普通高等教育“十一五”国家级规划教材

重 点 院 校 推 荐 教 材

分布式系统

FENBUSHI XITONG

李西宁 编著

科学出版社
www.sciencep.com

普通高等教育“十一五”国家级规划教材

重 点 院 校 推 荐 教 材

分 布 式 系 统

李西宁 编著

科学出版社

北 京

内 容 简 介

本书是作者在国外多年讲授分布式系统课程经验积累的基础上，结合国内现状编写的教材。

本教材以传统的分布式系统理论和经典算法为基础，重点介绍系统的构成模块、实现方案以及存在的问题。具体的构成模块包括网络、通信、并发计算、域名服务、同步与互斥、时间与协作、分布式事物处理、复制技术、容错机制以及安全机制。本教材还简要地介绍了该领域里的新进展，如移动软件代理、P2P 系统、网格计算、万维网服务等。作者试图采用浅显易懂的语言描述分布式系统的原理与实现，做到概念清晰，深入浅出。为方便学生复习、掌握书中所学知识，每章末附有丰富的习题。本书既保证内容的前后呼应，又力图做到每一章节自成体系，教师可根据学生的不同基础和需要，适当进行裁剪。本书配有中/英双语电子教案，有教学需求的教师可到科学出版社网站上下载（http://www.sciencep.com）。

本书可用作高等院校计算机及相关专业本科高年级或研究生一年级的教材，亦可供科研工作人员参考。

图书在版编目（CIP）数据

分布式系统/李西宁编著. —北京：科学出版社，2006

普通高等教育“十一五”国家级规划教材 重点院校推荐教材

ISBN 7-03-016760-0

Ⅰ. 分 … Ⅱ. 李 … Ⅲ. 分布式操作系统- 高等学校- 教材

Ⅳ. TP316.4

中国版本图书馆 CIP 数据核字（2006）第 141634 号

责任编辑：鞠丽娜/责任校对：赵 燕

责任印制：吕春珉/封面设计：三函设计

科学出版社 出版

北京东黄城根北街 16 号

邮政编码：100717

http://www.sciencep.com

铭浩彩色印装有限公司印刷

科学出版社发行 各地新华书店经销

*

2006 年 12 月第 一 版 开本：787×1092 1/16

2006 年 12 月第一次印刷 印张：20

印数：1—4 000 字数：474 200

定价：32.00 元

（如有印装质量问题，我社负责调换〈环伟〉）

销售部电话 010-62136131 编辑部电话 010-62138978-8002

序

起源于 20 世纪 70 年代中期的分布式系统已经历了近 30 年的开发与研究，从简单的多机文件共享到广义的资源共享，从单一的计算模型到多种多层次的计算模型，从封闭的局部网络到开放的全球网络，分布式系统已演化成近代计算机系统的基本组织结构，支持非常广泛的工业、商业应用。分布式系统自身也从学术界走入商业化，日益丰富完善，日益规范成熟。

概言之，分布式系统是一组协同工作的计算机。这组计算机被网络连接，用通信的手段进行协调同步，用合理的算法调度分配资源，从而达到高效可靠的计算。当然，分布式系统是一个泛指词。如果细分的话，我们可以有不同种类不同功能的分布式系统。例如，以分布计算为主的系统一般采用紧密耦合计算机系统，或者是共享内存的多处理器，或者是用高速网络相连的一组同构计算机。而另一方面，以网络服务为主的系统则面临多种多样的计算设备。这些设备可以是计算机、无线电话、传感器乃至家用电器。它们可以形成一个局域网，也可以开放到一个广域网。此外，现代分布式系统一般是在网络操作系统外层增加一层软件，亦称为“中间件”。用中间件实现的分布式系统易于标准化，使得不同厂商生产的软/硬件在用户面前呈现出友好的、一致的界面。

与单机系统相比，分布式系统具有几个显著的特征：共享性、开放性、并发性、可调节性、容错性以及透明性。资源共享是分布式系统的首要特征。资源可以是数据，可以是软件，也可以是硬件。例如，在客户/服务器模型里，共享的资源就是服务器所提供的各种服务。开放式的分布式系统遵循一套标准的协议和界面为用户提供服务，其主要优点是灵活性，在不影响现存服务的前提下，新的共享资源可被安全地扩充到系统中。顾名思义，并发性指的是在同一时间处理多个任务，其优点在于提高效率。可调节性的含义很广，不仅一个分布式系统的功能可调节，其网络规模亦可调节，管理结构也可调节。容错性关系到一个分布式系统能否可靠地运行，当事故或异常事件发生时，具备容错能力的系统可以自动排除故障并恢复运行。为了方便用户使用，一个分布式系统要尽量透明，要尽可能地隐藏系统的内部细节，使之作为一个整体展示在用户面前，这样才能易学易用易管理。

分布式系统并不是一种抽象的概念。要搞清楚分布式系统的原理及设计，我们不仅要探讨相关的理论基础，也要通过算法设计与分析来理解系统实现中的具体问题。在了解了分布式系统的目标和基本模型的基础下，我们将引入分布式系统的构成模块，重点介绍现存的解决方案以及存在的问题。具体的构成模块包括网络、通信、并发计算、域名服务、同步与互斥、时间与协作、分布式事务处理、复制技术、容错机制以及安全机制。通过学习，学生们应能掌握设计和实现分布式系统的基本知识和技能，并能把学到的知识运用到实践中。此外，本书最后一章还简要地介绍了分布式系统研究领域里的新进展，使得学生们在掌握基本原理的同时，进一步了解分布式系统研究

的新领域和新方向。

本书试图采用浅显易懂的语言描述分布式系统的原理与实现，力图减少对读者知识背景的要求。由于本书的内容颇为广泛，涉及程序设计、操作系统、数据库、网络通信、安全保密以及算法复杂性等，因此我们希望读者具备这些领域的基础知识。本书按照36（课堂教学）学时设计，可用作研究生或者本科高年级学生的教材。根据学生的不同基础，可适当进行删节。此外，本书每章之后都附有一组练习题，题目可分为三类：① 基本概念复习题；② 快速思考题；③ 作业题。为便于多媒体教学，本书配有中/英双语电子教案，有教学需求的教师可到科学出版社网站上下载（http://www.sciencep.com）。

南京大学孙钟秀教授，金志权教授，茅兵教授在百忙之中抽出时间阅读本书的初稿，并对书中内容和术语提出了许多有益的建议；科学出版社的鞠丽娜老师以及南京大学的徐洁磐教授、郑国梁教授在本书出版过程中给予了热情的帮助，在此由衷地向他们表示感谢。

由于作者的学识有限，书中难免有不当之处，恳请同行专家和广大读者提出宝贵意见。

作　者

2006年7月

目　录

第一章 引 论

计算机的发明给人类生活带来一场新的革命。如果说 19 世纪的机械革命把人们从繁重的体力劳动中解放出来，那么由计算机引发的信息革命则把人们从繁重的脑力劳动中解放出来。换言之，如果说机械是人手的外延，那么计算机就是人脑的外延。自从人类于 1945 年制造出第一台计算机以来，将近 40 年的时间里，计算机一直是神秘而昂贵的。大多数公司、科研机构或学校里的计算机都自成系统，独立工作。1965 年，英特尔公司的奠基人之一，戈登·莫尔（Gordon Moore）发现了一个有趣的现象：在大规模集成电路的芯片上，每平方英寸里集成的晶体管数量几乎每年翻一番。莫尔预言这种现象将会延续相当长的一段时间。到了 20 世纪 80 年代中期，个人计算机开始崭露头角。随着集成度日益增高，价格日益下降，如今计算机不仅进入寻常百姓家，而且其计算能力可与早年的大型计算机媲美。莫尔 60 年代的预言成了现实。尽管这种翻番的周期稍长了一些，从一年变成 18 个月，但人们已经普遍接受了莫尔的预言并称其为莫尔定律，包括莫尔在内的许多专家一致认为这种现象至少还会持续到 2020 年。正因为如此，计算机界广泛流传着这样一个幽默：如果汽车制造业的发展速度能像计算机工业的话，我们今天花几块钱就可以买到一辆高级轿车，而且只用一升汽油就可以跑上几千公里。虽然这种比喻并不太恰当，但可以从中体会到计算机工业惊人的发展速度。

计算机普及的另一个重要因素是因特网的发明。60 年代末美国国防部研制出第一个大型计算机网络 ARPANET，而真正意义上的计算机网络普遍应用却是出现在个人计算机崛起之后的局域网（LAN，local area network）。在一个相对较小的地理区域内，局域网把一组计算机连接在一起，于是这些计算机可以互相交换文件和信件，也可以共享文件服务器或打印机一类的设备。在局域网的基础上，广域网（WAN，wide area network）亦应运而生。广域网把四处分散的局域网连接成一个整体，从而达到更大范围里的信息共享。早年的 ARPANET 逐渐演化成世界上最大的广域网，也就是众所周知的因特网（Internet）。通过因特网，人们可以及时获取发生在世界各地的新闻，方便地查询各种图文并茂的资料，实时地订购物美价廉的物品，直接从事财经商业活动，即时与陌生的朋友网上聊天，快速下载多媒体的音乐录像……，不夸张地说，人类如今的生活已经和因特网息息相关，不可分离。

随着计算机技术和网络技术的发展，起源于 70 年代中期的分布式系统也得到充分的开发与研究，从简单的多机文件共享到广义的资源共享，从单一的计算模型到多种多层次的计算模型，从封闭的局部网络到开放的全球网络，分布式系统已演化成近代计算机系统的基本组织结构，支持非常广泛的工业、商业应用。分布式系统自身也从学术界走入商业化，日益丰富完善，日益规范成熟。

1.1 分布式系统的定义

什么叫做“分布式”系统呢？不同的教科书里有不同的定义，不同的文献里也各执

已见，众说纷纭。这里我们选取两个较为流行的定义，值得指出的是，对分布式系统的定义并不像数学定义那样严谨，只是用自然语言描绘出分布式系统的基本特征：

1）一个分布式系统是一组由网络连接的具有独立功能的计算机，在一套特殊软件的管理下，整个系统在用户面前呈现为一个透明的整体。

2）一个分布式系统是一组位于网络计算机上的并发构件，这些构件之间的通信以及任务协调都只能通过信件传递进行，其目的是实现资源共享。

这两个定义大同小异，实际上都隐含了对硬件和软件的描述。从硬件的角度来说，我们必须有一个计算机网络，网络中的每一台计算机都必须是自治的（autonomous），也就是能够独立运行的。从软件的角度说，我们必须要尽可能地隐藏硬件以及网络的细节，在用户面前整个系统就像是一台功能巨大而且易于管理使用的虚机器，从而实现系统范围内的资源共享。构成一个分布式系统的硬件成分可以多种多样。例如网络规模可大可小，可以是快速开关电路，可以是局域网，也可以是广域网。同样，连接在网络上的计算机可以具有不同的速度及外部设备配置，可以是同构计算机（即所有计算机采用相兼容的中央处理器），也可以是异构计算机。虽然分布式系统的软件一般与应用相关，但有几点设计原则是应该普遍遵循的。例如我们必须隐藏异构计算机之间的差异，必须隐藏计算机网络上系统内部的通信，这样才能在用户面前展示出一个统一的完整的界面。另外，一个分布式系统必须通过软件给用户提供一组应用程序设计接口（API，application programming interface），这样用户才能开发出各种各样的应用程序。

1.1.1　历史背景

计算机硬件的飞速发展起始于 70 年代中期。从那时候起，我们开始有了超小型和微型计算机。在这以前，计算机都需要安装在一种特殊的环境里，如带有空调的机房，稳定的供电系统等。而微型机对运行环境没有特殊要求，可以安装在任何人们日常工作的环境里。到了 80 年代末，微型计算机的性能/价格比越来越好，人们也逐渐用另外一个更确切的名字替代，称为个人计算机（PC）。与此同时，超小型计算机也变得体积更小，功能更强，人们也给这类计算机起了个名字，叫做工作站（workstation）。为了使不同的计算机用户之间彼此通信，计算机网络技术也在 70 年代有了长足的进展。IBM 公司在 1974 年研制出第一个分时用户网 SNA，数字设备公司于 1975 年推出 DECnet 网，西门子公司在 1978 年推出了 TRANSDATA 网，这些网络都是基于存储转发（store and forward）的传送协议。存储转发策略需要计算机之间有专门的点对点的通信线路，这样无疑使得网络价格颇为昂贵。70 年代初期，Xerox 公司的研究人员推出了一种基于共享通信电缆的网络，在这种网络中所有的计算机都连接到同一根同轴电缆上，这就是著名的以太网（EtherNet）。由于以太网通信速度高而且安装维护成本低，以以太网为主流的局域网很快就得到了迅速的普及。

随着计算机硬件和网络技术的发展，人们从中看出一种巨大的值得发掘的潜力，这就是如何使得连接在一起的计算机作为一个整体，共同协作，从而达到资源共享，发挥更大的计算效力。这便是提出分布式系统模型的初衷。30 年来，计算机科学家们开发研制了各种各样的分布式系统，解决了许多从理论到实践中的问题。这里我们从众多的分布式系统中选取几个里程碑式的典型系统（见表 1.1），并对这些系统做一简单描述。

表 1.1 分布式系统的里程碑

系统名称	组织机构	网络要求	计 算 机	研制日期
CM*	卡内基·梅隆大学	层次总线	PDP	1975
Cambridge DCS	剑桥大学	剑桥环	LSI-4	1979
Locus	加州大学洛杉矶分校	以太网	PC	1980
V System	斯坦佛大学	以太网	Sun	1982
Mach	卡内基·梅隆大学	以太网	Sun, PC	1985
CORBA	OMG	因特网	任何机器	1990
Distributed COM	微软公司	因特网	PC	1996
JINI	Sun Microsystems	因特网	任何机器	2000

1）CM*是卡内基·梅隆大学早期研制的具有代表性的分布式系统。严格地说，CM*本身是一个采用层次总线结构的多机系统，由大约 50 台左右的 PDP 构成。在该系统之上，覆盖了一层基于消息传送的多机操作系统 StarOS。这个操作系统是现代分布式操作系统的雏形，其目的是控制多个并发进程协同完成一个任务。该系统的特点是可靠性，一台机器的故障不会影响到系统内的其他机器。为了支持进程间快速通信，StarOS 引入了信箱概念，用以缓冲不同类型的信件。

2）早期分布式系统中另一个经典性系统是剑桥分布式计算机系统（Cambridge DCS）。剑桥大学计算机试验室的科学家们首先设计了分布式系统赖以生存的硬件基础，即计算机网络。这个网络是一个封闭环，所有的终端、服务器以及外部设备都连接到这个环状网络上，他们把这个网络命名为“剑桥环”。其后，他们又开发了分布式系统软件。在一些固定地址的服务器上，DCS 系统提供各种各样的系统服务，诸如打印管理、磁带管理、用户认证以及文件管理等。此外，DCS 还提供一个处理机库（processor bank）用以进行分布式计算。尽管剑桥分布式计算机系统没有得到普遍的商业应用，但该系统提出的模型和概念，如环形网、服务器等却被其他分布式系统所借鉴。

3）加州大学洛杉矶分校研制的 Locus 系统是一个面向分布式计算的系统。系统设计的主要目标是透明性，即在用户和应用程序面前隐藏分布式系统的内部细节。该系统以一个局域网（以太网）为运行环境，主要特点是提供一个完全透明的分布式文件系统来支持分布式进程。对这些分布式进程来说，存储在不同计算机上的文件就像是起源于同一个根目录的树形文件系统，任何一个文件都可以任意地改变其存储位置而不会影响到进程对它的存取访问。Locus 的另一个主要特点是灵活性，应用程序可以根据需要自由地创建生成分布式进程（近程或远程），而且进程可以在同构的计算机之间移动。这种进程迁移的思想无疑是如今“移动软件代理”模型的雏形。

4）V 系统是斯坦佛大学开发的分布式操作系统，它的运行环境是一组连接在以太网上的 SUN 工作站。该系统由一个相对较小的系统内核和一组服务模块所构成，同时还提供各种各样的应用程序库和一组控制管理命令。V 系统的关键在于高效率的进程迁

移，它引入了一种称作“预复制”的技术，可以在最短的时间内完成一个进程在计算机之间的迁移。V 系统还采用了轻量级进程（lightweight process）的概念，这些进程共享同一个地址空间，通过同步信件通信达到协同工作。V 系统是一个完整的分布式操作系统，对用户而言，系统的文件访问、进程并发以及计算机的物理位置等都是完全透明的。

5）卡内基·梅隆大学于 1985 年开始研制 Mach 分布式操作系统。这个系统的胃口很大，不仅要支持紧密耦合（closely coupled）的多处理机系统，也要支持松散耦合（loosely coupled）的多机系统。它还希望能够适用于各种各样的商业化工作平台。该系统的特点是它把进程间的通信机制和线程机制都设计在系统的内核中，从而达到高效的进程通信和线程管理调度。Mach 采用微内核结构，而把大量的服务工作交付给各种用户级的服务器。因此，这些用户级服务器需要提供丰富的应用程序设计接口（API）。虽然 Mach 本身只是学术界的科研项目，它的许多思想和模块却被商业化系统所利用，如 NeXT OS 和 IBM OS/2 就从中获益匪浅。

6）80 年代以来，面向对象（object oriented）技术得到了普遍认可。这种技术可通过封装、继承以及多态等机制为传统的程序设计提供一套全新的设计方法，而且大大改进了代码重用功能。随着面向对象技术的深入研究，分布式对象系统也脱颖而出。分布式对象系统的主要目的是提供一个标准的构件框架，使得不同厂家开发的 OO 软件能够通过不同的地址空间、网络和操作系统而交互访问。CORBA（common object request broker architecture）就是对象管理组织（OMG）所提出的一组标准规范，符合这个规范的对象可以互相交互，不论它们是用什么样的程序设计语言定义的，也不论它们运行于什么样的机器和操作系统。

7）微软公司的分布式组件（DCOM）扩展了组件对象模型，使得不同的 Windows 组件对象可以在不同的计算机上通过局域网、广域网或因特网上相互通信。目前通用的操作系统彼此屏蔽。当一个用户进程需要和一个远程进程中的组件进行通信时，它不可能直接发出调用，而必须遵循操作系统所定义的一套通信协议。DCOM 试图以一种透明的方式解决这个问题。它采用远程对象模型，每一个 DCOM 对象可以实现一组对外接口，通过这组接口支持对该对象的动态调用。尽管 DCOM 的实现不很规范而且比较复杂，也不像 CORBA 那样严谨，但世界上有千百万的 Windows 用户，DCOM 实际上被广泛使用，而 CORBA 和其他的分布式系统要想得到普遍应用，还有相当长的一段路要走。

8）JINI 是 SUN 公司推出的以 Java 为核心的分布式系统。它通过使用一个简易的即插即用（plug and play）模型，能够随时改变硬件或者软件的配置，从而提供了一个支持快速动态配置的分布式计算环境。JINI 的主要特点在于它能够使各种数字设备无需特殊安装或者人工干预，就能够在一个临时的称为服务联盟（federations of services）的环境中共同工作。联盟中的每一个成员都可以为其他成员提供资源或服务，同时又可以从其他成员那里获取自己所需的资源和服务。JINI 本身与硬件无关，安装 JINI 的数字构件不再受到所用软件、处理器、设备驱动器或传统网络协议的制约，其唯一要求只是一个能够运行 Java 的虚拟机。

1.1.2　分布式系统的应用

分布式系统的应用领域极为广泛，从通常意义上的分布式计算到电子商务（旅游、

定票、购物、个人银行等），分布式系统的应用几乎渗透到计算机应用的每一个角落。仅以分布式计算为例，我们可以看到分布式系统所蕴藏的巨大潜力。保守地说，任何一所现代化的大学里都拥有至少上百台的计算机。然而，90%的时间里这些计算机却无所事事。即便有人坐在计算机旁，大多数也都是在作文字处理，浏览网页或者下载文件。到了后半夜，这些计算机则完全空闲。这意味着每天都有无法估量的计算能力被白白浪费掉了。当真我们不需要这些计算能力吗？回答显然是否定。在人类探索自然的过程中，我们有太多的问题需要解决，而解决的方法则往往需要巨大的计算能力。分布式系统的应用之一就是通过网络技术把被浪费的计算资源充分利用起来。下面我们就看看几个实际的例子。

1. 搜索外星文明

世界上一个规模巨大的分布式计算项目是美国加州大学伯克利分校的搜索外星文明（SETI，search for extraterrestrial intelligence）的科学试验。这个试验的目的是通过对电磁波信号的分析来寻找其他星球上可能存在的具有文明智慧的生命，因为从目前的技术水平来看，探索“外星人”是否存在的有效手段是对来自遥远星球的电磁信号进行研究。当然，即便外星人向我们发出无线电信号，这些信号经过漫长的旅行到达地球时会变得非常微弱。为了接收各种微弱的无线电信号，科学家们发明了射电天文望远镜。目前世界上最大的射电望远镜是设置在波多黎各的 Arecibo。这台望远镜始建于 1963 年，其巨大的抛物面天线直径达 305m，可以接收来自太空的 400 万个波段的无线电信号。尽管伯克利大学拥有数台大型计算机来实时分析处理所获取的信号，但面对如此庞大的数据，这些计算机仍显得力不从心，只能选取比较强的并且具有代表意义的一小部分信号进行分析。显然，这样做势必会忽略掉某些真正有意义的无线电信号。为了能够分析所有微弱的信号以及不同的信号类型，SETI 号召分布于世界各地的计算机用户参与这个伟大的试验。参与的方法很简单，只要求参与者们下载一个类如屏幕保护程序那样的特殊软件。SETI 的专家们还设计了一套程序，将庞大的数据分割成细小的数据段，每个数据段都代表着一小块天空区域和某个波段的频率。SETI 把这些数据段发送给参与者的计算机，而所下载那个程序便自动地开始对数据进行分析。然而，这个程序并不抢占参与者的机器时间，当你工作时它自动停止，当你离开机器时它便出来利用这段空闲时间。实际上，当这种屏幕保护程序运行时，你的计算机已经加入到寻找外星人的行列中。

2. 分布式破译密码

另一个有趣的分布式应用来自于 distributed.net 的 RC5 项目。这个项目的组织者给出各种密码，要求参与这个项目的个人或单位在最短的时间里破译，获胜者可得不菲的奖金。要想破译密码，一般是先找到解读密码的“密钥”，通过密钥把密码翻译成人们能理解的信息。一种最为直截了当的方法就是把所有可能的密钥组合都拿来尝试计算一番，选取其中那把正确的密钥。可是这种计算的工作量却远非一台或数台计算器所能胜任。例如，假定一把密钥由 64 位二进制码构成，则我们需要尝试的密钥组合就高达 2^{64}，也就是说要有 18，446，744，073，709，551，616 个密钥的可能性。于是 RC5 的组织

者号召全世界的人们参与这个项目，下载 RC5 所开发的软件，利用参与者的计算机的空闲时间来破译密码。至今为止，RC5 已经发起 3 次密码破译，分别为 RC5_56，RC5_64，RC5_72。密码 RC5_56（56 位密钥）花费 250 天破译（1997 年）。其后，RC5_64 用了 1757 天，也就是将近 5 年的时间才被破译（2002 年）。现在，参与者们正在攻克 RC5_72。据该组织所公布的消息，目前 RC5 项目所拥有的计算能力相当于 16 万台奔腾计算机。

3. 高性能数据网格

在物理领域里，分布式计算也扮演着重要的角色。欧洲核研究组织（CERN）拥有世界上最大的粒子加速器。在那里，来自世界各地的物理学家们试图通过对粒子的研究而寻找宇宙的起源。目前他们正在建造一台大型强子碰撞机（large hadron collider）。一旦这台 LHC 投入使用（计划在 2007 年），必将使物理学家们面临大量的数据和信息。根据估计，大约每年产生的数据量高达 1000 万 GB。如果用光盘存储这些数据的话，则需要 250 万张 DVD 光盘。可想而知，CERN 一定是分布式系统的最大用户，因为无论多么先进的单机系统都无法胜任对这些数据的处理和分析。目前 CERN 正在和 IBM 联手研究利用分布式网格系统的可能性。据称，IBM 的分布式虚拟存储和文件管理技术将在这一合作中发挥重要作用。他们将最终创建一个空前庞大的高性能数据网格系统，这些数据可存放在不同地点，可基于不同操作系统，可动态扩展并且易于管理，藉以协助 CERN 的科学家探索有关物质和宇宙本质的基础问题。

在计算机领域里，当研究者提出一种新的思想、一种新的系统模型时，第一个需要回答的问题就是这种模型下是否存在杀手应用（killer application）。杀手应用的含义并不同于“杀手锏”，而相仿于“绝活儿”，表示就此一家，绝无分店。如果没有杀手应用，则这个模型或者没有存在的必要，或者没有太大的学术价值。不难看出上面的 3 个例子都是分布式系统的杀手应用，因为这些问题都是单机系统（甚至超级计算机）根本无法解决的。像这样的例子还有很多，如对人类基因的全面研究，对天气和气候的准确预报，对经济和财政的风险预测等。事实表明，分布式系统的开发与研究不仅是学术界里近 30 年来最活跃的领域之一，几乎在所有的计算机应用领域中，分布式系统都展现出其朝气勃勃的生命力。

1.2　分布式系统的显著特征

要想设计一个适用的分布式系统，则首先确定分布式系统具备哪些显著特征，以及在这些特征之下，我们有哪些用户需求必须得到满足。这两个问题实际上是任何系统设计中都要解决“共性”与“特性”的问题。显然，分布式系统的首要目的是资源共享（resource sharing），这是所有分布式系统所具备的共性。在这个大目的的前提下，我们希望分布在网络中的资源呈现出一种透明性（transparency），也就是说，用户无需了解系统内部的细节就可以享用可利用的资源。同时，我们还希望拥有一个开放性（openness）的系统，用户可以自由地选取各种各样的系统服务，而这些服务遵循一种一致的友好的用户协议。为了适应不同的应用，一个分布式系统还要具备可调节性（scalability），于是系统的规模可根据应用的要求进行裁剪，而且系统服务也可以自由地增加或删除。从

用户的角度来说，除了上述的共性之外，不同的用户或许还有自己特殊的要求。例如有的用户希望高度保密性，有的用户要求高性能计算，也有的用户需要数据的可靠性、完整性和一致性。

我们将在本节中详细讨论分布式系统的基本特征，即资源共享、透明性、开放性和可调节性。固然系统设计者应该满足客户的各种需求，但我们无法也不可能在教科书中列举所有的用户需求，这里我们只大致地讨论一下实践中经常遇到的比较重要的用户需求。

1.2.1 基本设计目标

分布式系统的主要目的是使用户方便简捷地访问远程资源，从而达到某种程度的资源共享。什么叫做“资源”？这个词看上去相当抽象，但却恰如其分地刻画了分布式系统中可被共享的那些东西。资源可以是任何事物：它可以是硬件，如磁盘、打印机、处理机、存储器、传感器或通信线路；也可以是软件，如进程、文件、视窗、网页或数据库。

1. 资源共享

1）资源共享的经济效益是不言而喻的。在一个公司里，雇员们共享一台或数台打印机显然要比给每人配置一台打印机要划算得多。同样，公司的数据库也不可能安装在每一个雇员的计算机里，那样将很难保证数据的一致性，只有通过资源共享才能使雇员们安全可靠地访问公司信息。

2）资源共享的另一个优点是便于协同工作。大多数公司，尤其是软件开发公司都依赖“团队”工作模式。在一个项目的开发周期里，要有数人、数十人乃至数百人同时合作。于是，一个项目里的设计人员不仅要访问其他设计人员的程序和数据，而且还享用同样的开发工具和管理工具。例如，整个系统只需维护一套编译程序、库程序和其他辅助工具，一旦对这些程序更新换代，所有的设计人员马上就能使用新的版本。

3）资源共享一般是通过一个称作“资源管理程序”的模块实现的。不同的资源可能需要不同的管理方法和访问认证策略。首先，我们对所管理的资源要有合理的命名，这样资源管理程序才能把资源的名字影射成资源所处的物理地址，并且协调对资源的并发访问。资源管理程序有两种常用的实现模型，一种是客户/服务器模型，另一种是面向对象模型。我们将在本书后面的章节中详细讨论这两种模型。

4）在实现资源共享时，我们必须要妥善地考虑系统的安全性。目前的许多系统都存在安全方面的弱点，导致资源被无端地破坏或被任意地滥用。例如，当用户访问资源时，用户名和口令都在网络上传输。如果这些信息没有加密，则很容易被别人窃取。分布式系统的安全性研究是一个极为重要的课题，本书亦将有专门的章节进行探讨。

2. 透明性

分布式系统的目的是资源共享，而实现资源共享的目标之一是系统的透明性。我们希望一个系统在用户面前呈现为一个透明的整体，而不是一组支离的构件。固然，由网络相连的一组相互分离的构件是分布式系统的基本属性。正是这种相互分离彼此独立的属性才

使我们能够进行并发计算、资源共享以及冗余容错。但是，我们不愿意让系统的用户看到或者觉察到这种内部的属性。通过各种隐蔽技术，使得一个分布式系统就像是一台功能完备的计算机，这才是透明性的真正含义所在。那么，透明性需要考虑哪些因素呢？根据国际标准化组织(ISO)于1995年所颁布的开放分布式处理参考模型(RM-ODP，reference model for open distributed processing)，分布式系统的透明性包括8种形式，见表1.2。

表1.2　ISO RM-ODP所定义的8种透明性形式

透明性	描　述
访问	隐蔽数据表达方法和资源访问方法的不同之处
位置	隐蔽资源所处的物理位置
迁移	隐蔽资源的物理移动
重定位	隐蔽正在使用的资源迁移
复制	隐蔽资源的复制
并发	隐蔽若干用户共享同一资源所产生的竞争
故障	隐蔽资源的故障与排错恢复
持续	隐蔽软件资源所处的存储空间：内存或磁盘

1）访问透明性试图隐蔽不同计算机之间数据表达方式的差异和访问方式的不同之处。例如，因特网上的信息有的是用ASCII代码，有的采用Unicode编码。目前许多浏览器都能够自动进行识别而使用户感觉不到信息编码的差异。另外，当我们访问一个网页时，我们只需要点击那个网页的URL，而无需顾及这个网页是如何得到访问的，它可能来自于一个基于Windows NT的服务器，也可能来自于一个基于Linux的服务器，这其中的访问细节完全被隐蔽了。

2）无论是何种资源必须有一个物理位置。位置透明性就是让用户根本不知道他所访问的资源位于何处。当然，每个可访问的资源必须有一个逻辑名字，而且这个逻辑名字也必须能够影射到该资源的物理位置。我们所要作的就是隐蔽这种影射过程。例如，众所周知的eBay是目前世界上最大的网上拍卖网站，它的URL是www.ebay.com。仅仅从这个URL，我们看不出这个名字和该网站的物理位置之间存在任何联系。当我们访问这个网站时，有一个我们看不见的“域名服务器”自动地把网站名影射到网站的物理地址，这种隐蔽的影射过程就保证了WWW的位置透明性。

3）迁移透明性和重定位透明性密切相关。在一个分布式系统中，如果我们允许任一个资源改变其所处的物理位置，同时又不影响用户对该资源的访问方式，则这个系统满足迁移透明性。

4）如果某个正在被用户使用的资源发生物理迁移，而系统能够自动地跟踪并且自动地做出新的地址影射，丝毫不影响正在访问该资源的用户，这样的系统便满足重定位透明性。目前，无线网络和移动计算系统越来越普及，对这类分布式系统而言，迁移透明性和重定位透明性就显得格外重要。

5）为了使一个分布式系统更可靠更高效，我们往往把某种资源（尤其是信息资源）复制成若干副本，同时把这些副本分布到不同的物理位置上。例如，一些著名的网站通常都会有若干服务器，每个服务器都携带一个资源副本，这样才能应付对该网站的密集