分布式系统技术内幕

FENBUSHI XITONG JISHU NEIMU

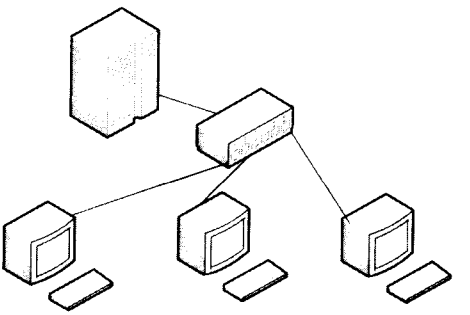
张军著

首都经济贸易大学出版社

分布式系统技术内幕

Inside Distributed Systems

张 军 著



首都经济贸易大学出版社
·北 京·

图书在版编目(CIP)数据

分布式系统技术内幕/张军著. —北京:首都经济贸易大学出版社,2006.7
ISBN 7-5638-1323-3

Ⅰ. 分… Ⅱ. 张… Ⅲ. 分布式操作系统 Ⅳ. TP316.4

中国版本图书馆 CIP 数据核字(2005)第 089567 号

分布式系统技术内幕
张 军 著

出版发行	首都经济贸易大学出版社
地 址	北京市朝阳区红庙(邮编 100026)
电 话	(010)65976483 65065761 65071505(传真)
网 址	http://www.sjmcb.com
E-mail	publish@cueb.edu.cn
经 销	全国新华书店
照 排	首都经济贸易大学出版社激光照排服务部
印 刷	北京泰锐印刷有限责任公司
开 本	787 毫米×960 毫米 1/16
字 数	391 千字
印 张	22.25
版 次	2006 年 7 月第 1 版第 1 次印刷
书 号	ISBN 7-5638-1323-3/TP·31
定 价	33.00 元

前　言

“我是做软件的”。我经常这样回答别人，这样的回答已经持续了二十多年。

做软件的人就有点像信教的，只有极少的人能修成正果，而大多数人却半信半疑，或弃之而投奔更为显赫的追求，或苦苦追索而不得其乐。

我经常游离在懵懵懂懂和大彻大悟之间。

在中国，程序员几乎成了白领阶层里下等人的同义词。在国外四五十岁的程序员司空见惯；而在中国，35 岁以上的程序员已经凤毛麟角。这种价值取向对中国软件业的发展无疑是致命的。

三年前的一个机会，我拿起了教鞭。

教软件的人更像传教士，在“冥顽不灵”的弟子面前经常束手无策。

我经常和学生说搞软件不需要大智慧，也不需要小聪明，勤奋是唯一捷径。也有些学生很努力，但不开窍，就是不明白。满足一知半解的人终会铸成大错。

一本好书往往是开启心灵的最重要的钥匙。现在软件的书很多，但好书太少，粗制滥造的太多。所以我一般不看国人写的书，虽然不绝对，但也是没有办法。国人大都是在“编”、“译”和“攒”书。这可能还是一个价值取向问题。

我一直想写一本关于软件的书。以前是没有时间，现在是无从下手。题目太大了，自己水平有限无法胜任；题目太小了，又觉得不“过瘾”，很多内容不能发挥。

连续给研究生开了三学年的分布式系统的选修课，觉得讲得挺过瘾，学生也挺捧场。于是，我决定写一本关于分布式系统的书，虽然冠以“技术内幕”还是有些牵强。

无数个日日夜夜，终于有了这本拙著。

本书的目的

分布式系统是计算机软件领域的一个重要研究分支，它涉及分布环境下软件中的通信、名字服务、事务处理、容错和安全等众多的理论和实践问题。从20世纪80年代以来，分布式系统在很多领域取得了突破性的进展，它以各种中间件软件的形式出现，并为现代信息系统建设提供了不可或缺的技术支持。但分布式系统的研究仍然面临着重大的技术挑战，受到很多未决问题的困扰。我国在分布式系统领域的研究起步较晚，基础薄弱，形成软件产品的能力与发达国家相比有很大差距。因此，分布式系统的研究不仅有重大的科学意义，而且将对软件产业的发展产生深远的影响。

目前国内关于分布式系统的图书主要是进口的影印本或翻译本，虽然它们代表了较高的水平，但因为语言或翻译的问题，这些图书一般晦涩难懂，这也是我下决心写这本书的一个原因。

本书的读者对象

本书的主要读者对象是计算机软件从业人员，包括大专院校计算机专业的本科生、研究生，本书站在国际和国内在相关领域的最前沿，试图把分布式系统的起源，理论基础，算法基础，分布、复制与集成相关的一些关键理论的技术背景全面地介绍给读者。尽管它不是一本讲解如何进行软件开发的书，但却试图成为所有软件从业人员必备的基础书籍。

本书的组织结构

现代分布式系统是一个以中间件方式组织的软件系统。它通过服务的形式来提供一个较完整的或不太完整的功能集合。这些服务有时候是独立的中间件，有时候是按照一定的技术标准被组织到一个更大的中间件系统中。分布式系统的研究在过去的20年里取得了很多重大进展，虽然目前尚未形成一个广泛共识的理论和学科架构，但其研究方向和研究方向之间的交叉、互动和统一的趋势越来越明显。有鉴于此，本书按照以下三个部分进行组织：

第一部分：导论

- 从第一台计算机的发明到今天，计算机系统经历了从基于主机的集中处理模式到基于网络互联的分布式处理模式的发展历程。分布式处理的出现曾经引发了关于集中处理和分布处理的激烈争论。今天我们看到的是：一方面基于主机的集中处理模式依然健在，它并没有像一些人预计的那样很快消亡；另一方面，基于网络互联的分布式处理逐渐成为信息系统开发的主流设计模式，但它并不是简单地排斥集中式的处理，而是致力于两种处理模式的集成（第1章）。

● 通常意义上的计算机系统是指一个普适性的计算环境，它既不是为某一编程语言、也不是为某一特定应用的需求而设计的。普适性的计算环境隐含了对计算机系统的两个方面的认识：第一，计算机系统作为可编程系统的概念；第二，计算机系统作为程序的执行系统的概念。深入理解计算机系统是理解分布式系统理论的必要准备（第 2 章）。

第二部分：分布式系统基础

● 缺乏完美同步的时钟和全局状态的问题是分布式系统的基本问题。在这个问题中，一个研究方向是关于如何获得精确的时钟，并设计和提供分布环境下物理时钟同步的解决方案。最新的时钟同步技术已经能够在大规模分布环境下达到毫秒数量级的精确度；另一个研究方向是关于逻辑时钟的定义、使用和同步问题。很多分布式应用中更关心事件发生的顺序关系，而不是事件发生的精确的物理时间。逻辑时钟为事件的全局排序以及获得系统的全局状态提供了理论和算法的支持（第 3 章）。

● 进程交互是分布式系统设计的核心内容。虽然不同的应用可以有不同的交互设计，但一些常用的、具有普遍性的进程交互模式开始引起人们的注意。例如，进程之间需要互斥性的访问一个资源；多个进程需要选举一个进程作为协调进程，或者进程之间需要对某些条件达成共识等。这些交互模式，或者说算法，开始于一些零星的实践，在经历了很多重复的设计之后，人们逐渐认识到有必要更加系统地研究分布式算法基础，为构造更为复杂的进程交互设计提供成熟的基本构件（第 4 章）。

● 系统分布的另一个需求就是复制。复制是提高分布式系统的性能、可用性、缩展能力和容错能力的一个重要技术手段。复制的负面影响是一致性问题。维持一致性需要附加的网络资源和计算资源，这个代价反过来会影响到系统的性能、可用性、缩展性等，因此复制是一把双刃剑。我们需要从应用需求出发，设计不同的复制和一致性模型，充分权衡各种因素，才有可能发挥复制技术的作用（第 5 章）。

● 在分布式系统中，进程的交互基础是基于通信通道的信息传递。通道和通信系统是分布式系统研究的重要组成部分。与更一般的通信研究不同的是，分布式系统关于通信的研究是建立在存在一个基本通信能力的假设之上的，这个基本通信能力是由网络操作系统提供的。分布式系统侧重研究在大规模分布和异构的环境下，客户服务的通信、组通信、对等通信、永久通信和流媒体通信等高级通信服务（第 6 章）。

● 分布式系统的分布本质上是进程的分布，进程的组织形式是重要的研究基础。进程可以组织为客户服务器的交互或者对等的交互（P2P）。进

程不仅是分布的，而且还具有动态属性，例如，迁移代码、移动主体等（第7章）。

- 不同的应用有不同的分布需求，在无法建立一个普适性的分布计算环境的条件下，更加务实的态度是从一些有共性的需求中抽取具有一定适用性的分布模型。这些模型在一些特定的应用领域起到了参考模型的作用。主要的分布模型有远程过程调用、分布式对象、分布式文档、分布式文件。一些复杂系统的开发可能涉及多个分布模型的使用，例如，使用分布式文档作用户界面，使用分布式对象做系统开发，分布式文件做数据存储等（第8章）。

第三部分：公共服务设计

- 分布式系统总是依赖名字和目录服务来提供分布资源的位置透明性。名字服务允许资源通过一个全局有效的名字被共享和检索，就像电话号码簿和黄页服务。虽然名字服务初看起来很简单，但在考虑到系统缩展性以后，服务的设计变得困难了。在一个大范围的系统中，为了能高效地通过名字访问到资源，我们一般有资源位置相对固定的假设。当资源发生频繁移动时，传统的名字服务就会不再适用，因此我们需要研究移动资源的管理问题（第9章）。

- 很多分布式系统都提供存储服务，也叫永久性。永久性的最简单的形式就是分布式文件系统，很多更先进的中间件系统还提供集成的数据库服务或者提供数据对象自动同步到数据库的能力。在一些数据的存储扮演重要作用的场合，我们需要为分布式事务处理提供服务。一个事务的重要属性是所有其中的多个读写操作原子般（atomically）地进行。原子性意味着如果事务处理成功，则所有的写操作必须完成；如果不成功，所有涉及的数据必须保持不受影响。遗憾的是，事务很难在多台计算机上扩展，更不用说物理位置相距的计算机上了（第10章）。

- 在一些有高可用性要求的分布式系统中，容错服务的设计扮演了很重要的角色。构成分布式系统的硬件和软件都有可能会发生故障，有时候这些故障不会造成什么后果，而有时候这些故障却是灾难性的。实现系统容错能力的关键技术是冗余设计，而冗余的本质是一些关键组件的复制。我们关心的一个问题是系统能否在部分组件故障的情况下不中断地服务，同时我们还需要关注在故障或灾难后系统的恢复问题（第11章）。

- 所有在非实验性环境下使用的分布式系统都必须提供安全服务。资源的分布同时也带来很多安全隐患：很多密码和隐私信息未经过任何加密就进行交换或/和存取在服务器里；一些个人隐私信息在当事人不知情

的情况下被采集和利用;虚假、欺诈和垃圾邮件随处可见等。同网络操作系统相比较,分布式系统中的安全问题遍及系统的方方面面。原则上分布式系统不能依赖底层网络操作系统去很好地支持整个网络的安全,分布式系统中一般有自己的安全服务设计。再加上系统的缩展性要求,实际上安全服务已经成为分布式系统最复杂的服务之一(第 12 章)。

致谢

在此诚挚感谢首都经济贸易大学的领导,信息学院的领导以及各位同仁和朋友。由于他们的支持与帮助本书得以从想法变为现实。

特别致谢首都经济贸易大学李宁副教授为本书的编写提供的宝贵意见。

同时,此书的写作参考了大量的文献,如果有些引用未能准确列出,在此谨表歉意并致谢。

我总想,一个学者一生有一两部自己满意、读者认可的作品足以,这也是我一直不敢动笔的原因。但是,目前的这本书离"完美之作"尚有很大差距。我希望读者能就本书内容展开批评,赐教于我,我会在以后各版中不断完善内容,实现我"出一本好书"的夙愿。我的邮箱是:zhangjun@ cueb. edu. cn

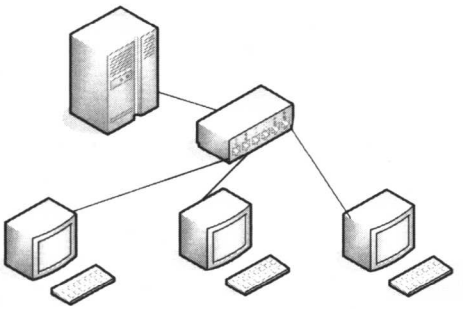
张 军

2005 年 7 月 北京

本书出版得到了国家自然科学基金项目资助,项目编号 60273027,项目负责人张军。

目　录

第一部分 导 论



第1章 分布式处理

计算机系统是一个集信息采集、存储、处理和分发为一体的信息系统。从第一台计算机的发明到今天，计算机系统经历了从基于主机的集中处理模式到基于网络互联的分布式处理模式的发展历程。分布式处理的出现曾经引发了关于集中处理和分布处理的激烈争论，其中有理论和技术上的分歧，但更多的是代表这两种不同技术潮流的利益集团之争。回顾历史总是让我们变得更加理性。今天我们看到的是：一方面基于主机的集中处理模式依然健在，它并没有像一些人预计的那样很快消亡；另一方面，基于网络互联的分布式处理逐渐成为信息系统开发的主流设计模式，但它并不是简单地排斥集中式的处理，而是致力于两种处理模式的集成。

1.1 集中式处理模式的演变

主机的英文单词为 host，原意是“主人”的意思，相对于客人而言。这个单词在计算机术语中延伸为“主机”的概念，是一个提供服务的设备概念。如果一个应用系统的所有处理都集中在一台计算机上时，我们称这种处理模式为基于主机的集中式处理模式。

集中式处理模式是计算机处理的固有模式，它随着计算机的发明而产生。集中式处理模式的演变经历了以下几个阶段。

1.1.1 计算机的原始模型

图 1-1 是最早发明的两台计算机的历史图片。它们和今天的计算机已经没有太多的可比性。计算机的原始模型更多是硬件设备的概念，加载程序、执行程序以及获得程序结果都需要用手工操作。在当时看来，这些

计算机的运算速度已经非常惊人;其中 1946 年发明的 ENIAC 达到了每秒 5 000次的运算速度。难怪当时有人惊叹地说:“世界上只需要两台这样的机器就可以完成所有的计算需求。”

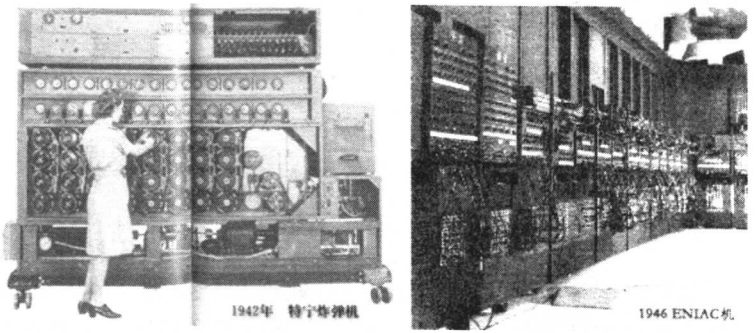


图 1-1 最早发明的两台计算机:特宁炸弹机和 ENIAC

1.1.2 带终端的计算机系统

终端设备的出现解放了对计算机的手工操作。终端是一个人机交互设备,早期的终端更像一台英文打字机。随着电子显示技术的发展,显示屏和键盘的组合构成了最常用的终端形式。从此用户通过终端来向计算机发号施令。终端的功能很简单,它负责将键盘的输入传送给主机,接收主机的输出并显示在屏幕上,如图 1-2(a)所示。在终端出现之前,计算机并没有今天意义上的用户的概念,因为计算机管理是通过操作员手工进行的,用户并不与计算机直接打交道。终端出现的意义不仅在于它改变了对计算机的操作方式,更重要的是它使得非专业计算机人员获得了直接使用计算机的能力。这是计算机从实验室步入商业应用的一个重要技术环节。

1.1.3 多用户计算机系统

将计算机从手工操作过程中解放出来,终端所解决的只是一个硬件设备问题。程序的加载、执行和显示结果的自动处理能力依赖于我们今天称为操作系统的系统软件。操作系统的一个重要进展就是允许计算机并发地执行多个程序,进而允许多个用户同时使用计算机。多用户系统允许计算机连接多个用户终端,操作系统的管理使得每个终端的用户好像在独立使用计算机。计算机与多终端的连接依赖于一个称为通信集中器(Concentrator)的设备如图 1-2(b)所示。通信集中器在小型多用户系统中有时以

多用户卡的形式出现。

允许多个程序并发执行的系统称之为多任务系统，多任务系统是多用户系统的技术基础，但多任务系统并不一定是多用户系统。例如，现代微机系统是多任务系统，用户可以同时执行多个程序，但它不是多用户系统，因为它不允许多个用户终端的连接。主机的多用户与网络的多用户也不是一个概念。主机的多用户是一个操作系统，多个终端；而网络的多用户是每个用户运行自己独立的操作系统。

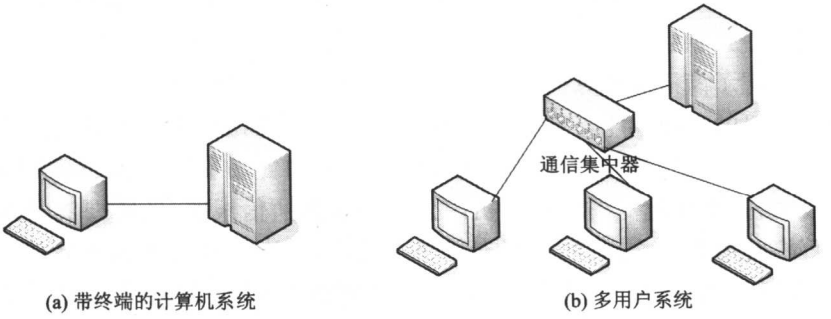


图 1－2　本地终端系统

1.1.4　远程多用户系统

图 1－2 所示的系统又称为本地终端主机系统。由于本地通信信号会随着距离而衰减，终端必须位于主机的一定距离内。如果不能很好解决这个问题，计算机的商业应用前景显然会大打折扣。借助广域网的通信技术来延伸终端分布就构成了远程终端系统，如图 1－3 所示。计算机与通信的结合开始于主机和终端的远程通信需求，这种通信网络早于计算机网络而出现。

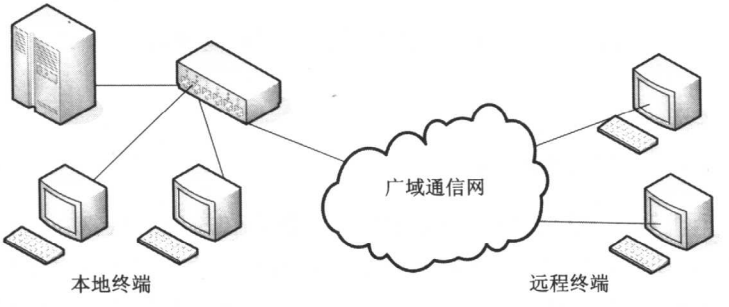


图 1－3　远程多用户系统

终端和主机的连接可以说是计算机网络的雏形。它区别于计算机网络在于以下两点:①终端不是完全意义上的计算机;②通信模式是点到点的和集中控制的。

伴随着基于主机的集中处理模式的演变,计算机系统的商业应用日趋成熟。到了20世纪70年代末和80年代初,主机能够稳定地、持续地提供服务的能力达到了相当的水平:一台IBM 3090计算机可以通过CICS系统同时为1万多个用户提供联机服务。国外几乎所有的大银行、保险公司、大型零售企业等都在集中式处理模式上完成了信息系统的建设。在这样一个演变过程中,信息化的大量资金投入成就了一批像IBM、DEC等一些很好把握了商机的计算机企业。

主机的卓越性能是有相应代价的。主机系统几乎成为了昂贵的代名词:因为需要建造满足苛刻条件的机房,需要购置昂贵的主机和必要的外部设备,需要建设专用的或者租用公共的数据通信线路,还需要高薪聘请计算机技术人员,同时还需要支付昂贵的培训、咨询和技术服务费用。当时即使很大的机构,也只有可数的几台计算机。这样沉重的准入条件让很多中小企业有心无力,望而却步。主机模式的另一个副作用是计算机技术和管理人员高度集中,集权和官僚化现象也日益严重,在某种意义上阻碍了信息化的进程。因为当用户新的需求涉及已有系统的维护或改动时,层层审批可能会导致项目中途夭折。

为了满足不同规模的应用系统的需求,主机在低端和高端的两个产品方向上同时发展,逐步形成了一个从小型机、中型机、大型计算机直到超级计算机的主机系列,价格和技术门槛高的问题得到了一定程度的缓解。正当主机系统如火如荼地发展、如日中天的时候,从20世纪80年代开始,微机和计算机网络出现在人们的视野中。很快,"大地震"开始了……

1.2 分布式处理产生的技术背景

1.2.1 微机的出现

小型机对个人和小企业而言依然过于庞大和昂贵。从20世纪70年代中期开始,计算机的设计在微型化的方向上取得了重大突破,那就是基于大规模集成电路技术(Large Integrated Circuit,LSI)的微处理器(Microprocessor)的成功开发以及采用微处理器的微型计算机的出现。微机首先以