第 1 章　绪　　论

计算机发展到今天，从个人计算机到巨型计算机系统，毫无例外都配置一种或多种操作系统。什么是操作系统，它具有什么样的功能等，我们将在这一章作一简要阐述。为了阐明这些问题，扼要地回顾一下操作系统的形成和发展过程是必要的。为便于今后的学习，我们要介绍一下操作系统的类型及其特点，研究操作系统的几种观点。最后，介绍几种常用操作系统。

1.1　操作系统概念

迄今，任何一个计算机系统都配置一种或多种操作系统。

计算机系统由两部分组成：硬件和软件。计算机硬件通常是由中央处理机（运算器和控制器）、存储器、输入设备和输出设备等部件组成，它构成了系统本身和用户作业赖以活动的物质基础和工作环境。

计算机软件包括系统软件和应用软件。系统软件如操作系统、多种语言处理程序（汇编和编译程序等）、连接装配程序、系统实用程序、多种工具软件等；应用软件是为应用编制的程序。

没有任何软件支持的计算机称为裸机（bare machine），它仅仅构成了计算机系统的物质基础，而实际呈现在用户面前的计算机系统是经过若干层软件改造的计算机。图 1.1 展示了这种情形。

由图 1.1 可看出，计算机的硬件和软件以及应用之间是一种层次结构的关系。裸机在最里层，它的外面是操作系统，经过操作系统提供的资源管理功能和方便用户的各种服务功能把裸机改造成为功能更强、使用更为方便的机器，通常称之为虚拟机（virtual machine）或扩展机（extended machine），而各种实用程序和应用程序运行在操作系统之上，它们以操作系统作为支撑环境，同时又向用户提供完成其作业所需的各种服务。

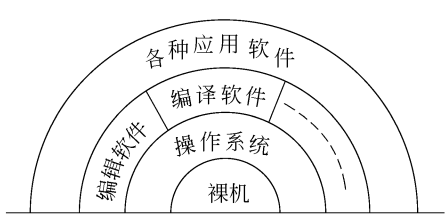


图 1.1　操作系统与硬件软件的关系

引入操作系统的目的可从三方面来考察。

（1）从用户的观点来看　计算机是为用户提供服务的，计算机所完成的任何工作，都是为了满足用户的计算或处理需求。因此，引入操作系统是让计算机为用户提供最好的服务，构建一个用

户和计算机之间的和谐交互环境。这要求计算机有一个良好的用户界面,使用户无须了解许多有关硬件和系统软件的细节,能够方便灵活地使用计算机。同时,计算机还能为用户提供一个可靠和安全的服务管理,以保证用户得到可靠安全的服务。

(2) 从系统管理人员的观点来看　引入操作系统是为了合理地组织计算机工作流程,管理和分配计算机系统硬件及软件资源,使之能为多个用户高效率地共享。因此,操作系统是计算机资源的管理者。

(3) 从发展的观点看　引入操作系统是为了给计算机系统的功能扩展提供支撑平台,使之在追加新的服务和功能时更加容易和不影响原有的服务与功能。

综上所述,我们可以非形式地把操作系统定义为:

操作系统是计算机系统中的一个系统软件,它是这样一些程序模块的集合——它们管理和控制计算机系统中的硬件及软件资源,合理地组织计算机工作流程,以便有效地利用这些资源为用户提供一个具有足够的功能、使用方便、可扩展、安全和可管理的工作环境,从而在计算机与其用户之间起到接口的作用。

操作系统的几个主要特点是:它是一个管理计算机软硬件资源的系统软件,它为用户提供尽可能多的服务,它的管理过程根据用户要求不同而有所不同,但主要是为了让用户高效率地共享计算机软硬件资源,但又要保证其可靠性和安全性以及可用性可管理性。

1.2　操作系统的历史

为了更好地理解操作系统的基本概念、功能和特点,让我们首先回顾一下操作系统形成和发展的历史过程。

操作系统是由于客观的需要而产生的,它伴随着计算机技术本身及其应用的日益发展而逐渐发展和不断完善。它的功能由弱到强,在计算机系统中的地位不断提高。至今,它已成为计算机系统中的核心,无一计算机系统是不配置操作系统的。

由于操作系统历来跟运行其上的计算机组成与体系结构休戚与共,因此我们考察各代计算机,看看它们的操作系统是什么样子,具有哪些功能和特征。

人们通常按照器件工艺的演变把计算机发展过程分为 4 个阶段。

1946 年至 20 世纪 50 年代末　第一代,电子管时代,无操作系统。

20 世纪 50 年代末至 20 世纪 60 年代中期　第二代,晶体管时代,批处理系统。

20 世纪 60 年代中期至 20 世纪 70 年代中期　第三代,集成电路时代,多道程序设计。

20 世纪 70 年代中期至 20 世纪末　第四代,大规模和超大规模集成电路时代,分时系统。

21 世纪初开始,以移动、分布和网络计算为代表,现代计算机正向着普适计算、网格计算以及巨型、微型、并行、分布、网络化、智能化和生物信息化几个方面发展着。

适应上述计算机发展过程,操作系统经历了如下的发展过程:手工操作阶段(无操作系统)、批处理、执行系统、多道程序系统、分时系统、实时系统、通用操作系统、网络操作系统、

分布式操作系统等。

1.2.1 手工操作阶段

在第一代计算机时期，构成计算机的主要元器件是电子管，计算机运算速度慢（只有几千次/秒），没有操作系统，甚至没有任何软件。用户直接用机器语言编制程序，并在上机时独占全部计算机资源。用户既是程序员，又是操作员。上机完全是手工操作：先把程序纸带（或卡片）装上输入机，然后启动输入机把程序和数据送入计算机，接着通过控制台开关启动程序运行。计算完毕，打印机输出计算结果，用户取走并卸下纸带（或卡片）。第二个用户程序上机，照此办理。这种由一道程序独占机器及人工操作的情况，在计算机速度较慢时是允许的，因为此时计算机所需时间相对较长，手工操作所占比例还不很大。

20世纪50年代后期，计算机的运行速度有了很大提高，从每秒几千次几万次发展到每秒几十万次上百万次。这时，由于手工操作的慢速度和计算机的高速度之间形成矛盾，手工操作与机器有效运行时间之比将大大地加大，这种矛盾已经到了不能容忍的地步。唯一的解决办法是摆脱人的手工操作，实现作业的自动过渡。这样就出现了批处理。

1.2.2 早期批处理

在计算机发展的早期阶段，用户上机时需要自己建立和运行作业，并做结束处理。由于没有任何用于管理的软件，所有的运行管理和具体操作都由用户自己承担。每个作业都由许多作业步组成，任何一步的错误操作都可能导致该作业从头开始。在当时，计算机的价格是极其昂贵的，计算机（CPU）的时间是非常宝贵的，尽可能提高CPU的利用率成为十分迫切的任务。

解决的途径有两个：首先配备专门的计算机操作员，程序员不再直接操作机器，减少操作机器的错误。另一个重要措施是进行批处理，操作员把用户提交的作业分类，把一批中的作业编成一个作业执行序列。每一批作业将有专门编制的监督程序（monitor）自动依次处理。

早期的批处理可分为两种方式。

1. 联机批处理

慢速的输入输出（I/O）设备是和主机直接相连。作业的执行过程为：

（1）用户提交作业：作业程序、数据，用作业控制语言编写的作业说明书；

（2）作业被做成穿孔纸带或卡片；

（3）操作员有选择地把若干作业合成一批，通过输入设备（纸带输入机或读卡机）把它们存入磁带；

（4）监督程序读入一个作业（若系统资源能满足该作业要求）；

（5）从磁带调入汇编程序或编译程序，将用户作业源程序翻译成目标代码；

(6) 连接装配程序把编译后的目标代码及所需的子程序装配成一个可执行程序;

(7) 启动执行;

(8) 执行完毕,由善后处理程序输出计算结果;

(9) 再读入一个作业,重复(5)~(9)各步;

(10) 一批作业完成,返回到(3),处理下一批作业。

这种联机批处理方式解决了作业自动转接,从而减少作业建立和人工操作时间。但是在作业的输入和执行结果的输出过程中,主机 CPU 仍处在停止等待状态,这样慢速的输入输出设备和快速主机之间仍处于串行工作,CPU 的时间仍有很大的浪费。

2. 脱机批处理

这种方式的显著特征是增加一台不与主机直接相连而专门用于与输入输出设备打交道的卫星机,如图 1.2 所示。

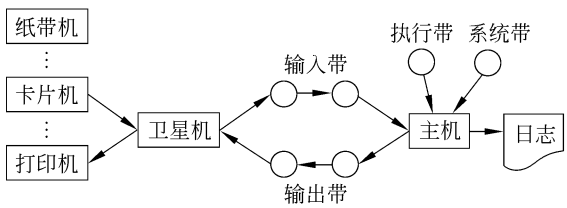


图 1.2 早期脱机批处理模型

卫星机的功能是:

(1) 输入设备通过它把作业输入到输入磁带;

(2) 输出磁带将作业执行结果输出到输出设备。

这样,主机不是直接与慢速的输入输出设备打交道,而是与速度相对较快的磁带机发生关系。主机与卫星机可以并行工作,二者分工明确,以充分发挥主机的高速度计算能力。因此脱机批处理和早期联机批处理相比大大提高了系统的处理能力。

批处理出现于 20 世纪 50 年代末到 60 年代初,它是为了提高主机的使用效率,在解决人机矛盾(主机高速度和输入输出设备的慢速度的矛盾)的过程中逐步发展起来的。它的出现促使了软件的发展。还有重要的是监督程序,它管理作业的运行——负责装入和运行各种系统处理程序,如汇编程序、编译程序、连接装配程序、程序库(如输入输出标准程序等);完成作业的自动过渡,同时也出现程序覆盖等程序设计技术。

批处理克服了手工操作的缺点,实现了作业的自动过渡,改善了主机 CPU 和输入输出设备使用情况,提高了计算机系统的处理能力。但仍有些缺点:磁带需人工拆装,既麻烦又易出错;而另一个更重要的问题是系统的保护。让我们来回忆一下在监督程序管理下的解题过程,如图 1.3 所示。

在进行批处理过程中,监督程序、系统程序和用户程序之间存在着一种调用关系,任何一个环节出问题,整个系统都会停顿;用户程序也可能会破坏监督程序和系统程序,这时,只有操作员进行干预才能恢复。20 世纪 60 年代初期,硬件获得了两方面(即通道和中断技术)的进展,导致操作系统进入执行系统阶段。

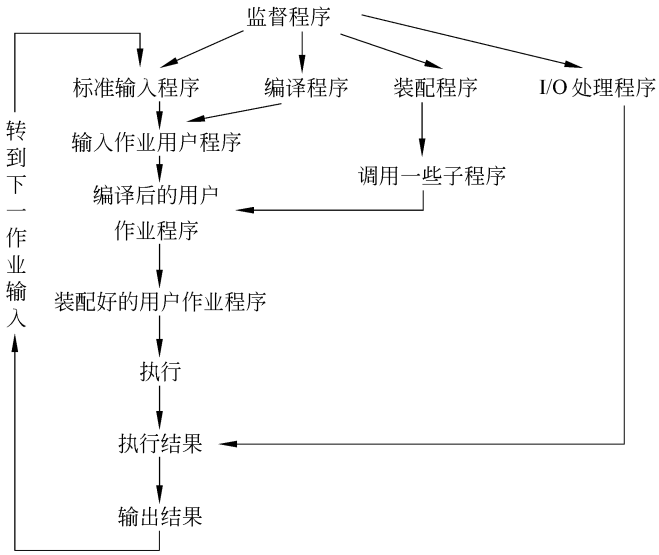


图 1.3 监督程序管理下的解题过程

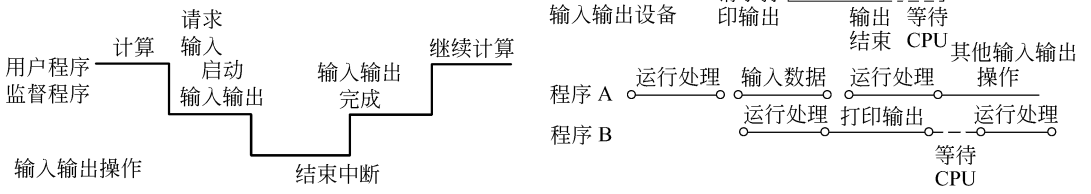
通道是一种专用处理部件，它能控制一台或多台输入输出设备工作，负责输入输出设备与主存之间的信息传输。它一旦被启动就能独立于 CPU 运行，这样可使 CPU 和通道并行操作，而且 CPU 和多种输入输出设备也能并行操作。中断是指当主机接到外部信号（如输入输出设备完成信号）时，马上停止原来工作，转去处理这一事件，处理完毕后，主机回到原来的断点继续工作。

借助于通道、中断技术和输入输出可在主机控制下完成批处理。这时，原来的监督程序的功能扩大了，它不仅要负责作业运行的自动调度，而且还要提供输入输出控制功能。这个发展了的监督程序常驻内存称为执行系统（executive system）。执行系统实现的也是输入输出联机操作，和早期批处理系统不同的是：输入输出工作是由在主机控制下的通道完成的。主机和通道、主机和输入输出设备都可以并行操作。用户程序的输入输出工作都是由系统执行而没有人工干预，由系统检查其命令的合法性，以避免不合法的输入输出命令造成对系统的影响，从而提高系统的安全性。此时，除了输入输出中断外，其他中断如算术溢出和非法操作码中断等可以克服错误停机，而时钟中断可以解决用户程序中出现的死循环等。

许多成功的批处理系统在 20 世纪 50 年代末和 60 年代初出现，典型的操作系统是 FMS（Fortran Monitor System）即 FORTRAN 监督系统和 IBM/7094 机上的 IBM 操作系统 IBSYS。执行系统实现了主机、通道和输入输出设备的并行操作，提高了系统效率，方便用户对输入输出设备的使用。但是，这时计算机系统运行的特征是单道顺序地处理作业，即用户作业仍然是按一道一道作业顺序处理。那么可能会出现两种情况：对于以计算为主的作业，输入输出量少，外围设备空闲；然而对于以输入输出为主的作业，又会造成主机空闲。这样总的来说，计算机资源使用效率仍然不高。因此操作系统进入了多道程序阶段：多道程序合理搭配交替运行，充分利用资源，提高效率。

1.2.3 多道程序系统

上述批处理系统,每次只调用一个用户作业程序进入内存并运行,称为单道运行。图 1.4(a)给出了单道程序工作示例。



(a) 单道程序工作示例　　(b) 多道程序工作示例

图 1.4　程序工作示例

而图 1.4(b)给出了多道程序工作示例。在单处理机系统中,多道程序运行的特点是:

(1) 多道　计算机内存中同时存放几道相互独立的程序。

(2) 宏观上并行　同时进入系统的几道程序都处于运行过程中,即它们先后开始了各自的运行,但都未运行完毕。

(3) 微观上串行　实际上,各道程序轮流使用 CPU,交替执行。

在批处理系统中采用多道程序设计技术,就形成了多道批处理系统。要处理的许多作业存放在外部存储器中,形成作业队列,等待运行。当需要调入作业时,将由操作系统中的作业调度程序对外存中的一批作业,根据其对资源的要求和一定的调度原则,调几个作业进入内存,让它们交替运行。当某个作业完成,然后再调入一个或几个作业。这种处理方式,在内存中总是同时存在几道程序,系统资源得到比较充分的利用。

多道程序系统中,要解决这样一些技术问题。

(1) 并行运行的程序要共享计算机系统的硬件和软件资源,既有对资源的竞争,但又须相互同步。因此同步与互斥机制成为操作系统设计中的重要问题。

(2) 随着多道程序的增加,出现了内存不够用的问题,提高内存的使用效率也成为关键。因此出现了诸如覆盖技术、对换技术和虚拟存储技术等内存管理技术。

(3) 由于多道程序存在于内存,为了保证系统程序存储区和各用户程序存储区的安全可靠,提出了内存保护的要求。

多道程序系统的出现标志着在操作系统渐趋成熟的阶段先后出现了作业调度管理、处理机管理、存储器管理、外部设备管理、文件系统管理等功能。

1.2.4 分时操作系统

批处理方式下,用户以脱机操作方式使用计算机,用户在提交作业以后就完全脱离了自

己的作业，在作业运行过程中，不管出现什么情况都不能加以干预，只有等该批作业处理结束，用户才能得到计算结果。根据结果再作下一步处理，若有错，还得重复上述过程。它的好处是计算机效率高。不过，用户十分留恋手工操作阶段的联机工作方式，独占计算机，并直接控制程序运行。但独占计算机方式会造成资源效率低。既能保证计算机效率，又能方便用户使用，成为一种新的追求目标。20 世纪 60 年代中期，计算机技术和软件技术的发展使这种追求成为可能。由于 CPU 速度不断提高和采用分时技术，一台计算机可同时连接多个用户终端，而每个用户可在自己的终端上联机使用计算机，好像自己独占机器一样。

所谓分时技术，就是把处理机的运行时间分成很短的时间片，按时间片轮流把处理机分配给各联机作业使用。若某个作业在分配给它的时间片内不能完成其计算，则该作业暂时中断，把处理机让给另一作业使用，等待下一轮时再继续其运行。由于计算机速度很快，作业运行轮转得很快，给每个用户的印象是好像他独占了一台计算机。而每个用户可以通过自己终端向系统发出各种操作控制命令，完成作业的运行。

多用户分时操作系统是当今计算机操作系统中最普遍使用的一类操作系统。

1.2.5 实时操作系统

20 世纪 60 年代中期计算机进入第三代，计算机的性能和可靠性有了很大提高，造价亦大幅度下降，导致计算机应用越来越广泛。计算机由于用于工业过程控制、军事实时控制等形成了各种实时系统。实时操作系统是以在允许时间范围之内做出响应为特征的。它要求计算机对于外来信息能以足够快的速度进行处理，并在被控对象允许时间范围内做出快速响应，其响应时间要求在秒级、毫秒级甚至微秒级或更小。实时操作系统在嵌入式计算得到了越来越广泛的应用。特别是移动计算等非 PC 机、PDA（个人数字助理）和手机等新设备的出现，更加强了这一趋势。

例如，随着移动通信进入 3G 时代，诺基亚等公司研制的 Symbian 手机操作系统、微软公司研制的 Windows Mobile、近年崛起的操作系统新秀 Linux 等都已有了巨大的市场和用户群体。

1.2.6 通用操作系统

多道批处理系统和分时系统的不断改进、实时系统的出现及其应用日益广泛，致使操作系统日益完善。在此基础上，出现了通用操作系统。它可以同时兼有多道批处理、分时、实时处理的功能，或其中两种以上的功能。例如，将实时处理和批处理相结合构成实时批处理系统。在这样的系统中，它首先保证优先处理任务，插空进行批作业处理。通常把实时任务称为前台作业，批作业称为后台作业。将批处理和分时处理相结合可构成分时批处理系统。在保证分时用户的前提下，没有分时用户时可进行批量作业的处理。同样，分时用户和批处理作业可按前后台方式处理。

从 20 世纪 60 年代中期开始，国际上开始研制大型通用操作系统。这些系统试图达到功能齐全、可适应各种应用范围和操作方式变化多端的环境的目标。但是这些系统本身很

庞大,不仅付出了巨大的代价,而且由于系统过于复杂和庞大,在解决其可靠性、可维护性、可理解性和开放性等方面都遇到很大的困难。相比之下,UNIX 操作系统却是一个例外。这是一个通用的多用户分时交互型的操作系统。它首先建立的是一个精干的核心,而其功能却足以与许多大型的操作系统相媲美,在核心层以外可以支持庞大的软件系统,它很快得到应用和推广并不断完善,对现代操作系统有着重大的影响。目前广泛使用的各种工作站级的操作系统,例如 SUN 公司的 Solaris,IBM 公司的 AIX 等都是基于 UNIX 的操作系统。即使 Microsoft 公司的 Windows 系列操作系统,其主要原理也是基于 UNIX 系统的。另外,目前广为流传的 Linux 系统也是从 UNIX 演变成的。

至此,操作系统的基本概念、功能、基本结构和组成都已形成并渐趋完善。

1.2.7 操作系统的进一步发展

进入 20 世纪 80 年代,大规模集成电路工艺技术的飞跃发展,微处理机的出现和发展,掀起了计算机大发展大普及的浪潮。一方面迎来了个人计算机的时代,同时又向计算机网络、分布式处理、巨型计算机和智能化方向发展。操作系统有了进一步的发展,例如:

(1) 个人计算机上的操作系统,例如 Windows 操作系统系列;
(2) 嵌入式操作系统,例如 Symbian 操作系统;
(3) 网络操作系统;
(4) 分布式操作系统;
(5) 智能化操作系统。

1.3 操作系统的基本类型

通过上一节的讨论,我们已知,随着计算机技术和软件技术长期发展,已形成了各种类型的操作系统,以满足不同的应用要求。根据其使用环境和对作业处理方式,操作系统的基本类型有:

(1) 批处理操作系统(batch processing operating system);
(2) 分时操作系统(time sharing operating system);
(3) 实时操作系统(real time operating system);
(4) 个人计算机操作系统(personal computer operating system);
(5) 网络操作系统(network operating system);
(6) 分布式操作系统(distributed operating system)。

下面对它们作一概要的说明。

1.3.1 批处理操作系统

批处理操作系统是一种早期的大型机用操作系统。不过,现代操作系统大都具有批处

理功能。图 1.5 给出了批处理系统中作业处理步骤及状态。

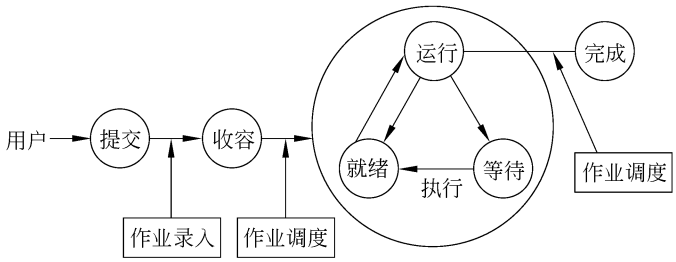


图 1.5 批处理系统中作业处理及状态

批处理系统的主要特征是：

(1) 用户脱机使用计算机 用户提交作业之后直到获得结果之前就不再和计算机打交道。作业提交的方式可以是直接交给计算中心的管理操作员，也可以是通过远程通信线路提交。提交的作业由系统外存收容成为后备作业。

(2) 成批处理 操作员把用户提交的作业分批进行处理。每批中的作业将由操作系统或监督程序负责作业间自动调度执行。

(3) 多道程序运行 按多道程序设计的调度原则，从一批后备作业中选取多道作业调入内存并组织它们运行，成为多道批处理。

多道批处理系统的优点是由于系统资源为多个作业所共享，其工作方式是作业之间自动调度执行。并在运行过程中用户不干预自己的作业，从而大大提高了系统资源的利用率和作业吞吐量。其缺点是无交互性，用户一旦提交作业就失去了对其运行的控制能力；而且是批处理的，作业周转时间长，用户使用不方便。

值得一提的是不要把多道程序系统(multiprogramming)和多重处理系统(multiprocessing)相混淆。一般讲，多重处理系统配制多个 CPU，因而能真正同时执行多道程序。当然，要想有效地使用多重处理系统，必须采用多道程序设计技术。反之不然，多道程序设计原则不一定要求有多重处理系统的支持。多重处理系统比起单处理系统来说，虽增加了硬件设施，却换来了提高系统吞吐量、可靠性、计算能力和并行处理能力等好处。

1.3.2 分时系统

分时系统一般采用时间片轮转的方式，使一台计算机为多个终端用户服务。对每个用户能保证足够快的响应时间，并提供交互会话能力。因此它具有下述特点。

(1) 交互性 交互会话工作方式给用户带来了许多好处。首先，用户可以在程序动态运行情况下对其加以控制，从而加快调试过程，提供了软件开发的良好环境。其次，用户上机提交作业方便。特别对于远程终端用户，不必将其作业交给机房，在自己的终端上就可以提交、调试、运行其程序。第三，分时系统还为用户之间进行合作提供方便。他们可以通过文件系统、电子邮件或其他通信机制彼此交换数据和信息，共同完成某项任务。

(2) 多用户同时性 多个用户同时在自己的终端上上机，共享 CPU 和其他资源，充分发挥系统的效率。

(3) 独立性　由于采用时间轮转方式使一台计算机同时为多个终端服务,对于每个用户的操作命令又能快速响应,因此,客观效果上用户彼此之间都感觉不到有别人也在使用该台计算机,如同自己独占计算机一样。

分时操作系统是一个联机的(on-line)多用户(multi-user)交互式(interactive)的操作系统。UNIX 是当今最流行的一种多用户分时操作系统,但 CTSS(compatible time sharing system)和 MUTICS(multiplexed information and computing service)这两个系统也是值得一提的。前者是一个实验性的分时系统,在 1963 年由 MIT 研制成功的。后者是由 MIT、Bell 实验室和 GE 公司联合在 1965 年开始设计的,尽管它并没有取得最后成功,但对 UNIX 的研制是有影响的。

1.3.3　实时系统

实时系统是另外一类联机的操作系统。它主要随着计算机应用于实时控制和实时信息处理领域中而发展起来的。

实时系统的主要特点是提供即时响应和高可靠性。系统必须保证对实时信息的分析和处理的速度比其进入系统的速度要快,而且系统本身要安全可靠,因为像生产过程的实时控制、武器系统的实时控制、航空订票、银行业务等实时事务系统,信息处理的延误或丢失往往会带来不堪设想的后果。实时系统往往具有一定的专用性,它大多用于嵌入式计算中。与批处理系统、分时系统相比,实时系统的资源利用率可能较低。

设计实时操作系统要考虑如下一些因素。

(1) 实时时钟管理(定时处理和延时处理)。

(2) 连续的人-机对话,这对实时控制往往是必需的。

(3) 过载保护在实时系统中进入系统的实时任务的时间和数目有很大的随意性,因而在某一时刻有可能超出系统的处理能力,这就是所谓过载问题,要求采取过载保护措施。例如对于短期过载,把输入任务按一定的策略在缓冲区排队,等待调度;对于持续性过载,可能要拒绝某些任务的输入;在实时控制系统中,则及时处理某些任务,放弃某些任务或降低对某些任务的服务频率。

(4) 高度可靠性和安全性需采取冗余措施。双机系统前后台工作,包括必要的保密措施等。

1.3.4　通用操作系统

批处理系统、分时系统和实时系统是操作系统的三种基本类型,在此基础上又发展了具有多种类型操作特征的操作系统,称为通用操作系统。它可以同时兼有批处理、分时、实时处理和多重处理的功能,或其中两种以上的功能。

1.3.5　个人计算机上的操作系统

个人计算机上的操作系统是联机的交互式单用户操作系统,它提供的联机交互功能与