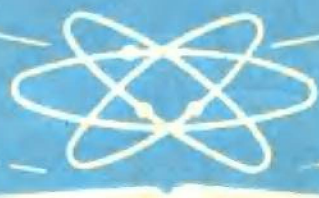
中等专业学校教材

操作系统

楼星 盛建 编



西北电讯工程学院出版社

中等专业学校教材

操 作 系 统

楼 星 盛 建 编

西北电讯工程学院出版社

1 9 8 6

内容简介

本书主要阐述操作系统的基本概念和基本原理。全书共分为八章。第一章讲述了操作系统的概况；第二章至第六章分别介绍了操作系统对各种资源的管理；第七章简要介绍了操作系统的结构及其设计要点；第八章介绍了UNIX操作系统的主要数据结构和特点。

本书在文字上力求通俗易懂，可作为中等专业学校计算机专业教科书，也可作为一般科技人员学习计算机的参考书。

中等专业学校教材

操作系统

楼星盛建编

西北电讯工程学院出版社出版
西北电讯工程学院印刷厂印刷
陕西省新华书店发行　各地新华书店经售
开本 787×1092 1/16　印张 14 2/16　字数 342 千字
1986 年 6 月第一版　1986 年 6 月第一次印刷　印数 1—13,000

统一书号：15322·52　　定价：2.10元

出 版 说 明

根据国务院关于高等学校教材工作分工的规定，我部承担了全国高等学校工科电子类专业课教材的编审、出版的组织工作。从一九七七年底到一九八二年初，由于各有关院校，特别是参与编审工作的广大教师的努力和有关出版社的紧密配合，共编审出版了教材159种。

为了使工科电子类专业教材能更好地适应社会主义现代化建设培养人才的需要，反映国内外电子科学技术水平，达到“打好基础、精选内容、逐步更新、利于教学”的要求，在总结第一轮教材编审出版工作经验的基础上，电子工业部于一九八二年先后成立了高等学校《无线电技术与信息系统》、《电磁场与微波技术》、《电子材料与固体器件》、《电子物理与器件》、《电子机械》、《计算机与自动控制》，中等专业学校《电子类专业》、《电子机械类专业》共八个教材编审委员会，作为教材工作方面的一个经常性的业务指导机构，并制定了一九八二～一九八五年教材编审出版规划，列入规划的教材、教学参考书、实验指导书等共217种选题。在努力提高教材质量，适当增加教材品种的思想指导下，这一批教材的编审工作由编审委员会直接组织进行。

这一批教材的书稿，主要是从通过教学实践、师生反映较好的讲义中评选择优和从第一轮较好的教材中修编产生出来的。广大编审者，各编审委员会和有关出版社都为保证和提高教材质量作出了努力。

这一批教材，分别由电子工业出版社、国防工业出版社、上海科学技术出版社、西北电讯工程学院出版社、湖南科学技术出版社、江苏科学技术出版社、黑龙江科学技术出版社和天津科学技术出版社承担出版工作。

限于水平和经验，这一批教材的编审出版工作肯定还会有许多缺点和不足之处，希望使用教材的单位、广大教师和同学积极提出批评建议，共同为提高工科电子类专业教材的质量而努力。

电子工业部教材办公室

前言

本教材系由中专电子类教材编审委员会计算机编审小组审定，并推荐出版。

本教材由北京无线电工业学校楼星、盛建编写。贵州无线电工业学校黄大胜担任主审。

本书共分八章。第一章通过对批处理系统、分时系统和实时系统的介绍，使读者对操作系统的基本内容和概念有一概括的了解。第二章对作业的组织及作业的管理和调度方法作了详细介绍。第三章详细讲述了进程的基本概念、进程的调度方法及死锁等问题。第四章系统地介绍了存贮管理的基本概念及各种存贮管理方法。第五章介绍了对设备的管理方式，并详细介绍了缓冲技术在设备管理中的作用。第六章详细介绍了以文件方式管理计算机系统中信息的方法。第七章对操作系统的结构及其设计要点作了简单的介绍。第八章介绍了UNIX操作系统的主要数据结构和特点。

本书是按80学时编写的。

本书第一、二、三、四章由楼星编写，第五、六、七、八章由盛建编写。本书在编写过程中得到过许多同志的帮助，特在此表示感谢。由于编者水平有限，书中错误和不妥之处在所难免，恳请读者批评指正。

编者 1985年9月

第一章　操作系统概述

自1946年世界上第一台电子计算机问世以来，计算机科学技术得到了迅速发展，出现了许多速度快、容量大、功能强、外部设备齐全、软件丰富的计算机系统。高性能计算机系统的应用，推动了现代科学技术的发展，开创了信息处理的新时代。尤其是七十年代初期，微型计算机的出现，使计算机的应用进入了社会各个领域，对国民经济的发展，人们的生活方式和工作方式都产生了深远的影响。

现代计算机系统是由许多硬件设备和软件设备组成的，前者是指中央处理机(CPU)、存贮器和输入/输出(I/O)设备等物理资源；后者是为了方便用户和充分发挥计算机效能的各种程序和数据的总称，包括汇编程序、编译程序、控制程序、操作系统、诊断程序、专用程序包、数据管理系统以及各种应用程序等。计算机能够广泛应用于各行各业，除了硬件的迅速发展外，是和计算机软件的发展分不开的，尤其是配置了各种功能的操作系统以后，把计算机的应用水平提高到一个崭新的阶段。

操作系统是计算机系统中极为重要的系统软件，它是在硬件基础上利用计算机自身的处理能力设计出来的一整套系统程序。它对计算机系统中各种硬、软资源进行合理的管理和控制，大大提高了资源的利用效率，并为用户提供了一种简单而有效的使用方法。

操作系统和其它事物一样，也有它的产生和发展过程。本章主要介绍操作系统的基本概念和基本功能，并通过对批量处理系统、多道程序系统、分时系统和实时系统的简单介绍，使读者对操作系统的产生和发展，对几种类型的操作系统的基本特点有一初步的了解。

§1　操作系统的概念和功能

§1.1　操作系统的概念

操作系统(Operating System)是指用以控制和管理计算机硬件资源和软件资源的程序集合。它是扩充计算机功能、方便用户的系统软件。

通常我们使用计算机，从宏观上看，是使用整个计算机系统，但从微观上看，是使用各种硬件资源和软件资源。怎样组织管理这些资源，使之既能相互协调，又高效率地完成各种计算任务呢？操作系统就对计算机的全部资源科学地进行分配、控制、调度和回收等工作。

如果从软件的角度静态地看操作系统，它是各种系统程序和数据结构(包括各种表格)的集合。它是由指挥和管理系统运行的各种程序及数据结构组成的一个大型系统软件。

计算机有了操作系统之后，用户就不再直接使用裸机(不包含软件的硬件机器)，而是通过操作系统来控制和使用计算机。因此，可以称它是介于用户和裸机之间的一个接口。

从用户的角度看，操作系统把计算机扩充为功能更强、服务质量更高、使用更加方便灵活的计算机系统。例如，在没有操作系统的情况下，若用户程序所要求的内存空间大于实际容量时，就无法运行，操作系统可以把实际内存扩充成为若干个能够满足用户要求的虚拟存贮器，使每个用户程序都能投入运行。同样，操作系统还可以把处理机扩充成为多个用户共享的

虚拟处理机。

为了提高计算机系统的利用率，往往有几道作业，甚至几十道作业同时在计算机系统内并行运行。要组织控制这些作业的正常运行是一个很复杂的问题，是任何一个操作员都难以胜任的。然而，在操作系统的控制下，却可以实现计算机算题过程自动化。例如，我们要求计算机完成某程序的计算工作，用户只要按规定编写好作业说明书，把它和程序一起交给操作员，操作员只要做必要的准备工作，将其输入计算机系统，该程序就完全在操作系统的控制下运行，直至完成。如果有几道、甚至几十道作业要同时运行，亦如上所述，不需要操作员一一控制，而是在操作系统控制下自动完成。

由以上讨论可知，操作系统是现代计算机系统的重要组成部分，它提供的各种强有力的服务功能，使得用户能够更加方便灵活地使用计算机。特别在当今，具有高度并行处理能力的计算机系统发展很快，各种类型的操作系统的发展亦很快，因此一个从事计算机专业的科技工作者，如果不了解操作系统，就不可能透彻了解计算机系统的动态工作过程，亦无法真正掌握一个计算机系统，并充分发挥其处理能力，更不可能从事计算机系统的设计和调试工作。

§1.2 操作系统的基本功能

研究一个操作系统，可以从不同的角度来观察和剖析，因此就有不同观点。归纳起来主要有三种，即资源管理观点、进程观点和分层虚拟机观点。这些观点之间是有密切联系的，了解每一种观点都有助于全面理解和分析操作系统。在这里，我们先从资源管理程序的观点来说明操作系统的基本功能。对其它两种观点将在第三章和第七章中分别阐述。

我们应该明确资源管理的目的：(1)方便用户；(2)提高各种资源的利用率；(3)保证安全可靠。为实现这些目的，对每一种资源管理而言，操作系统应当解决的是一些普遍性问题和方法，即：(1)掌握资源的使用状况；(2)确定资源的分配和调度原则；(3)执行分配这种资源的算法和操作；(4)回收这些资源的算法和操作。

操作系统要管理的资源很多，通常可以将它们分为四大类，即处理机、存贮器、I/O设备及信息(程序和数据等)。正是这些资源使操作系统本身及用户作业有了活动的物质基础和工作环境。对一个操作系统而言，资源使用方法和管理策略的确定是很重要的，它是决定一个操作系统的规模、类型、功能及其实现过程的重要因素。所以，从资源管理观点要求操作系统至少应有这四方面的管理程序，即：控制和管理处理机的程序；管理存贮器的程序；控制管理I/O设备的程序和管理信息的程序。正是这一整套控制管理程序，使计算机能有条不紊地完成各种资源管理和控制工作，因此，资源管理观点亦可称为资源管理程序的观点。

按照资源管理观点，可以把操作系统的基本功能归纳为下列五个方面：

1. 处理机管理。处理机是计算机系统中关键性资源。为了提高处理机的利用率，在现代计算机系统中常常有多道作业同时运行。因此，处理机管理的主要任务是如何根据一定原则，作好处理机的调度工作。这在多道程序设计的情况下是很复杂的，必须设计一些专门性程序来实现这一功能。

2. 存贮管理。存贮管理的主要任务是对内存资源进行合理分配。因此，必须随时掌握存贮空间的分配情况，并进行存贮空间的分配和回收工作。由于多道程序设计的引入，给存贮管理带来复杂性，即要让多个做业共享内存，保证它们之间互不侵犯；还要解决内存容量不足的问题，即把一个实际存贮器扩充为比它本身容量大得多的虚拟存贮系统。

3. 设备管理。设备管理要求充分发挥通道和主机、通道和通道、设备和设备并行工作的能力，还要提高使用外设的灵活性，使用户避免直接引入设备名。也就是说用户程序的输入输出操作与实际设备无关，仅由操作系统决定。这样，有些独占设备（如穿孔机、纸带输入机等）就可以搞成虚设备。例如，当作业需要穿孔机穿孔输出操作时，就可以改为“写入”磁盘，以后再由一个专门子程序把这些信息输出，对此设备管理的任务，就应记录好每一设备的状态，以便确定一种有效的办法把该设备分给某作业，并在完成I/O操作后及时收回该资源。

4. 文件管理。在现代计算机系统中，操作系统不仅把程序、数据、各种信息，甚至把外设都当作文件来管理。通过文件系统的管理，把存贮容量很大，但不易使用的外存贮器改造成为可以按名存取，方便灵活，又可共享的文件空间。所以文件管理必须记载好文件所在的位置及使用情况，决定文件存取的权限和保护要求，并实现文件的打开、关闭等一系列操作。

5. 作业管理和控制。可以说，计算机的一切工作都是为了完成作业。当有多道作业在计算机系统中运行时，操作系统如何控制计算机的作业流程呢？通过作业管理和控制，可以把一个实际的联机控制台改造成为每道作业都能使用的若干个虚拟控制台，并顺利地控制完成各个作业流程。

从以上讨论可知，一台计算机系统经过操作系统的一系列改造和扩充，可以成为内存容量充分大，外部设备足够多，外存很好用，控制台相当多，对用户更加方便灵活，功能更强的虚拟计算机系统。本书的第二章到第六章将进一步详细叙述上述各种功能。

§2 批量处理系统

§2.1 作业、作业步和作业流

“作业”这一名词大家都很熟悉。在这里，为了能全面准确地描述用户的一次上机算题工作，我们引用了“作业”这一术语。通常把用户一次提交给计算机的一个具有独立性的计算任务称为作业。从计算机系统角度看，作业也可定义为计算机系统为用户一次上机算题所要做的全部工作。它是各种类型操作系统管理用户算题任务的独立单位。如果静态地观察一个作业内容，则应由用户程序及其所需要的数据结构、有关命令和说明等构成。

计算机要完成一个作业应做的工作可分为若干有序的处理步骤。例如输入、编译、装配、运行、输出等，其中每一步骤称为一个作业步。作业步是由完成作业过程中某一相对独立事件的程序和数据构成，并用一定的命令定义之。例如上述为用户作业所做的工作即可分为编译步、装配目标程序步及运行目标程序步等。如果把一批作业有次序的排列在一起让计算机依次逐个地去完成，使作业一批批地进入系统，经过处理后又一批批退出系统，形成源源不断的作业流。在批量处理系统中，就是把作业按一定的原则组成作业流，由系统监督程序自动控制完成。

§2.2 批量处理系统

五十年代，以电子管为代表的第一代计算机由于其容量小，速度慢，软件亦刚开始发展，只有汇编语言和少量服务性程序，操作系统尚未出现，计算机的使用基本上是手工操作方式，即操作人员把预先穿好的某作业纸带（或卡片）装进纸带输入机，将程序和有关数据

存入磁带，然后通过控制台开关调入程序并启动其运行。在运行过程中，若出现各种问题也是由操作员利用控制台开关进行控制的。计算机的全部资源亦被该用户所独占，直到该作业完成，才能让下一个用户作业上机操作。可见，这种人工操作方式有如下特点：

（1）用户独占全机资源。一个用户上机算题，计算机的全部资源均为该用户独占。这对用户而言当然是很方便的，然而计算机各部分资源的利用率却很低，因为在某一时刻，只允许计算机的某些部分去执行某一作业步，其他资源则处在空闲状态，有些资源该作业根本不需要，也被用户独占。所以在手工操作时期，机器的利用率是很低的。

（2）中央处理机等待人工操作。算题的一切操作都是联机的，致使CPU常常处在等待人工操作的空闲状态。由于用户纸带（或卡片）只能在该作业上机时才允许装入，而人工操作速度又不可能太快，所以在装纸带时机器必须等待。同样当作业完成后需要卸带取卡，或运行中需要人工干预时，机器也只能空等，尤其当运行的作业比较短时，这样的等待更为频繁。

可见，人工操作方式严重地影响机器利用率的提高，这在计算机运算速度比较慢的初期还不十分突出，随着计算机科学技术的发展，计算机的规模和运算速度都不断增大，这种矛盾就突出起来。例如某作业在速度为1万次/s机器上需运行1h；而在速度为60万次/s的计算机上就只要行运1min；如果在速度为600万次/s的机器上则只需要运行0.1 min；而人工操作时间不可能有太大的变化，例如3 min，其人工操作时间和机器运行时间之比各为1:20、3:1、30:1，这个比值的变化充分说明该作业在600万次/s的计算机上运行时，主机仅运行0.1 min就要等待3 min，如果机器速度再增大，则会导致计算机的绝大部分时间都处在空闲等待状态。这种严重的“人机矛盾”，就使得资源利用率迅速下降。

早期，为了改善这种情况，人们把若干个作业合成一批，制成纸带或卡片，通过输入机存入磁带，以后就由监督程序把该批的第一个作业调入运行，当计算完成后就自动再调入该批的第二个作业，这样就可以不间断地处理各个作业直到该批作业全部处理完毕才转去输入第二批作业，从而实现了作业间转换的自动化，缩短了作业建立和人工操作时间，较好地解决了“人机矛盾”。这种从一个作业自动过渡到另一作业的工作方式，就是早期的批量处理。早期批量处理系统的组织如图1-1所示。显然在早期批处理系统中，作业的输入输出都是联机的。作业信息从纸带输入到磁带，从磁带调入内存运行以及输出计算结果等工作都是在CPU直接控制下进行的。随着计算机运算速度的提高，它与I/O设备的运行速度之间的差距就愈来愈大，因而CPU和I/O设备间速度不匹配的矛盾就越来越明显。为了解决这一矛盾，在批量处理中，引进脱机输入输出的技术，即在主机之外另设一台小型卫星机，用于专门处理输入输出工作，如图1-2所示。输入机上的作业通过卫星机输入到磁带上（输入带），主机

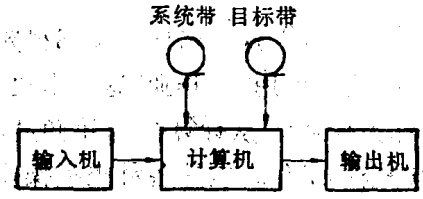


图1-1　早期批处理系统示意图

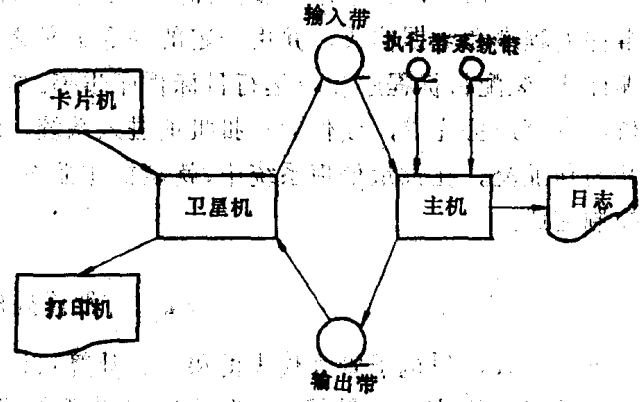


图1-2　脱机批量处理模型

从输入带上把作业调入内存运行；作业运行完以后，主机把结果记录在输出带上，由卫星机负责输出工作。这样，主机不仅摆脱了慢速的输入输出工作，而且还可以通过卫星机使慢速的外围设备和主机并行工作，即在主机调入作业运行的同时，卫星机亦控制外围设备进行输入输出工作，从而提高了主机的利用效率。

现代批处理系统的基本思想是：在内存中同时存放经过合理搭配的几道作业，例如把运算量大的和输入输出量大的作业搭配在一起，以便主机和外设都能不停地并行工作，其特点是多道和成批。多道是指内存中同时存放多个作业，辅存中存放大量后备作业，以便随时调度执行；成批是指作业可一批批输入系统，但作业一旦进入系统，用户就不能直接干预，直到该作业运行完毕（包括正常运行完成和出错），才允许用户根据结果去分析作业的运行情况，做相应地处理。显然，批处理系统大大提高了系统的利用率和作业的吞吐能力。

§2.3 批量处理系统的功能

为了使批处理系统能有效地工作，必须为系统配置相应的软件，即配置具有各种功能的系统程序。监督程序（Monitor）就扮演了“总管”角色。它负责装入需要运行的各种系统处理程序，并完成从一个作业到另一作业的自动过渡工作。

当系统接到用户通过某种形式发来的命令后，首先应判断命令的正确性，若无误则通过查“命令表”找该命令处理程序的入口地址。若查不到则认为该命令是非法的，转出错处理。因此在系统中必须配置相应的命令解释程序来完成上述工作。

批量处理系统中作业是成批输入的，因此系统必须具备识别一个作业（或作业步）的开始和结束标志的能力，作业定序程序就执行这个工作。一个作业完成时，它便自动检索“作业描述表”（该表中记录作业标识符和有关作业的某些限制等），从中找到下一个作业并把它调入内存投入运行。在监督程序中，就是由上述的命令解释程序和作业定序程序来实现作业控制功能的。

系统中内存容量是有限的，不可能把系统程序和用户程序全部装入内存，而是把暂时不用的程序存放在后援存贮器中，仅当需要时才把它们调入内存，在调入之前还必须找到一个适当的内存空间。另外，编译后的源程序还只是单个程序的集合。尚待把它们装配成一个完整的目标程序，系统还应该找到一定的内存空间来存放装入程序，然后才能运行。所以监督程序必须具有分配管理内存的功能。

我们知道，系统在完成作业过程中首先必须通过输入机将作业输入，当作业运算完成后，又要把结果由输出机输出，在作业运行过程中，系统还要不断地把执行中有关的信息告诉操作员。因此，在监督程序中还应具有读写控制的功能。

在具体实现I/O操作时，必须启动指定的I/O设备，处理来自I/O设备的中断请求等，通常是由设备处理程序来实现的。

为便于管理，系统还把一些数据、信息，甚至低速的I/O设备都看成文件来管理，在批处理系统中，是用读写控制、I/O设备处理以及文件管理程序来完成I/O设备管理功能的。

综上所述，为使批量处理系统能有效地工作，其监督程序应具备作业控制、内存管理、设备管理及文件管理等功能。

批量处理系统的出现，缩短了手工操作时间，改善了主机和I/O设备之间速度不匹配的矛盾，亦促进了软件的发展，但仍存在不少缺点。如上述的监督程序、系统处理程序和用户

程序之间的地位是平等的，它们之间是一种互相调用关系，而不是控制和被控制的关系。因此，如果某用户程序出现一条非法指令或用户程序陷入死循环时，就会使系统瘫痪，更严重的是无法防止用户程序破坏监督程序和系统程序，一旦此类事件发生，就可能使整个批处理系统产生混乱。

六十年代初期，由于通道和中断技术的出现，为主机运算和输入输出工作提供了充分并行的可能。通道是一种硬件机构，它能控制一台或多台外设独立地完成输入输出工作。它与主机的同步问题，最初是用一种询问指令。由主机不断发询问指令以询问通道工作完成与否，若未完成则继续询问，直到完成为止。这种循环询问，显然要浪费主机不少时间。因此，引入了中断技术，即在输入输出结束或发生某些故障时，相应的硬件机构就向主机发出中断请求，主机接到信号后马上停止它原来的工作，转去处理中断请求，待处理完毕再回到原来中断的地方继续工作。为提高主机利用率，使通道工作和主机运算能充分并行，又引入了缓冲技术，即把主机所需的输入信息由通道提前输入到缓冲区，主机把输出信息先送到缓冲区，由通道独立输出，以此减少主机的等待时间。为管理好中断和控制输入输出工作，必须增加中断处理程序和输入输出控制程序(IOCS)，由于它们对很多程序都起着指挥和控制作用，需要常驻内存。在系统中，为保证内存空间足够大，把一些系统程序，如编译程序、装配程序等放在外存，而把经常用到的起控制指挥作用的系统程序放在内存，具有这种特点的系统，称为执行系统。而这些常驻内存的程序总称为执行程序。

执行程序和监督程序的显著区别是：执行程序对其它程序拥有指挥控制权，它和用户程序之间的关系不是平等的调用关系，而是控制、被控制的关系，其他程序是在其指挥和控制下工作的。这样，系统就可以对不合法的要求进行检查，因而提高了系统的安全性；另一方面，由于执行系统发挥了通道和主机的并行性，亦提高了系统的效率。但执行系统仍然存在不少缺点，例如批处理系统和执行系统都是脱机操作方式，用户不能对其作业进行控制或修改；虽然主机和通道可以并行操作，但由于只有单道作业运行，主机仍有较多的时间浪费在等待输入输出工作上，尤其当用户程序的输入输出工作量较大时，主机就要等待更多的时间。为了克服这些缺点，促使了多道成批系统和分时系统的产生。

§3 多道程序系统

六十年代初期，执行系统使用不久，人们就发现：如果能在内存中同时存放几道用户程序，并允许它们交替运行将是十分有利的。当一道程序因某种原因而不得不暂停时，系统就马上让另一道作业投入运行，从而使处理机得到充分利用。在这种思想指导下，出现了多道程序设计和多道程序系统。

§3.1 多道程序的特征

多道程序设计的思想是，把若干个作业同时放在内存中，并且都处在运行状态，但在某一给定时刻，真正在处理机上执行的却只有一道程序(当计算机只有一个处理机时)，其他程序则可能因某种原因被暂时挂起，也可能是在进行I/O操作。例如，当处理机对第一道程序进行处理后，需要输出时，CPU在处理完它的I/O请求后就转去执行第二道程序的处理工作，这就使第一道程序的I/O操作和第二道程序的处理工作并行，当第二道程序需要输出

操作时，CPU在处理完它的I/O请求后又转向第三道程序，使第三道程序的处理工作和第一、二道程序的I/O操作并行……，可见多道程序设计的特征是"多道程序并行"。从宏观上看，多道程序都处在运行状态，它们之间是并行的；但从微观角度看，每一道程序又是交替地在处理机上运行，它们分时地占有处理机。所谓多道程序系统就是能够控制多道程序并行的程序系统。

§3.2 多道程序系统的功能

多道程序设计巧妙地利用CPU的等待时间来处理其他程序，为提高计算机资源的利用率提供了极好的途径，但由于有多道作业同时运行，增加了系统的复杂程度。例如，多道程序同时在内存，就有如何分配和管理内存的问题；多道程序都要投入运行，系统应如何调配处理机和I/O设备呢？运行中各道程序之间又怎样保证它们不互相干扰，尤其是由于多道程序同时运行，它们所需求的资源总和往往要大于系统中实有资源，操作系统又如何解决资源不足的矛盾呢？这些都是多道程序系统应当解决的新课题。

为使多道程序能有条不紊地运行，系统中必须增加各种管理程序，以便各负其责把各种资源科学地管理起来。按照系统中资源的类型可以把管理程序的功能分为下列四类：

1. 存贮管理的功能。由于有多道程序同时在内存，存贮管理的任务首先是按一定的原则分配内存，使各道程序不仅能得到足够的存贮空间，且不会产生重叠和丢失。为保证各用户程序之间不互相破坏，当然也不能破坏管理程序，要求存贮管理程序具有保护功能。当多道程序要求的存贮容量大于实际内存容量时，存贮管理还应具有内存扩充的能力。

2. 处理机管理的功能。在单道程序系统中，处理机为一道程序独占，而在多道程序运行时则不同，处理机为几道程序所共享。因此处理机管理的首要任务是如何科学地分配处理机，使各道程序都有机会得到处理机，并保证它们有条不紊地运行。

3. 文件管理的功能。为使各道程序能顺利运行，常常需要调用各种子程序和数据等，怎样确保这些信息存取无误呢？一种行之有效的方法是把各种信息以文件的形式存放在存贮器中，文件管理的功能就是使这些信息不仅存取方便，而且安全可靠，便于共享。

4. I/O设备管理的功能。在多道程序运行中，常常会发生几道程序争夺I/O设备的情况，因此I/O设备管理的功能是如何按一定的原则，把I/O设备和有关通道等分给申请的作业，并启动指定的设备进行操作。当该作业用完时，能及时收回有关设备。

多道程序系统显著地提高了资源的利用率，增加了系统对作业的吞吐能力，因此被现代计算机系统广泛地采用。但多道程序系统在运行过程中是不允许用户和机器进行交互式对话的，用户一旦把作业交给系统，就离开他的作业，直到该作业运行完毕。这使某些用户（例如调试程序时）感到很不方便。他们希望在程序运行过程中能随时进行干预，以便更快地完成调试任务。因而，促使分时系统的产生。

§4 分时系统（Time Sharing System）

最初的手工操作是一种联机操作方式，作业运行的每一步都依赖人工控制，批量处理系统否定并代替了手工操作，使一批作业之间的转换由系统自动完成，因而提高了效率，但处理机的利用率仍不理想。多道程序系统发展了批量处理和执行系统的优点，并使多道作业并行运行、作业运行过程更加自动化、各类资源的利用率也更高了，因而进入了更高级的脱机

操作阶段。但后来发现，在不少情况下仍有联机的必要。例如几个系统程序员同时调试某个新的系统程序，在调试过程中常常需要相互配合和人工干预，因为有些情况是无法预测的，需要根据程序运行情况随时进行修改或控制。用户在调试某些新程序时也有类似情况，也就是说，在用户与系统，用户和自己作业之间应具有交互对话的能力。另外，在批处理情况下，用户必须把作业送到计算机房(或计算中心)，由那里的操作员去完成，用户则可能要等待几小时，甚至几天才能得到所需要的结果。如果在本单位(甚至本地)没有计算机的情况下，就更不方便。倘若用户能通过自己的终端设备，直接将作业发至计算中心，经计算后又直接将结果发回，那将是很理想的。这种远程信息进入是批处理系统向分时系统发展的第一步，它是通过通讯方式向计算中心发送作业(或接收结果)的批量型终端。分时系统是一种联机操作方式，不过它是一种比早期手工操作更高级的联机操作方式。

§4.1 分时概念和分时系统

在日常生活中，常常会碰到一个设备有好几个人都要用它，那么最简单的处理办法就是让他们错开时间使用；也可以有另一种情况，一个人要同时完成好几件工作，他可以先干某件工作，在他干的过程中抽空去干第二件、第三件工作……，如此并行地干几件工作，有可能在较短的时间里完成全部工作。前者可以说是把一个设备的使用时间分配给几个人共享；后者则是分配一个人的工作时间去干几件事情。这两种情况都包含通常所说的分时概念。这如同纺织女工同时看管多台织布机，她分时处理各台织布机工作中出现的问题，使各台织布机能顺利地并行工作，大大地提高了生产效率。在计算机领域中为了提高资源利用率，不少地方都引用了分时概念。例如，CPU 与通道分时使用内存、只读存贮器；通道与通道分时使用 CPU、内存、通道公用控制器等；同一通道中的 I/O 设备又分时使用内存、通道等。在分时系统中，分时概念又有了进一步发展，例如，CPU 在运行中是按一定的时间片(例如 50ms)轮流为各道程序服务的。如果在计算机系统中每道程序都联系着一个用户终端，那么每道程序每次运行一个时间片，就相当于 CPU 每隔一定时间为用户终端服务一次，使每个用户都可以在自己的终端上共享计算机系统的软、硬资源。只要时间片选择的恰当，各用户之间并不感到有别的用户存在，好象整个系统都为他所独占一样，具有这种特点的计算机系统称为分时系统。

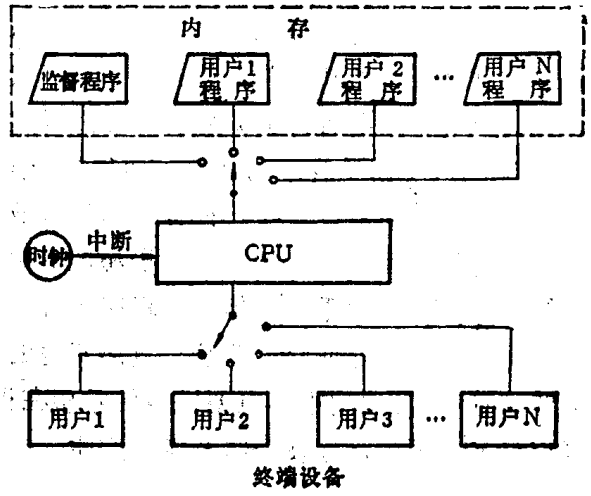


图 1-3 分时系统示意图

一个分时系统可以带几十个甚至上百个终端，每个用户都可以在自己的终端上操作或控制他的作业运行。而分时系统则按一定的时间片为他们服务，所以分时系统也被称为多终端系统或多路系统。如图 1-3 所示。

§4.2 分时系统的特征

分时系统的工作流程一般可简述如下：每个终端上的用户都可以随时通过终端向系统提

出命令，请求完成某项工作；系统接到命令后就检查分析发来的命令是否正确；若无误，则应为用户做好准备工作并回答(例如在字符终端上显示“ENTER IS READY”)，即告知终端已激活，允许用户打入程序；如果打入程序有错误，系统马上指出错误，并允许当场修改；当用户打入“程序输入完毕”，要求运行时，可打入运行命令，系统接到命令后便立即投入运行，并根据用户要求将运行结果输出；用户还可根据运行结果再向系统提出要求，直到完成预定工作为止。

从以上工作流程可见，分时系统显著的特点是多路调制性和交互性，即一个分时系统可以有若干用户多路使用，而不会被混淆和破坏，每个用户还可以和系统交互对话(即通常说的人-机对话)。分时系统的这一特点，给使用者带来很大方便。

由于终端上用户的工作速度，相对于计算机的运算速度是很慢的，因此，系统能够在很短时间内(例如几秒钟)同时响应所有用户的要求。例如一台有二十个终端的分时系统，只要把时间片划分为 50 ms，系统就可以每隔一秒钟，轮流为每个终端服务一次。如某作业在一个时间片结束前还未运算完毕，那么，该作业就被暂时中断运行，等待下一轮时间片到来，直到该作业运算完成。这样完成一个作业的运算可能要若干时间片，但每个用户的每次要求都能得到及时响应，因而每个用户都好象他自己独占计算机一样方便如意。所以，分时系统的另一特点是独占性。总之，分时系统为用户提供的工作环境是一种随时可供多个用户使用，通用性很强的计算机系统。它具有如下优点：

第一，分时系统可以节省人力、物力开支，有利于推广计算机的应用。在分时系统支持下，各用户可以不配置计算机，只要有一台价格便宜的终端设备，就可以方便地在自己的终端上使用该计算机系统。例如一个大型企业只要有一个分时系统就可以让本单位各车间、科室通过自己的终端来使用计算机。这对学校的教学工作更为有利，几十名学生在调试程序时，可以同时上机操作，这为提高学生上机解题能力，为普及推广使用计算机创造了有利条件。

第二，加快检查调试程序的速度。由于用户能在自己的终端上，在较短时间内采用交互会话方式及时地编写、调试、修改、运行自己的程序，从而缩短了调试程序的周期，加快了解题的速度。

第三，进一步提高系统资源的利用率。一个大型的分时系统，尤其是巨型机，由于其功能强大、价格昂贵，在一个单位，甚至在一个地区都难以充分发挥其效能，只有通过大量的远地用户分时使用，才能更好地提高系统利用率。在分时用户之间亦可以通过计算机的文件系统彼此交流信息和计算结果，促使更快地完成共同关心的任务。

第四，它是加快和提高科研、工程设计质量的有力工具。在科研和工程设计中，常常借助一些数学模拟方法来寻找或论证最佳方案，设计者就可通过分时系统的交互作用，在自己的终端上及时发现问题，不断修改，以求得到最佳方案。这就缩短了程序的调试周期，加快了科研和工程设计的速度。

§4.3 影响响应时间的因素

在分时系统中，当用户向系统提出请求后，系统能否及时响应是分时系统很重要的性能。如果要用户等很久才响应，就会使用户不满。通常，把用户在终端上提出请求之后到系统给

出响应之间的时间间隔称为响应时间*。从用户观点看，当然希望系统能尽快响应，但从系统角度说，则要统一考虑多方面因素。那么响应时间的大小究竟和哪些因素有关呢？

1．用户终端数目的多少直接影响响应时间。如果系统中有 n 个终端，其时间片为 q s，则所有终端轮转一次就要 nq s。而每个用户在 nq s 时间内，就能获得 q s的处理机工作时间。可以把 nq 近似看成响应时间，那么当 q 一定时，n 的大小就直接影响响应时间，即终端数目增多时，响应时间就增大，如果响应时间太长，用户就不满意了。因此，终端数应控制在适当范围内。顺便指出，在上述情况中，对某一用户 CPU 处理作业的速度只有实际机器的 $1/n$ 了。可见，终端数越多，对用户作业的处理速度就愈低。

2．时间片的大小直接影响响应时间。例如一台有二十个终端用户的系统，当其时间片为 1 s时，则响应时间可达 20 s 之久，这样长的等待时间，是用户无法容忍的。如果把时间片减少为 0.01s，响应时间就相应地减到 0.2s，这就缩短了用户等待时间，但是由于时间片的减小，就会使系统消耗在信息交换的时间所占比例增大，即系统为之付出的代价亦增大了。

3．信息交换速度愈高，交换的信息量愈少，可以减少响应时间。信息交换所需时间主要是对外存贮器的访问时间和信息传递时间两部分。如果内存和外存之间信息交换速度愈快，则处理机就有更多时间来处理各终端作业，响应时间自然就可以减小。当然，在信息交换速度一定时，如果交换的信息量愈少，交换所需时间就愈短；如果时间片也一定时，要完成一个作业需要交换信息的次数就愈少，这就缩短了作业在系统中的停留时间，减少了系统中作业的数目，从而改善了系统的响应时间。

由以上讨论可知，系统的响应时间是和用户终端数、时间片大小、信息交换速度以及交换信息量大小有关。在确定响应时间时，必须根据系统设计目标，权衡诸因素而定。

多道程序系统和分时系统的出现，标志着操作系统的正式形成，它在提高机器利用率、方便用户等方面都是令人满意的。但由于计算机的应用日益扩大，上述系统的功能在某些领域，如实时控制、实时采样等方面已不能满足要求，因此，又出现了实时系统。

§5 实时系统（Real-Time System）

六十年代中期，计算机进入第三代，固体组件代替了分立元件，计算机的性能和可靠性都有很大的提高，价格也大幅度下降。因此，计算机的应用范围迅速扩大，不仅在科学技术研究领域离不开它，商业上的数据处理，工业生产上的自动控制、仓库管理，甚至医疗诊断、飞机订票等方面都广泛应用它。由于各行业情况复杂，对计算机的要求也各不相同，上述各种计算机系统已不能满足实际需要，因此又出现了实时操作系统。

所谓实时，就是及时，不失时宜。实时系统是指能及时响应外部事件的请求，在规定时间内迅速作出处理，在时间上能及时控制有关实时设备和实时任务协调一致工作的计算机系统。如果计算机是用来控制飞机飞行、导弹发射，以及控制冶炼、石油化工等生产过程时，都要求计算机能把测量系统测得的数据及时加工处理，以便不误时机地实现控制目的，这种系统称为实时控制系统。而用于企业管理和查询工作的计算机系统，例如航空公司的查询系

*在分时系统中，响应时间亦可指从用户发出数据块起至显示结果时所经历的全部时间；在批处理系统中，响应时间是指用户作业进入计算机系统到输出结果之间的时间间隔。

统，要求能及时回答顾客提出的有关航班、航线、票价、订座等问题。大银行的管理业务是一个复杂而细致的工作，要求计算机系统能及时分析处理大量数据，它是具有实时处理特性的计算机系统。通常把实时控制系统和实时处理系统统称为实时系统。所以实时操作系统是一种响应时间很快、可靠性很高的一种操作系统。

由于实时系统大部分都是专用性的，因此其规模大小相差甚大，各种实时系统的特点和功能亦各不相同。一般较大的实时系统，除了具有一般操作系统的基本功能外，还有如下特点和功能：

1. 实时系统对响应时间的要求比分时系统更高，一旦外部信号向系统提出要求，系统应能立即响应处理。根据不同控制对象要求，通常是在秒的数量级，但在某些实时系统中响应时间可达毫秒、甚至微秒数量级。由于实时系统要在很短时间内对外来信号做出判断和处理，因此系统应具有对实时任务进行处理和控制的功能。例如要求系统根据用户规定时间启动执行某任务，并按照一定的时间不断重复执行该任务。这种严格的时间管理，通常是由实时时钟所产生的脉冲数来计量的，因此必须设置一个实时时钟和相应的时钟管理程序，为系统的计时，或定时任务(延时任务)提供依据，并控制其按一定的时间关系和逻辑关系协调工作。

2. 对系统的可靠性要求很高。实时系统常常用在实时控制方面，在设计实时系统时最重要的是要保证高可靠性。例如在军事上用来控制导弹发射、核爆炸、导航等，如果产生错误和丢失信息就会造成严重的后果。所以在不少的实时系统中，都采用双工体制。在严格的双工系统中，有两台完全相同的计算机，其中一台为主机用于实时控制或实时信息处理，另一台做为后备，但必须和主机同时并行。要求两台机器的运行状态在任何瞬间都完全相同。如果主机发生故障，后备机器便可立即代替主机继续工作，从而保证系统的高可靠性。

3. 具有连续的人机对话能力。实时系统没有分时系统那样强的交互会话能力，它仅允许终端访问有限数量的专用程序，不能书写程序或修改一组已存程序，但它必须具有连续的人机对话能力。例如，在较复杂的人机对话中，当终端发送一消息并收到计算机系统回答后，终端又发送该问题的补充信息，这就要求计算机能记住上次从终端发来的消息，并根据补充信息形成第二次回答的能力。类似地，用户可以发来第三次、第四次……有关该问题的信息，计算机都能做出相应的回答。

4. 系统的整体性很强。实时系统所管理的联机设备和资源，必须按一定的时间关系和逻辑关系，保证协调一致工作，不然就会造成不堪设想的后果。

5. 具有过载防护能力。在实时系统中，任务进入系统往往具有很大的随机性，有时就会超过系统的处理能力，因而产生过载。所以，必须为系统设计某种防护机构来保证一旦发生过载，系统仍能正常运行。

以上讨论的批量处理系统、分时系统和实时系统是目前操作系统的三种基本类型。但对某一个具体的操作系统而言，为了使系统应用范围广，处理能力强些，又可能不属于其中某一类型，而是同时兼有这三类或其中某两类的特点，也可能以某一类为主兼有其他类型的特点，我们称这样的操作系统为通用操作系统。不同类型操作系统之间的这种差异，主要是由于各个系统所侧重的目标不同，或者说赋予系统的主要任务不同而造成。例如重点要求系统对作业吞吐量大，机器利用率高，则可以由批处理系统来完成；如果还希望多终端又有较强的交互能力，就应把批处理系统和分时系统的特点都兼容起来。由于实际上对操作系统的要求是很复杂的，所以操作系统的种类也很多。例如DJS-1000系列机的操作系统就有：SOS

独立操作系统；RTOS 实时操作系统；DOS 磁盘操作系统；RDOS 实时磁盘操作系统；MRDOS 带内存管理和保护的实时磁盘操作系统等。PDP-11是一个多挡广泛兼容的小型机系列，其操作系统则更多，如RT-11是小型单用户前/后台操作系统。MUMPS-11为小型到大型的分时系统；RSX-11D是一个大型的实时多道程序的操作系统；RSX-11M是从小型到中型的实时多道操作系统；IAS是一个大型多用户分时系统等。在微型机中则更为繁多，其中比较普遍的操作系统有CP/M单用户操作系统，MP/M多用户操作系统，MS-DOS，PC DOS V2.0，UNIX及其变种ONIX等。

§6 本章小结

本章首先讨论了操作系统的概念及其功能，进而揭示操作系统的发展是随着生产、科学技术发展，特别是有关学科的发展而发展起来的。在操作系统的发展过程中，我们又重点介绍各发展阶段具有代表性的操作系统类型、特点和功能，单道批处理系统是在解决人机矛盾和CPU与I/O速度矛盾的过程中，在手工操作基础上发展起来的。由于脱机批处理否定了最初的手工联机操作，因此计算机的利用率得到了提高。多道程序的引入是为了进一步提高CPU、内存和I/O设备的利用率。分时系统使用户能和计算机发生交互作用，每个用户都可以在自己的终端上方便地使用计算机，它不仅使系统的利用率更高，而且功能也更趋完善。多道程序和分时系统的出现标志着操作系统的真正形成。实时系统大多数是为特定的实时任务而设计的，所以特别要求系统能很快响应并不误时机地进行控制，或对输入信息能及时做出处理，而且要求安全可靠，整体性强，能保证各部分协调一致地工作。通常把批处理系统、分时系统、实时系统称为操作系统的三大基本类型。

多道程序设计和程序交换是实现各种类型操作系统的方法。单道指的是一次执行一个作业，即不论作业大小，整个系统一次只执行一个作业，多道程序系统克服了这一缺点。多道程序设计由于内存中同时有多道作业程序，因此当一道作业因等待某事件而不能运行时，处理机就可以执行另一道作业程序，从而大大提高了处理机的利用率。所谓程序交换法，即在内存中只放一道可运行的作业程序，其他作业程序则存放在后备存贮器中，作为后备作业，当内存中程序因某种原因运行不下去时，系统就把它交换到后备存贮器中，并从后备作业中选出一个可运行的作业交换到内存中来运行。在具体实现一个操作系统时，设计者可根据设计目标选用两种方法之一，或把这两种方法结合起来，以使系统的性能更好，效率更高。

复习思考题

1. 何谓操作系统？它具有哪些基本功能？
2. 试比较多道程序系统、分时系统、实时系统各有何特点？
3. 为什么说多道程序系统和分时系统的出现，标志着操作系统的形成？
4. 何谓响应时间？影响响应时间的因素有哪些？应怎样确定一个系统的响应时间？
5. 随着计算机科学技术的发展及其推广应用，操作系统经历了哪些发展阶段？