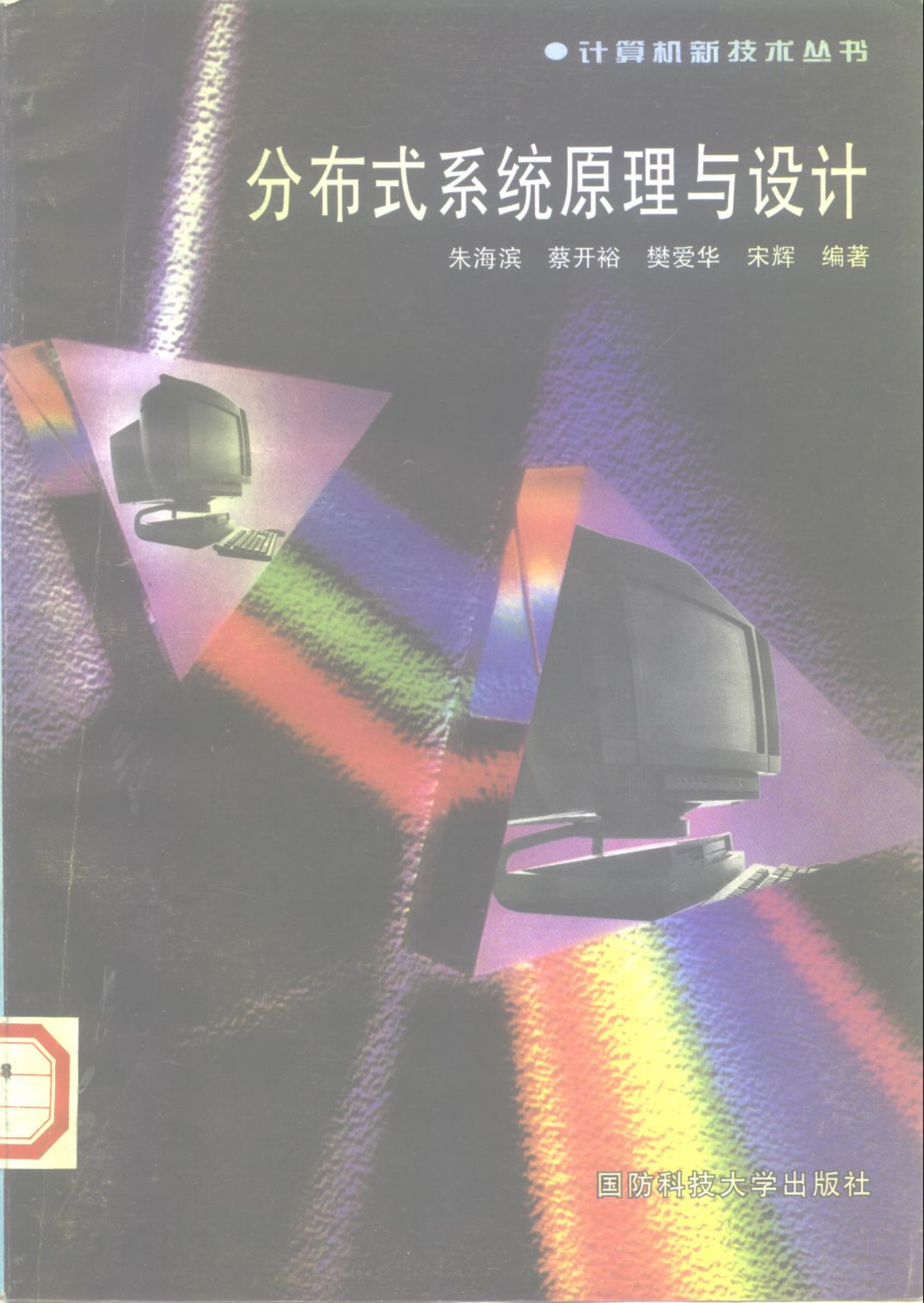
● 计算机新技术丛书

# 分布式系统原理与设计

朱海滨　蔡开裕　樊爱华　宋辉　编著



国防科技大学出版社

# 分布式系统原理与设计

朱海滨 蔡开裕 樊爱华 宋 辉 编著

国防科技大学出版社

**图书在版编目(CIP)数据**

分布式系统原理与设计/朱海滨，蔡开裕，樊爱华，宋　辉. — 长沙：国防科技大学出版社，1997.9

ISBN 7 — 81024 — 460 — 4

Ⅰ 分布式系统原理与设计
Ⅱ 朱海滨 蔡开裕 樊爱华 宋　辉
Ⅲ 分布式计算机系统
Ⅳ TP338

责任编辑：何　晋
责任校对：卢天贶
封面设计：陆荣斌



国防科技大学出版社出版发行
电话:(0731)4555681　邮政编码:410073
E-mail:gfkdcbs@public.cs.hn.cn
新华书店总店北京发行所经销
湖南大学印刷厂印装

开本:787 × 1092　1/16　印张:20.5　字数:486千
1997年9月第1版第1次印刷　印数:　1—3000册

**ISBN　7—81024—460—4**
**TP · 96　　定价:24.00元**

# 内 容 提 要

本书主要介绍基于计算机网络的松耦合分布式计算机系统的基本概念、基本原理和基本设计方法。

全书共分十二章，首先介绍分布式计算机系统的基本概念和设计问题；然后根据自底向上的路线，深入讨论网络与通信、分布式文件系统（包括共享文件、合作文件服务器和文件备份等）、实时与容错系统、分布式共享存储器、安全与保护机制的设计与实现；最后两章分别介绍当前典型的分布式系统、当前最新的分布式应用CSCW和自行设计研制的一个CSCW系统原型。

本书内容翔实，由浅入深，循序渐进，全面反映了分布式系统的基本原理和设计技术，可用作高年级本科生或研究生的教材和参考书，也可用作操作系统和计算机网络课程补充教材，对于分布式系统的设计和开发人员也有很好的参考价值。

# 前　言

20 世纪 80 年代以来，随着计算机技术特别是微处理器和网络技术的发展，计算机已经逐步成为大众化的工具。不久的将来，计算机将会成为人们生活中的必需品，出现在商店的电器部柜台上，人们将坐在家中操作个人工作站或微机，通过网络购物、订票。而所有这些都需要分布式计算机系统的支持和帮助。

分布式系统是现在和今后一段时期内计算机领域的研究重点。可以这样说，计算机及其相关专业的学生如果不懂得分布式系统或者分布式处理技术，将难于立足于计算机研究和应用领域。所以，国外的著名大学从80年代便开设了关于分布式系统的课程，我国各个大学也相继开设了相关课程，国内也出版了相关的教材。但由于该领域的研究工作十分活跃，知识技术发展日新月异，教材首当其冲地也必须得到相应的更新，否则将无法适应分布处理技术的发展，我在讲授《分布处理技术》课程时采用过1988年版的外文原版教材，也采用了1994年和1995年版的教材，就发现其中的大量内容得到了更新。

本书便是本着不断更新知识的宗旨，在许多国内版现有教材的基础上，参考国外最新经典教材和国内外最新研究成果写成的。当然，不可避免地，本书同样也将面临着知识老化而被淘汰的危险，比如，为了适应新发展，不得不放弃了本书1996年初稿中将近三分之一的内容。虽竭尽全力，力图从原理、概念和方法上总结、归纳原理性的基本知识，力争使本书的寿命能够相对较长些，但能否达到目的，只有让读者和可能将该书作为教材的老师们去评说了。

本书假设读者已经具备计算机体系结构、操作系统、程序设计和计算机网络的基本知识。当然，如果没有上述这些知识，本书也给出了必要的介绍和铺垫。

本书主要介绍设计和建立分布式系统的基本原则和基本概念。重点讨论分布式系统的设计，包括分布式系统的关键特征、网络与通信基础、分布式系统内核技术、分布式文件系统、分布式共享存储器、实时与容错、安全与保护等，辅助讨论新型分布式应用（CSCW系统）的设计。通过对于本书的学习，可以使学生掌握设计分布式系统所需的必要知识和方法。

本书是应国防科技大学计算机系96教学计划的要求、主要根据给硕士研究生授课时的讲义内容而写成的，兼顾研究生和本科生教学的需要。参考了近两年来国际上流行的分布式系统的教材和相关论文，同时结合我们自己的科研实践，介绍了应用实例系统。本书可作为研究生和大学高年级学生相关课程的教材或参考书，也可供具有一定背景的技术人员自学。本书可以作为传统体系结构和操作系统后续课程的专用教材，也可以用作为操作系统和计算机网络课程的补充教材。本书可以根据需要分别作为40学时和60学时的教程使用，内容可以根据教学时数和教学对象进行适当取舍。比如，如果给修过《计算机网络》课程的本科生讲授40学时，第二、九、十、十二章和11.5、11.6节可以不讲。

本书第一、五、六、七、八、九、十一、十二章和第2.5节由朱海滨执笔，第二章（2.5节除外）和第三章（3.5节除外）由蔡开裕执笔，第十章由樊爱华执笔，第3.5节和第四章由宋辉执笔，全书由朱海滨统稿。

本书的完成得到了多方支持和帮助。感谢国防科技大学教材出版基金的支持，感谢邹鹏教授、王志英教授对本教材出版工作给予的大力支持与鼓励，感谢张晨曦教授、张春元副教授提供了大量最新资料。邓胜兰副研究员为我们提供了许多有关Mach操作系统的工程资料，进一步丰富了本书的实例。感谢王朴教授、沈清教授、王峰助研、博士生吴飞、王克非和王恺硕士等同志在CSCW研究中给予的支持与合作。感谢在美国Oregon Graduate Institute攻读博士学位的张卉小姐，专门从美国寄来最新的英文版教科书，对本书的编著有很好的参考价值。感谢连续几届所有选修《分布与处理技术》课程的研究生，在为他们授课的同时，我每次都有新的收获。

感谢蔡开裕、樊爱华、宋辉同志的合作与支持，没有他们的参与，本书不可能这样及时地与读者见面。

最后，要特别感谢我的妻子张晶多年来的全力支持，没有她的支持，我将既无时间，又无精力来完成本书的编著工作。

虽倾注全力，力争编著出高水平的教材，但毕竟水平有限，另外成书时间仓促，书中谬误之处在所难免，敬请读者不吝指教，以便进一步改进。

朱海滨

1997年6月于长沙

# 目　录

## 第四章　分布式系统核心技术



## 第五章　分布式文件服务



## 第六章　文件共享

# 第一章 导 论

计算机系统的发展非常迅猛，从 1945 年计算机出现以来，计算机经历了几代的更新和变化。由于最初的计算机非常昂贵、庞大，一般都由计算中心建立专用机房放置，并配备专门人员管理，计算机用户一般通过计算机管理员使用计算机。

然而，自 80 年代以来，随着技术的不断发展，特别是微处理器的发展和网络技术的出现，计算机已经逐步成为大众化的工具，计算机逐渐成为人们生活中的必需品。它将出现在商店的电器部柜台上，人们将坐在家中操作个人工作站或微机，通过网络购物、订票。而所有这些都需要分布式计算机系统的支持和帮助。

分布式系统已经用于社会中几乎各个领域，每当开发出一个新的应用系统时，都能发现其巨大的应用价值。计算机网络使新设计的系统可以为远程用户服务，而分布处理技术则为人们设计新的应用系统提供更为先进的环境和系统基础(多个计算机)。

分布式系统这一术语应用非常广泛，既可以是多计算机系统，又可以是采用不同设计方法或面向不同设计目标的多处理机系统，因此，该术语使用的内涵和外延随着讨论场合的不同变化很大。所以，我们首先要确定一下本书重点讨论的问题和本书认定的分布式系统的基本含义。

## 1.1 什么是分布式系统

分布式与集中式系统的区别并没有严格定义，分布式系统的定义主要依据这类系统的特征来描述。国内外学者都做过相应的描述，最简单的描述就是“分布式系统就是一组独立计算机构成的系统，在用户看来好像是一个计算机系统一样”。这句话有两个方面的意义，一是从硬件方面考虑，强调独立、自治的计算机；二是从软件方面考虑，强调系统的整体性。

在这里我们给出以下定义：

分布式系统是由多个相互连接的处理资源组成的计算机系统，这些资源可以合作执行一个共同的任务，最少依赖于集中的程序、数据和硬件等资源。

对于这个定义，需要强调以下几个要点：

第一，分布式系统是由多个处理机或多个计算机组成。

第二，这些计算机或处理机可以物理相邻，用机器内部总线或开关连接，通过共享主存进行通信；也可以在地理上分散，用计算机网络互连，采用消息(或称报文)通信。

第三，这些计算机或处理机组成一个整体，对用户是透明的，即用户使用任何资源时不必知道这些资源在哪里。

第四，一个程序可分散到多个计算机或处理机上运行。

第五，系统的表现与单一系统一样。

## 1.2 硬件观点

虽然所有的分布式系统都含有多个 CPU，但有多种不同的组织方式，特别是这些 CPU 的互连和通信方式更是变化多端。自从具有多 CPU 的计算机系统提出以来，形成了许多不同的硬件组织模式。按 Flynn 的体系结构分类法，这类计算机具有两个重要特征，即单指令流多数据流(SIMD)和多指令流多数据流(MIMD)。

在SIMD计算机系统中，阵列处理机是一种典型结构，它由一个指令部件取得指令，然后将指令同时发往多个数据操作部件并行操作。阵列处理机类似于传统计算机，具有大量的算术与逻辑部件，以规则的阵列连接起来。它们可以用来执行矩阵运算和向量运算等规则操作。其特点是，整个处理单元阵列都服从于一个单一的指令流控制，某些指令可作用于分布于所有处理单元阵列上的数据项。图1-1表示了一个具有少量处理单元的阵列机，一般的阵列机都含有许多单元，如美国的ILLIAC具有256个单元，英国ICL公司的DAP含有1024个单元。这种计算机适用于对大量数据完成相同操作的计算任务，在处理大量有规则的数据时可以获得很高的处理速度，但在处理分布式问题或处理需要并发许多独立任务等问题时效率不高。例如，ILLIAC、ICL DAP和Connection Machine等都是典型的SIMD计算机系统。

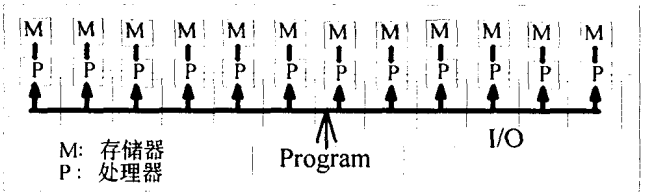



图1-1 阵列处理机

在MIMD计算机中，由独立的处理机执行各自得到的指令对各自的数据进行操作。我们讨论的分布式系统均属于这种计算机系统，这类系统又可以分为紧耦合系统和松耦合系统。

所谓紧耦合的多处理机系统主要在于其是共享主存的，即多个 CPU 共享同一物理地址空间。在紧耦合系统中，一组处理器集成在一起由一个操作系统管理，操作系统负责为用户任务分配处理器和存储空间，并能使这些用户任务并行运行。在具有共享存储器的紧耦合系统中，由于大量的处理单元共享单个存储器或同一地址空间，所以处理器的个数一般受到存储器带宽的限制。

在松耦合的多计算机系统中，每个 CPU 都具有自己的局部存储器，构成计算机，每个计算机都有自己独立的操作系统，操作系统通过通信和协作完成用户任务的并行执行。

在这类分布式系统中，硬件还包括多 CPU 与共享存储器，或高速互连的分离、但具有统一的虚拟地址空间的多处理器/存储器。这样可以使用户任务像在传统的单机中一样通过共享变量或共享表格与操作系统或其它用户任务通信。在分布式系统中，访存带宽竞争的问题可以通过高速缓存(Cache)技术缓解。

这两类分布式系统又可分为基于总线的结构和基于交换的结构(图 1-2)。所谓总线结构，是指系统中存在多个 CPU 共享的总线，用于传输数据。而在交换结构中，没有公共

总线，而是在 CPU 中之间建有专用的数据通路。

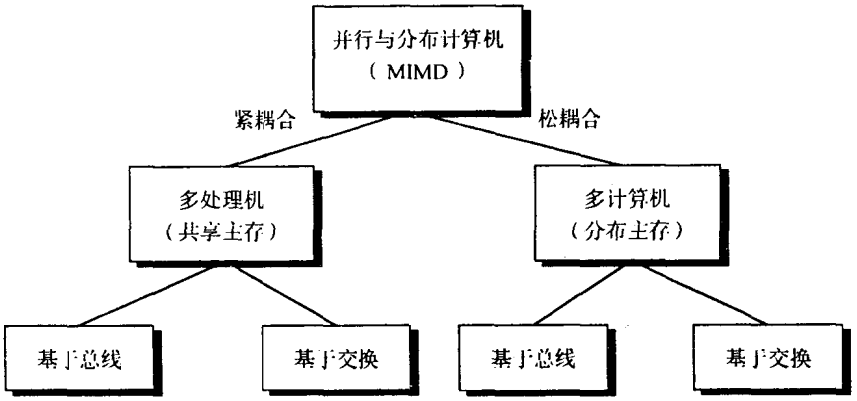



图 1-2 分布式计算机系统的分类

1.基于总线的多处理机

在这种结构的多处理机中，每个 CPU 都与总线直接相连，存储器也与总线相连，即多个 CPU 通过总线共享存储器。典型的总线有 32 位总线和 64 位总线，总线又分为地址总线、数据总线和控制总线，各位之间并行处理。如果一个 CPU 要从存储器中读一个字，首先要将该字的地址放到地址总线上，然后，将适当的控制信号放到控制总线上表示读，存储器将该字的值放到数据总线上，让发出请求的 CPU 读取该字；对于写操作，CPU 要同时将地址和数据分别放到地址总线和数据总线上，并在控制总线上施加写信号，存储器在写信号的控制下完成数据写入。

由于只有一个存储器，如果一个 CPU(A)要向存储器中写一个字， 另一个 CPU(B)在此之后要读取该字，那么， CPU(B)应该读取 CPU(A)刚刚写进去的值。具有这一特征的存储器可称为一致的存储器，一致性在分布式系统中是一个非常重要的原则，并且有许多不同的表现方式，如读写一致、写读一致和写写一致等。

在总线模式下，主要的问题是总线上只能挂少量(几个)的 CPU ，如果挂得太多，性能将会明显下降。其解决办法是利用高速缓存 Cache ，将 Cache 置于总线之间，如图 1-3 所示。 Cahce 中存放最近访问的字，所有访存请求均先由 Cache 响应，如果待访问的字在 Cache 中，则可直接响应，不必再申请总线，如果 Cache 很大，则成功的可能性(称为命中率)将很高，可使总线通信量大大减少，这样可以在总线上多挂 CPU 。一般而言， Cache 在 64K 字到 1M 字之间，可以达到 90 % 以上的命中率。

但是， Cache 的引入也带来了新的、并且是比较严重的问题，即一致性问题。例如，若不加控制，两个 CPU(A 和 B)都将同一字读到自己的 Cache 中，然后， CPU(A)改写了这个字，当 CPU(B)读该字时，便从自己的 Cache 中读取了未修改的字，出现不一致性。因此，针对这一问题进行了许多研究，也提出了许多解决办法，一种常见方法就是写穿透(Write Through)方法，即当每次写 Cache 时，同时也写存储器，在这种方法中，写仍然需要申请总线，读命中时，则不需要占用总线。仅仅这样处理这还不够，每个 Cache 还需要监控总线，当一个 Cache 发现其中的字被修改时，它还需要清除其中的副本，或者读取新修改的存储器中的内容。这种 Cache 叫做监听 Cache(Snoopy Cache)。

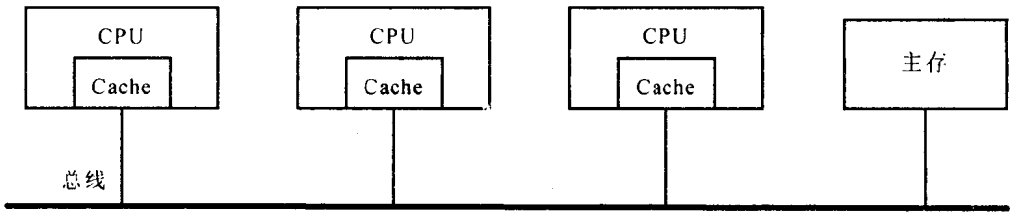
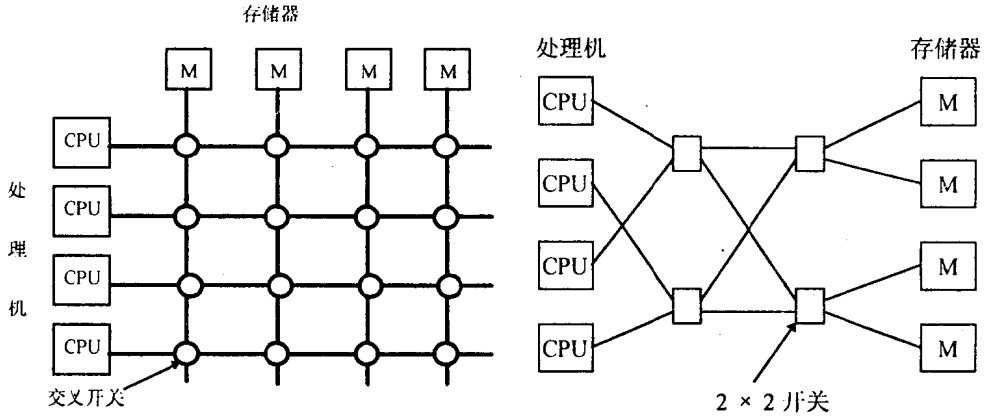



图 1-3　基于总线的多处理机

具有监听 Cache 、并采用写穿透方式的多处理机中，可以保证存储器的一致性，其组织结构对于程序员透明，减轻了程序员的负担。目前，几乎所有基于总线的多处理机结构都采用类似的 Cache 机制，在这种机制的支持下，一条总线上可以挂上几十个 CPU 。

2.基于交换的多处理机

要采用更多(如上百个)的 CPU 构成多处理机，必须采用不同的组织方法来连接 CPU 和存储器，一种方法是将存储器分成模块，然后用交叉开关互连(图 1-4(a))，每个交叉点是一个电子开关，每个 CPU 与每个存储器通过开关都可以直接相连。当 CPU 要访问一个确定的存储器模块时，相应的交叉开关立即合上，使它们直接相连，然后直接访问。交叉开关的实质就是许多 CPU 可以同时访问存储器。当然，如果有两个以上的 CPU 要访问同一个存储器，仍需要等待。




(a)交叉开关　　(b)Omega 开关网络

图 1-4　基于交换的多处理机

假如有 n 个 CPU 和 n 个存储器，其交叉点便是 $n^2$，随着 n 的增大，交叉开关的代价相当昂贵，完全采用这种方法并不一定能获得较高的性能价格比。于是，有了一种采用交换开关的分级网络形式。图 1-4(b)表示了一个 2 × 2 的 Omega 网，其中，每个开关可以将每个输入连接到每个输出，通过这些开关，每个 CPU 可以访问到所有存储器，开关状态可以在纳秒时间量级内切换。

一般情况下，若有 n 个 CPU 和 n 个存储器，则 Omega 网需要 $\log_2 n$ 级，每一级包含 n/2 个交换开关，总共需要$(n\log_2 n)/2$ 个，比起交叉开关方案来，其开关数要少得多，但数量仍然较大。进一步看，还有一个延迟问题。例如，对于 n = 1024 ， CPU 与存储器之间

将通过 10 级开关互连，则从请求到返回需要 20 级的开关延迟。假设 CPU 是当今的 RISC 芯片，主频为 100MIPS，那么，指令执行时间为 10ns，如果存储器请求在 10ns 中完成，那么开关延迟便成了瓶颈；否则必须将开关限定在 0.5ns 之内，那么，整个多处理机就需要 5120 个 0.5ns 的开关，价格将更昂贵。

为了节省开支，便推出了层次性的系统，在这种结构的系统中，有些存储器与每个 CPU 直接相连，每个 CPU 都可以快速地访问自己的局部存储器，但是，若要访问其它 CPU 的局部存储器则比较慢。这种结构叫做 NUMA(非一致存储器访问—— NonUniform Memory Access)结构。

虽然 NUMA 结构的机器比 Omega 网的平均访问时间要短，但它又带来新的复杂性，使程序和数据的放置(Placement)问题成为关键，也就是要求优化放置，尽可能得到较高的局部存储器访问率。

总的来说，基于总线的多处理机受限于其通信能力，最多挂接几十个 CPU 。若要增加 CPU 个数，必须利用开关网络，如交叉开关或 Omega 交换开关。大量的交叉开关是非常昂贵的，而大量的 Omega 交换开关则又贵又慢。 NUMA 计算机则需要复杂的算法来保证良好的数据和软件放置。因此，建立大量的紧耦合的共享存储器的多处理机虽然是可能的，但非常困难且又造价昂贵。

3.基于总线的多计算机

相对于多处理机而言， 多计算机的建立比较容易，每个 CPU 都与自己的局部存储器直接相连，唯一的问题是 CPU 之间如何通信。很明显，这里需要一种互连模式，但由于仅仅是 CPU 之间的通信，通信量会明显少于同时用做 CPU 与 CPU 之间、 CPU 与存储器之间的互连网。在图 1-5 中，其拓扑结构与基于总线的多处理机类似，但其通信量明显减少，它不必使用高速总线，实际上， 10 ~ 100Mbps 的局域网已相当实用，比起 300Mbps 的高速总线要求低得多。图中表明，构成多计算机只需要不同的工作站通过局域网互连即可，不必利用 CPU 板通过高速总线互连。

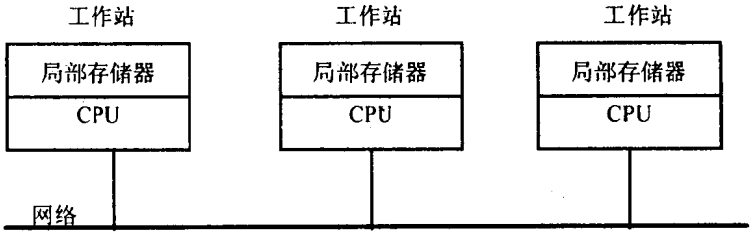



图 1-5 由局域网和工作站组成的多计算机系统

4.基于交换的多计算机

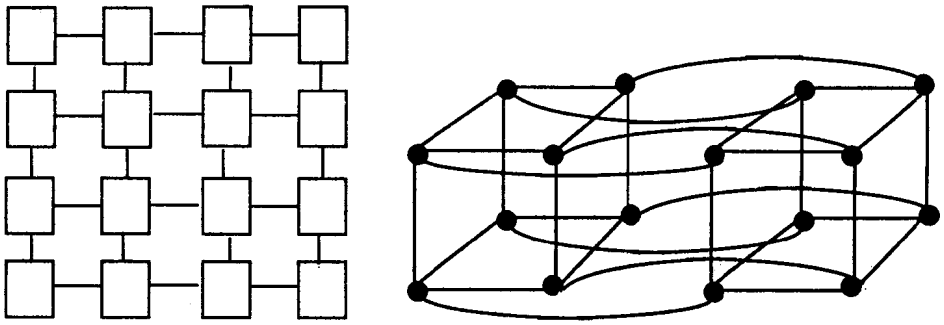
在基于交换的多计算机结构中，仍要保持每个 CPU 只与特定的局部存储器直接相连，互连结构仍是 CPU 之间的互连。图 1-6 中表示了两种连接方式。一种叫做栅格结构，一种叫做超立方体结构。

栅格结构容易理解，就像在一块插件板上有多个 CPU，相邻 CPU 之间互连，这种结构适用于具有二维特性的计算任务，如图论和视觉问题等。

超立方体结构是一个 n 维的立方体，图中表示了四维超立方体。可以这样认为，它是

由两个普通的立方体(由 8 个结点和 12 条边形成)构成，每条边表示两个 CPU 之间直接相连，两个立方体中对应结点相互连接。

若要构成五维的立方体，我们必须再加一组四维立方体，然后将对应的结点相连。对于 n 维立方体，每个 CPU 都与其它 n 个 CPU 相连，因此，线路的复杂性与结点数成对数形式增长。由于只有相邻的结点之间相连，所以，许多消息必须要通过多个结点转发才能到达目的结点。但是，这种结构中，最长路径与结点数也只以对数方式增长，而栅格结构则以结点数的平方根方式增长。例如，具有 1024 个结点的超立方体结构已经使用多年，现在具有 16384 个结点的结构也已实用。



(a)栅格结构　　(b)超立方体结构

图 1-6　基于交换的多计算机

## 1.3 软件观点

对于分布式系统，软件和硬件是两个不可分割的整体，从程序员或应用的角度来看，软件的概念更为重要。虽然操作系统的结构不像硬件体系那样规整，但仍然可以划分成两种操作系统类型，即紧耦合的系统和松耦合的系统，这两种系统分别对应于两类硬件系统结构。

在松耦合的系统中，机器和用户独立工作，在必要的时候可以进行通信。例如，一组个人计算机(PC)通过局域网互连共享资源，如激光打印机或数据库等，每个 PC 都具有自己的 CPU 、存储器、硬盘和操作系统，这组 PC 构成的系统便是松耦合的系统，每台 PC 都是独立的，各自完成自己的任务。如果需要，可以通过网络访问共享资源，如果网络出问题，每台 PC 仍然可以继续工作。

在紧耦合的系统中，各台机器合作完成同一个任务。例如，一个多处理机系统运行一个下棋程序，每个 CPU 都设有棋盘来记录对奕过程并评判，每个 CPU 都要做评判工作，所有其它 CPU 的棋盘都可能通过该棋盘来生成，当评判棋局结束时，所有 CPU 上的棋盘都要重新复位。这一系统中的软件，包括系统软件和应用软件，都要支持这一要求，从而形成非常紧密耦合的系统。

我们从硬件概念上可以分为四种类型的系统，从软件概念上分为两种不同类型的系统，那么从理论上讲，硬件与软件结合而成的分布式系统类型应当为八种。但实际上，只有四种类型有实际意义，因为多处理机硬件无论使用总线还是使用交换开关都只能配备紧耦合的软件系统。下面我们讨论最常见的软硬件的组织情况。

1.网络操作系统

这是一种典型的松耦合的软件与松耦合的硬件相结合形成的系统。典型的组成是一组工作站由局域网互连在一起，其中，每台工作站上安装网络软件。在这种系统中，用户可以利用有盘或无盘工作站工作，所有的命令和程序均在工作站上运行。同时，用户也可以根据需要进行远程登录，利用其它工作站工作。命令如下:

rlogin 机器名

该命令的效果是让用户当前使用的工作站成为远程工作站的仿真终端，程序运行和命令执行均由远程工作站完成。若要使用另一个远程工作站，必须首先从现用的远程工作站上退出(logout)，再用远程登录命令使用这个远程工作站。任何时候只能使用一台远程机器，机器的选择由用户确定。

网络软件还支持远程拷贝命令，使用户可以在不同的工作站之间拷贝文件。命令格式如下:

rcp 机器名 1：文件名 1 机器名 2：文件名 2

表示将“机器名 1”代表的工作站上的“文件名 1”表示的文件拷贝到“机器名 2”表示的工作站中名为“文件名 2”的文件中。在此，用户必须非常清楚地了解文件拷贝中的源机器和目机器。

从上面可以看出，网络操作系统给用户提供的支持是非常低级和初步的，应当提供更为有效的手段，一种方法就是提供为所有工作站可访问的全局共享文件系统，该文件服务器可以从工作站(称为客户机)上的用户程序中接受文件请求，并检查、执行该请求，最后返回结果(图 1-7)。

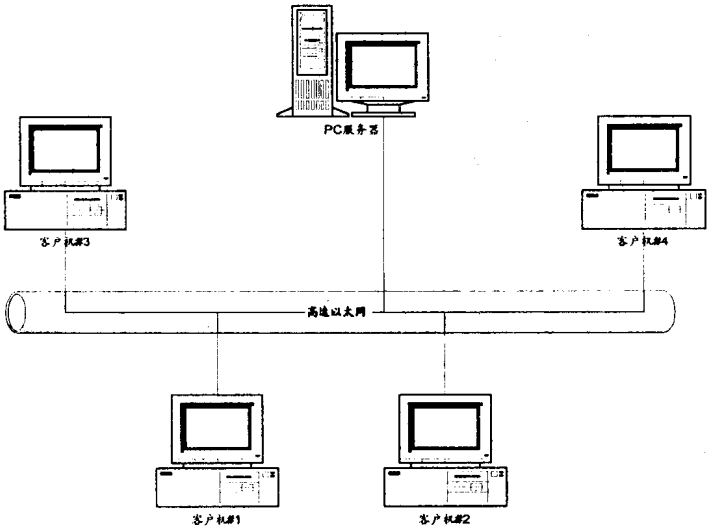



图 1-7 四台客户机和一个服务器通过网络互连

文件服务器一般管理一个层次化的文件系统，该系统都有一个根目录，根目录中包含子目录或文件，工作站可以将其中的子目录安装到它们自己的本地文件系统中。例如，图 1-8 中表示有两个文件服务器，一个具有名为 games 的目录，另一个具有名为 work 的目录。每个目录都包含多个文件，每个客户机都安装了这两个服务器，但用不同的安装方式。客户机 1 将这两个目录都安装到了根目录上，从而可以这样访问两个目录，即\games 和\work 。而客户机 2 将 games 目录安装到了根目录上，而将 work 目录安装为 games 的子目录，要这样访问两个目录，即\games 和\games\work 。

虽然对于客户机而言，如何安装服务器的目录并不重要，但要注意不同的客户机在这