第 2 章　操作系统基础知识

大纲要求：

- 操作系统基础知识，包括操作系统的类型、功能。
- 处理机管理，包括进程的基本概念、进程的控制、进程间的通信、进程调度、信号量与 P、V 操作、高级通信原语、死锁和线程的基本概念等。
- 存储管理，包括主存保护、分区存储管理、分页存储管理、分段存储管理和虚存管理等。
- 设备管理，包括设备的类型、设备分配有关的调度算法、通道、DMA 与缓冲技术、假脱机和磁盘调度等。
- 文件管理，包括文件与文件系统的概念、文件的结构和组织等。
- 作业管理，包括作业管理的基本概念、作业调度及调度算法、评价作业调度算法应用的目的及对系统性能的影响。
- 图形用户界面和操作方法。

2.1　操作系统概述

2.1.1　考点辅导

1. 操作系统的定义

操作系统是计算机系统中最重要的系统软件，其他所有的软件都是建立在操作系统之上的，并在操作系统的统一管理和支持下运行。任何用户都是通过操作系统使用计算机的。

操作系统的定义为：操作系统(Operating System，OS)是计算机系统中的一个系统软件，它管理和控制计算机系统的硬件和软件资源，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功能强大、使用方便的工作环境，从而在计算机与用户之间起到接口的作用。

操作系统的主要任务是使硬件所提供的能力得到充分的利用，支持应用软件的运行并提供相应的服务。由于操作系统在计算机系统中占据着重要地位，所以它已经成为现代计算机系统中一个必不可少的关键组成部分。

2. 操作系统的作用

(1)　通过资源管理，提高工作效率。

操作系统的主要作用就是通过 CPU 管理、存储管理、设备管理和文件管理，对各种资源进行合理的分配，改善资源的共享和利用程度，最大限度地发挥计算机系统的工作效率，提高计算机系统的“吞吐量”(即系统在单位时间内处理工作的能力)。

(2)　改善人机界面，提供友好的工作环境。

操作系统既是计算机硬件和各种软件之间的接口，又是用户与计算机之间的接口。安

装操作系统后，用户面对的不再是笨拙的裸机、由 0 和 1 组成的代码及一些难懂的机器指令，而是操作便利、服务周到的操作系统，操作系统明显地改善了用户界面，提高了用户的工作效率。

3. 操作系统的特征

操作系统主要有并发性(concurrency)、共享性(sharing)、虚拟性(virtual)和不确定性(non-determinacy)4 个基本特征。

1)　并发性

并发性是指在计算机系统中存在着许多同时进行的活动。对计算机系统而言，并发是指宏观上看系统内有多道程序同时运行，微观上看实际上是串行运行。

2)　共享性

共享性是指系统中各个并发活动要共享计算机系统中的各种软、硬件资源，因此操作系统必须解决在多道程序间合理地分配和使用资源。

3)　虚拟性

虚拟性是操作系统中的重要特征，所谓虚拟是指把物理上的一台设备变成逻辑上的多台设备。例如我们将在本章后面介绍的假脱机(spooling)技术，就是利用快速、大容量、可共享的磁盘作为中介，模拟多个非共享的低速的输入输出设备，这样的设备称为虚拟设备。

4)　不确定性

通常一个程序的初始条件相同时，无论何时运行，结果都应该相同。但由于操作系统并发执行系统内的各种进程，与这些进程有关的事件如：从外部设备来的中断、输入输出请求、各种运行故障、发生的时间等都不可预测，如果处理不当，将导致系统出错，这种不确定性所带来的错误是很难查找的。

4. 操作系统的功能

1)　处理机管理

处理机是计算机系统的心脏，在单用户系统或单道系统中，处理机为一个用户或一个作业服务，其管理简单，但资源利用率低。为提高系统资源的利用率，引入了多道程序技术，即多个程序(作业)同时运行。在多道程序或多用户的情况下，要组织多个作业同时运行，对多个用户进行响应，就需要解决对处理机的分配、调度和资源回收等问题。处理机管理负责解决如何把 CPU 时间合理地、动态地分配给程序运行的基本单位——进程，使处理机得到充分的利用。许多操作系统是以作业和进程的方式进行管理的，实现作业和进程的调度，分配处理机，控制作业和进程的执行。现代的操作系统还引入了线程(thread)作为分配处理机的基本单位。

由于操作系统对处理机的管理策略不同，其提供的作业处理方式也就不同，如批处理方式、分时处理方式和实时处理方式，从而呈现在用户面前的就有不同的操作系统。在操作系统中，最重要的资源是处理机，最重要的管理是处理机管理。

2)　存储管理

计算机系统中，存储器(一般称为主存或内存)是运行程序和存放工作数据的部件，存储管理的工作主要是对内存储器进行分配、扩充和保护。

- 内存分配：在内存中除了操作系统和其他系统软件外，还要有一个或多个用户程序。如何分配内存，以保证系统及各用户程序的存储区互相不冲突，是内存分配

所要解决的问题。

- 存储保护：系统中有多个程序在运行，如何保证一道程序在执行过程中不会有意或无意地破坏另一道程序？如何保证用户程序不会破坏系统程序？这些就是存储保护问题。
- 内存扩充：当用户作业所需要的内存量超过计算机系统所提供的内存容量时，如何把内部存储器和外部存储器结合起来管理，为用户提供一个容量比实际内存大得多的虚拟存储器，使这个虚拟存储器和内存一样方便使用，这就需要使用内存扩充。

存储器是计算机系统中最重要的资源之一，因为任何程序和数据，以及各种控制用的数据结构，都必须占有一定的存储空间，因此，存储管理的目的就是尽量提高内存的使用效率。存储管理的好坏直接影响着系统性能。

3) 设备管理

现代计算机系统常常配置很多种类的输入输出设备，它们的输入输出速度差别很大。计算机系统常常采用通道、控制器和设备 3 级控制方法管理这些设备。设备管理的任务就是监视这些资源的使用情况，根据一定的分配策略，把通道、控制器和设备分配给请求输入输出操作的程序，并启动设备完成所需的操作。为了发挥设备和处理机的并行工作能力，常常采用缓冲技术和虚拟技术。

由于输入/输出设备种类很多，使用方法各不相同，因此，设备管理应为用户提供一个良好的界面，使具体的设备特性透明化，以便用户能方便、灵活地使用这些设备。

4) 文件管理(信息管理)

文件管理是对系统软件资源的管理。对用户来说，文件系统是操作系统中最直观的部分。我们把程序和数据统称为信息或文件。当一个文件暂时不用时，就把它放到外部存储器(如磁盘、磁带和光盘等)上保存起来。对这些文件如果不能很好地进行管理，就会引起混乱，甚至使其遭受破坏。这就是文件管理需要解决的问题。

文件管理的功能包括：建立、修改和删除文件；按文件名进行访问；决定文件信息的存放位置、存放形式及存取权限；管理文件间的联系及提供对文件的共享、保护和保密等，允许多个用户协同工作又不引起混乱。

5) 用户接口(作业管理)

上述 4 项功能是操作系统对软、硬件资源的管理。除此以外，操作系统也必须为用户提供一个友好的用户接口——命令接口和图形接口。一般来说，用户通过两种命令接口请求操作系统的服务。一种接口是作业一级的接口，即提供一组控制操作命令，如 UNIX 的 Shell 命令语言或作业控制语言(JCL)让用户组织和控制自己作业的运行。作业控制又分成两类：联机控制和脱机控制。另一种用户接口是程序一级的接口(编程接口)，即提供一组广义指令(或称系统调用、程序请求)供用户程序和其他系统程序调用。当这些程序要求进行数据传输、文件操作或有其他资源要求时，通过这些广义指令向操作系统提出申请，并由操作系统代为完成。

操作系统对计算机的资源进行全面管理，它的基本特征是多任务并行和多用户资源共享。多任务并行是指操作系统可以支持用户同时提交多项任务，同时工作；资源共享是指系统中的资源为多个用户共同使用。

5. 操作系统的类型

根据操作系统的使用环境和对作业的处理方式来划分，操作系统主要有以下几种基本类型。

1)　批处理操作系统

在批处理操作系统(Batch Processing Operating System)中，系统操作员将作业成批提交，由操作系统选择作业调入内存加以处理，最后由操作人员将运行结果交给用户。

批处理系统的特点：一是“多道”，指系统内可同时容纳多个作业；二是“成批”，指系统成批自动运行多个作业。批处理系统的目标是提高资源利用率和实现作业执行的自动化。

批处理操作系统分为单道批处理和多道批处理两种。

(1)　单道批处理操作系统：一次可提交多个作业，而不是单个作业。当一个作业运行结束后，随即自动调入同批的下一个作业运行，从而节省了作业之间的人工操作时间，提高了资源的利用率。早期单道批处理系统解决了作业自动转换问题，从而减少了作业建立和人工操作的时间。单道批处理存在的主要问题是：CPU 和 I/O 设备使用忙闲不均(取决于当前作业的特性)，对以计算为主的作业，外设空闲；对以 I/O 为主的作业，CPU 空闲。

(2)　多道批处理操作系统：正是为了解决单道批处理操作系统存在的问题而产生了多道批处理操作系统。它除了保持作业自动转换的功能外，还能支持同一批中的多道用户程序在一个 CPU 上同时运行。作业调度程序从后备作业中选取多个作业进入主存，在任意一个时刻，每当运行中的一个作业因输入输出操作而需要调用外部设备时，就把 CPU 及时交给另一道等待运行的作业，从而将主机与外部设备的工作方式由串行改变为并行，进一步避免了因主机等待外设完成任务而白白浪费宝贵的 CPU 时间的情况。

2)　分时操作系统

分时操作系统(Time Share Operating System)是指一台计算机连接多个终端，系统把 CPU 时间分为若干时间片，采用时间片轮转的方式处理用户的服务请求，对每个用户能保证及时响应，并提供交互会话能力。

分时操作系统具有下述特点。

- 多用户同时性：允许多个用户同时联机使用计算机。
- 交互性：每个用户可随时通过终端向系统提出服务请求，系统也可随时通过终端响应用户，从而加快了调试过程。
- 独立性：由于采用时间片轮转方式使一台计算机同时为多个用户服务，对于每个用户的操作命令又能快速响应，因此，用户彼此之间都感觉不到别人也在使用同一台计算机，如同自己独占计算机一样。
- 及时性：系统对用户的响应非常及时，不会让用户等待执行命令的处理时间过长。

分时操作系统的主要目标是保证用户响应的及时性。通常，计算机系统中往往同时采用批处理和分时处理方式来为用户服务，即时间要求不强的作业放入“后台”(批处理)处理，需频繁交互的作业放在“前台”(分时)处理。

3)　实时操作系统

实时操作系统(Real Time Operating System)是随着计算机应用于实时控制和实时信息处理而发展起来的。实时操作系统是指系统能够及时响应事件，并以足够快的速度完成对该

事件的处理。实时操作系统包括实时控制系统和实时处理系统。实时控制是指生产过程控制(如炼钢、电力生产和数控机床)及武器控制等；实时处理是指实验数据采集和订票系统等。

实时操作系统的主要特点是及时性和高可靠性。

4) 网络操作系统

网络操作系统(Network Operating System)开发是在原来各自计算机操作系统的基础上，按照网络体系结构的协议、标准进行开发的，包括计算机网络管理、通信、资源共享、系统安全和多种网络应用服务等。其功能主要包括高效、可靠的网络通信；对网络中共享资源的有效管理；电子邮件、文件传输、共享硬盘、打印机等服务；网络安全管理；互操作能力。

5) 分布式操作系统

分布式操作系统(Distributed Operating System)与网络操作系统都是工作在一个由多台计算机组成的系统中，这些计算机之间可以通过一些传输设备进行通信和系统资源共享。分布式操作系统更倾向于任务的协同执行，并且各系统之间无主次之分，系统之间也无须采用标准的通信协议进行通信。分布式操作系统基本上废弃(或改造)了各单机的操作系统，整个网络设有单一的操作系统，由这个操作系统负责整个系统的资源分配和调度，为用户提供统一的界面。用户在使用分布式操作系统时不需要像使用网络操作系统那样，指明资源在哪台计算机上，因此分布式操作系统的透明性、稳固性、统一性及系统效率都比网络操作系统要强，但实现起来难度也大。分布式操作系统对于多机合作和系统重构、稳固性和容错能力有更高的要求，希望分布式操作系统有更短的响应时间、更大的吞吐量和更高的可靠性。

分布式操作系统与网络操作系统最大的差别是：网络操作系统的用户必须知道网址，而分布式系统用户则不必知道计算机的确切地址；分布式操作系统负责全系统的资源分配，通常能很好地隐藏系统内部的实现细节，如对象的物理位置、并发控制、系统故障处理等对用户都是透明的。

6) 微机操作系统

微机操作系统(Microcomputer Operating System)是指配置在微型计算机上的操作系统。常用的微机操作系统有 DOS、Windows、OS/2、UNIX 和 Linux 等。其中，Microsoft 公司开发的单用户单任务操作系统 DOS 是首先在 IBM-PC 上使用的微机操作系统。MS-DOS 操作系统是16位微机单用户单任务操作系统的标准。多任务操作系统Windows 98/NT/2000/XP是 Microsoft 公司开发的一系列图形用户界面的多任务、多线程的操作系统。

7) 嵌入式操作系统

嵌入式操作系统(Embedded Operating System)运行在嵌入式智能芯片环境中，对整个智能芯片及其控制的各种部件和装置等资源进行统一协调、处理、指挥和控制。

6. 研究操作系统的观点

研究和分析操作系统，可以从资源管理观点和虚拟机观点出发。

1) 资源管理观点

引入操作系统是为了合理地组织计算机的工作流程，管理和分配计算机系统硬件和软件资源，使资源能为多个用户共享。因此，操作系统是计算机资源的管理者。

这里的资源是指计算机系统进行数值计算和数据处理所需的物质基础，通常分为系统

硬件资源和软件资源。硬件资源是组成计算机和计算机操作所需的物理实体，它们是看得见摸得着的设备，如处理机、存储器及输入/输出设备(键盘、显示器、打印机和磁盘等)。软件资源是依赖于一定的物理实体才能被人们所感知的一类资源，如程序和数据等，它们可经显示器或打印机等设备展现给用户。操作系统是控制和管理计算机系统资源的一组程序，其工作是当用户程序和其他程序争用这些资源时提供有序的和可控的分配。

我们通常将操作系统分为 CPU 管理、存储管理、设备管理、文件管理、用户与操作系统接口 5 个主要部分。主要研究资源的使用情况、资源的分配策略及分配和回收资源。

2)　虚拟机观点

从服务用户的机器扩充的观点来看，操作系统为用户使用计算机提供了许多服务功能和良好的工作环境。用户不再直接使用硬件机器(称为裸机)，而是通过操作系统来控制和使用计算机，从而把计算机扩充为功能更强、使用更方便的计算机系统(称为虚拟计算机)。操作系统的全部功能，如系统调用、命令、作业控制语言等，称为操作系统虚拟机。

虚拟机观点从功能分解的角度出发，考虑操作系统的结构，将操作系统分成若干层次，每一层次完成特定的功能，从而构成一个虚拟机，并为上一层次提供支持，构成它的运行环境。这样，通过逐个层次的功能扩充最终完成操作系统虚拟机，从而向用户提供各种服务，完成用户的各项任务。

2.1.2　典型例题分析

例：嵌入式操作系统的主要特点是微型化、__(25)__。(2013 年下半年试题 25)

A. 可定制、实时性、高可靠性和易移植性

B. 可定制、实时性和易移植性，但可靠性差

C. 实时性、可靠性和易移植性，但不可定制

D. 可定制、实时性和可靠性，但不易移植

分析：本题考查操作系统的基础知识。

嵌入式操作系统运行在嵌入式智能芯片环境中，对整个智能芯片以及它所操作、控制的各种部件装置等资源进行统一协调、处理、指挥和控制。其主要特点：

①微型化。从性能和成本角度考虑，希望占用资源和系统代码量少，如内存少、字长短、运行速度有限、能源少(用微小型电池)。

②可定制。从减少成本和缩短研发周期考虑，要求嵌入式操作系统能运行在不同的微处理器平台上，能针对硬件变化进行结构与功能上的配置，以满足不同应用需要。

③实时性。嵌入式操作系统主要应用于过程控制、数据采集、传输通信、多媒体信息及关键要害领域需要迅速响应的场合，所以对实时性要求高。

④可靠性。系统构件、模块和体系结构必须达到应有的可靠性，对关键要害应用还要提供容错和防故障措施。

⑤易移植性。为了提高系统的易移植性，通常采用硬件抽象层(Hardware Abstraction Level，HAL)和板级支持包(Board Support Package，BSP)的底层设计技术。

答案：A

2.1.3 同步练习

_____支持网络系统功能，并具有透明性。

A. 批处理操作系统　　B. 分时操作系统

C. 实时操作系统　　D. 分布式操作系统

2.1.4 同步练习参考答案

D

2.2 进 程 管 理

2.2.1 考点辅导

2.2.1.1 基本概念

在计算机系统上运行的程序是指令的集合，每一个程序完成特定的任务。在只允许一个程序运行的系统(称为单道系统)中，这个程序独占系统资源，而系统按程序的指令顺序运行，程序的顺序执行有两个基本特征：程序的封闭性和程序的可再现性。

- 封闭性：指程序运行时独占系统资源，只有程序本身能改变系统的状态。
- 可再现性：指程序运行不受外部因素的影响，只要初始条件相同，运行结果就相同。

多道程序系统让多个程序在系统中轮流运行，当一个程序不用处理机时，另一个程序就使用。也就是说，处理机在程序间来回切换，从而获得宏观上的并行(微观上的串行)，以提高处理机的利用率。这种切换，通常是由中断引起的。由于中断以不可预测的次序发生，即程序的指令执行序列也以不可预测的次序前进，这样就会产生操作系统的另一个特性——不确定性。即在多道程序系统中，顺序程序的封闭性和可再现性消失了，需要采用一个新的概念——进程来描述程序的执行。进程是运行中的程序，是系统进行资源分配和调度的独立单位。

1. 进程及其组成

进程是一个程序关于某个数据集的一次运行。进程是一个动态的概念，而程序是静态的概念，是指令的集合。因此，进程具有动态性和并发性。

进程通常由程序、数据和进程控制块(PCB)组成。程序是进程运行所对应的运行代码，一个进程对应于一个程序，一个程序可以同时对应于多个进程，代码在运行过程中不会被改变的程序，常称为纯码程序或可重入程序，这类程序是可共享的程序。

进程控制块是进程动态特性的集中反映，也是进程存在的唯一标志。在操作系统中，进程是进行系统资源分配、调度和管理的最小单位。现代操作系统中还引入了线程，线程是比进程更小的、能独立运行的基本单位，在引入线程的操作系统中，线程是进程中的一个实体，是 CPU 调度和分派的基本单位，是处理机分配的最小单位。

2. 进程的状态及其转换

在多道系统中，进程的运行是走走停停的，在处理机上的交替运行，使它的运行状态不断变化。进程的状态主要有三态模型和五态模型。三态模型中最基本的状态有 3 种：运行、就绪和阻塞。

- 运行(running)：正占用处理机。
- 就绪(ready)：只要获得处理机即可运行。
- 阻塞(blocked)：也称等待或挂起状态，正等待某个事件(如 I/O 完成)的发生。

在进程运行的过程中，由于自身进展情况及外界环境的变化，这 3 种基本状态可以在一定的条件下相互转换，进程的状态及转换如图 2-1 所示。

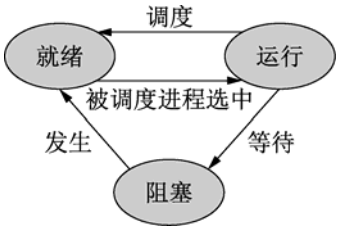

图 2-1　进程状态及其转换图

五态模型比三态模型更加复杂，在三态模型的基础上增加了新建态和终止态。新建态对应于进程刚刚被创建还没有被提交时的状态，此时应在等待系统完成创建进程的所有必要信息。创建进程时分两个阶段：第一个阶段为一个新进程创建必要的管理信息；第二个阶段让该进程进入就绪状态。有了新建态，操作系统往往因系统的性能和内存容量的限制推迟新建态进程的提交。进程的终止态也可分为两个阶段：第一个阶段等待操作系统进行善后处理；第二个阶段释放内存。

2.2.1.2　进程的控制

进程的控制就是对系统中所有进程从创建到消亡的全过程实施有效的控制。不仅要控制正在运行的进程，而且还要能创建新的进程，撤销已完成的进程。进程的控制机构是由操作系统内核实现的。通常将与硬件密切相关的模块放在紧挨硬件的软件层中，并使它们常驻内存，以便提高操作系统的运行效率，通常将这部分称为操作系统的内核，它为系统对进程进行控制和对存储器进行管理提供了有效的控制机制。

1. 支撑功能

1)　中断处理

操作系统的各种重要活动最终都依赖于中断。例如，各种类型的系统调用、键盘命令的输入、设备驱动及文件系统等都依赖于中断。通常内核只对中断进行“有限次处理”，然后转入有关进程继续处理。这不仅可以减少中断处理的时间，还可以提高程序的并发性。

2)　时钟管理

操作系统的许多活动要用到时钟管理。如在分时系统时间片调度算法中，当时间片用完时，由时钟管理产生一个中断信号，通知调度程序重新调度。在实时系统中的截止时间控制、批处理系统中的最长运行时间的控制等都要用到时钟管理。

3) 原语操作

内核在执行某些基本操作时，往往是通过原语操作来实现的。原语是由若干条机器指令构成的，用于完成特定功能的一段程序。原语在执行的过程中是不可分割的。进程控制原语主要有：创建原语、撤销原语、挂起原语、激活原语、阻塞原语以及唤醒原语。

2. 资源管理功能

资源管理功能包括：进程管理、存储器管理和设备管理。

2.2.1.3 进程间的通信

1. 同步与互斥

在操作系统中，多个进程并发执行，因此进程间必然存在资源共享和相互合作的问题。

1) 进程间的同步

一般情况下，一个进程相对于另一个进程的速度是不可预测的，也就是说，进程之间是异步运行的。为了成功地协同工作，有关进程在某些确定的点上应当保持同步：一个进程到达了这些点后，除非另一进程已经完成了某个活动，否则就停下来，等待该活动结束。

同步是指进程之间的一种协同工作关系，使这些进程相互合作，共同完成一项任务。进程间的直接相互作用构成进程的同步。同步机制应满足的基本要求是：有描述能力、可以实现、效率高、使用方便。

2) 进程间的互斥

在多道系统中，各进程可以共享各类资源，但有些资源一次只能供一个进程使用。这种资源称为临界资源，如打印机、公共变量、表格等。互斥是要保证临界资源在某一时刻只被一个进程访问。

3) 临界区管理的原则

临界区是进程中对临界资源实施操作的那段程序。对互斥临界区管理的原则是：有空即进、无空则等、有限等待、让权等待。

2. 信号量机制

信号量机制是一种有效的进程同步与互斥工具。目前主要有：整型信号量、记录型信号量、信号量集机制。

1) 整型信号量与 P 操作和 V 操作

信号量是一个整型变量，根据控制对象的不同被赋予不同的值。信号量分为两类：公用信号量，实现进程间的互斥，初值为 1 或资源的数目；私用信号量，实现进程间的同步，初值为 0 或某个正整数。

信号量 S 的物理意义：$S \geqslant 0$ 表示某资源的可用数，若 $S<0$，则其绝对值表示阻塞队列中等待该资源的进程数。

除了设置初值外，对信号量只能进行特殊的操作：P 操作和 V 操作。P 操作和 V 操作都是不可分割的原子动作，也称为原语，其中 P 操作表示申请一个资源，V 操作表示释放一个资源。

P 操作和 V 操作都是原语。利用信号量 S 的取值表示共享资源的使用情况。在使用时，把信号量 S 放在进程运行的环境中，赋予其不同的初值，并在其上实施 P 操作和 V 操作，以实现进程间的同步与互斥。

P 操作和 V 操作的定义如下。

P(*S*)：①*S*=*S*−1；②若 *S*<0，则该进程进入 *S* 信号量的队列中等待。

V(*S*)：①*S*=*S*+1；②若 *S*≤0，则释放 *S* 信号量队列上的一个等待进程，使之进入就绪队列。

当 *S*>0 时，表示还有资源可以分配；当 *S*<0 时，其绝对值表示信号量等待队列中进程的数目。每执行一次 P 操作，意味着要求分配一个资源；每执行一次 V 操作，就意味着释放一个资源。

2)　利用 P 操作和 V 操作实现进程的互斥

令信号量 mutex 的初值为 1，进入临界区时执行 P 操作，退出临界区时执行 V 操作，于是临界区就改写成下列形式的代码段：

```
P(mutex);
临界区
V(mutex);
```

由于 mutex 初值为 1，P、V 是原子操作，可以实现互斥。

3. 高级通信原语

P 操作和 V 操作是用来协调进程间关系的，编程较困难、效率低，而且没有信息交换，故常称为低级通信原语。交换的信息量多时要引入高级通信原语，进程高级通信的类型主要有如下几种。

(1)　共享存储系统：相互通信的进程共享某些数据结构或存储区，以实现进程之间的通信。

(2)　消息传递系统：进程间的数据交换以消息为单位，程序员直接利用系统提供的一组通信命令(原语)来实现通信，如 Send(A)、Receive(A)。

(3)　管道通信：所谓管道，是指用于连接一个读进程和一个写进程，以实现它们之间通信的共享文件(pipe 文件)。向管道(共享文件)提供输入的发送进程(即写进程)，以字符流的形式将大量的数据送入管道；而接收进程可从管道接收大量的数据。由于通信是采用管道的方式，所以叫管道通信。

2.2.1.4　进程调度

进程调度即处理机调度，它的主要功能是确定在什么时候分派处理机，并确定分给哪一个进程。在一些操作系统中，一个作业从提交到完成需要经历高、中、低 3 级调度。

- 高级调度：又称“长调度”“作业调度”或“接纳调度”，它决定处于输入池中的哪个后备作业可以调入主系统做好运行的准备，成为一个或一组就绪进程。系统中一个作业只需经过一次高级调度。
- 中级调度：又称“中程调度”或“对换调度”，它决定处于交换区中的哪个就绪进程可以调入内存，以便直接参与对 CPU 的竞争。在内存资源紧张时，为了将进程调入内存，必须将内存中处于阻塞状态的进程调至交换区，以便为调入进程腾出空间。
- 低级调度：又称“短程调度”或“进程调度”，它决定处于内存中的哪个就绪进程可以占用 CPU，是操作系统中最活跃、最重要的调度程序，对系统的影响很大。

1. 调度方式

调度方式是指当有更高优先级的进程到来时如何分配 CPU。调度方式分为可剥夺式和不可剥夺式两种。可剥夺式是指当有更高优先级的进程到来时，强行将正在运行的进程所占用的 CPU 分配给高优先级的进程；不可剥夺式是指当有更高优先级的进程到来时，必须等待正在运行的进程自动释放占用的 CPU，然后将 CPU 分配给高优先级的进程。

2. 进程调度算法

常用的进程调度算法有：先来先服务、时间片轮转、优先级调度和多级反馈调度算法。

1) 先来先服务

先来先服务(FCFS)是按照作业提交或进程变为就绪状态的先后次序，分配 CPU。即每当进入进程调度时，总是将就绪队列队首的进程投入运行。FCFS 的特点比较有利于长作业，而不利于短作业；有利于 CPU 繁忙的作业，而不利于输入/输出繁忙的作业。

2) 时间片轮转

FCFS 算法主要用于宏观调度，时间片轮转算法主要用于微观调度，通过时间片轮转，提高进程并发性和响应时间，从而提高资源利用率。

时间片轮转的实现过程是将系统中所有的就绪进程按照 FCFS 原则，排成一个队列。每次调度时将 CPU 分派给队首进程，让其执行一个时间片。时间片的长度从几毫秒到几百毫秒。在一个时间片结束时，发生时钟中断，调度程序据此暂停当前运行进程的执行，将其送到就绪队列的末尾，并通过上下文切换执行当前的队首进程。进程可以未使用完一个时间片，就出让 CPU(如阻塞)。

时间片长度的确定主要考虑以下 4 个方面。

- 时间片长度变化的影响：时间片过长，退化为 FCFS 算法，进程在一个时间片内都执行完，造成响应时间长；时间片过短，用户的一次请求需要多个时间片才能处理完，上下文切换次数增加，系统效率降低，同样造成响应时间增长。
- 对响应时间的要求：T(响应时间)=N(进程数目)×q(时间片)。
- 就绪进程的数目：数目越多，时间片越小。
- 系统的处理能力：应当使用户输入在一个时间片内能处理完，否则会使响应时间、平均周转时间和平均带权周转时间延长。

3) 优先级调度

优先级调度分为静态优先级和动态优先级两种。

- 静态优先级：进程的优先级是在创建时就已确定好了的，直到进程终止都不会改变。确定优先级的依据主要有：进程类型(系统进程优先级较高)、对资源的需求(对 CPU 和内存需求较少的进程优先级较高)、用户要求(紧迫程度和付费多少)。
- 动态优先级：在创建进程时赋予一个优先级，在进程运行过程中还可以改变，以便获得更好的调度性能。进程每执行一个时间片，就降低其优先级，从而一个进程持续执行时，其优先级可能会降低到出让 CPU 为止。

4) 多级反馈调度

多级反馈调度算法是时间片轮转算法和优先级算法的综合与发展。其优点是：照顾了短进程、提高了系统吞吐量、缩短了平均周转时间；照顾输入/输出型进程，获得较好的输入/输出设备利用率和缩短响应时间；不必估计进程的执行时间，动态调节优先级。

2.2.1.5　死锁

1. 死锁的基本概念

当若干进程竞争使用资源时，可能每个进程要求的资源都已被另一进程占用，于是也就没有一个进程能继续运行，这种情况称为死锁。例如，P1 进程占有资源 R1，P2 进程占有资源 R2，这时，P1 又需要资源 R2，P2 也需要资源 R1，它们在等待对方占有的资源时，又不会释放自己占有的资源，因而使双方都进入了无限等待状态。死锁是系统的一种出错状态，不仅浪费大量的系统资源，甚至会导致整个系统的崩溃，所以死锁是应该尽量预防和避免的。

系统发生死锁时，死锁进程的个数至少为两个；所有死锁进程都有等待资源，其中至少有两个进程已占有资源。产生死锁的情况主要有：进程推进顺序不当；同类资源分配不当；PV 操作使用不当。

2. 产生死锁的 4 个必要条件

产生死锁的原因：一是系统提供的资源数量有限，不能满足每个进程的使用；二是多道程序运行时，进程推进顺序不合理。发生死锁必须同时具备下述 4 个条件。

- 互斥：进程互斥使用资源，任意时刻一个资源只为一个进程所独占，其他进程若请求一个已被占用的资源，只能等待占用者释放后才能使用。
- 不可剥夺(不可抢占)：进程所获得的资源在未使用完毕之前，不能被其他进程强行剥夺，而只能由获得该资源的进程自己释放。
- 请求保持：进程每次申请它所需要的一部分资源，在申请新的资源的同时，继续占用已分配到的资源。零星地请求资源，即已获得部分资源后再次请求资源时被阻塞。
- 循环等待：在进程资源有向图中存在一个进程环路，环路中每一个进程已获得的资源同时被下一个进程所请求。

进程资源有向图由方框、圆圈和有向边 3 个部分组成。其中，方框表示资源，圆圈表示进程。请求资源：○→□，箭头由进程指向资源；分配资源：○←□，箭头由资源指向进程。

3. 解决死锁的方法

解决死锁的方法如下。

- 死锁的预防：根据产生死锁的 4 个必要条件，只要使其中之一不能成立，死锁就不会出现。
- 死锁的避免：最著名的死锁避免算法是 Dijkstra 提出的银行家算法。
- 死锁的检测：采用合理的死锁检测算法确定死锁的存在，并识别出与死锁有关的进程和资源，以供系统采用适当的解除死锁的措施。
- 死锁的解除：检测到死锁发生后，常采用资源剥夺法和撤销进程法解除死锁。

2.2.1.6　线程

1. 线程的基本概念

线程是比进程更小的能独立运行的基本单位。在引入线程的操作系统中，线程是进程

中的一个实体，是 CPU 调度和分派的基本单位。线程自己基本上不占用系统资源，只占用一点儿在运行中必不可少的资源(如程序计数器、一组寄存器和栈)，但它可与同属一个进程的其他线程共享该进程所占用的全部资源。相应地，线程也同样有就绪、等待和运行 3 种基本状态。在有的系统中线程还有终止状态。

2. 线程的属性

线程的属性如下。

- 每个线程都有一个唯一的标识符和一张线程描述表。
- 不同的线程可以执行相同的程序。
- 同一进程中的各个线程共享该进程的内存地址空间。
- 线程是处理机的独立调度单位，多个线程是可以并发执行的。
- 线程在生命周期内会经历等待状态、就绪状态和运行状态等各种状态变化。

3. 引入线程的好处

传统的进程有两个基本属性：可拥有资源的独立单位、可独立调度和分配的基本单位。由于在进程的创建、撤销和切换中，系统必须为之付出较大的时空开销，因此在系统中所设置的进程数目不宜过多，进程切换的频率不宜太高，这就限制了并发程度的提高。引入线程后，将传统进程的两个基本属性分开，将线程作为调度和分配的基本单位，而将进程作为独立分配资源的单位。用户可以通过创建线程来完成任务，以减少程序并发执行时付出的时空开销。

引入线程的好处主要有如下几个。

- 创建一个新线程花费的时间少。
- 两个线程间切换花费的时间少。
- 由于同一进程内的线程共享内存和文件，线程之间相互通信无须调用内核，故不需要额外的通信机制，使通信更简便，信息传送速度也更快。
- 线程能独立执行，能充分利用和发挥处理机与外围设备并行工作的能力。

2.2.2 典型例题分析

例 1：已知有 5 个进程共享一个互斥段，如果最多允许 2 个进程同时进入互斥段，则相应的信号量的变化范围是__(26)__。(2015 年下半年试题 26)

A. −5～1　　B. −4～1　　C. −3～2　　D. −2～3

分析：本题考查处理机管理。

信号量是一个整型变量，根据控制对象的不同被赋予不同的值。信号量分为两类：公用信号量，实现进程间的互斥，初值为 1 或资源的数目；私用信号量，实现进程间的同步，初值为 0 或某个正整数。信号量 S 的物理意义：$S\geq 0$ 表示某资源的可用数，若 $S<0$，则其绝对值表示阻塞队列中等待该资源的进程数。

系统中有 5 个进程共享一个互斥段，如果最多允许 2 个进程同时进入互斥段，则信号量 S 的初值应设为 2，当第一个进程进入互斥段时，信号量 S 减 1 等于 1；当第二个进程进入互斥段时，信号量 S 减 1 等于 0；……；当第 5 个进程进入互斥段时，信号量 S 减 1 等于 −3。可见，信号量的变化范围是−3～2。

答案：C

例 2：进程的三态模型如图 2-2 所示，其中的 a、b 和 c 处应分别填写__(27)__。(2015 年下半年试题 27)

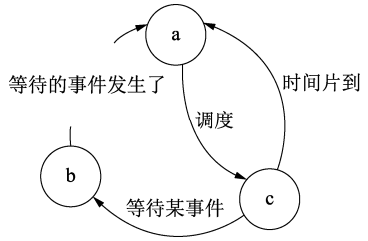

图 2-2　进程的三态模型

A. 就绪、阻塞和运行　　B. 就绪、运行和阻塞
C. 阻塞、就绪和运行　　D. 运行、就绪和阻塞

分析：本题考查处理机管理。

进程的状态有 3 种：运行、就绪和阻塞。运行(running)：正占用处理机。就绪(ready)：只要获得处理机即可运行。阻塞(blocked)：也称等待或挂起状态，正等待某个事件(如 I/O 完成)的发生。

答案：A

例3：假设有5个进程共享一个互斥段X，如果最多允许2个进程同时进入互斥段X，则信号量*S*的变化范围是__(25)__；若信号量*S*的当前值为–3，则表示系统中有__(26)__个正在等待该资源的进程。(2015年上半年试题25、26)

(25) A. –5～1　　B. –1～3　　C. –3～2　　D. 0～5
(26) A. 0　　B. 1　　C. 2　　D. 3

分析：本题考查处理机管理知识。

信号量是一个整型变量，根据控制对象的不同被赋予不同的值。信号量分为两类：公用信号量，实现进程间的互斥，初值为 1 或资源的数目；私用信号量，实现进程间的同步，初值为 0 或某个正整数。信号量 *S* 的物理意义：*S*≥ 0 表示某资源的可用数，若 *S*<0，则其绝对值表示阻塞队列中等待该资源的进程数。

如果最多允许 2 个进程同时进入互斥段 X，那么信号量初值应为 2。如果 5 个进程进入互斥段 X，那么此时信号量值应为−3。因此信号量的变化范围是−3～2。如果 5 个进程进入互斥段 X，而互斥段 X 最多允许 2 个进程同时进入，则系统中有 3 个正在等待该资源的进程。

答案：(25)C　　(26)D

例 4：假设系统有 *n* 个进程共享资源 R，且资源 R 的可用数为 2，那么该资源相应的信号量 *S* 的初值应设为__(27)__。(2014 年下半年试题 27)

A. 0　　B. 1　　C. 2　　D. *n*

分析：本题考查操作系统进程管理中信号量与同步互斥方面的基本知识。

本题中已知有 *n* 个进程共享 R 资源，且 R 资源的可用数为 2，所以，信号量的初值应设为 2。

答案：C

例 5：若进程 P1 正在运行，操作系统强行撤下 P1 进程所占用的 CPU，让具有更高优先级的进程 P2 运行，这种调度方式称为 (26) 。(2014 年上半年试题 26)

A. 中断方式　　B. 抢占方式　　C. 非抢占方式　　D. 查询方式

分析：本题考查操作系统进程管理方面的基础知识。

在操作系统进程管理中，进程调度方式是指某进程正在运行，当有更高优先级的进程到来时如何分配 CPU。调度方式分为可剥夺和不可剥夺两种。可剥夺式是指当有更高优先级的进程到来时，强行将正在运行进程的 CPU 分配给高优先级的进程；不可剥夺式是指当有更高优先级的进程到来时，必须等待正在运行进程自动释放占用的 CPU，然后将 CPU 分配给高优先级的进程。

答案：B

例 6：假设系统有 6 个进程共享一个互斥段，如果最多允许 3 个进程同时进入互斥段，则信号量的初值为 (26) ，信号量 *S* 的变化范围是 (27) 。(2013 年下半年试题 26、27)

(26) A. 0　　B. 1　　C. 3　　D. 6

(27) A. 0～6　　B. −3～3　　C. −4～2　　D. −5～1

分析：本题考查操作系统进程管理中信号量与同步互斥方面的基础知识。

本题中已知有 6 个进程共享一个互斥段，而且最多允许 3 个进程同时进入互斥段，这意味着系统有 3 个单位的资源，所以，信号量的初值应设为 3。

当第一个申请该资源的进程对信号量 *S* 执行 P 操作，信号量 *S* 减 1 等于 2，进程可继续执行；当第二个申请该资源的进程对信号量 *S* 执行 P 操作，信号量 *S* 再减 1 等于 1，进程可继续执行；当第三个申请该资源的进程对信号量 *S* 执行 P 操作，信号量 *S* 再减 1 等于 0，进程可继续执行；当第四个申请该资源的进程对信号量 *S* 执行 P 操作，信号量 *S* 再减 1 等于 −1，进程申请的资源得不到满足，处于等待状态；当第五个申请该资源的进程对信号量 *S* 执行 P 操作，信号量 *S* 再减 1 等于−2；当第六个申请该资源的进程对信号量 *S* 执行 P 操作，信号量 *S* 再减 1 等于−3。所以信号量 *S* 的变化范围是−3～3。

答案：(26)C　　(27)B

2.2.3　同步练习

1. 某企业有生产部和销售部，生产部负责生产产品并送入仓库，销售部从仓库取产品销售。假设仓库可存放 *n* 件产品。用 PV 操作实现他们之间的同步过程如图 2-3 所示。

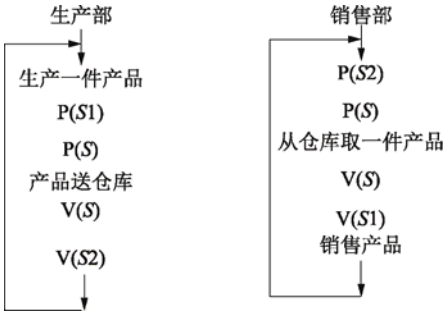

图 2-3　PV 操作实现同步过程

其中，信号量 *S* 是一个互斥信号量，初值为 (1) ；*S*1 是一个 (2) ；*S*2 是一个 (3) 。

(1) A. 0　　B. 1　　C. *n*　　D. −1

(2) A. 互斥信号量，表示仓库的容量，初值为 *n*

B. 互斥信号量，表示仓库是否有产品，初值为 0

C. 同步信号量，表示仓库的容量，初值为 *n*

D. 同步信号量，表示仓库是否有产品，初值为 0

(3) A. 互斥信号量，表示仓库的容量，初值为 *n*

B. 互斥信号量，表示仓库是否有产品，初值为 0

C. 同步信号量，表示仓库的容量，初值为 *n*

D. 同步信号量，表示仓库是否有产品，初值为 0

2. 在操作系统的进程管理中，若系统中有 8 个进程要使用互斥资源 R，但最多只允许 2 个进程进入互斥段(临界区)，则信号量 *S* 的变化范围是 (1) ；若信号量 *S* 的当前值为 4，则表示系统中有 (2) 个进程正在等待该资源。

(1) A. −2～0　　B. −2～1　　C. −6～2　　D. −8～1

(2) A. 1　　B. 2　　C. 3　　D. 4

3. 若计算机系统中某时刻有 5 个进程，其中 1 个进程的状态为“运行”，2 个进程的状态为“就绪”，2 个进程的状态为“阻塞”，则该系统中并发的进程数为 (1) ；如果系统中的 5 个进程都要求使用两个互斥资源 R，那么该系统不产生死锁的最少资源数 R 应为 (2) 个。

(1) A. 2　　B. 3　　C. 4　　D. 5

(2) A. 5　　B. 6　　C. 8　　D. 9

2.2.4　同步练习参考答案

1. (1)B　(2)C　(3)D

2. (1)C　(2)D

3. (1)D　(2)B

2.3　存 储 管 理

2.3.1　考点辅导

2.3.1.1　基本概念

现代计算机系统中的存储系统通常是多级存储体系，至少有主存(内存)和辅存(外存)两级，有的系统有更多级。系统中主存的使用一般分成两部分：一部分为系统空间，存放操作系统本身及相关的系统数据；另一部分为用户空间，存放用户的程序和数据。提高主存的利用率，对主存信息实现有效保护是存储器管理的主要任务。

1. 存储器的结构

存储器的功能是保存数据，存储器的发展方向是高速度、大容量和小体积。一般存储器的结构有“寄存器－主存－外存”结构或“寄存器－缓存－主存－外存”结构。下面介绍几个与存储器相关的概念。

(1) 虚拟地址：数据的存放地址是由符号决定的，故又称符号名地址，或者称为名地址，而把源程序的地址空间叫作符号名地址空间或者名空间。它从 0 号单元开始编址，并顺序分配所有的符号名所对应的地址单元，所以它不是主存中的真实地址，故称为相对地址、程序地址、逻辑地址或虚拟地址。