

新世纪计算机类本科系列教材



计算机操作系统

(修订版)

汤子瀛 哲凤屏 汤小丹 编著



西安电子科技大学出版社
<http://www.xduph.com>

新世纪计算机类本科系列教材

计算机操作系统

(修订版)

汤子瀛 哲凤屏 汤小丹 编著

西安电子科技大学出版社

内 容 简 介

本教材介绍了计算机系统中的一个重要系统软件——操作系统(OS)。全书共分10章,第1章介绍OS的发展过程、基本特征、功能以及OS的结构设计;第2、3章详细地阐述了进程和线程的基本概念、同步与通信、调度与死锁;第4章介绍连续式、离散式存储器的管理方式及虚拟存储器;第5、6、7章分别介绍设备管理、文件管理和用户接口;第8章介绍了计算机网络系统、网络OS所提供的功能和服务,以及Internet和Intranet;第9章对保障系统安全的访问控制、认证、数据加密和防火墙四大技术作了较详细的阐述;第10章介绍了一个OS的实例——UNIX系统V的内核结构。

本教材可作为计算机科学与工程和计算机应用专业本科生的教科书,也可作为从事计算机工作的科技人员学习OS的参考书。

图书在版编目(CIP)数据

计算机操作系统/汤子瀛等编著. —2版

—西安:西安电子科技大学出版社,2001.8

新世纪计算机类本科系列教材

ISBN 7-5606-0496-X

I. 计… II. 汤… III. 操作系统(软件)—高等学校—教材 N. TP316

中国版本图书馆CIP数据核字(2001)第028905号

责任编辑 李惠萍

出版发行 西安电子科技大学出版社(西安市太白南路2号)

电 话 (029) 8227828 邮 编 710071

http://www.xduph.com E-mail: xdupfxb@pub.xaonline.com

经 销 新华书店

印 刷 西安文化彩印厂

版 次 1996年12月第1版 2001年8月第2版 2002年9月第10次印刷

开 本 787毫米×1092毫米 1/16 印张 23.125

字 数 546千字

印 数 78 001~86 000

定 价 24.00元

ISBN 7-5606-0496-X/TP·0232

XDUP 0766A02-10

*** 如有印装问题可调换 ***

本书封面贴有西安电子科技大学出版社的激光防伪标志,无标志者不得销售。

再版前言

操作系统(OS)是最重要的计算机系统软件,同时也是最活跃的学科之一,其发展极为迅速。为使本教材内容能及时地紧跟时代潮流,从1981年至今,我们已对本教材做过多次修改。2000年我们又对1996年出版的《计算机操作系统》教材进行了重写。为了适当压缩篇幅,我们调整了该教材的结构,从原来的15章改为10章,即将原来的第2、3章合并为“进程的描述与控制”一章;原来的第5、6章合并为“存储器管理”一章;第8、9章合并为“文件管理”一章;第11、12章合并为“网络操作系统”一章;另外,还考虑到在大学低年级的实践中,学生已经学习过Windows OS的使用,故本书将原第15章删掉。

我们在本教材中,介绍了许多在90年代引入或广泛使用的技术,如微内核OS结构、线程的控制与通信、数据一致性、系统容错技术等,又因为20世纪90年代是计算机网络特别是Internet大发展的年代,故我们对网络操作系统一章做了较大的修改。还应强调说明的是,随着网络的广泛应用,系统安全性问题提到了头等重要的地位。事实上,若不能确保系统(网络)的安全性,则系统(网络)是难以被人接受的,故在国内外的OS教科书中,大多增加了一章或几章内容用于介绍系统的安全性保障。我们在第9章中,对系统安全性做了较全面的阐述。

本次再版的《计算机操作系统》一书共分10章。第1章仍为OS引论,介绍OS的发展过程、基本特征和功能,新增加了OS的结构设计;第2、3章详细地阐述了进程和线程的基本概念、进程控制、同步与通信以及调度与死锁,增加了线程的控制、线程的同步与通信;第4章为存储管理,内容有连续分配、离散式分配存储管理方式和虚拟存储器;第5、6章分别为设备管理和文件管理;第7章介绍操作系统接口,其中,增加了UNIX系统的Shell语言和系统调用的实现方法;第8章为网络操作系统,扼要地介绍了计算机网络的基本概念、网络OS的工作模式、功能和提供的服务,以及Internet/Intranet;第9章对保障系统和网络安全的存取控制、认证、数据加密和防火墙四大技术,做了较详细的阐述;第10章介绍了当前广泛使用的OS实例——UNIX系统V的内核结构。

本教材在编写过程中,得到了西安电子科技大学出版社的大力支持与合作。此外,汤蓓莉、王侃雅等同志在整理、校对、绘图等工作中,都付出了艰辛的劳动,使本教材能如期地与读者见面。在此谨向以上各位表示衷心感谢。

本教材虽经多次反复修改,突出了操作系统的基本概念,反映了当代操作系统的新技术,但限于编者水平,在本次编写的教材中,仍难免会有错误和不当之处,恳请读者批评指正。

编 者

2000年12月

目 录

第一章 操作系统引论

1.1 操作系统的目标和作用	1	1.3.3 虚拟(Virtual)	13
1.1.1 操作系统的目标	1	1.3.4 异步性(Asynchronism)	13
1.1.2 操作系统的作用	2	1.4 操作系统的主要功能	14
1.1.3 推动操作系统发展的主要动力	3	1.4.1 处理机管理功能	14
1.2 操作系统的发展过程	4	1.4.2 存储器管理功能	15
1.2.1 无操作系统的计算机系统	4	1.4.3 设备管理功能	16
1.2.2 单道批处理系统	5	1.4.4 文件管理功能	17
1.2.3 多道批处理系统	6	1.4.5 用户接口	18
1.2.4 分时系统	8	1.5 操作系统的结构设计	19
1.2.5 实时系统	10	1.5.1 软件工程的基本概念	19
1.3 操作系统的基本特性	11	1.5.2 传统的操作系统结构	20
1.3.1 并发(Concurrence)	11	1.5.3 微内核 OS 结构	22
1.3.2 共享(Sharing)	12	习题	25

第二章 进 程 管 理

2.1 进程的基本概念	26	2.4.2 哲学家进餐问题	48
2.1.1 程序的顺序执行及其特征	26	2.4.3 读者—写者问题	49
2.1.2 前趋图	27	2.5 管程机制	51
2.1.3 程序的并发执行及其特征	28	2.5.1 管程的基本概念	51
2.1.4 进程的特征与状态	29	2.5.2 利用管程解决生产者—消费者问题	52
2.1.5 进程控制块	32	2.6 进程通信	54
2.2 进程控制	34	2.6.1 进程通信的类型	54
2.2.1 进程的创建	34	2.6.2 消息传递通信的实现方法	55
2.2.2 进程的终止	35	2.6.3 消息传递系统实现中的若干问题	57
2.2.3 进程的阻塞与唤醒	36	2.6.4 消息缓冲队列通信机制	58
2.2.4 进程的挂起与激活	38	2.7 线程	60
2.3 进程同步	38	2.7.1 线程的基本概念	60
2.3.1 进程同步的基本概念	38	2.7.2 线程间的同步和通信	63
2.3.2 信号量机制	41	2.7.3 内核支持线程和用户级线程	64
2.3.3 信号量的应用	44	2.7.4 线程控制	65
2.4 经典进程的同步问题	46	习题	68
2.4.1 生产者—消费者问题	46		

第三章 处理机调度与死锁

3.1 处理机调度的基本概念	70	3.4.1 多处理器系统的类型	85
3.1.1 高级、中级和低级调度	70	3.4.2 进程分配方式	86
3.1.2 调度队列模型	72	3.4.3 进程(线程)调度方式	87
3.1.3 选择调度方式和调度算法的 若干准则	73	3.5 产生死锁的原因和必要条件	90
3.2 调度算法	75	3.5.1 产生死锁的原因	90
3.2.1 先来先服务和短作业(进程) 优先调度算法	75	3.5.2 产生死锁的必要条件	92
3.2.2 高优先权优先调度算法	77	3.5.3 处理死锁的基本方法	92
3.2.3 基于时间片的轮转调度算法	79	3.6 预防死锁的方法	93
3.3 实时调度	80	3.6.1 预防死锁	93
3.3.1 实现实时调度的基本条件	80	3.6.2 系统安全状态	94
3.3.2 实时调度算法的分类	82	3.6.3 利用银行家算法避免死锁	95
3.3.3 常用的几种实时调度算法	83	3.7 死锁的检测与解除	98
3.4 多处理机系统中的调度	85	3.7.1 死锁的检测	98
		3.7.2 死锁的解除	100
		习题	101

第四章 存储器管理

4.1 程序的装入和链接	103	4.5 虚拟存储器的基本概念	125
4.1.1 程序的装入	104	4.5.1 虚拟存储器的引入	125
4.1.2 程序的链接	105	4.5.2 虚拟存储器的实现方法	126
4.2 连续分配方式	106	4.5.3 虚拟存储器的特征	127
4.2.1 单一连续分配	107	4.6 请求分页存储管理方式	128
4.2.2 固定分区分配	107	4.6.1 请求分页中的硬件支持	128
4.2.3 动态分区分配	108	4.6.2 内存分配策略和分配算法	129
4.2.4 可重定位分区分配	110	4.6.3 调页策略	132
4.2.5 对换(Swapping)	112	4.7 页面置换算法	133
4.3 基本分页存储管理方式	113	4.7.1 最佳置换算法和先进先出置换算法	133
4.3.1 页面与页表	114	4.7.2 最近最久未使用(LRU)置换算法	134
4.3.2 地址变换机构	115	4.7.3 Clock 置换算法	136
4.3.3 两级和多级页表	116	4.7.4 其它置换算法	137
4.4 基本分段存储管理方式	119	4.8 请求分段存储管理方式	138
4.4.1 分段存储管理方式的引入	119	4.8.1 请求分段中的硬件支持	138
4.4.2 分段系统的基本原理	120	4.8.2 分段的共享与保护	140
4.4.3 信息共享	122	习题	142
4.4.4 段页式存储管理方式	123		

第五章 设备管理

5.1 I/O 系统	144	5.1.4 总线系统	150
5.1.1 I/O 设备	144	5.2 I/O 控制方式	151
5.1.2 设备控制器	146	5.2.1 程序 I/O 方式	151
5.1.3 I/O 通道	148	5.2.2 中断驱动 I/O 控制方式	152

5.2.3 直接存储器访问 DMA I/O 控制方式	153	5.4.5 SPOOLing 技术	166
5.2.4 I/O 通道控制方式	154	5.5 设备处理	167
5.3 缓冲管理	155	5.5.1 设备驱动程序的功能和特点	168
5.3.1 缓冲的引入	155	5.5.2 设备驱动程序的处理过程	169
5.3.2 单缓冲和双缓冲	156	5.5.3 中断处理程序的处理过程	170
5.3.3 循环缓冲	158	5.6 磁盘存储器管理	171
5.3.4 缓冲池(Buffer Pool)	159	5.6.1 磁盘性能简述	172
5.4 设备分配	161	5.6.2 磁盘调度	173
5.4.1 设备分配中的数据结构	161	5.6.3 磁盘高速缓存(Disk Cache)	176
5.4.2 设备分配时应考虑的因素	162	5.6.4 提高磁盘 I/O 速度的其它方法	178
5.4.3 设备独立性	163	5.6.5 廉价磁盘冗余阵列	179
5.4.4 独占设备的分配程序	165	习题	181

第六章 文件管理

6.1 文件和文件系统	182	6.4.2 目录结构	200
6.1.1 文件、记录和数据项	182	6.4.3 目录查询技术	204
6.1.2 文件类型和文件系统模型	183	6.5 文件存储空间的管理	205
6.1.3 文件操作	185	6.5.1 空闲表法和空闲链表法	205
6.2 文件的逻辑结构	186	6.5.2 位示图法	206
6.2.1 文件逻辑结构的类型	187	6.5.3 成组链接法	207
6.2.2 顺序文件	187	6.6 文件共享与文件保护	209
6.2.3 索引文件	189	6.6.1 基于索引结点的共享方式	209
6.2.4 索引顺序文件	190	6.6.2 利用符号链实现文件共享	210
6.2.5 直接文件和哈希文件	191	6.6.3 磁盘容错技术	211
6.3 外存分配方式	191	6.7 数据一致性控制	213
6.3.1 连续分配	192	6.7.1 事务	214
6.3.2 链接分配	193	6.7.2 检查点	215
6.3.3 索引分配	195	6.7.3 并发控制	215
6.4 目录管理	198	6.7.4 重复数据的数据一致性问题	216
6.4.1 文件控制块和索引结点	198	习题	219

第七章 操作系统接口

7.1 联机命令接口	221	7.3.1 系统调用的基本概念	233
7.1.1 联机命令的类型	221	7.3.2 系统调用的类型	234
7.1.2 键盘终端处理程序	223	7.3.3 系统调用的实现	236
7.1.3 命令解释程序	225	7.4 UNIX 系统调用	238
7.2 Shell 命令语言	227	7.4.1 UNIX 系统调用的类型	238
7.2.1 简单命令	227	7.4.2 被中断进程的环境保护	241
7.2.2 重定向与管道命令	230	7.4.3 系统调用陷入后需处理的若干公共问题	241
7.2.3 通信命令	231	7.5 图形用户接口	243
7.2.4 后台命令	232	7.5.1 桌面、图标和任务栏	243
7.3 系统调用	233		

7.5.2 窗口	245	习题	249
7.5.3 对话框	247		

801

第八章 网络操作系统

8.1 计算机网络概述	250	8.4 网络操作系统提供的服务	267
8.1.1 计算机网络的拓扑结构	250	8.4.1 电子邮件服务	268
8.1.2 计算机广域网	253	8.4.2 文件传输服务	269
8.1.3 计算机局域网	255	8.4.3 目录服务	270
8.1.4 开放系统互连参考模型	256	8.5 支持 Internet 与 Intranet 的功能和服务	272
8.2 客户/服务器模式	258	8.5.1 Internet 简介	273
8.2.1 客户/服务器模式的形成及其优点	258	8.5.2 Internet 提供的信息服务	275
8.2.2 两层结构的客户/服务器模式	259	8.5.3 Intranet 及其特征	277
8.2.3 三层结构的客户/服务器模式的引入	260	8.6 Windows NT	279
8.2.4 两层 C/S 与三层 C/S 的比较	261	8.6.1 Windows NT 的发展过程	279
8.3 网络操作系统的功能	262	8.6.2 Windows NT 的优良性能	280
8.3.1 数据通信功能	263	8.6.3 网络文件/打印服务	281
8.3.2 资源共享功能	264	8.6.4 目录服务	283
8.3.3 网络管理功能	265	8.6.5 数据安全策略	286
8.3.4 应用互操作功能	267	习题	287

第九章 系统安全性

9.1 引言	289	9.3.2 基于物理标志的认证技术	304
9.1.1 系统安全性的内容和性质	289	9.3.3 基于公开密钥的认证技术	305
9.1.2 对系统安全威胁的类型	290	9.4 访问控制技术	307
9.1.3 对各类资源的威胁	291	9.4.1 访问矩阵(Access Matrix)	307
9.1.4 信息技术安全评价公共准则	293	9.4.2 访问矩阵的修改	309
9.2 数据加密技术	294	9.4.3 访问控制矩阵的实现	310
9.2.1 数据加密的基本概念	294	9.5 防火墙技术	312
9.2.2 对称加密算法与非对称加密算法	297	9.5.1 包过滤防火墙	313
9.2.3 数字签名和数字证明书	298	9.5.2 代理服务技术	315
9.2.4 网络加密技术	300	9.5.3 规则检查防火墙	317
9.3 认证技术	302	习题	317
9.3.1 基于口令的身份认证技术	302		

第十章 UNIX 系统内核结构

10.1 UNIX 系统概述	319	10.2.2 进程状态与进程映像	325
10.1.1 UNIX 系统的发展史	319	10.2.3 进程控制	327
10.1.2 UNIX 系统的特征	321	10.2.4 进程调度与切换	329
10.1.3 UNIX 系统的内核结构	322	10.3 进程的同步与通信	330
10.2 进程的描述和控制	323	10.3.1 sleep 与 wakeup 同步机制	330
10.2.1 进程控制块 PCB	323	10.3.2 信号(signal)机制	331

10.3.3	管道机制	332	10.5.4	磁盘驱动程序	345
10.3.4	消息机制	333	10.5.5	磁盘读、写程序	346
10.3.5	共享存储区机制	334	10.6	文件管理	348
10.3.6	信号量集机制	335	10.6.1	UNIX 文件系统概述	348
10.4	存储器管理	336	10.6.2	文件的物理结构	349
10.4.1	请求调页管理的数据结构	337	10.6.3	索引结点的管理	351
10.4.2	换页进程	338	10.6.4	空闲磁盘空间的管理	353
10.4.3	请求调页	340	10.6.5	文件表的管理	354
10.5	设备管理	340	10.6.6	目录管理	355
10.5.1	字符设备缓冲区管理	341	习题		357
10.5.2	块设备缓冲区管理	342	357	参考文献	358
10.5.3	内核与驱动程序接口	344			

第一章 操作系统引论

计算机系统由硬件和软件两部分组成,操作系统 OS(Operating System)是配置在计算机硬件上的第一层软件,是对硬件系统的首次扩充。它在计算机系统中占据了特别重要的地位;而其它的诸如汇编程序、编译程序、数据库管理系统等系统软件,以及大量的应用软件,都将依赖于操作系统的支持,取得它的服务。操作系统已成为现代计算机系统(大、中、小及微型机)中都必须配置的软件。

1.1 操作系统的目标和作用

在计算机系统上配置操作系统的主要目标,与计算机系统的规模和操作系统的应用环境有关。通常,对于配置在大、中型计算机系统上的 OS,都有着较高的要求,相应地,其 OS 就具有较强的功能;又如,对应用于实时工业控制和武器控制环境下的 OS,则要求其 OS 具有实时性和高度可靠性。

1.1.1 操作系统的目标

目前存在着多种类型的 OS,不同类型的 OS,其目标各有所侧重。通常在计算机硬件上配置的 OS,其目标有以下几点:

1. 方便性

配置 OS 后可使计算机系统更容易使用。一个未配置 OS 的计算机系统是极难使用的,因为计算机硬件只能识别 0 和 1 这样的机器代码。因此,用户要在计算机上运行自己所编写的程序,就必须用机器语言书写程序;用户要想输入数据或打印数据,也都必须自己用机器语言书写相应的输入程序或打印程序。如果我们在计算机硬件上配置了 OS,用户便可通过 OS 所提供的各种命令来使用计算机系统。比如,用编译命令可方便地把用户用高级语言书写的程序,翻译成机器代码,大大地方便了用户,从而使计算机变得易学易用。

2. 有效性

在未配置 OS 的计算机系统中,诸如 CPU、I/O 设备等各类资源,都会因经常处于空闲状态而得不到充分利用;内存及外存中所存放的数据由于无序而浪费了存储空间。配置了 OS 后,可使 CPU 和 I/O 设备由于能保持忙碌状态而得到有效的利用,且由于可使内存和外存中存放的数据有序而节省了存储空间。此外,OS 还可以通过合理地组织计算机的工作流程,而进一步改善资源的利用率及提高系统的吞吐量。

方便性和有效性是设计操作系统时最重要的两个目标。在过去的很长一段时间内，由于计算机系统非常昂贵，因而使其有效性比方便性更为重要。正因如此，现在的大多数操作系统其理论上都着重于如何提高计算机系统的资源利用率和系统的吞吐量问题。但是，近十年来在微机上所配置的 OS，则更重视其方便性。

3. 可扩充性

随着 VLSI 技术和计算机技术的迅速发展，计算机硬件和体系结构，也随之得到迅速发展，相应地，它们也对 OS 提出了更高的功能和性能要求。此外，计算机网络、特别是 Internet 的发展，也对 OS 提出了一系列更新的要求。因此，OS 必须具有很好的可扩充性，方能适应发展的要求。这就是说，OS 应采用层次化结构，以便于增加新的功能层次和模块，并能修改老的功能层次和模块。

4. 开放性

80 年代以来，由于计算机网络的发展，尤其是 LAN 的迅速发展，使计算机操作系统的应用环境，已逐步由单机环境转向网络环境。为使来自不同厂家的计算机和设备能通过网络加以集成化，并能正确、有效地协同工作，实现应用的可移植性和互操作性，必须具有统一的开放环境，进而要求 OS 具有开放性。

开放性是指系统能遵循世界标准规范，特别是遵循开放系统互连 OSI 国际标准。凡遵循国际标准所开发的硬件和软件，能彼此兼容，可方便地实现互连。开放性已成为 90 年代计算机技术的核心问题，也是一个新推出的系统或软件能否被应用的重要因素。

1.1.2 操作系统的作用

可以从不同的观点(角度)来观察 OS 的作用。从一般用户的观点，可把 OS 看作是用户与计算机硬件系统之间的接口；从资源管理观点上看，则可把 OS 视为计算机系统资源的管理者。

1. OS 作为用户与计算机硬件系统之间的接口

OS 作为用户与计算机硬件系统之间接口的含义是：OS 处于用户与计算机硬件系统之间，用户通过 OS 来使用计算机系统。或者说，用户在 OS 帮助下，能够方便、快捷、安全、可靠地操纵计算机硬件和运行自己的程序。应注意，OS 是一个系统软件，因而这种接口是软件接口。图 1-1 是 OS 作为接口的示意图。由图可看出，用户可通过以下三种方式使用计算机。

(1) 命令方式。这是指由 OS 提供了一组联机命令(语言)，用户可通过键盘输入有关命令，来直接操纵计算机系统。

(2) 系统调用方式。OS 提供了一组系统调用，用户可在自己的应用程序中通过相应的系统调用，来操纵计算机。

(3) 图形、窗口方式。用户通过屏幕上的窗口和图标来操纵计算机系统和运行自己的

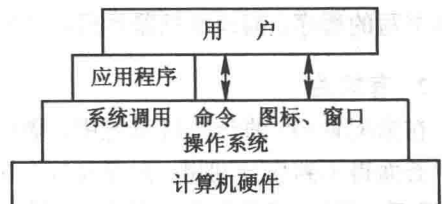


图 1-1 OS 作为接口的示意图

程序。

2. OS 作为计算机系统资源的管理者

在一个计算机系统中,通常都含有各种各样的硬件和软件资源。归纳起来可将资源分为四类:处理器、存储器、I/O 设备以及信息(数据和程序)。相应地,OS 的主要功能也是针对这四类资源进行有效的管理,即:处理机管理,用于分配和控制处理机;存储器管理,主要负责内存的分配与回收;I/O 设备管理,负责 I/O 设备的分配与操纵;文件管理,负责文件的存取、共享和保护。可见,OS 确是计算机系统资源的管理者。事实上,当今世界上广为流行的一个关于 OS 作用的观点,正是把 OS 作为计算机系统的资源管理者。

3. OS 用作扩充机器

对于一台完全无软件的计算机系统(即裸机),即使其功能再强,也必定是难于使用的。如果我们在裸机上覆盖上一层 I/O 设备管理软件,用户便可利用它所提供的 I/O 命令,来进行数据输入和打印输出。此时用户所看到的机器,将是一台比裸机功能更强、使用更方便的机器。通常把覆盖了软件的机器称为扩充机器或虚机器。如果我们又在第一层软件上再覆盖上一层文件管理软件,则用户可利用该软件提供的文件存取命令,来进行文件的存取。此时,用户所看到的是一台功能更强的虚机器。如果我們又在文件管理软件上再覆盖一层面向用户的窗口软件,则用户便可在窗口环境下方便地使用计算机,形成一台功能更强的虚机器。

由此可知,每当人们在计算机系统上覆盖上一层软件后,系统功能便增强一级。由于 OS 自身包含了若干个层次,因此当在裸机上覆盖上 OS 后,便可获得一台功能显著增强、使用极为方便的多层扩充机器或多层虚机器。

1. 1. 3 推动操作系统发展的主要动力

在出现 OS 后的短短 40 年中,操作系统取得了重大的发展,其主要动力可归结为下述四个方面。

1. 不断提高计算机资源利用率

在计算机发展的初期,计算机系统特别昂贵,人们必须千方百计地提高计算机系统中各种资源的利用率,这就成为最初发展的动力。由此形成了能自动地对一批作业进行处理的批处理系统。

2. 方便用户

当资源利用率不高的问题得到基本解决后,用户在上机、调试程序时的不方便性便成为主要矛盾。于是人们又想方设法改善用户上机、调试程序时的条件,这又成为继续推动 OS 发展的主要(推动)因素。随之便形成了允许进行人机交互的分时系统,或称为多用户系统。

3. 器件的不断更新换代

计算机器件的不断更新,使得计算机的性能不断提高、规模急剧扩大,从而推动了 OS 的功能和性能也迅速增强和提高。例如,当微机由 8 位发展到 16 位,进而又发展到 32 位

时,相应的微机 OS 也就由 8 位微机 OS 发展到 16 位,进而又发展到 32 位微机 OS,此时相应 OS 的功能和性能,也都有显著的增强和提高。

4. 计算机体系结构的不断发展

计算机体系结构的发展,也不断推动着 OS 的发展并产生新的操作系统类型。例如,当计算机由单处理机系统发展为多处理机系统时,相应地,操作系统也就由单处理机 OS 发展为多处理机 OS。又如,当计算机继续发展而出现了计算机网络后,相应地,也才有了网络操作系统。

1.2 操作系统的发展过程

OS 的形成迄今已有 40 年的时间。在 20 世纪 50 年代中期出现了第一个简单的批处理操作系统。到 20 世纪 60 年代中期产生了多道程序批处理系统;不久又出现了基于多道程序的分时系统。20 世纪 80 年代至 90 年代是微型机、多处理机和计算机网络大发展的年代,同时也是微机 OS、多处理机 OS 和网络 OS 的形成和大发展的年代。

1.2.1 无操作系统的计算机系统

1. 人工操作方式

从第一台计算机诞生(1945 年)到 50 年代中期的计算机,属于第一代,这时还未出现 OS。这时的计算机操作是由用户(即程序员)采用人工操作方式直接使用计算机硬件系统,即由程序员将事先已穿孔(对应于程序和数据)的纸带(或卡片)装入纸带输入机(或卡片输入机),再启动它们将程序和数据输入计算机,然后启动计算机运行。当程序运行完毕并取走计算结果后,才让下一个用户上机。这种人工操作方式有以下两方面的缺点:

(1) 用户独占全机。此时,计算机及其全部资源只能由上机用户独占;

(2) CPU 等待人工操作。当用户进行装带(卡)、卸带(卡)等人工操作时,CPU 及内存等资源是空闲的。

可见,人工操作方式严重降低了计算机资源的利用率,此即所谓的人机矛盾。随着 CPU 速度的提高和系统规模的扩大,人机矛盾变得日趋严重。此外,随着 CPU 速度的迅速提高而 I/O 设备的速度却提高缓慢,又使 CPU 与 I/O 设备之间速度不匹配的矛盾更加突出。为了缓和此矛盾,曾先后出现了通道技术、缓冲技术,但都未能很好地解决上述矛盾,直至后来又引入了脱机输入/输出技术,才获得了较为令人满意的结果。

2. 脱机输入/输出(Off-Line I/O)方式

为了解决人机矛盾及 CPU 和 I/O 设备之间速度不匹配的矛盾,20 世纪 50 年代末出现了脱机输入/输出技术。该技术是指事先将装有用户程序 and 数据的纸带(或卡片)装入纸带输入机(或卡片机),在一台外围机的控制下,把纸带(卡片)上的数据(程序)输入到磁带上。当 CPU 需要这些程序和数据时,再从磁带上高速地调入内存。

类似地,当 CPU 需要输出时,可由 CPU 直接高速地把数据从内存送到磁带上,然后再在另一台外围机的控制下,将磁带上的结果通过相应的输出设备输出。图 1-2 示出了

脱机输入/输出过程。由于程序和数据的输入和输出都是在外围机的控制下完成的，或者说，它们是在脱离主机的情况下进行的，故称为脱机输入/输出方式；反之，在主机的直接控制下进行输入/输出的方式称为联机输入/输出(On-Line I/O)方式。这种脱机 I/O 方式的主要优点如下：

(1) 减少了 CPU 的空闲时间。装带(卡)、卸带(卡)以及将数据从低速 I/O 设备送到高速磁带(或盘)上，都是在脱机情况下进行的，都不占用主机时间，从而有效减少了 CPU 的空闲时间，缓和了人机矛盾。

(2) 提高 I/O 速度。当 CPU 在运行中需要数据时，是直接从高速的磁带或磁盘上将数据调入内存的，不再是从低速 I/O 设备上输入，从而大大缓和了 CPU 和 I/O 设备速度不匹配的矛盾，进一步减少了 CPU 的空闲时间。

1.2.2 单道批处理系统

1. 单道批处理系统(Simple Batch Processing System)的处理过程

早期的计算机系统非常昂贵，为了能充分地利用它，应尽量让该系统连续运行，以减少空闲时间。为此，通常是把一批作业以脱机方式输入到磁带上，并在系统中配上监督程序(Monitor)，在它的控制下使这批作业能一个接一个地连续处理。其自动处理过程是：首先，由监督程序将磁带上的第一个作业装入内存，并把运行控制权交给该作业。当该作业处理完成时，又把控制权交还给监督程序，再由监督程序把磁带(盘)上的第二个作业调入内存。计算机系统就这样自动地一个作业一个作业地进行处理，直至磁带(盘)上的所有作业全部完成，这样便形成了早期的批处理系统。由于系统对作业的处理都是成批地进行的，且在内存中始终只保持一道作业，故称为单道批处理系统(Simple Batch System)。图 1-3 示出了单道批处理系统的处理流程。

由上所述不难看出，单道批处理系统是在解决人机矛盾和 CPU 与 I/O 设备速度不匹配的矛盾的过程中形成的。换言之，批处理系统旨在提高系统资源的利用率和系统吞吐量。但这种单道批处理系统仍然不能很好地利用系统资源，故现已很少使用。

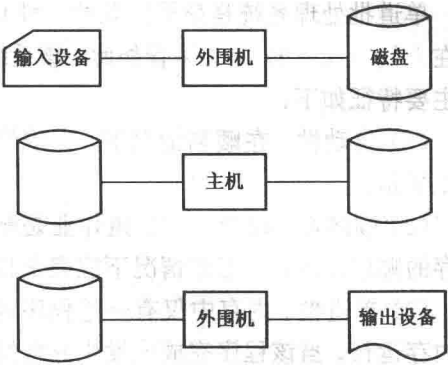


图 1-2 脱机 I/O 示意图

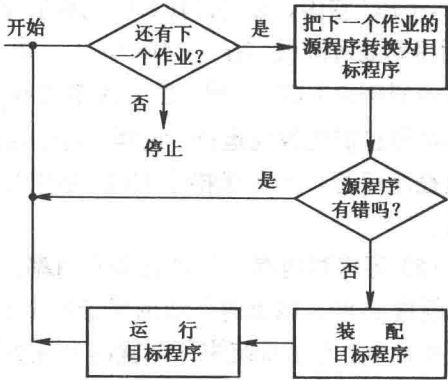


图 1-3 单道批处理系统的处理流程

2. 单道批处理系统的特征

单道批处理系统是最早出现的一种 OS，严格地说，它只能算作是 OS 的前身而并非是目前人们所理解的 OS。尽管如此，该系统比起人工操作方式的系统已有很大进步。该系统的主要特征如下：

(1) 自动性。在顺利情况下，在磁带上的一批作业能自动地逐个地依次运行，而无须人工干预。

(2) 顺序性。磁带上的各道作业是顺序地进入内存，各道作业的完成顺序与它们进入内存的顺序之间，在正常情况下应完全相同，亦即先调入内存的作业先完成。

(3) 单道性。内存中仅有一道程序运行，即监督程序每次从磁带上只调入一道程序进入内存运行，当该程序完成或发生异常情况时，才换入其后继程序进入内存运行。

1.2.3 多道批处理系统

1. 多道程序设计的基本概念

在单道批处理系统中，内存中仅有一道作业，它无法充分利用系统中的所有资源，致使系统性能较差。为了进一步提高资源的利用率和系统吞吐量，在 60 年代中期又引入了多道程序设计技术，由此而形成了多道批处理系统 (Multiprogrammed Batch Processing System)。在该系统中，用户所提交的作业都先存放在外存上并排成一个队列，称为“后备队列”；然后，由作业调度程序按一定的算法从后备队列中选择若干个作业调入内存，使它们共享 CPU 和系统中的各种资源。具体地说，在 OS 中引入多道程序设计技术可带来以下好处：

(1) 提高 CPU 的利用率。当内存中仅有一道程序时，每逢该程序在运行中发出 I/O 请求后，CPU 空闲，必须在其 I/O 完成后才继续运行；尤其因 I/O 设备的低速性，更使 CPU 的利用率显著降低。图 1-4(a) 示出了单道程序的运行情况，从图可以看出：在 $t_2 \sim t_3$ 、 $t_6 \sim t_7$ 时间间隔内 CPU 空闲。在引入多道程序设计技术后，由于同时在内存中装有若干道程序，并使它们交替地运行，这样，当正在运行的程序因 I/O 而暂停执行时，系统可调度另一道程序运行，从而保持了 CPU 处于忙碌状态。图 1-4(b) 示出了四道程序时的运行情况。

(2) 可提高内存和 I/O 设备利用率。为了能运行较大的作业，通常内存都具有较大容量，但由于 80% 以上的作业都属于中小型，因此在单道程序环境下，也必定造成内存的浪费。类似地，对于系统中所配置的多种类型的 I/O 设备，在单道程序环境下也不能充分利用。如果允许在内存中装入多道程序，并允许它们并发执行，则无疑会大大提高内存和 I/O 设备的利用率。

(3) 增加系统吞吐量。在保持 CPU、I/O 设备不断忙碌的同时，也必然会大幅度地提高系统的吞吐量，从而降低作业加工所需的费用。

2. 多道批处理系统的特征

在批处理系统中引入多道程序设计技术后，会使系统具有以下特征：

(1) 多道性。在内存中可同时驻留多道程序，并允许它们并发执行，从而有效地提高

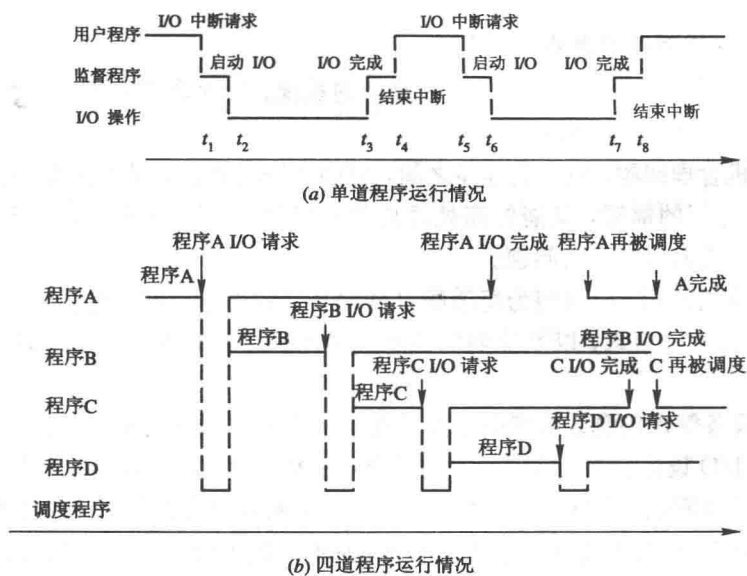


图 1-4 单道和多道程序运行情况

了资源利用率和系统吞吐量。

(2) 无序性。多个作业完成的先后顺序与它们进入内存的顺序之间，并无严格的对应关系，即先进入内存的作业可能较后完成甚至是最后完成；而后进入内存的作业又可能先完成。

(3) 调度性。作业从提交给系统开始直至完成，需要经过以下两次调度：首先是作业调度，这是指按一定的作业调度算法，从外存的后备作业队列中，选择若干个作业调入内存；其次是进程调度，指按一定的进程调度算法，从已在内存的作业中选择一个作业，将处理机分配给它，使之执行。

3. 多道批处理系统的优缺点

虽然早在 60 年代就已出现了多道批处理系统，但至今它仍是三大基本操作系统类型之一。在大多数大、中、小型机中都配置了它，说明它具有其它类型 OS 所不具有的优点。多道批处理系统的主要优缺点如下：

(1) 资源利用率高。由于在内存中驻留了多道程序，它们共享资源，可保持资源处于忙碌状态，从而使各种资源得以充分利用。

(2) 系统吞吐量大。系统吞吐量是指系统在单位时间内所完成的总工作量。能提高系统吞吐量的主要原因可归结为：第一，CPU 和其它资源保持“忙碌”状态；第二，仅当作业完成时或运行不下去时才进行切换，系统开销小。

(3) 平均周转时间长。作业的周转时间是指从作业进入系统开始，直至其完成并退出系统为止所经历的时间。在批处理系统中，由于作业要排队，依次进行处理，因而作业的周转时间较长，通常需几个小时，甚至几天。

(4) 无交互能力。用户一旦把作业提交给系统后，直至作业完成，用户都不能与自己的作业进行交互，这对修改和调试程序是极不方便的。

4. 多道批处理系统需要解决的问题

多道批处理系统是一种有效但又十分复杂的系统。为使系统中的多道程序间能协调地运行，必须解决下述一系列问题：

(1) 处理机管理问题。在多道程序之间，应如何分配被它们共享的处理机，使 CPU 既能满足各程序运行的需要，又能提高处理机的利用率，以及一旦把处理机分配给某程序后，又应在何时收回等一系列问题。

(2) 内存管理问题。应如何为每道程序分配必要的内存空间，使它们“各得其所”且不致因相互重叠而丢失信息，以及应如何防止因某道程序出现异常情况而破坏其它程序等问题。

(3) I/O 设备管理问题。系统中可能具有多种类型的 I/O 设备供多道程序所共享，应如何分配这些 I/O 设备，如何做到既方便用户对设备的使用，又能提高设备的利用率。

(4) 文件管理问题。在现代计算机系统中，通常都存放着大量的程序和数据(以文件形式存在)，应如何组织这些程序和数据，才能使它们既便于用户使用，又能保证数据的安全性和一致性。

(5) 作业管理问题。对于系统中的各种应用程序，其中有的属于计算型，即以计算为主的程序；有的属于 I/O 型，即以 I/O 为主的程序；又有些作业既重要又紧迫；而有的作业则要求系统能及时响应，这时应如何组织这些作业。

为此，应在计算机系统中增加一组软件，用以对上述问题进行妥善、有效地处理。这组软件应包括：能控制和管理四大资源的软件、合理地对各类作业进行调度的软件，以及方便用户使用计算机的软件。正是这样一组软件构成了操作系统。据此，我们可把操作系统定义为：操作系统是一组控制和管理计算机硬件和软件资源，合理地对各类作业进行调度，以及方便用户使用的程序的集合。

1.2.4 分时系统

1. 分时系统(Time-Sharing System)的产生

如果说，推动多道批处理系统形成和发展的主要动力，是提高资源利用率和系统吞吐量，那么，推动分时系统形成和发展的主要动力，则是用户的需求。或者说，分时系统是为了满足用户需求所形成的一种新型 OS。它与多道批处理系统之间，有着截然不同的性能差别。用户的需求具体表现在以下几个方面：

(1) 人一机交互。每当程序员写好一个新程序时，都需要上机进行调试。由于新编程序难免有些错误或不当之处需要修改，因而希望能像早期使用计算机时一样地对它进行直接控制，并能以边运行边修改的方式，对程序中的错误进行修改，亦即，希望能进行人一机交互。

(2) 共享主机。在 60 年代计算机非常昂贵，不可能像现在这样每人独占一台微机，而只能是由多个用户共享一台计算机，但用户在使用机器时应能够像自己独占计算机一样，不仅可以随时与计算机交互，而且应感觉不到其他用户也在使用该计算机。

(3) 便于用户上机。在多道批处理系统中，用户上机前必须把自己的作业邮寄或亲自送到机房。这对于用户尤其是远地用户来说是十分不方便的。用户希望能通过自己的终端