算机为分布式计算创建了一个理想环境。

2）资源共享。分布式计算体系反映了计算结构的现代组织形式。每个组织在面向网络提供共享资源的同时，独立维护本地组织内的计算机和资源。采用分布式计算，可非常有效地汇集资源。

3）可伸缩性。在单机计算中，可用资源受限于单台计算机的能力。相比而言，分布式计算有良好的伸缩性，对资源需求的增加可通过提供额外资源来有效解决。例如，将更多支持电子邮件等类似服务的计算机增加到网络中，可满足对这类服务需求增长的需要。

4）容错性。由于可以通过资源复制维持故障情形下的资源可用性，因此，与单机计算相比，分布式计算提供了容错功能。例如，可将数据库备份并维护到网络的不同系统上，以便在一个系统出现故障时，还有其他副本可以访问，从而避免服务瘫痪。尽管不可能构建一个能在故障情况下提供完全可靠服务的分布式系统，但实现系统的最大化容错能力，是开发者的职责。

然而，无论何种形式的计算，都有其利与弊的权衡。分布式计算发展至今，仍然有很多

需要解决的问题。分布式计算主要的缺点如下：

1）多点故障。分布式计算存在多点故障的可能。由于涉及多个计算机，且都依赖于网络通信，因此一台或多台计算机的故障及一条或多条网络链路的故障都会导致分布式系统出现问题。

2）安全性低。分布式系统为非授权用户的攻击提供了更多机会。在集中式系统中，所有计算机和资源通常只受一个管理者控制，而分布式系统的非集中式管理机制包括许多独立组织。非集中式管理使安全策略的实现和增强变得更为困难，因此，分布式计算在安全攻击和非授权访问防护方面较为脆弱，并可能会影响系统内的所有参与者。

1.1.3　分布式计算的相关计算形式

与分布式计算类似的计算形式有很多，下面讨论单机计算、并行计算、网络计算、网格计算和云计算这五种形式，以便更好地区分和理解分布式计算的概念。

1. 单机计算

单机计算是最简单的计算形式，即利用单台计算机（如 PC）进行计算，此时计算机不与任何网络互连，因而只能使用本计算机系统内可被即时访问的所有资源。在最基本的单用户单机计算模式中，一台计算机在任何时刻只能被一个用户使用。用户在该系统上执行应用程序，不能访问其他计算机上的任何资源。在 PC 上使用的诸如文字处理程序或电子表格处理程序等应用就是单用户单机计算的计算形式。

多用户也可参与单机计算。在该计算形式中，并发用户可通过分时技术共享单台计算机中的资源，我们称这种计算方式为集中式计算。通常将提供集中式资源服务的计算机称为大型机（mainframe computing）。用户可通过终端设备与大型机系统相连，并在终端会话期间与之交互。

如图 1-1 所示，与单机计算模式不同，分布式计算包括在通过网络互连的多台计算机上执行的计算，每台计算机都有自己的处理器及其他资源。用户可以通过工作站完全使用与其互连的计算机上的资源。此外，通过与本地计算机及远程计算机交互，用户可访问远程计算机上的资源。WWW 是该类计算的最佳例子。当通过浏览器访问某个 Web 站点时，一个诸如 IE 的程序将在本地系统运行并与运行于远程系统中的某个程序（即 Web 服务器）交互，从而获取驻留于另一个远程系统中的文件。

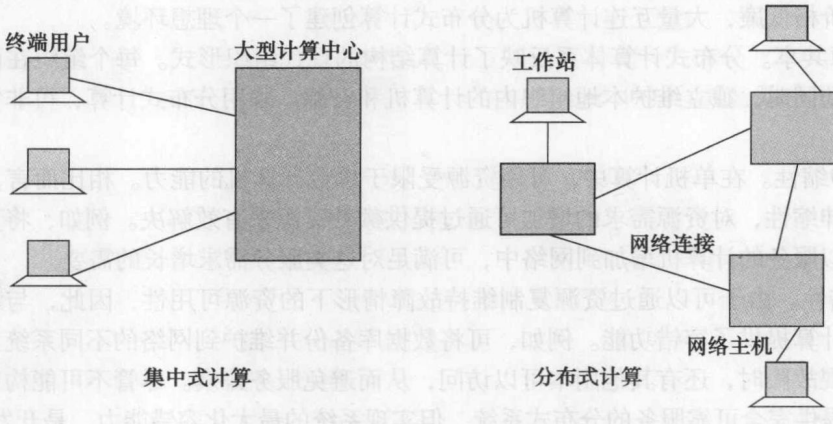


图 1-1　集中式计算与分布式计算

2. 并行计算

并行计算（或称并行运算）是相对于串行计算的概念（如图 1-2 所示），最早出现于 20 世纪六七十年代，指在并行计算机上所做的计算，即采用多个处理器来执行单个指令。通常并行计算是指同时使用多种计算资源解决计算问题的过程，是提高计算机系统计算速度和处理能力的一种有效手段。它的基本思想是用多个处理器来协同求解同一问题，即将被求解的问题分解成若干个部分，各部分均由一个独立的处理机来并行计算。

并行计算可分为时间上的并行和空间上的并行。时间上的并行指流水线技术，而空间上的并行指用多个处理器并发地执行计算。与分布式计算的区别是，分布式计算强调任务的分布执行，而并行计算强调任务的并发执行。

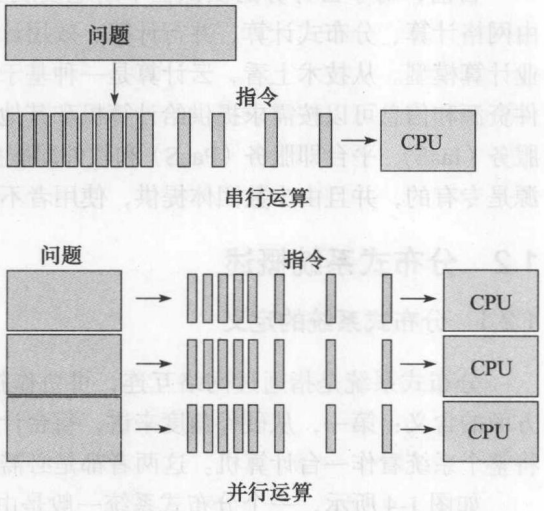


图 1-2　串行运算与并行运算

3. 网络计算

网络计算是一个比较宽泛的概念，随着计算机网络而出现。网络技术的发展，在不同的时代有不同的内涵。例如，有时网络计算指分布式计算，有时指云计算或其他新型计算方式。总之，网络计算的核心思想是，把网络连接起来的各种自治资源和系统组合起来，以实现资源共享、协同工作和联合计算，为各种用户提供基于网络的各类综合性服务。网络计算在很多学科领域发挥了巨大作用，改变了人们的生活方式。

4. 网格计算

网格计算是指利用互联网把地理上广泛分布的各种资源（计算、存储、带宽、软件、数据、信息、知识等）连成一个逻辑整体，就像一台超级计算机一样，为用户提供一体化信息和应用服务（计算、存储、访问等）。网格计算强调资源共享，任何结点都可以请求使用其他结点的资源，任何结点都需要贡献一定资源给其他结点。

网格计算侧重并行的计算集中性需求，并且难以自动扩展。云计算侧重事务性应用、大量的单独的请求，可以实现自动或半自动的扩展。

5. 云计算

云计算这个概念最早由 Google 公司提出。2006 年，Google 高级工程师克里斯托夫·比希利亚第一次提出“云计算”的想法，随后 Google 推出了“Google 101 计划”，该计划的目的是让高校的学生参与云的开发，为学生、研究人员和企业家提供 Google 式的无限计算处理能力，这是最早的“云计算”概念，如图 1-3 所示。云计算概念包含两个层次的含义：一是商业层面，即以“云”的

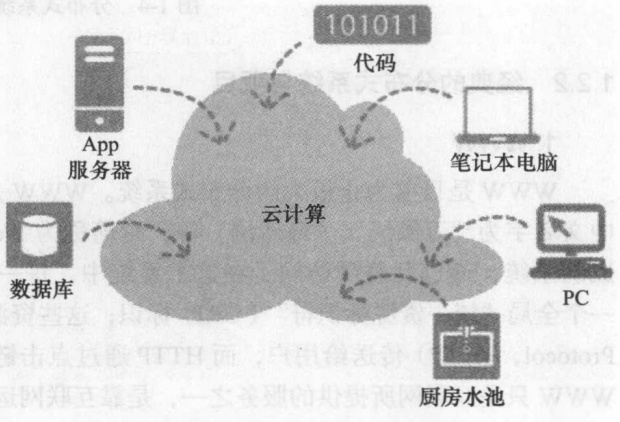


图 1-3　云计算概念示意图

方式提供服务；二是技术层面，即各种客户端的“计算”都由网络负责完成。通过把云和计算相结合，说明 Google 在商业模式和计算架构上与传统的软件和硬件公司不同。

目前，对于云计算的认识在不断地发展变化，云计算仍没有普遍一致的定义。通常是指由网格计算、分布式计算、并行计算、效用计算等传统计算机和网络技术融合而形成的一种商业计算模型。从技术上看，云计算是一种基于互联网的计算方式，通过这种方式，共享的软硬件资源和信息可以按需求提供给计算机和其他设备。当前，云计算的主要形式包括基础设施即服务（IaaS）、平台即服务（PaaS）和软件即服务（SaaS）。云计算强调专有，即请求或获取的资源是专有的，并且由少数团体提供，使用者不需要贡献自己的资源。

1.2　分布式系统概述

1.2.1　分布式系统的定义

分布式系统是指通过网络互连，可协作执行某个任务的独立计算机集合。这个定义有两方面的含义：第一，从硬件角度来讲，每台计算机都是自主的；第二，从软件角度来讲，用户将整个系统看作一台计算机。这两者都是必需的，缺一不可。

如图 1-4 所示，一个分布式系统一般是由多个位于不同位置的网络上计算机组成的系统，这些计算机通过网络传递消息与通信，从而完成一个共同的目标（项目）。

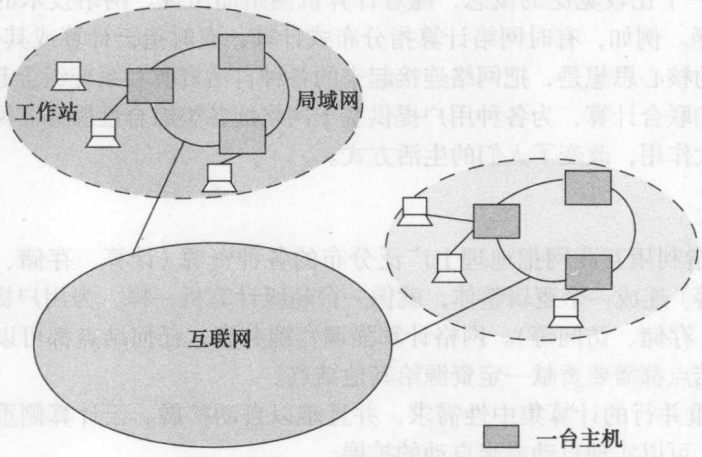


图 1-4　分布式系统示意图

1.2.2　经典的分布式系统与项目

1. WWW

WWW 是目前为止最大的分布式系统。WWW 是环球信息网（World Wide Web）的缩写，中文名字为“万维网”、“环球网”等，常简称为 Web。它是一个由许多互相链接的超文本组成的系统，通过互联网访问。在这个系统中，每一个有用的事物称为一样“资源”，并且由一个全局“统一资源标识符”（URI）标识；这些资源通过超文本传输协议（Hypertext Transfer Protocol，HTTP）传送给用户，而 HTTP 通过点击链接来获得资源。WWW 并不等同互联网，WWW 只是互联网所提供的服务之一，是靠互联网运行的一项服务。

WWW 是建立在客户机 / 服务器（Client/Server，C/S）模型之上的。WWW 以超文本标记

语言（标准通用标记语言下的一个应用）与 HTTP 为基础，能够提供面向 Internet 服务的、一致的用户界面的信息浏览系统。其中 WWW 服务器采用超文本链路来链接信息页，这些信息页既可放置在同一主机上，也可放置在不同地理位置的主机上；本链路由统一资源定位器（URL）维持，WWW 客户端软件（即 WWW 浏览器）负责信息显示与向服务器发送请求。

2. SETI@home

SETI@home（Search for Extra Terrestrial Intelligence at Home，寻找外星人）是一个利用全球联网的计算机共同搜寻地外文明的项目，本质上它是一个由互联网上的多个计算机组成的处理天文数据的分布式计算系统。SETI@home 是由美国加州大学伯克利分校的空间科学实验室开发的一个项目，它试图通过分析阿雷西博射电望远镜采集的无线电信号，搜寻能够证实地外智能生物存在的证据，该项目参考网站为 http://setiathome.berkeley.edu/index.php。

SETI@home 是目前 Internet 上参加人数最多的分布式计算项目。SETI@home 程序在用户的 PC 上，通常在屏幕保护模式下或后台模式运行。它利用的是多余的处理器资源，不影响用户正常使用计算机。SETI@home 项目自 1999 年 5 月 17 日开始正式运行，至 2004 年 5 月，累积进行了近 5×10^{21} 次浮点运算，处理了超过 13 亿个数据单元。截至 2005 年关闭之前，它已经吸引了 543 万用户，这些用户的计算机累积工作 243 万年，分析了大量积压数据，但是项目没有发现外星文明的直接证据。SETI@home 是迄今为止最成功的分布式计算试验项目。

3. BOINC

BOINC（Berkeley Open Infrastructure for Network Computing，伯克利开放式网络计算平台）是由美国加利福尼亚大学伯克利分校于 2003 年开发的一个利用互联网计算机资源进行分布式计算的软件平台。BOINC 最早是为了支持 SETI@home 项目而开发的，之后逐渐成为主流的分布式计算平台，为众多的数学、物理、化学、生命科学、地球科学等学科类别的项目所使用。如图 1-5 所示，BOINC 平台采用了传统的客户端 / 服务端架构：服务端部署于计算项目方的服务器，一般由数据库服务器、数据服务器、调度服务器和 Web 门户组成；客户端部署于志愿者的计算机，一般由分布在网络上的多个用户计算机组成，负责完成服务端分发的计算任务。客户端与服务端之间通过标准的互联网协议进行通信，实现分布式计算。

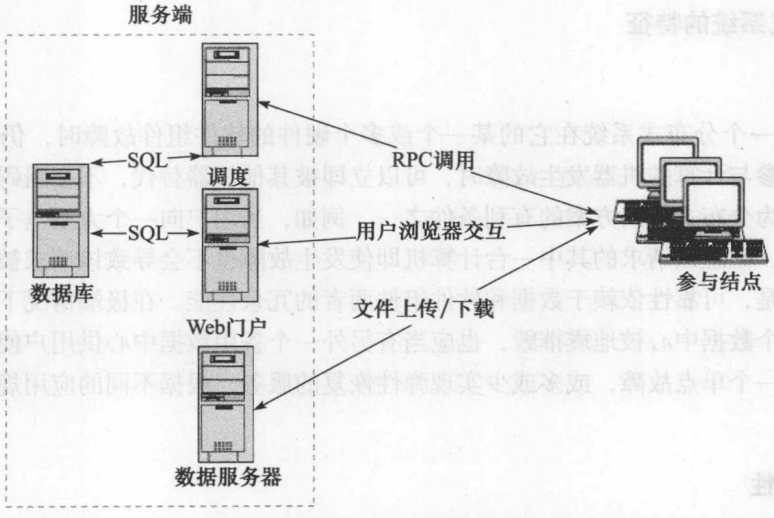


图 1-5　BOINC 的体系结构

BOINC 是当前较为流行的分布式计算平台，提供了统一的前端和后端架构，一方面大大简化了分布式计算项目的开发，另一方面对于参加分布式计算的志愿者来说，参与多个项目的难度也大大降低。目前已经有超过 50 个分布式计算项目基于 BOINC 平台，BOINC 平台上的主流项目包括 SETI@home、Einstein@Home、World Community Grid 等。更详细的介绍请参考该项目网站 http://boinc.ssl.berkeley.edu。

4. 其他分布式计算项目

除了以上 3 个经典的分布式系统外，还有很多其他的分布式计算项目，它们通过分布式计算来构建分布式系统和实现特定项目目标。

- Climateprediction.net：模拟百年以来全球气象变化，并计算未来地球气象，以应对未来可能遭遇的灾变性天气。
- Quake-Catcher Network（捕震网）：借由日渐普及的笔记本电脑中内置的加速度计，以及一个简易的小型 USB 微机电强震仪（传感器），创建一个大的强震观测网；可用于地震的实时警报或防灾、减灾等相关的应用上。
- World Community Grid（世界社区网格）：帮助查找人类疾病的治疗方法，进行改善人类生活的相关公益研究，包括艾滋病、癌症、流感病毒等疾病及水资源复育、太阳能技术、水稻品种的研究等。
- Einstein@Home：2005 年开始的项目，用于找出脉冲星的引力波，验证爱因斯坦的相对论预测。
- FightAIDS@home：研究艾滋病的生理原理和相关药物。
- Folding@home：了解蛋白质折叠、聚合以及相关疾病。
- GIMPS：寻找新的梅森素数。
- Distributed.net：2002 年 10 月 7 日，以破解加密术而著称的 Distributed.net 宣布，在经过全球 33.1 万名计算机高手共同参与，苦心研究了 4 年之后，他们于 2002 年 9 月中旬破解了以研究加密算法而著称的美国 RSA 数据安全实验室开发的 64 位密钥——RC5-64 密钥。目前正在进行的项目是破解 RC5-72 密钥。

1.2.3　分布式系统的特征

1. 可靠性

可靠性指一个分布式系统在它的某一个或多个硬件的软件组件故障时，仍能提供服务的能力。当一个参与计算的机器发生故障时，可以立即被其他机器替代，不会阻碍请求任务的完成，这无疑成为分布式解决方案的有利条件之一。例如，当用户向一个大型电子网站发送一个普通的请求时，处理该请求的其中一台计算机即使发生故障也不会导致该请求被取消。一个显而易见的结论是，可靠性依赖于数据和软件组件两者的冗余性能。在极端情况下，即使一个购物车系统的整个数据中心被地震摧毁，也应当有另外一个备用数据中心供用户使用。显然，这种通过消除每一个单点故障，或多或少实现弹性恢复的服务，根据不同的应用规模是有相应的成本的。

2. 可扩展性

可扩展性是指一个系统为了支持持续增长的任务数量可以不断扩展的能力。由于数据容

量不断增加或者工作量不断增加，如交易的数量，一个系统会超出预期的规模，我们可能需要在不损失系统性能的情况下完成扩展。基于上例，可通过增加服务器数量的方式实现横向扩展，但是也可以考虑通过给每台服务器增加更多系统资源的方式实现纵向扩展。

为了区别两者，假设已经将一个应用程序的工作量分配给 100 个服务器。在理想的情况下，每台服务器持有 1/100 的数据资源，处理 1/100 的查询，现在假如增加了 20% 的数据资源，或者增加 20% 的查询数量，我们可以简单地增加 20 台服务器，这就是横向扩展，对并行处理程序几乎没有限制。我们也可以给这 100 台服务器增加额外的磁盘容量（为了存储增加的数据资源），增加额外的内存，或者更换更快的处理器（为了处理增加的查询数量），这是纵向扩展，通常对机器的限制比较高。

3. 可用性

使用单机处理任务时，当处理器出现问题或者关闭时会造成任务暂停，直到处理器被修复或者被替换，任务才能得以继续进行。可用性是一个系统尽可能地限制这种潜在风险发生的能力，会涉及两种不同的机制：快速检测错误机制和快速启动恢复程序机制。这种建立一个能够迅速发现并解决结点故障的保护系统的过程通常称为故障转移。

快速检测错误机制的关键在于定期检测每个服务器的状态，通常将此任务分配给任务管理者结点。如果没有一个特殊的管理者结点，那么通过分布式系统实现这种机制更加困难。P2P 结构的网络将其中一个结点定义为超级结点，专用于负责后台检测。P2P 认为，当其中一个结点离开网络，与之相关的结点应得到一个友好通知。这种假设有利于系统的设计。用这种方式解决类似的错误是可行的，但是对于大多数硬件上的错误不实用。

快速启动恢复程序机制通过复制（将数据复制到多台服务器）和冗余（每个实例连接多台服务器）来实现。在基础设施级别上提供错误管理服务是不够的。在这种环境上运行的服务必须通过采用适当的恢复技术来保存易失性存储的内容。

4. 高效性

我们如何估算分布式系统的效率呢？假设通过分布式的方式运行一个操作，系统会得出一个结果集合。有两种方式可以测算出它的效率，第一种是反应时间（时延），表示系统得到第一个结果的延迟。第二种是吞吐量（带宽），表示在一个给定的单位时间内所能交付的结果项的数目。这两种方式有利于证明一个系统在实际行为中是否合格，表现为一个网络流量的函数。这两种方式与下列单位成本变量有关：

1）消息的总数量：系统的所有结点所能发送的全部的消息的数量，不考虑单个消息的大小。

2）消息的总大小：代表数据交换量。

分布式数据结构支持的复杂的操作（例如，在一个分布式索引中搜索一个具体的键）可以表示为其中一个单位成本的一个函数。

一般来说，对一个分布式结构的分析简化为统计消息的数量，这种方式太简单，忽略了很多方面的影响，包括网络拓扑结构、网络负载及其变化，以及硬件和软件在参与数据处理和路由时可能的不统一性等。然而，开发一个精确的开销模型，准确地考虑所有这些性能因素是一个困难的任务。

5. CAP 理论

在构建分布式系统时，如何使一个处理大量事务的大型数据仓库系统所能提供的服务是