计.算.机.系.列.教.材

COMPUTER

操作系统概论

许曰滨 孙英华 韩新霞 编著

6-43

人民邮电出版社

计.算.机.系.列.教.材

COMPUTER

封面设计／ 董志桢

ISBN 7-115-07950-1

9 787115 079503 >

ISBN7-115-07950-1/TP·1202

定价：19.80 元

人民邮电出版社

操作系统概论

许曰滨 孙英华 韩新霞 编著

人民邮电出版社

内容提要

本书作为计算机系统软件方面的教科书，主要介绍计算机操作系统的基本原理与结构。为了配合上机实验，书中对目前较为流行的操作系统 DOS、Windows 98、Windows NT 和 UNIX 也作了简要介绍。

本书可作为大、中专科学生的教材，也可作为计算机爱好者的自学参考书。

计算机系列教材

操作系统概论

◆ 编　　著　许曰滨　孙英华　韩新霞

◆ 人民邮电出版社出版发行　北京市崇文区夕照寺街 14 号

邮编　100061　　电子函件　315@ pptph.com.cn

网址　http://www.pptph.com.cn

北京朝阳展望印刷厂印刷

新华书店总店北京发行所经销

◆ 开本：787 × 1092　1/16

印张：15.5

字数：376 千字　　1999 年 7 月第 1 版

印数：10 101 – 14 100 册　　2000 年 7 月北京第 2 次印刷

ISBN 7-115-07950-1/TP·1202

定价：19.80 元

前　言

在计算机系统的不断发展中，操作系统一直是人们研究的中心课题。尤其是在计算机软硬件功能日益强大的今天，操作系统的位置更显得无比重要。可以说，计算机上如果没有操作系统，简直就是一堆废铁。

操作系统的根本任务是管理计算机的软硬件资源，为用户提供安全可靠、方便有效的运行环境。因此，它既是计算机与用户联系的桥梁，又是软件和硬件连接的纽带。鉴于它在计算机系统中的无可替代的作用，操作系统课程也被列为计算机专业的主要课程。

我们在参考了国内外有关技术资料和论著的基础上编写了这本教材，主要向读者介绍操作系统的基本知识，为进一步学习计算机的其他软硬件课程打下基础。为了增加读者的感性认识，本书介绍了目前常见的几种操作系统并编写了部分上机实验题。

本书共分 10 章。

前 6 章主要介绍操作系统原理。其中，第 1 章概要地介绍了操作系统的发展过程及有关的一些基本概念；第 2 章介绍作业管理的主要功能与实现方法；第 3 章首先介绍进程的基本概念，然后介绍进程调度和进程控制算法，并兼顾了进程互斥与同步的实现方法；第 4 章介绍内存储器的结构和管理技术，主要包括存储空间的分配与回收，存储保护和保密等；第 5 章介绍设备、通道及中断概念，并介绍设备分配、启动及缓冲区的管理过程；第 6 章介绍外存空间的管理，并介绍文件、目录的组织与管理方法。在前 6 章中，每章都附有一定数量的习题供读者课外练习用。

后 4 章介绍目前较为流行的几种操作系统，旨在配合读者上机实验。其中，第 7 章介绍磁盘操作系统 DOS；第 8 章介绍视窗操作系统 Windows 98；第 9 章介绍多任务操作系统 UNIX；第 10 章介绍网络操作系统 Windows NT。每章后面都有 5 至 6 个实验题目，可供 8 至 10 个学时上机实验用。对于不同专业和不同实验环境的读者，可以有选择地使用其中的部分内容。

由于作者水平有限，书中肯定存在不少缺点和错误，敬请读者批评指正。

作　者

1999 年 5 月

第1章　绪论

电子计算机系统分为硬件和软件两部分，其中，硬件部分又称“裸机”，它是由若干物理设备连接而成的。现代计算机的硬件功能很强，用途很多。然而，硬件本身提供给外界的界面却是十分粗糙的，许多信息都令人费解、难以使用。为此，人们研制了一种能够管理和控制这种裸机的软件，这就是“操作系统”（Operating System）。

1.1 操作系统的形成与发展

计算机从 1946 年问世至今，已有半个多世纪的发展时期。它的应用从单纯的数值运算，发展到过程控制、数据处理和网络通讯等各个方面。在工农业生产、国防建设、文化教育及社会生活的许多领域，发挥着重要的作用。

计算机的发展大体经历 4 个发展阶段。从 1946 年至 50 年代末，计算机处于以电子管为特征的“第 1 代”发展时期。在这一时期，计算机的运算速度慢，存储容量小，软件功能低。计算机的应用主要集中在数值计算方面。当时的操作基本上采用手工方式。进入 60 年代，计算机发展到以晶体管为特征的“第 2 代”发展时期。在这一时期内，除了硬件上的功能得到扩充外，软件方面出现了高级程序设计语言和用于管理计算机的软件——“管理程序”，比如，批处理系统、分时系统等。这一时期，操作系统已基本形成。从 60 年代中期开始，计算机又进入到以集成电路为特征的“第 3 代”发展时期。计算机开始向许多应用领域渗透，因而出现了大量适用于某些专业的高级语言。一些用于过程控制的实时系统和通用的操作系统开始在计算机上运行。70 年代以后，计算机进入“第 4 代”发展时期。超大规模集成电路的出现使得计算机的体积越来越少，功能越来越强，价格越来越低，极大地促进了计算机应用领域的扩展。特别是通讯技术的发展以及大型数据库管理系统、远程处理系统和计算机网络的出现，操作系统的功能也越来越强大。

下面我们对操作系统的产生与发展，分阶段作以介绍。

1.1.1 手工操作阶段

早期的计算机由于运算速度低、外部设备少，加之没有配置操作系统，用户使用计算机是相当困难的。用户需要用机器语言，也就是由二进制数形式构成的指令系统来设计程序，编制好的程序被输入到机内后，直接通过控制台上的开关和指示灯来监督和控制程序的运行，其繁琐程度可想而知。

在这一阶段，计算机的硬件刚刚发展为第 1 代。当时，一个程序的运行，一般需要以下

3 个步骤：

1. 通过控制台开关设定计算机为初始状态，将存有程序和数据的卡片或者其他介质装入输入设备上。设定程序的存放位置后，按下“作业装入”开关，使程序和数据由输入机上输入。
2. 再用控制台开关设定程序的开工地址，按下“作业运行”开关，开始运行作业，根据控制台上的状态显示灯，确定运行是否正常，若出现错误则停机处理。
3. 运行结束后，将结果存于某一存储地址上。用开关设定结果的存储地址，按下“结果输出”开关，将该地址的数据从输出机（比如，卡片机）输出。用户卸下卡片，取走结果。

这种操作方式的缺点是：

- 操作麻烦：操作员要了解计算机各部分的技术细节才能准确地控制作业运行，而这样做极容易出错，维护工作非常复杂。
- 资源浪费：计算机的所有软硬件资源全部被一个作业独占，不用的资源也不能给其他作业使用。特别是，当一个作业完成，另一个作业开始，作业切换过程中的上机、下机将浪费大量时间。

1.1.2 管理程序的发展阶段

到 50 年代末期，计算机进入了第 2 代发展时期，计算机的处理能力有了显著的提高，主要表现在：CPU 的运行速度明显的加快；内存的容量有了很大增长；外部设备数量增多。计算机硬件的提高为软件的发展奠定了基础。在程序设计方面，人们研制出了一些高级程序设计语言及编译软件，如著名的算法语言“Fortran”、“Algol”等。在这一时期，同时出现了一种监督和控制计算机运行的软件，即“管理程序”。

管理程序，可以看作是操作系统的前身，它的研制目的主要是为了向用户提供方便的操作接口，尽可能高效率地利用计算机。

在这一阶段中，用户的应用程序可以使用高级语言进行设计，然后由管理程序负责调用相关的编译程序将其编译为机器语言程序再投入运行。程序设计人员不必再使用繁琐的机器语言来编写程序。这意味着程序设计人员可以摆脱计算机上的操作，将自己程序的运行交给计算站的操作员来完成。

由于不同的管理程序依赖不同的计算机硬件，特别是各自所追求的使用方式有所不同，这一时期的管理程序，在处理方式上存在着明显的差异，大体可归类为：批处理（batch processing）方式、分时处理（time sharing）方式和实时处理（real time processing）方式等。具有代表性的系统有：50 年代中、后期投入运行的 IBM 704 FORTRAN 监控程序（批处理）系统、美国军用地面防空实时控制系统 SAGE 和 60 年代初期运行的美国 MIT 计划项目 MULTICS 分时系统等。

1.1.3 操作系统的发展阶段

从 60 年代中期开始，计算机系统进入了第 3 代发展时期。以集成电路为中心的硬件设计与制造工艺不断提高，一大批功能完善、集成度高的微处理器涌入市场。由于其价格便宜，因而被广泛引入到输入输出接口设备、终端及外部设备的设计中。这一时期的计算机除了 CPU 的运行速度加快，内存容量大大增加之外，还出现了输入输出通道、中断装置、大容量的外存储器——磁盘等。操作系统也开始由早期的管理程序发展成为包括批处理、联机、分时或实时等功能的通用系统。

操作系统，是管理计算机系统内全部资源的程序模块和数据集合的总称。它作为用户和计算机硬件之间的连接软件，主要是响应应用程序的各种要求，解决各种要求之间的矛盾，使系统的利用率提高，以便用户能够简便而高效地使用计算机。

从采用新技术方面上看，操作系统对计算机的管理水平比前期的管理程序有了进一步提高。它开始使用“虚拟化”技术，以提高系统运行大程序的能力，比如，虚拟机（Virtual Machine）、虚拟存储器（Virtual Storage）和虚拟设备（Virtual Units）等。此外，在输入输出方面，操作系统提供了“Spooling”（又称“假脱机输入输出”）技术，使系统的输入输出得到改善。这一时期具有代表性的操作系统有：IBM 的 OS 360 系统、UNIVAC 的 EXEC 系统、FACOM 的 MONITOR V 系统，以及 Bell 实验室于 70 年代初推出的 UNIX 系统等。这些系统在操作系统的理论研究方面都起了十分重要的作用。

进入 70 年代中期以来，随着通讯技术的进步，加上通讯软件的支持，将多个独立的计算机系统连接成计算机网络成为可能。操作系统随之得到进一步发展，于是出现了以信息通讯为基础，将作业分布在不同地域上进行分散处理的“网络操作系统”。

网络操作系统，作为为计算机网络研制的一种操作系统，把网络中的各台计算机有机地联合起来，达到网络资源的管理与共享。由于网络中的各台计算机上可能运行着不同的单机操作系统，因此网络操作系统在实现资源共享的信息通信中，需要对不同系统的信息表示进行转换，并将传输信息作一致化处理，比如打包和拆包，实施必要的信息校验和差错处理等。

随着超大规模集成电路制造业的飞速发展，一种以内存储器为中心，用紧密耦合方式设计的多机系统（Multiple-Processor System）显示出非凡的处理能力。它的机内管理者就是所谓的“分布式操作系统”（Distributed Operating System）。

分布式操作系统是一种适合于多机系统管理的操作系统。它让系统中的多台计算机相互协作，共同完成一个大任务。它的主要职责是，将一个计算任务分解为若干子任务，然后分布到多台计算机上进行运算，最后把各子任务的解合成起来，形成一个最终解。分布式操作系统的主要目的是通过多台计算机的并行运算提高系统处理大任务的能力。

从上面的介绍可以看出，操作系统发展的每一步都离不开计算机硬件技术的发展。目前，随着计算机通讯网络和多媒体技术的研究，以及用户对计算机可靠性的要求不断提高，人们开始注重以下几方面的研究：

1. 系统容错功能

当硬件的某些部分发生故障后，计算机仍能不同程度地完成一些例行的运算，这主要依赖于故障诊断、纠错及双工操作技术来实现。

2. 核心软件固化

随着硬件价格的下降和固化技术的提高，由硬件直接实现对操作系统的某些核心软件进行固化，可以提高系统的性能和抗干扰能力。

3. 分布式多媒体

通过通讯手段将分布于网络中的多媒体设备资源管理起来，实现大范围的多媒体共享。

1.2 系统管理方式

上一节已经论及了操作系统中采用的几种主要管理方式，即批处理、分时处理和实时处理等，这是计算机应用日趋多样化的结果。其中，批处理又分为单道批处理和多道批处理两种。我们将各种处理方式实现的系统分别称作单道批处理系统、多道批处理系统、分时系统和实时系统。

用户的一个计算任务设计好后，提交给计算机，计算机便按照系统要求的某种处理方式进行运算，最后把计算结果返给用户。通常，我们将一个计算任务称作一道“作业”，按照系统同时能够处理的道数划分，操作系统又可分为单道系统和多道系统。

其中，多道系统的程序设计，又称“多道程序设计”，是一项比较复杂的技术，需要解决多道作业的资源共享问题、CPU 分配问题、因资源竞争冲突造成的死锁问题以及有关的互斥与同步问题等。

下面我们对这些处理方式作进一步的介绍，以求读者对不同处理方式有一个概要地了解。

1.2.1 单道批处理方式

单道批处理是早期的一种计算机管理方式。它的管理可以简单地描述为：一次运行一个作业、不间断地运行，直到一批作业全部运行结束为止。

图 1-1 是典型的单道批处理系统硬件配置图。

用户将应用程序和初始数据制作成一叠卡片，交给计算站操作人员在计算机上进行运算。

操作员将若干个作业收集在一起，装入卡片读入机，读入用户的源程序和初始数据，然后，编译成可以运行的目标程序存储在目标磁带上。接下来，操作员在控制台上按下“开始”开关，计算机从磁带上读出第 1 道作业存入内存，开始运行。当运行完成以后，接着运行第 2 道作业，第 3 道作业……。

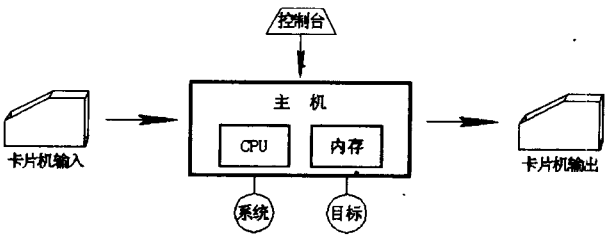

图 1-1 单道批处理系统的硬件配置

这是一种封闭式的管理方式。用户和操作员的职责分离，避免了因用户在计算机前的思索造成的 CPU 空闲时间和低效率的人工操作时间，与用户直接操作裸机相比有了很大的进步。

单道批处理系统的引入，解决了作业之间的自动切换问题，提高了资源的利用率。同时，由于这种系统需要有“标准输入输出程序”、“汇编程序”、“编译程序”、“连接程序”等支持，这种要求在一定程度上推动了软件技术的发展。

批处理的重点在于“批”，由于这种系统可以将一批作业一次性地装入系统，因此用户的响应时间比较长，一般要等作业运行结束后才通知用户。这就要求用户在提交作业之前，针对作业运行中可能出现的各种情况，预先规定好处理的措施，并写入操作说明书中一并提交给系统。

从系统效率上看，单道批处理系统比人工操作提高了很多。但是，由于 CPU 的运算速度与外部设备的输入输出速度差距很大，CPU 仍需要花费大量时间等待低速外部设备的输入输出，浪费仍然是很大的。这是单道批处理方式的不足之处。

1.2.2 多道批处理方式

多道批处理是为了提高 CPU 的利用率而设计的一种管理方式，它是单道批处理方式的一种改进形式。它允许多道作业同时进入内存，共同使用 CPU 进行运算。这里突出了一种全新的设计思想——多道程序设计的思想，即允许多个作业按交替方式或者并行方式运行。通常，将采用这种方式运行的程序称作“并发”程序。

并发（Concurrent）程序，是既可以并行运行，也可以交替运行的程序。在单处理机的系统中，它们的运行只能是交替地进行，但是从宏观上看，它们的运行是并行的，只有在多处理机系统中，这些并发程序才能够真正地并行运行。

在多道程序设计的系统中，当一个运行的作业推进到某个位置，因需要等待某种事件（比如输入输出）运行不下去时，就将处理机切换给其他作业运行，以避免处理机的闲置。当它的输入输出完成后，又可以将处理机转给它，使它恢复运行。这种做法使处理机和系统的其他资源都能得到比较充分的利用。

这样一来，CPU 必然扮演一个十分活跃的角色。它一会儿运行管理程序，一会儿运行某道应用程序，过一会儿又运行另一道应用程序，十分忙碌。只有当所有应用程序都等待某种事件出现（比如，输入输出操作）时，CPU 才会空闲。当然，我们并不希望这种情况出

现。

多道批处理方式区别于单道批处理方式的一个重要方面是，CPU 运行程序与输入输出操作实现了并行化，提高了资源的利用率。

多道系统中的一个很突出的问题是资源竞争。比如，多道程序并发运行，每个程序都要使用一些系统资源，由于系统的资源总是有限的，因而，这些程序必然产生资源竞争。操作系统应当制定一种方案将有限的资源提供给它们共享使用，避免产生竞争与冲突。另外，系统内的多个进程有时需要相互配合、协调工作，因此操作系统也应当提供一种机制，使它们达到运行的同步。

在操作系统的设计中引入多道程序设计思想，标志着操作系统的发展趋于成熟，同时，也引发了软件理论的研究。在此基础上，逐步形成了操作系统的 5 大处理功能——作业管理、处理机管理、存储管理、设备管理和文件管理等。

1.2.3 分时处理

分时处理，又称会话型处理，是在多道程序设计基础上发展起来的一种处理方式。它把时间分隔技术应用到 CPU 的调度上，形成了一种新的操作系统。

分时，指的是若干并发程序对 CPU 分时使用。即将多道用户程序装入内存，让它们轮流在 CPU 上运行，每一道程序使用 CPU 的时间长度都有限制，不能超过一个固定的时间片。任何程序如果在分给的时间片内未能处理完时，应当放弃 CPU，等到下一次分得 CPU 时再继续处理。

时间片（time slicing），是多个并发程序分享 CPU 的时间单位，通常为几十毫秒。

比如，系统中设定的时间片是 50 毫秒，假定并发的程序有 40 个，那么，操作系统对每个程序的平均响应时间为 50×40=2（秒）。在一般的分时系统中，用户的响应时间通常控制在 3 秒以内。

这种处理方式需要一个关键设备——“时钟”。时钟在其漫长的计时过程中，每隔一个时间片就产生一次中断，CPU 接到中断信号后，立即转入管理程序，经过一番处理后将 CPU 切换给下一个程序，使之投入运行。

分时系统与多道批处理方式的一个重要区别在于，用户可以在机房里直接操作计算机，使管理方式由封闭型变为开放型。一般来说，一台计算机可以挂上几十台甚至上百台终端机，每个用户可使用一台。他们各自通过自己的终端机向系统输入命令，操作系统则根据用户的要求进行运算。用户运行自己的程序包括 3 部分内容：输入数据部分、运算部分和输出数据部分，该处理方式就是让用户的每项数据输入，都自己来做而不是操作员代劳，给用户极大的自由。

这样，用户就有了充分的机会与自己运行的程序进行信息交流。不过应当注意，系统中的每个用户在占用终端机期间，其大部分时间是在思考问题，系统真正响应用户的时间是比较少的。操作系统并不使 CPU 等待某个输入信息的完成就启动其他程序了。当输入操作结束后才可能再次恢复其运行。由于计算机上连接的终端很多，因而操作系统可以在用户思考的时间里以轮流方式为其他用户服务。

这样一来，计算机与外部设备的并行度可以得到极大的提高。CPU 获知输入操作结束的方法也有许多，比如：

1. 定期查询方式

每次查询，如果有输入就接收下来，如果没有就继续处理别的程序。等轮流运行程序一遍后再来查询输入过程结束与否。

2. 中断方式

CPU 可以不再过问输入输出问题，当用户的信息输入输出结束后，将中断信号通知 CPU，CPU 接到该信息后便暂停当前的工作，处理输入输出信息。有关中断的基本原理将在第 5 章中阐述。

分时处理的主要优点是，

- 当用户需要与程序进行信息交流时，CPU 不是在程序中等待，而是循环运行其他程序。
- 当多个用户同时在一台计算机上操作时，各用户好像都在独立地使用自己的计算机，用户一边与计算机通话，一边让计算机来运行自己的程序。

1.2.4 实时处理

实时处理突出了系统处理的即时性或响应性，它通常能对随机发生的事件进行随时处理，并立即送回处理结果，其响应时间绝对能够满足对象系统的业务要求。

实时处理有着非常广泛的应用领域。每一个实时系统都有固定的处理对象，根据作业的固有特征，实时系统可分为 3 种处理类型：

1. 过程控制

让计算机时刻监视受控对象的状态，实施即时控制。常见的有化学过程、生产管理、电力输送等资源分配控制。过程控制的主要特点是，所需信息通过传感器输入，系统的响应条件十分严格，整个过程无人介入。

2. 指令控制

在系统运行中可输入人的决策信息，以便随时对控制策略进行修正。常见于列车运行控制、道路交通管理、航空调度管理及军事应用等大规模控制系统。这种控制要解决的主要问题是人-机接口问题。

3. 询问响应

系统备有一个综合信息库，供多台终端上的用户进行联机信息查询。对于各台终端输入的处理要求，立即处理完并送回处理结果。常见的有飞机票或火车票订票、股票交易、情报检索、公共信息服务、银行业务、库存管理等。由于在这种处理中，大多数用户要求的响应时间都较长，因而系统的实时性比较弱，所以也简单地称之为“联机处理”。

实时系统与分时处理系统的区别是：

- 分时系统的交互性较强，而实时系统的交互性较差。

通常，分时系统是具有较强通用性的计算机系统，用户与计算机之间的会话比较频繁，因而系统提供的交互能力也比较强；而实时系统大都是具有特殊用途的专用系统，它只允许用户访问有限数量的程序，不允许用户书写程序或者修改系统内的程序。

- 分时系统对时间的响应比较弱，而实时系统的响应比较强。

分时系统的响应时间一般是以人所能接受的时间来确定的。响应时间的数量级通常为“秒”；实时系统的响应时间较为严格，它是以控制过程或者信息处理过程所能接受的延迟时间来确定的，响应时间有时可达到毫秒甚至微秒量级。

在实时系统中，任务提交给系统的时间和数量可能存在很大的随机性，因此在一个较短的时期内，有可能超出系统的处理能力，使系统出现“过载”现象。当出现过载现象时，系统要有一定的防护能力。比如，在一些过程控制系统中，当发生短期超载时，抛弃一部分不重要的任务，或者降低这些不重要任务的服务频率等，以保证某一重要任务的及时响应。

由于实时处理对响应时间的要求十分严格，因而系统的安全性成了一个重要问题。比如为了防止由于信息破坏造成系统瘫痪，需采用信息热备份，使出现故障后能及时得到切换。此外，在硬件设计中常常加入容错技术，使得一般性故障不影响系统的正常运行。

1.3 操作系统的组成

初期的计算机管理程序是由一个大程序构成的。随着计算机系统的规模变大，特别是系统中同时融入多种处理方式，仅用一个程序来实现整个系统是极不现实的。于是，按照不同的处理方式及功能将系统划分为若干个功能模块，通过相互间的调用方式将它们连接起来，从而组成了操作系统。

操作系统，作为位于计算机硬件和应用程序之间的一层软件，它是加到裸机上的一层“服饰”。有了它，计算机的硬件可以得到有效的扩充，因而呈现给用户的是一台比裸机功能强大的计算机——“虚拟机”。

虚拟机，是由计算机硬件和运行于其上的操作系统组成的计算机系统。

由于操作系统的若干程序模块之间存在着调用和被调用的关系，因而，一台虚拟机本身可以看成是经过多层次的功能扩充实现的。这种由多个模块按不同层次构造的操作系统，从设计、调试、运行到维护都十分方便，为开发大规模的软件系统开创了先河。这就是所谓的“结构程序设计”方法。

对于不同的操作系统，其组成结构和各模块功能差异很大。每个系统应当由多少模块组成，各模块的层次结构及它们之间的调用关系如何，都没有固定的模式。不过，从资源管理的角度看，这些模块的功能应归结为：进程管理、存储管理、设备管理和文件管理，各种管理自成系统。

在当今的计算机系统中，操作系统已经把硬件资源和软件资源有效地组织起来，构成了完美的整体。但是作为计算机系统的用户，应当怎样将自己的作业控制计划通知给这个系统

呢？操作系统接收了用户的要求，又怎样并将这些要求转交给相应的管理模块进行加工呢？任务加工完成后，又怎样把结果回送给用户呢？

为了解决这些问题，操作系统必须有一层接口软件，将用户与操作系统内核部分连接起来。这样，操作系统除了具有上述 4 个资源管理功能外，还应增加一个接口功能——“作业管理”功能。

（在一些教科书中把作业管理与进程管理合并在一起，称为“处理机管理”。）

作业运行之前，用户需要规划好作业的操作过程，并使用系统提供的作业控制命令写一份作业说明书，或者按作业要求输入一条操作命令，操作系统则按照用户的意图调用相应的管理模块进行处理。

图 1-2 是用户与操作系统及计算机硬件的层次关系图。

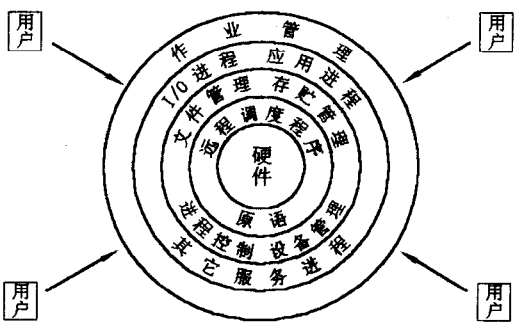

图 1-2　用户与操作系统及计算机硬件的层次关系图

下面是各个管理系统的功能：

1. 作业管理

主要任务是，进行作业执行前的各种准备和作业结束的清理工作，以及为确保作业运行，为它们申请所必要的各种资源等。

在大多数操作系统中，作业管理是作为一个过程来执行的。用户要把自己的处理工作归纳成为作业的形式才能输入系统。当然，不同的操作系统所规定的作业形式是有很大差异的。作业管理模块中有一个专门用来接收用户命令的处理程序，对用户的每一次键盘操作给予响应。此外，负责作业调度的程序将按照一定的算法从已经输入的若干个作业中，选择一个或多个作业，为它（们）申请内存和外部设备，做好在 CPU 上运行的准备。当作业运行结束时，一个进行“后处理”的程序实现撤消作业的工作，即调用“存储管理系统”和“设备管理系统”将作业占用的资源收回，将该作业从系统中删除。

2. 进程管理

在单道系统中，处理机被一个作业所独占，其分配和使用不发生资源竞争，管理比较简单。但在多道系统中，要组织多个作业同时运行，不可避免地会产生对处理机的竞争。进程管理就是要解决处理机的合理使用问题。

在多道程序设计环境中，通常，进程管理中包括一个负责进程调度的程序，该程序按照特定的策略将 CPU 分配给一个具有运行资格的程序，使它投入运行。此外，系统还应包括

进程控制程序，用于登记和管理各个进程的状态，实现进程的互斥操作和同步运行等，并协助进程调度程序来管理作业对 CPU 的竞争。

3. 存储管理

存贮器分为内存和外存。因为多道程序运行时所竞争的存储资源是内存，所以这里所说的存储管理仅对内存而言。主要包括：用某种数据结构记录内存的使用情况，按照一定的策略对主存储器进行分配、内存保护和内存扩充等。

操作系统对存储器的管理有多种方式，如分区式、分页式、分段式等。

4. 设备管理

管理各类外围设备，根据一定的分配策略，将设备分配给某一应用程序，并在适当时候将设备回收，以备分配给其他应用程序。另外，设备管理还包括输入输出程序 Spooler、设备驱动程序及中断处理程序等。

通常，因为外部设备的处理速度远远低于 CPU，所以如何避免主机时间的浪费、尽可能的发挥外设和主机的并行工作能力，是设备管理中必须考虑的问题。

5. 文件管理

文件是计算机的软件资源，程序和数据都是以文件的形式存储在存储设备上的。文件管理功能主要包括外存空间的管理与回收，对文件进行存取、检索、更新，及有效地实现文件共享。

文件管理应当是一个文件组织简单、方便用户使用、具有安全保护措施、易于用户维护的独立系统。

1.4 通用操作系统

上面讲的“多道程序设计”技术，最初是针对多道用户程序的。后来，这一技术很快被应用于操作系统本身的设计中。一个通用操作系统应当分为两部分：基本部分和常规部分。其中，基本部分是操作系统的核心，又称作“内核”；而常规部分是内核的用户。常规部分通常以“进程”的形式与应用进程并发运行。为了将它们与应用进程相区别，通常称它们为“系统进程”。

1.4.1 操作系统的基本特征

引入多道程序设计技术以后，操作系统将呈现如下基本特性：

1. 并发性（Concurrency）

主要是指多个系统进程之间、系统进程和应用进程之间可以并行执行。它们可以分别以不同的速度向前推进。比如，操作系统将管理输入输出通道的程序创建为系统进程，那么，当遇到需要输入输出操作时，就激活该进程，启动通道实现数据传送。必然地，在它运行期

间，还有其他进程也在并发运行。

这一特性主要表现在：1）系统进程之间、系统进程与用户进程之间实现 CPU 的自动切换；2）对于具有相互依赖关系的进程，操作系统提供一些信息交换手段，以使它们相互协调、同步运行；3）保护每个独立进程的运行不受其他进程运行的干扰。

2. 共享性（Sharing）

操作系统允许多个进程共享系统的软、硬件资源。采用的方法有：1）资源合理分配与回收；2）对资源进行互斥使用；3）对资源进行共享使用；4）进行安全保护等。

操作系统的这一特性，可以让系统满足多进程对资源的随机性需求，使有限的资源得到有效地利用。为了防止由于进程竞争资源形成系统“死锁”的局面，系统应具有一种预防和化解死锁的手段。

3. 不确定性（Nondeterminacy）

这一特性是并发性和共享性的必然结果。由于系统内的进程很多，每个进程所引发的事件都不是预先安排的，因此，进程的推进速度是不可预知的。即并发进程所处的状态是不确定的。从另一方面说，为了达到进程的并发和共享，操作系统应及时、准确地响应各个进程的需求。由于它们的推进速度不可预知，因而它们在某一时刻的资源拥有情况和系统资源的共享情况也是不确定的。

正是由于操作系统具有上述 3 种特性，使得它必然是一个非常复杂的系统，这也给系统的设计、维护和理解造成很大的困难。下面我们通过分析一个用户程序的运行过程，看一下操作系统各模块间的转换机制。

1.4.2 用户程序运行过程

下面我们将通过一个用户程序的运行过程，来叙述操作系统的整体构造。首先假定计算机上使用的设备是一台具有输入输出功能的卡片机。

当用户将程序的代码编制出来后，用户将代码交给计算站或计算中心，由操作人员制作一些适当的控制卡片插入用户代码中间，形成了计算机能够识别的计算机程序，这就是我们说的“作业”（Job）。

一个作业从提交到运行结束，大体经历如下过程：

(1) 当作业装入卡片输入机上以后，就进入作业“提交”（Submit）状态，由此开始了作业的计算过程。这里，我们使用“提交”一词，仅仅是为了表明作业计算过程中的一个阶段的状态，使用其他词汇来表示也是可以的。

(2) 作业提交到计算机上，接下来由作业管理模块中的“命令处理程序”负责接受操作员的命令，通过解释后，调用设备管理模块中的一个负责输入输出的程序（即 Spooling）将作业从卡片机上读入。经制作，使作业成为能够被作业调度程序处理的形式，并放到外存的“输入井”中，这时的作业处于等待调度的状态——“后备”状态。

由于输入井属于外存的一个存储空间，所以运行时，命令处理程序需要调用文件管理系统，获得适当的外存空间位置。

(3) 处于后备状态的作业已完全处于操作系统的管理之下，由作业管理摸块中的“作业调度程序”负责对后备作业进行调度。在这一阶段，需要调用存储管理和设备管理，为它申请所需的内存空间和外部设备，申请成功后，将作业装入内存，再调用一个称作“进程创建原语”的程序使作业做好运行准备。

(4) 作业被装入内存后，就进入“执行”状态，处于执行状态的作业称作“进程”，由进程管理模块实施管理。进程管理按照一种既定的原则调度各个进程投入运行。

为了提高处理机效率，在多道系统中，往往使若干作业同时处于执行状态，让它们共享处理机。由此，进程又被细划为多个状态，比如，“就绪”状态、“运行”状态和“等待”状态等。

在进程调度中，系统需要从“就绪”状态的进程中挑选出一个或部分进程，让它（们）使用 CPU，投入运行。

(5) 作业运行完毕后，作业由“执行”状态转为“完成”状态。在这期间，要调用设备管理模块和存储管理模块，将作业占用的外部设备和内存空间释放。

(6) 作业完成后的处理，也称“后处理”，由作业管理模块完成。作业管理模块需要调用设备管理模块将运算结果在卡片机上输出，由它的“作业卸出程序”调用文件管理系统，将作业在输入输出井上的空间收回。图 1-3 是一个作业运行状态图。

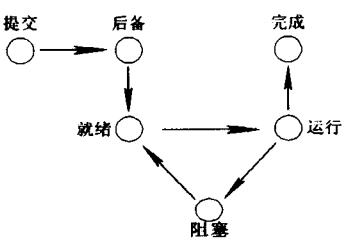

图 1-3 作业运行状态图

其中，“提交”状态是用户任务变成作业的第 1 个状态；“就绪”是作业变成进程的第 1 个状态。“完成”是作业生存周期的最后一个状态。

进程由“运行”转为“阻塞”可能是由种种原因引起的（比如，进程需要启动低速设备进行输入输出等操作）；进程由“阻塞”转为“就绪”，表明自己等待的事件已经出现（输入输出操作已完成），具备了运行的条件。只有当进程处于“运行”状态时，进程才能获得 CPU，得以向前推进。