计 算 机 科 学 丛 书

分布式系统设计

Distributed System Design

(美) Jie Wu 著
(佛罗里达大西洋大学)
高传善 等译

机械工业出版社
China Machine Press
COURSE TECHNOLOGY

计算机科学丛书

分布式系统设计

(美) Jie Wu 著

（佛罗里达大西洋大学）

高传善 等译

机械工业出版社
China Machine Press

本书较为全面地介绍了分布式系统领域的一些基本概念，提出了分布式系统的各种问题，如互斥问题、死锁的预防和检测、处理机间的通信机制、可靠性问题、负载分配问题、数据管理问题及其可能的解决方案，并讨论了分布式系统设计在操作系统、文件系统、共享存储器系统、数据库系统和异构型处理中的应用。

本书适用于学习分布式系统设计的高年级本科生、研究生和从事分析、设计分布式系统的计算机专业人员。

本书版权登记号：图字：01-2001-0267

图书在版编目(CIP)数据

分布式系统设计/（美）吴杰（Jie Wu）著；高传善等译 -北京：机械工业出版社，2001.2

(计算机科学丛书)

书名原文：Distributed System Design

ISBN 7-111-08574-4

Ⅰ.分… Ⅱ.①吴 ②高 Ⅲ 分布式操作系统-系统设计 Ⅳ TP316 4

中国版本图书馆CIP数据核字(2000)第57510号

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：赵阿玲

北京忠信诚胶印厂印刷 · 新华书店北京发行所发行

2001 年 2 月第 1 版第 1 次印刷

787mm × 1092mm 1/16 · 19.25印张

印数：0 001-5 000册

定价：30.00元

作 者 简 介

吴杰（Jie Wu）于1982年在上海科技大学获得计算机工程学士学位，1985年获得该校计算机科学硕士学位，1989年在佛罗里达大西洋大学 (Florida Atlantic University，FAU）获得计算机工程博士学位。1985～1987年期间他在上海科技大学从事教学工作。从1989年8月开始，他在FAU的计算机科学与工程系担任教授和系研究生部主任。

吴先生在各种出版刊物与会议文集上发表或联合发表了100多篇技术论文，包括《IEEE软件工程学报》(IEEE Transactions on Software Engineering)、《IEEE计算机学报》(IEEE Transactions on Computers)、《IEEE并行与分布式系统学报》(IEEE Transactions on Parallel and Distributed Systems)、《并行与分布式计算杂志》(Journal of Parallel and Distributed Computing)、《计算机杂志》(The Computer Journal）以及《并发性：实践与经验》(Concurrency: Practice and Experience)等等。他的研究兴趣在于容错计算、并行/分布式处理、互连网络、佩特里（Petri）网应用以及软件工程等方面。他是1996～1997年FAU最佳研究年度奖的获得者。

吴先生是Upsilon Pi Epsilon和美国计算机学会 (ACM) 的成员、IEEE的高级成员。他在亚、欧和北美洲的各大学和学院举办讲座和研讨会。目前，他是《国际计算机与应用杂志》的编委。他还是《IEEE并行与分布式系统学报》(IEEE Transactions on Parallel and Distributed Systems)“设计网络容错路由的挑战”特别专题的客座编辑。吴先生是1999年第12届ISCA并行与分布式计算系统国际会议程序委员会的两主席之一，他还是1996年和1998年IEEE分布式计算系统国际会议程序委员会的成员。

译 者 序

在计算机发展进入了网络计算的新阶段中，分布式系统已得到了越来越广泛的研究和应用。许多高等院校也开设了相关课程。但是，目前在国内尚缺少系统地讲述分布式系统领域涉及问题的书籍。本书译自1999年CRC Press出版的《Distributed System Design》。作者Jie Wu（吴杰）现为美国佛罗里达州Florida Atlantic University（FAU）教授，长期从事分布式系统相关领域的研究工作。他是IEEE高级会员，担任了多种学术刊物的编辑，并是1999年第12届ISCA并行与分布式计算系统国际会议程序委员会两主席之一。在本书中，作者提出了分布式系统领域的一些重要问题，包括基本概念、问题和可能的解决方案。本书共12章，大致可分成三部分：介绍和基础（第1～3章），分布式系统领域的各种问题（第4～11章），分布式系统的应用（第12章），并着重于分布式系统设计的软件部分。本书不仅可用于大学本科高年级和研究生的分布式系统设计课程或高级操作系统课程的教材或参考书，对于从事分布式系统设计与应用的专业人员也是十分有用的。

本书由复旦大学计算机科学系高传善组织和主持翻译，并由他校阅了全书。黄晓霖翻译第1～4章，并负责统一术语。岑志伟翻译第5～8章，周鑫翻译第9～12章。翻译中我们力求忠实于原著，但限于时间及水平，不当之处在所难免，欢迎批评指正。

译 者

2000年8月

高传善，1942年出生，1963年毕业于复旦大学。1981～1983年在美国伊利诺大学作访问学者。现为复旦大学计算机科学与工程学院副院长、计算机科学系教授和博士生导师；教育部科技委信息学部委员；纽约科学院院士和IEEE计算机学会会员。主要研究方向：网络与分布系统及其应用。科研成果：曾获省部级科技进步一等、二等和三等奖分别二次、二次和三次。

前　言

显然，未来对计算速度、系统可靠性和成本实效性的要求必将促使发展另外的计算机模型来替代传统的冯·诺依曼结构的计算机。随着计算机网络的出现，一个新的梦想成为可能——分布式计算。当用户需要完成任何任务时，分布式计算提供对尽可能多的计算机处理能力和数据的透明访问，同时实现高性能与高可靠性的目标。在过去的10年里，人们对分布式计算系统的兴趣迅猛增加。分布式计算的主题是多种多样的，许多研究人员正在研究有关分布式硬件结构和分布式软件设计的各方面问题以开发利用潜在的并行性和容错性。

分布式计算系统（或分布式系统）多种多样并涉及不同的系统体系结构。对一些用户来说，一个分布式系统是为解决单个问题而紧密结合在一起工作的多处理机的集合。对另一些用户来说，一个分布式系统可能意味着一个由地理上分散的各自独立的处理机组成的计算机网络，这些处理机连接在一起以实现对不同资源的共享。然而，分布式系统这个词在计算机系统中被如此广泛应用以至于它的使用变得有点贬值。许多这方面的混乱来源于缺乏对物理的分布和逻辑的分布的区分。通过区分这两个概念，就可以更准确地描述一个分布式系统的属性。

对于分布式系统，我们使用以下定义：一个分布式系统是一个对用户看起来像普通系统，然而运行在一系列自治处理单元（PE）上的系统，每个处理单元有各自的物理存储器空间并且消息的传输延迟不能忽略不计。在这些处理单元间有紧密的合作。系统必须支持任意数量的进程和处理单元的动态扩展。

目的

建立一个分布式系统的主要目的在于：

- 固有的分布式应用。分布式系统以一种很自然的方式开始存在，例如，在我们的社会中，人群在地理上是分布式的并且分布式地共享信息。一方面，一个分布式数据库系统中的信息产生于不同的分支机构（子数据库），所以能够快速地完成本地访问。另一方面，系统也提供了全局视图来支持各种全局操作。
- 性能/成本。分布式系统的并行性降低了处理的瓶颈，提供了全面改进的性能，也就是说，分布式系统提供了更好的性能价格比。
- 资源共享。分布式系统能有效地支持不同位置的用户对信息和资源（硬件和软件）的共享。
- 灵活性和可扩展性。分布式系统可以增量扩展，并能方便地修改或扩展系统以适应变化的环境而无需中断其运行。
- 实用性和容错性。依靠存储单元和处理单元的多重性，分布式系统具有在系统出现故障的情况下继续运行的潜力。

• 可伸缩性。分布式系统能容易地扩大以包括更多的资源（硬件和软件）。

概述和读者

这本书尝试着提出了分布式系统领域的一些重要问题，包括基本概念、问题和一些可能的解决方案，可用于研究生的分布式系统设计课程，也可用于高年级本科生和研究生的高级操作系统课程。它向学生介绍了分布式系统特有的有关设计方面的一些内容。本书着重于设计的软件部分，因为大部分相应的硬件部分在许多有关计算机网络和并行计算机的教科书中都已得到很好的介绍。

这本书中的所有高级设计和算法都使用建议的类CSP分布式控制描述语言（DCDL）表示（CSP代表通信顺序进程）。虽然这本书不可能覆盖分布式计算系统的所有问题，但我们的目标在于给出有关每个涉及到的问题的基本方面。我们鼓励学生通过学期项目、硕士和博士论文在这些问题上作更多的研究。我们假设学生至少熟悉一门高级程序设计语言，熟悉操作系统和计算机体系结构的基本概念以及离散数学的基础。

这本书的大部分材料来自原始资料、当代文献中的研究论文和作者自己在这方面的研究成果。分布式处理的广阔题材在本书的组织上得以体现。本书共12章，大致可分成三部分：介绍和基础（第1～3章），分布式系统的各种问题（第4～11章）以及应用（第12章）。一些相关主题没有包括进来，比如分布式实时系统和分布式系统软件。我们尽量包括足够一个学期课程的材料。

内容

第1章介绍一些基本概念，讨论分布式计算系统的目的，提出分布式计算系统的范围，同时还提供了本书的简介。

第2章概述一般的分布式程序设计语言，介绍类CSP分布式控制描述语言（DCDL）。这个语言用于描述一些控制问题，比如并行的表示进程间的通信与同步和容错设计。附录列出了在DCDL中通用的符号。

第3章正式涉及分布式系统，介绍了一些概念，如时钟、事件和状态以及描述一个分布式系统的两种方法：时空视图和交叉视图。

第4章讨论对于分布式系统设计十分重要的互斥问题。互斥保证了相互冲突的并发进程能共享资源。我们还讨论了有关互斥的三个问题：选举、投标和自稳定。

第5章研究分布式系统中死锁的预防与检测。分布式系统一般具有高度的资源和数据共享，在这种情况下可能导致死锁的发生。这一章讨论了几个分布式系统特有的死锁问题的解决方案。

第6章研究对于分布式系统的性能至关重要的处理机间高效通信的机制。这一章研究了三种类型的通信：一对一（单播）、一对多（组播）和一对所有（广播）以及它们的性能。

第7章讨论没有特别约束的处理机间的通信机制，这些约束包括自适应性、无死锁和容错性。这一章还介绍了用于不同目的的虚通道和虚网络的概念。

第8章涉及分布式系统的可靠性问题。使用分布式系统的一个重要目的就是高度的可依赖性，包括可靠性、安全性和保密性。一个基本问题就是检测和处理系统中可能出现的故障。在这一

章中我们研究了处理分布式系统中的节点与通信故障、拜占庭式故障和软件故障的各种方法。

第9章和第10章包括分布式系统中的负载分配问题。负载分配是分布式系统的资源管理部分，它在处理机间公平透明地重新分配系统的负载，使系统的总体性能最佳。第9章研究静态负载分配，即利用预知的有关系统的知识做出负载分配的决策，而且负载在运行期间不能重新分配。第10章涉及动态负载分配算法，即利用（至少部分地利用）系统状态信息（节点上的负载）来做出负载分配的决策。

第11章描述分布式数据管理问题，包括两个特别问题：（a）对共享数据访问的同步并同时支持高度的并发性；（b）可靠性。

第12章包括分布式设计在操作系统、文件系统、共享存储器系统、数据库系统和异构型处理中的应用，同时列出将来可能的研究方向。

附录包括了DCDL中的通用符号列表。

目　录

第1章 概 论

显然，未来对计算速度、系统可靠性和成本实效性的要求必将促使发展另外的计算机模型来取代传统的冯·诺依曼型结构的计算机。随着计算机网络的出现，一个新的梦想成为可能——分布式计算。当用户需要完成任何任务时，分布式计算提供对尽可能多的计算机能力和数据的透明访问，同时实现高性能与高可靠性的目标。在过去的10年里，人们对分布式计算系统的兴趣迅猛发展。有关分布式计算的主题是多种多样的，许多研究人员正在研究关于分布式硬件结构和分布式软件设计的各方面问题以开发利用其潜在的并行性和容错性。在这一章里，我们将考虑一些基本概念以及与分布式计算相关的一些问题，并列出了本书所覆盖的主题。

1.1 推动因素

计算机技术的发展可以通过使用计算机的不同方式来描述。在50年代，计算机是串行处理机，一次运行一个作业直至完成。这些处理机通过一个操作员从控制台操纵，而对于普通用户则是不可访问的。在60年代，需求相似的作业作为一个组以批处理的方式通过计算机运行以减少计算机的空闲时间。同一时期还提出了其他一些技术，如利用缓冲、假脱机和多道程序等的脱机处理。70年代产生了分时系统，不仅作为提高计算机利用率的手段，也使用户离计算机更近了。分时是迈向分布式系统的第一步：用户可以在不同的地点共享并访问资源。80年代是个人计算的10年：人们有了他们自己专用的机器。由于基于微处理器的系统所提供的出色的性能/价格比和网络技术的稳步提高，90年代是分布式系统的10年。

分布式系统可以有不同的物理组成：一组通过通信网络互连的个人计算机，一系列不仅共享文件系统和数据库系统而且共享CPU周期的工作站（而且在大部分情况下本地进程比远程进程有更高的优先级，其中一个进程就是一个运行中的程序），一个处理机池（其中终端不隶属于任何一个处理机，而且不论本地进程还是远程进程，所有资源得以真正的共享）。

分布式系统是无缝的，也就是说网络功能单元间的接口很大程度上对用户不可见。分布式计算的思想还被应用在数据库系统[16、38、49]，文件系统[4、24、33、43、54]，操作系统[2、39、46]和通用环境[19、32、35]。

另一种表示同样思想的说法是用户把系统看成一个虚拟的单处理机而不是不同处理机的集合。向分布式系统发展的主要推动因素在于：

- 固有的分布式应用。分布式系统以一种很自然的方式开始存在，例如，在我们的社会中，人群在地理上是分布式的并且分布式地共享信息。一方面，一个分布式数据库系统中的信息产生于不同的分支机构（子数据库），因此本地访问可以很快进行；另一方面，系统也提供了全局视图来支持各种全局操作。
- 性能/成本。分布式系统的并行性减少了处理瓶颈，全方位提高了性能，也就是说，分布式系统提供了更好的性能价格比。

- 资源共享。分布式系统能有效地支持不同地方的用户对信息和资源(硬件和软件)的共享。
- 灵活性和可扩展性。分布式系统可以增量扩展，并能方便地修改或扩展系统以适应变化的环境而无需中断其运行。
- 实用性和容错性。依靠存储单元和处理单元的多重性，分布式系统具有在系统出现故障的情况下继续运行的潜力。
- 可伸缩性。分布式系统容易扩大规模以包括更多的资源（硬件和软件）。

LeLann[123]讨论了分布式系统的目的与目标，并通过区分物理的分布和逻辑的分布解释了其中一些与众不同的特征。可扩展性、逐渐增加的实用性和更好的资源共享被认为是最重要的目标。

目前对分布式系统[5, 23]的兴趣主要有两种刺激因素：技术上的变化和用户的需求。技术上的变化有两方面：微电子技术的进步生产出快速而廉价的处理器；通信技术的进步使得高效的计算机网络进入实用阶段。

计算机间长距离而且相对慢速的通信链路长期以来存在着，然而就是最近出现了快速、廉价而且可靠的局域网（LAN）技术。这些局域网通常以10～100Mbps（兆比特每秒）的速率运行。与此同时，城域网（MAN）和广域网（WAN）也变得越来越快和更加可靠。通常情况下，局域网跨越的地域直径不超过几公里，城域网可覆盖的直径达几十公里，广域网可扩展到整个世界。最近异步传输模式（ATM）被认为是未来的新兴技术，它可以为局域网和广域网提供高达1.2Gbps（千兆比特每秒）的数据传输速率。

在用户需求上，很多企业实际上都是相互合作的，例如办事处、跨国公司、大学计算中心等等，它们都要求共享资源和信息。

1.2 基本计算机组成

冯·诺依曼型的计算机包括CPU、存储部件和I/O。CPU是计算机的大脑。它通过读取、检查并逐条执行指令来执行存储在存储器中的程序。存储器单元存储指令（指令的序列叫做程序）和数据。I/O负责指令和数据进出处理器。

两个基本的计算机组织结构如下：

- 物理共享式存储器结构（图1-1a）有一个被所有CPU共享的单一存储器地址空间。这样一个系统叫做紧耦合系统。在一个物理共享式存储器系统中，CPU间的通信通过读和写共享存储器的操作进行。
- 物理分布式存储器结构（图1-1b）没有共享存储器，每个CPU有自己的本地存储器。CPU和本地存储器对被称做处理单元（PE）或简称处理机。这样一个系统有时被称做松耦合系统。在一个物理分布式存储器系统中，处理机间的通信通过互连网络上的消息传递来进行，发送处理机发送命令，接收处理机接收命令。

在图1-1中，互连网络用于连接系统的不同部件并支持数据和指令的传送。I/O部件没有画出来，但并不说明它不重要。

然而，通信模型的选择无需和物理系统联系在一起。我们可以区分硬件呈现的物理共享（图1-1）和编程模型呈现的逻辑共享。图1-2表示了4种可能的共享组合。

图1-2中标有“共享存储器”的方格代表了单处理机并行程序设计和基于共享存储器模型的

共享存储器多处理机程序设计。进程间必须同步以保证对共享数据访问的一致。在并发执行的进程间提供同步的编程结构包括信号量[12]、管程[17]、临界区[11]和屏蔽[20]。标有“消息传递”的方格代表了分布式存储器系统，其中通信是通过消息的传递进行的。显式发送和接收消息的命令经常被用到。“模拟消息传递”用来在一个物理共享存储器体系结构上提供逻辑分布式编程模型。消息通过共享存储器的缓冲区传递。图1-2中标有“分布式共享存储器（DSM）”的方格也被叫做共享虚拟存储器或分布式存储器系统。这个模型通过使分布式存储器系统对程序员看来就像共享存储器系统一样，使得在分布式存储器环境下编程尽可能如在共享存储器系统那样容易。

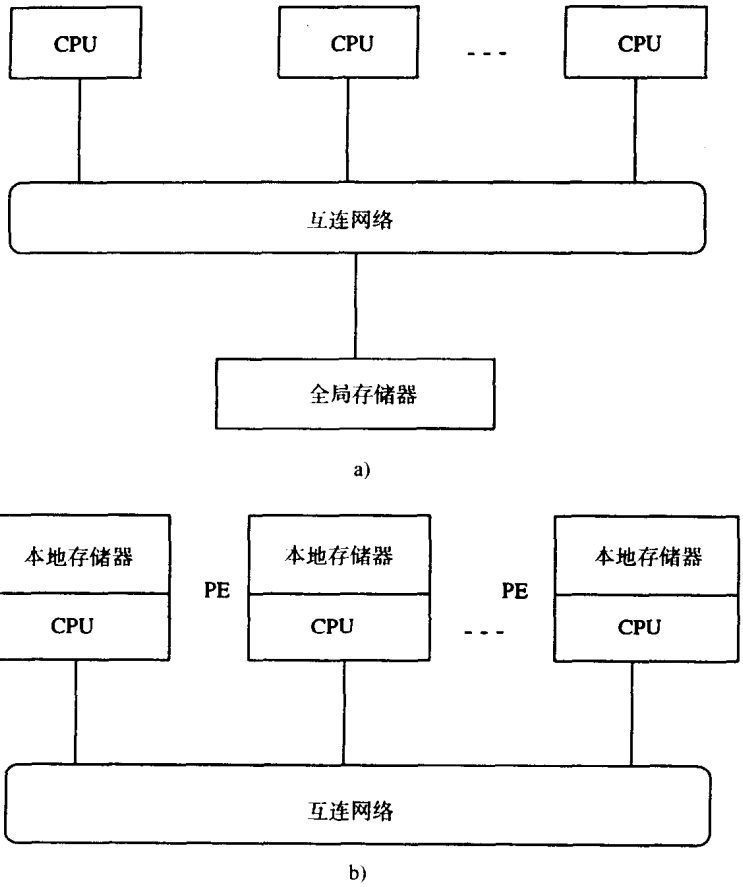

图1-1 两种基本的计算机结构：a）物理共享式存储器结构，b）物理分布式存储器结构

	逻辑共享式	逻辑分布式
物理共享式	共享存储器	模拟消息传递
物理分布式	分布式共享存储器 (DSM)	消息传递

图1-2 物理和逻辑共享式/分布式存储器的比较

1.3 分布式系统的定义

当讨论分布式系统时，我们面临许多以下这些形容词所描述的不同类型：分布式的、网络的、并行的、并发的和分散的。分布式处理是一个相对较新的领域，所以还没有一致的定义。与顺序计算相比、并行的、并发的和分布式的计算包括多个PE间的集体协同动作。这些术语在范围上相互覆盖，有时也交换使用。在[44]中，Seitz给出了每一个的定义来区分它们之间的不同含义：

- “并行的”意味着从一个单一控制线程对数据集的锁步（lockstep）动作。在并行计算机级别上，单指令多数据（SIMD）计算机就是一个使用多个数据处理单元在许多数据项上同时进行相同或相似操作的例子。
- “并发的”意味着某些动作可以以任意次序执行。例如，在更高级别上和在多指令多数据（MIMD）并行计算机上进行部分独立的操作。
- “分布式的”意味着计算的成本或性能取决于数据和控制的通信。

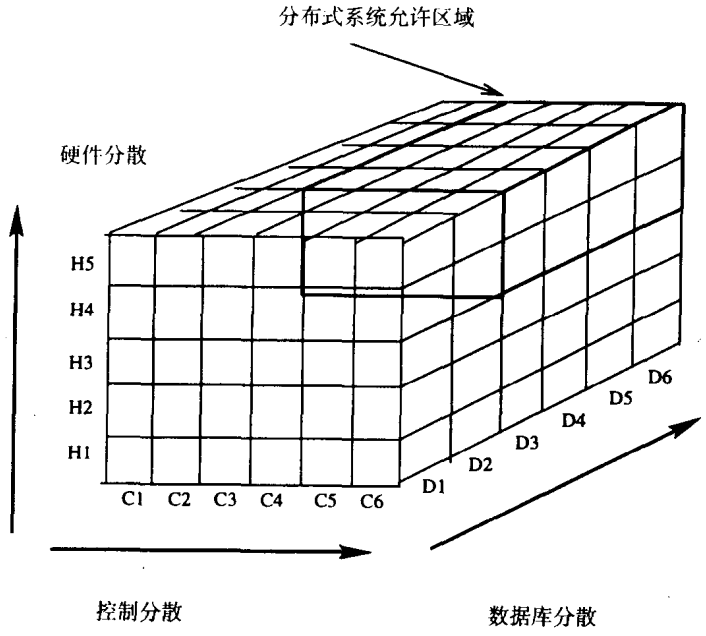

图1-3　Enslow的分布式系统模型

如果一个系统的部件局限在一个地方，它就是集中式的；如果它的部件在不同地方，部件之间要么不存在或仅存在有限的合作，要么存在紧密的合作，它就是分散式的。当一个分散式系统不存在或仅存在有限的合作时，它被称做网络的；否则它就被称做分布式的，表示在不同地方的部件之间存在紧密的合作。在给出分布式系统具体定义的模型中，Enslow[113]建议分布式系统可以用硬件、控制、数据这三个维度加以检验。

分布式系统 = 分布式硬件 + 分布式控制 + 分布式数据

Enslow的定义同时还要求资源的分布必须对用户透明。图1-3给出了从Enslow的模型[113]改编

而来的形式，如果系统所有三个维度（硬件、控制和数据）都达到一定程度的分散，系统可以被归类为分布式系统。在硬件组成这一维上的一些点表示如下：

H1 只有一个控制单元的单个CPU。

H2 有多个ALU（算术逻辑单元）的单个CPU，只有一个控制单元。

H3 分开的专用功能单元，比如带一个浮点协处理器的CPU。

H4 带多个CPU的多处理机，但只有一个单独的I/O系统和一个全局存储器。

H5 带多个CPU的多计算机，多个I/O系统和多个本地存储器。

类似地，在控制维上的点按分散程度的递增排序如下：

C1 单个固定控制点。注意，系统物理上可能有也可能没有多个CPU。

C2 单个动态控制点。在多个CPU的情况下，控制器不时地在CPU间切换。

C3 固定的主/从结构。例如，在一个只有一个CPU和一个协处理器的系统中，CPU就是固定的主结构（master），协处理器则是固定的从结构（slave）。

C4 动态的主/从结构。主/从的角色可以通过软件改变。

C5 使用同一控制器副本的多个同类控制点。

C6 使用不同控制器的多个异类控制点。

数据库中有两个部件可以是分布式的：文件和记录这些文件的目录。可以使用以下两种方式之一或结合使用它们来实现分布：复制和分区。如果一个数据库有多个副本在不同的地点,就称做被复制。如果一个数据库被分成位于不同地点的子数据库，就称做被分区。这一维上的点包括：

D1 含有文件和目录的单一拷贝的集中式数据库。

D2 含有单一集中式目录并且没有本地目录的分布式文件。

D3 每个站点都有文件和目录拷贝的复制数据库。

D4 有一个主结构的分区数据库，主结构保留所有文件的一个完全副本。

D5 有一个主结构的分区数据库，主结构仅保留一个完整的目录。

D6 无主结构文件或目录的分区数据库。

Schroeder[34]，给出了分布式系统特征的列表。如果一个系统具有以下所有特征，它很可能就是一个分布式系统：

- 多处理单元（PE）。
- 互连硬件。
- 处理单元的故障无关。
- 共享状态。

一些研究人员还认为计算机网络和并行计算机是分布式系统的一部分[9、21、48、55]。图1-4和图1-5分别表示了计算机网络和并行计算机的分类。在非冯·诺依曼的模型中，数据流模型是基于贪心求值（greedy evaluation）的，一个函数（程序）只要它的输入数据一就绪就被马上执行。缩减模型是基于懒惰求值（lazy evaluation）的，只有需要函数的结果时才执行求值。大部分并行计算机是建立在冯·诺依曼的模型上的。Flynn的分类法是被最广泛地用来在冯·诺依曼模型范围内分类系统。Flynn的分类法是基于指令流和数据流的多重性：