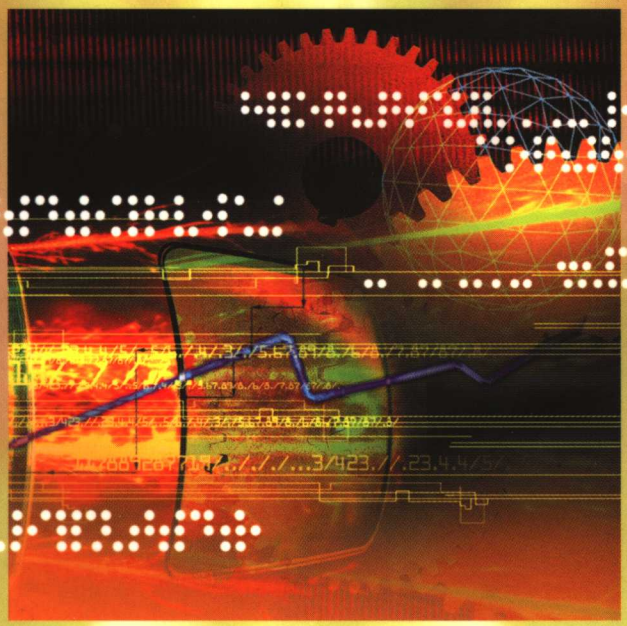
高等学校研究生系列教材

分布计算系统

Distributed Computing Systems

徐高潮　胡　亮　鞠九滨



高等教育出版社
HIGHER EDUCATION PRESS

高等学校研究生系列教材

分布计算系统

徐高潮　胡　亮　鞠九滨

高等教育出版社

内 容 提 要

本书介绍由计算机网络构成的分布计算系统的结构和实现技术，侧重于基本概念、基本原理和基本方法的讲授。全书共分为12章：分布计算系统的基本概念和体系结构，分布计算系统的进程通信，分布式程序设计语言，命名与保护，分布式同步和互斥机构，死锁问题及其处理技术，容错技术，分布式数据管理，分布式文件系统的设计问题与实现方法，分布式调度，分布式共享存储器技术以及基于对象的分布式系统。

本书比较全面和系统地反映了国际上在这一领域取得的主要成果。特别是最近几年的成果，强调了通信、容错、死锁、同步和互斥、并发控制、分布式调度及分布式对象方面的内容。

本书叙述详细、由浅入深，具有丰富的图示，力求做到既通俗易懂，又具有一定的理论深度。本书可作为高等院校本科高年级学生和研究生教材，也可供有关科技人员参考。

图书在版编目(CIP)数据

分布计算系统/徐高潮，胡亮，鞠九滨．—北京：
高等教育出版社，2004．1
ISBN 7-04-013309-1

Ⅰ．分…　Ⅱ．①徐…　②胡…　③鞠…　Ⅲ．分布式计算机系统　Ⅳ．TP338.8

中国版本图书馆CIP数据核字（2003）第113171号

出版发行	高等教育出版社	购书热线	010-64054588
社　　址	北京市西城区德外大街4号	免费咨询	800-810-0598
邮政编码	100011	网　　址	http://www.hep.edu.cn
总　　机	010-82028899		http://www.hep.com.cn
经　　销	新华书店北京发行所		
印　　刷	北京二二〇七工厂		
开　　本	787×1092　1/16	版　　次	2004年1月第1版
印　　张	27.25	印　　次	2004年1月第1次印刷
字　　数	530 000	定　　价	33.80元

本书如有缺页、倒页、脱页等质量问题，请到所购图书销售部门联系调换。

前　言

由于计算机技术和网络技术的快速发展以及分布计算系统的应用领域日益扩大，人们对分布计算系统的研究兴趣日益增加，现在，分布计算系统已经是一个非常热门的研究领域。在过去的10年中，分布计算系统的理论体系和相关技术更加成熟，许多研究者在此领域进行了大量的研究工作，并取得了大量的研究成果。在此期间，作者在教学过程中也积累了许多新的素材，同时也积累了自己在分布计算系统方面的一些研究成果，本书力求反映这些最新的研究成果。分布式应用已涉及到社会生活的各个方面，越来越多高等院校的计算机系开设了分布计算系统这门课程。

本书所选用的参考文献主要有：《Distributed Systems: Principles and Paradigms》(Andrew S. Tanenbaum 著，清华大学出版社2002年影印版)、《分布式系统设计》(吴杰著，机械工业出版社2001年中译本)、《Distributed Operating Systems: Concepts and Practice》(Doreen L. Galli 著，人民邮电出版社影印版)、《机群计算》(鞠九滨等著，吉林大学出版社1999年)等。为了保证本书内容的正确性、准确性和时代性，作者选用参考资料时，尽量选用了最近出版的有权威性的原著和最新的研究成果，包括文献和专著。在此，我们对所引用著作的作者和出版者深表感谢。

本书介绍了用计算机网络组成的分布计算系统的结构和实现技术，侧重于基本概念、基本原理和基本方法。全书共分为12章：分布计算系统的基本概念和体系结构、分布计算系统的进程通信、分布式程序设计语言、命名与保护、分布式同步和互斥机构、死锁问题及其处理技术、容错技术、分布式数据管理、分布式文件系统的设计问题与实现方法、分布式调度、分布式共享存储器技术以及基于对象的分布式系统。

读者应该在学过计算机网络、计算机组成原理和操作系统之后学习本课程。如果读者已经学过网络程序设计课程，则2.1节和2.2节可以略去不学。目录中带有*标记的部分可选择讲授。

清华大学郑纬民教授审阅了全稿，对本书提出了不少宝贵意见，作者在此表示诚挚的谢意。

衷心希望读者指出错误和提出改进意见。

作者

吉林大学计算机科学与技术学院

2003年9月

策划编辑 刘建元
责任编辑 武林晓
封面设计 王凌波
责任印制 宋克学

郑重声明

目　录

第一章 绪 论

1.1 为什么需要分布计算系统

20世纪90年代以来,两大技术的快速进步导致了人们对分布计算系统的兴趣迅速增加,使得分布计算系统成为一个热点研究领域。这两大技术进步分别是:第一,计算机硬件技术和软件技术的发展;第二,高性能计算机网络技术的发展。这两大技术的进步不仅极大地改变了人们使用计算机的方式,同时使得人们对尽可能多的计算能力、数据的透明访问、高性能和高可靠性的目标的追求不再是梦想。

计算机技术的发展可以通过人们使用计算机的方式的改变来描述。在20世纪50年代,程序设计人员使用计算机前必须预约上机时间,他们使用计算机时,会占用全部计算机资源。60年代出现了批处理技术,人们提交作业并排队等候处理,计算机每次运行一个作业,用户晚些时候来取结果。70年代产生了分时系统,人们可同时使用一台计算机,而每个用户都觉得他在单独使用整个计算机。80年代是个人计算机的时代,每个用户有一台属于他自己专用的计算机。现在,不仅每个用户有自己专用的计算机,而且通过计算机网络可同时使用多台计算机。

使用计算机的方式由预约上机方式转变为批处理方式的首要技术条件是必须有足够大的计算机存储器。只有存储器做得充分大,才可以装得下整个操作系统以及待处理的应用程序。在批处理方式中,缓冲、假脱机和多道程序的脱机处理等技术得到了广泛应用。需要批处理技术的原因是它可以有效地利用昂贵的计算机机时,减少了计算机的空闲时间。

出现分时系统的主要技术条件是计算机变得更为便宜而且功能更为强大。使用分时系统的原因是因为它可以使程序设计人员获得较高的效率,也更进一步地提高了计算机的利用率。分时系统是迈向分布计算系统的第一步,用户可以在不同的地点共享和访问资源,但分时系统给用户提供的是一个集中的共享环境,允许很多设备如打印机、存储器以及软件和数据共享。

超大规模集成电路和高性能计算机网络技术的发展使得个人计算机代替了分时系统。由于这时每个用户有自己专用的计算机,并且可通过计算机网络共享和访问其他的计算机资源,使得用户在任何时间都能保证使用所需的计算能力成为可能,但此时提供给用户的是

一个基于网络的、分散的共享环境。

与分时系统的共享环境相比,在基于网络的分散的共享环境中,用户在系统的使用和管理上会遇到新的问题。在使用上,个人计算机的操作系统软件允许在网络上的个人计算机之间复制文件和远程登录,要求用户必须知道本地对象和远程对象之间的差别,以及对象放在哪个远程机器上。如果网络的规模很大,这个问题将变得很严重。

在系统管理上,分时系统的操作人员每天进行文件系统备份操作,系统管理员把可用的处理机分配给最需要处理机的用户,系统编程人员只需简单地安装新的或已改进的软件。但在由个人计算机互连的网络环境中,个人计算机的用户必须既是一个操作员,又是系统管理员和系统编程人员,在拥有成百上千台个人计算机的建筑物中,操作人员不可能四处奔走进行文件备份操作。系统编程人员也不再能很简单地把新的软件放到文件系统中。

针对上述使用和管理上出现的新问题,虽然也有一系列的解决方法,但没有一个能像使用分时系统共享环境那样令人满意。例如,最常用的一种方法是在个人计算机系统中增加一个网络复制命令,使用此命令把文件从一个计算机复制到另一个计算机上。另一种稍好的方法是使用网络文件系统,它允许文件在某种程度上实现真正共享。但在所有的方法中,除极少数以外,用户总要知道本地操作和远程操作的差别。出现这些问题的最主要原因在于,传统的操作系统(它仍然是现代个人计算机软件的基础)从来都不是为具有多个处理机和很多文件系统的环境所设计的,这种环境需要分布式操作系统。

在分布计算系统中,多台计算机构成一个完整的系统,其行为类似一个单机系统。即登录到系统的用户不必了解系统中有多少台机器,它们的位置在哪里,它们的功能是什么,文件在哪里,作业在哪一台机器上运行等任何有关硬件物理分布的细节。分布式操作系统是实现分布计算系统的核心。

以上是从人们使用计算机方式的变化这个角度上分析了为什么需要分布计算系统。其实人们需要分布计算系统还在于它和单机系统相比具有许多优点。

1.2 分布计算系统的相关概念

1.2.1 什么是分布计算系统

分布计算系统又叫分布式计算机系统,简称分布式系统。对于分布计算系统,不同的研究者给出了不同的定义,但没有一个是完全令人满意的,也没有一个被所有的研究者所承认的。例如,Andrew S. Tanenbaum 教授给出的定义如下:分布计算系统是由多个独立的计算机系统相互连接而成的计算系统,从用户的角度来看,它好像是一个集中的单机系统[7]。

总结不同研究者给出的不同定义，可以概括如下：

分布计算系统是由多个相互连接的处理资源组成的计算系统，它们在整个系统的控制下可合作执行一个共同的任务，最少依赖于集中的程序、数据和硬件。这些处理资源可以是物理上相邻的，也可以是在地理上分散的。

现在对这一定义进行说明：

(1) 系统是由多个处理器或计算机系统组成的。

(2) 这些计算资源可以是物理上相邻的、由机器内部总线或开关连接的处理器，通过共享主存进行通信；这些计算资源也可以是在地理上分开的、由计算机通信网络(远程网或局域网)连接的计算机系统，使用报文(Message)进行通信。

(3) 这些资源组成一个整体，对用户是透明的，即用户使用任何资源时都不必知道这些资源在哪里。

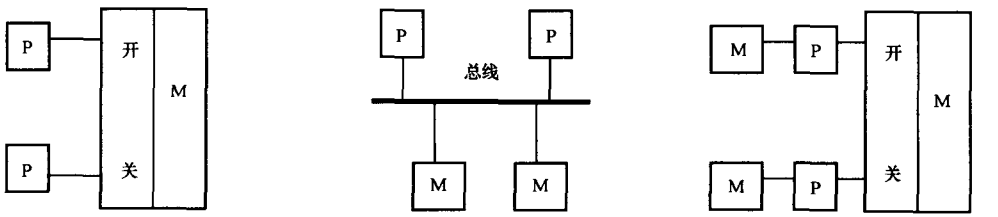
(4) 一个程序可分散到各个计算资源上运行。

(5) 各个计算机系统地位平等，除了受全系统的操作系统控制外，不存在主从控制和集中控制环节。

分布计算系统属于多指令流多数据流(MIMD)结构。

1.2.2 松散耦合与紧密耦合分布计算系统

从结构上分，有两大类分布式系统：紧密耦合分布式系统和松散耦合分布式系统。紧密耦合分布式系统由多个处理机经机器内部总线或机器内互连网络(如 mesh、cube 等)连接而成，因此也叫多处理机系统。紧密耦合分布式系统中的多个处理机有共享的主存储器(如图 1.2.1(a)和 1.2.1(b))，也可以既有共享的主存储器又有自己专用的主存储器(如图 1.2.1(c))。紧密耦合分布式系统一般局限于一个局部区域，各部分相距很近，可以说在物理上是分散的，但不是在地理上分散的。各处理机通过共享主存交换信息，并具有较高的通信速



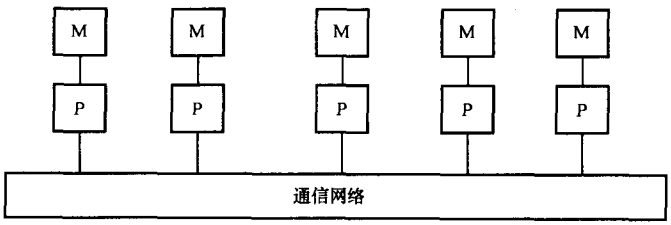
(a) 开关互连，只有共享存储器　(b) 总线互连，只有共享存储器　(c) 开关互连，既有共享存储器，又有专用存储器

P：处理机，　M：主存储器

图 1.2.1　紧密耦合分布式系统

度。本书不讨论这种系统。

松散耦合分布式系统由多个计算机系统(包括处理机和相应的主存储器)经过通信网络连接而成,是多计算机系统(如图 1.2.2)。计算机之间或处理机之间使用报文交换方法通信。松散耦合分布式系统没有专门用于各处理机共享的主存,但是,各计算机的主存可以共享,构成所谓分布式共享存储器系统。通信网络可以是远程网,也可以是局域网。通常的"分布式系统"一词指的是松散耦合分布式系统,本书以后各部分的分布式系统除特别说明外,均指松散耦合分布式系统。前面 Andrew S. Tanenbaum 教授对分布计算系统的定义,其实就是指松散耦合分布式系统。



P: 处理机, M: 主存储器

图 1.2.2 松散耦合分布式系统

1.2.3 同构型与异构型分布计算系统

按照组成分布计算系统的计算机种类和所使用的计算机网络的种类,分布计算系统可分成同构型分布式系统和异构型分布式系统。对于同构型分布式系统而言,组成该系统的计算机的硬件和软件是相同的或非常相似的,同时,组成该系统的计算机网络的硬件和软件也是相同的或非常相似的。对于异构型分布式系统而言,组成该系统的计算机的硬件或软件是不同的,或者组成该系统的计算机网络的硬件或软件也是不同的[3]。

分布计算系统的异构性包括计算机系统的异构性和通信网络的异构性。计算机系统的异构性主要有三个方面的表现,即:硬件异构性、操作系统异构性和程序设计语言异构性。

计算机系统的硬件异构性主要有以下三个方面的表现:

(1) 计算机的指令系统不同。这意味着一种机器上的程序模块不能在另一种不兼容的机器上执行,很显然,一种机器上的可执行代码程序也不能在另一种不兼容的机器上执行。

(2) 数据表示方法不同。例如,不同类型的计算机虽然都是按字节编址的,但是高字节和低字节的规定可能恰好相反。浮点数的表示方法也常常不一样。M6800 CPU 的字长是 16 位,并且可按字节单位寻址,它是高字节在前,低字节在后;而 Intel 80286 CPU 也是按字节单位寻址,但它是低字节在前,高字节在后,因此,在 Intel 80286 CPU 上运行的程序不能直接操

作在 M6800 CPU 上运行的程序所产生的数据。

(3) 机器的配置不同。尽管机器的类型可能相同,其硬件配置也可以互不兼容。例如,一台机器可能比另一台机器具有更多的内存或外存,而另一台则可能多配置了某种外部设备等。

即使硬件配置完全相同,如果运行的操作系统不同,则由这些机器组成的分布计算系统仍为异构型。操作系统异构性的三个主要表现方面为:

(1) 操作系统所提供的功能可能大不相同。例如,不同的操作系统提供了不同的命令集。

(2) 操作系统所提供的系统调用在语法、语义和功能方面不相同。

(3) 文件系统不同。文件系统的不同主要表现在三个方面。第一,不同的系统采用不同的文件命名方式。例如,UNIX 系统和 MS - DOS 系统的文件系统都是层次型的,但是 UNIX 用符号"/"分隔路径名的各部分,而 MS - DOS 文件系统中的文件名的第一部分是驱动器名,后跟一个":"符号,最后的部分采取 ****** . *** 的形式。这样,UNIX 程序可以产生和使用任何 MS - DOS 文件名,反过来则不行。所以如果把两个系统容纳在一个系统中共享文件,那么许多操作将会失败。第二,文件的保护方法和程度也不相同。例如,UNIX 系统把用户分成三类:文件的所有者、同组用户和其他用户,不同用户对某个文件有不同的访问权限。而 MS - DOS 则根本无文件保护措施。因此,很难要求一个文件系统在共享另一个文件系统的文件时能够按那个文件系统的保护方法对文件进行保护。第三,不同的文件系统有不同的文件模式。例如,UNIX 只提供了字节流模式,VMS 则提供多种记录结构的文件模式。

程序设计语言的异构性表现在不同的程序设计语言用不同的方法在文件中存储数据。例如,Pascal 程序以十进制字符串的形式在文件上存放整数,而 UNIX 系统中,C 语言则以二进制形式在文件中存放整数。因此,用不同语言编写的程序之间有可能不能直接共享数据文件。

通信网络的异构性在硬件方面主要表现在,不同的计算机网络和同一种计算机之间的接口硬件可能不同,连接方法和通信方法也可能不同。通信网络的异构性在软件方面主要表现在不同计算机网络的通信协议不相同。

随着计算机技术的发展和计算机系统的广泛应用,异构性是不可避免的。即使用户的初衷是建立一个同构型分布计算系统,但往往最终也会自然地演变为异构型分布计算系统。这是因为:

(1) 分布计算系统已成为资源共享的重要形式。随着分布计算系统资源的增多,其他用户也希望加入系统,共享其资源。这些新的系统往往有着与原有系统不同的硬件和软件。

(2) 由于硬件性能的提高和其价格的下降,当扩充一个分布计算系统时,人们往往会选