

第 3 章 计算系统的基本思维

【问题引出】用二进制数来表示各类数据的目的是便于计算机进行数值计算和数据处理，现代计算机是一个完整的计算系统，由硬件和软件组成。那么，如何使软件和硬件能相互“理解”“支持”，并在数值计算和数据处理中“齐心协力”地高效工作呢？这就是本章所要讨论的问题。

【教学重点】计算系统的基本构成与层次关系、各硬件系统的工作原理、操作系统对各硬件系统及磁盘文件的组织与管理、硬件设备的相互连接、操作系统对信息交换的控制与管理等。

【教学目标】熟悉处理器系统、存储器系统、总线系统、输入/输出接口等的基本结构组成；了解操作系统在计算系统中的作用和地位，操作系统对各硬件系统和文件系统的管理方法。

§ 3.1 计算系统的基本概念——“系统组成”的基本思维

计算系统是对计算机的抽象。相应地，计算机是计算系统的具体体现。具体地说，计算机是一种设备，计算系统则是一种动态实体，由硬件系统、软件系统和它们管理的数据构成，用于解决计算问题以及与其所处的环境进行交互。计算系统全面体现了计算思维和系统科学思想，即以“计算”为目的，探索问题求解的过程抽象和工作方式；以“系统”为研究和应用对象，探索系统的存在方式和变化规律。例如，计算机硬件系统和软件系统的定义与划分、操作系统的功能作用和管理模块的定义与划分，就是计算思维和系统科学研究的典型实例。

3.1.1 计算机硬件系统

计算机硬件是指那些看得见、摸得着的部件，它们是构成计算机的物理装置。我们把构成计算机的所有部件称为硬件，并将这些硬件的整体称为硬件系统，它是计算系统的物理实现。如果按功能抽象，则可将硬件系统划分为如图 3-1 所示的 4 个系统。

1. 处理器系统

处理器系统是计算系统的核心部件，由 CPU 及其指令系统组成，完成指令控制、操作控制、时序控制和数据处理。指令系统发出控制指令，指挥各有关部件有条不紊地执行各项操作。

2. 存储器系统

存储器系统是计算系统中存放程序和数据的各种存储设备（内存和外存）、控制部件、信息调度设备和算法的总称。其中，内存包括随机存储器（Random Access Memory, RAM）、只读存储器（Read-Only Memory, ROM）、缓冲存储器（Flash）等；外存包括磁盘、磁带、U 盘（USB Flash）等。

3. 输入/输出系统

输入/输出系统由输入设备、输出设备，以及连接所有外部设备的接口电路组成。接口电路也

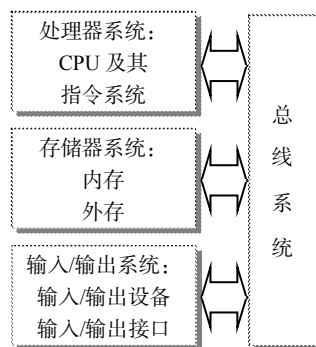


图 3-1 硬件系统

称为输入/输出接口（Input/Output Interface, I/O 接口），它是实现输入/输出管理的控制机构。所有的输入/输出设备和外存都需要利用输入/输出接口，才能通过总线实现信息交换。

4. 总线系统

总线系统是多个功能部件之间进行数据传送的公共通路，是构成计算系统的互联机构。总线系统的主要特征是多个部件共享传输介质，如一个部件发出的信号可以被连接到总线上的其他部件接收。采用总线传输方式可使各系统部件间实现地址、数据和控制信息的传递与交换，从而大大减少了信息传输线的数量，增强了系统的灵活性，提高了系统的运行效率。

3.1.2 计算机软件系统

计算机软件是相对于硬件而言的，它是在操作系统的基础上逐渐发展形成的。为了便于与硬件区分，我们把计算机中使用的各种程序称为软件，并将计算机中所有程序的集合称为软件系统，它是计算系统的软件支撑，以拓展硬件功能和方便用户操作。

1. 软件的定义

软件是为使计算机高效地工作所配置的各种程序及相关的文档资料的总称。其中，程序是经过组织的计算机指令序列，指令是组成计算机程序的基本单位；文档资料包括软件开发过程中的需求分析、方案设计、编程方法等文档，以及使用说明书、用户手册、维护手册等。

1983 年，IEEE 对软件给出了一个较为新颖的定义：软件是计算机程序、方法、规范及其相应的文稿，以及其在计算机上运行时必需的数据。这个定义在学术上有重要的参考价值，它将程序与软件开发方法、程序设计规范及其相应的文档联系在一起，将程序与其在计算机上运行时必需的数据联系在一起，这是在考虑了软件生存周期中的各项主要因素之后提出的。

软件在用户和计算机硬件之间架起了联系的桥梁，用户只有通过软件才能使用计算机。同时，计算机软件是对硬件功能的扩充与完善。软件有多种类型，但其基本功能、任务和特点是一致的。

（1）**软件的基本功能**。使用户能根据自己的意图来指挥计算机工作，并使计算机硬件系统能高效发挥作用。

（2）**软件的主要任务**。一是为用户操作计算机提供方便和提高使用效率；二是充分发挥计算机硬件系统的功能作用；三是扩大计算机的用途。

（3）**软件的主要特点**。软件是一种逻辑实体，而不是具体的物理实体；软件是看不见、摸不着的纯智力产品，其研制成本远远大于其生产成本；软件没有老化问题，但其维护比硬件复杂得多。我们可把软件的特点概括为独创性、无形性、复制性、永久性、复杂性以及非价格的创新竞争。

2. 软件的分类

与硬件一样，计算机软件也在不断发展。从冯·诺依曼计算机开始，软件的发展过程经历了计算机语言、翻译程序、操作系统、服务程序、数据库管理系统、应用程序等。根据软件的功能作用，可将其分为系统软件（System Software）、支撑软件（Supporting Software）和应用软件（Application Software）3 类。

（1）**系统软件**是计算机厂家为实现对计算系统的管理、调度、监视、服务、维护，以及扩充硬件功能等而提供给用户的软件。操作系统、翻译程序、服务程序都属于系统软件。系统软件的显著特点是与具体应用领域无关。

（2）**支撑软件**是随着数据库应用系统开发和网络应用的不断拓展而逐渐形成的软件。随着计

算机技术的发展,软件的开发、维护与运行的代价在整个计算系统中所占的比重越来越大,远远超过硬件系统。目前,常用的支撑软件有数据库管理系统、各类工具软件(如系统诊断软件、图像处理软件)、网络软件、软件开发环境、中间件(独立的系统软件或服务程序)等。

(3) **应用软件**是相对于系统软件而言的,是用户针对各种具体应用问题而开发的一类专用程序或软件的总称。例如,计算机辅助设计软件、计算机辅助测试软件、计算机辅助制造软件、计算机辅助教学软件、专家系统(Expert System, ES)、各类信息管理系统、科学计算软件等,这些应用软件在各有关领域大显神通,给传统的产业部门注入了新的活力。

【提示】虽然翻译程序和服务程序与操作系统都是由计算机厂商提供的系统软件,但使用却完全不同:操作系统在系统启动时由“引导程序”(Boot Program)从外存中将其装入内存;而翻译程序和服务程序在需要时,由操作系统将它们从外存调入内存,任何其他可执行程序都是如此。

3.1.3 计算机操作系统

一台没有任何软件支持的计算机称为裸机,用户直接面向裸机进行应用操作是相当困难的,几乎无法实现。因为计算机硬件系统只能识别由0和1组成的二进制代码信息,不能识别人类的日常表达,所以用户直接操作、使用、管理和维护计算机时,总是觉得机器“太硬”了(不理解人意),而机器总是觉得用户“太笨”了(动作缓慢)。因此,迫切需要解决让计算机理解用户的操作意图和因为用户的操作速度太慢,致使计算机无事可做等问题。为了摆脱人的“高智低能”和发挥计算机的“低智高能”,必须让计算机来管理自己。于是,操作系统应运而生。

1. 操作系统的定义

操作系统是随着硬件和软件的不断发展而逐渐形成的一套大型程序。从功能上讲,它是用户与计算机硬件之间的接口,是用户和其他软件、计算机硬件之间的桥梁。从作用上讲,它为用户操作和使用计算机提供一个良好的操作与管理环境,使计算机的使用效率成倍地提高,并且为用户提供了方便的使用手段和令人满意的服务质量。因此,可将其定义为: 操作系统是有效地组织和管理计算机系统中的硬件和软件资源,合理地组织计算机工作流程,控制程序的执行,并提供多种服务功能及友好界面,方便用户使用计算机的系统软件。

2. 操作系统的作用

现在呈现在用户面前的计算机,是经过若干次抽象的计算系统,我们可把整个计算系统按功能划分为4个层次,即硬件系统、操作系统、支撑软件和应用软件。这4个层次表现为一种单向服务关系,即外层可以使用内层提供的服务,反之则不行,其层次关系如图3-2所示。

操作系统由进程管理、存储管理、文件管理(File Management)和设备管理4个模块组成,它在计算机中的作用体现在以下3个方面。

(1) **操作系统是用户与计算机硬件之间的接口。**用户需要通过操作系统来使用计算机,操作系统使用户在不需要过多了解硬件的情况下,方便地使用计算机中的资源,并且能够根据用户需求对硬件进行改造和扩充。因此,可把操作系统看作人机交互的接口或界面。

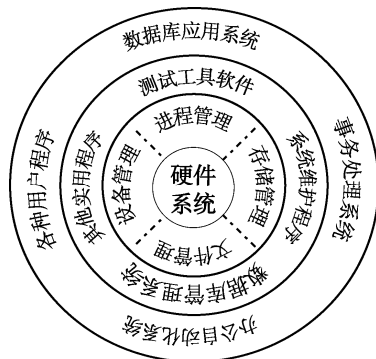


图 3-2 计算系统的层次关系

(2) **操作系统是计算系统的资源管理者**。操作系统的基本任务是管理计算系统中的软硬件资源,是软件和硬件之间的协调者。一方面,它控制和管理着硬件,向上层的实用程序和用户应用程序提供一个屏蔽硬件工作细节的良好使用环境,把一个裸机变成可操作且方便灵活的计算机系统;另一方面,计算机中的程序和数据以文件形式存放在外存中,如何与内存进行数据交换由操作系统来实现,最大限度地提高资源的利用效率和实现资源共享,操作系统使得计算机成为一台“海纳百川”的机器。

(3) **操作系统为用户提供了一个虚拟机**。计算机硬件系统的功能是有限的,但通过操作系统及其相关软件,可以完成多种多样、复杂多变的任务,很多需要硬件实现的功能可以由软件实现,许多硬件实现不了的功能也可以由软件来实现。为了方便用户开发更为复杂的程序,操作系统提供了更容易理解的、任务相关的、控制硬件的指令,通常称之为应用程序接口(Application Program Interface, API),它将对硬件进行控制的具体细节封装起来,通过在裸机上加上一层又一层软件来组成一个完整的计算机系统,为用户提供了一台扩展基本功能、使用更为方便的机器。虚拟机是一个功能强大、安全可靠、效率极高的计算机系统。

【提示】进程管理、存储管理、文件管理和设备管理是操作系统的四大功能,其工作原理结合各硬件系统一起讨论,这样有利于深刻了解计算系统的工作机制。

3. 常用操作系统

操作系统的研发与硬件系统的功能及其性能是相对应的。随着硬件系统的不断发展,研发出了许多不同类型的操作系统。其中,影响最大、目前使用最广的计算机主流操作系统有以下3种。

(1) **UNIX 操作系统**是1969年由美国AT&T公司的Bell实验室开发出来的,也是能在笔记本计算机、个人计算机(又称为微型计算机)、巨型计算机等多种硬件环境下运行的多任务、多用户操作系统。由于UNIX操作系统可满足各行业的应用需求,因此其已成为重要的企业级操作平台,也是操作系统中的常青树。

UNIX操作系统的特点主要体现为技术成熟、结构简练、功能强大、可移植性和兼容性好、伸缩性和互操作性强,是当今世界最流行的多用户、多任务操作系统之一,被认为是开放系统的代表。总体上,UNIX操作系统的主要发展趋势是统一化、标准化和不断创新。

(2) **Windows 操作系统**是美国Microsoft公司于1985年在磁盘操作系统(MicroSoft Disk Operating System, MS-DOS)的基础上发展而来的。MS-DOS是为个人计算机研制的单用户、单任务命令行界面操作系统,曾经被广泛安装在个人计算机上,它对个人计算机的普及应用是功不可没的。虽然如今MS-DOS已退出历史舞台,但它的很多概念在Windows操作系统中仍然是重要的,而且MS-DOS的命令行方式在网络中仍被采用。

MS-DOS的特点是简单易学,系统规模小(占用的内存空间小),因而对硬件要求低,并且运行速度快。但由于它提供的是一种以字符为基础的用户接口,如果不熟悉硬件和MS-DOS的操作命令,便难以高效使用个人计算机,人们期盼个人计算机变成一个直观、易学、好用的工具。Microsoft公司响应MS-DOS用户的愿望,研制开发了基于图形用户接口(Graphical User Interface, GUI)方式的新型操作系统——Windows操作系统。在图形用户界面中,Windows操作系统所支持的应用软件都用一个图标表示,用户只要把鼠标指针移到某图标上并双击,即可进入该软件。这种界面方式为用户提供了极大的方便,从此把计算机的操作使用提高到了一个崭新的阶段,并为今天计算机网络的操作和多媒体系统的操作使用创造了极为有利的条件,这也是Windows操作系统受到用户欢迎并得以迅速发展的重要原因之一。目前,个人计算机中最常使用的操作系统是具有多用户、多任务功能的Windows XP、Windows 10等操作系统,以及基于Windows操作系

统的网络操作系统。

(3) **Linux 操作系统**是由芬兰赫尔辛基大学的研究生 Linus Benedict Torvalds 于 1991 年主导开发的,是一种能运行于多种平台、源代码公开、免费、功能强大、遵循可移植操作系统接口 (Portable Operating System Interface of UNIX, POSIX) 标准、与 UNIX 操作系统兼容的操作系统。Linux 操作系统继承了自由软件的优点,是最成功的开源软件之一。Linux 操作系统的源程序能完整地上传到 Internet,允许自由下载,因而不仅被众多高校、科研机构、军事机构和政府机构广泛采用,也被越来越多的行业所采用。目前, Linux 操作系统是一种得到广泛应用的多用户、多任务操作系统,许多计算机公司(如 IBM、Intel、Oracle、SUN 等)大力支持 Linux 操作系统,各种常用软件也纷纷移植到 Linux 平台。

Linux 操作系统的特点主要体现为开放性、多用户、多任务、良好的用户界面、支持多个虚拟控制台、可靠的系统安全、共享内存页面、支持多种文件系统、强大的网络功能、良好的可移植性等。Linux 操作系统非常适合应用于需要运行各种网络应用程序,并提供各种网络服务的计算机。正是 Linux 操作系统的开源才使得它可以根据自身的需要做专门的开发。因此,它更适合需要自行开发应用程序的用户和那些需要学习 UNIX 命令工具的用户。

【提示】随着移动互联网 (Mobile Internet) 和智能手机的普及,近年来推出了许多用于移动终端用户 (智能手机) 的操作系统,它们功能强大,性能优越。

4. 操作系统的用户接口

用户接口 (User Interface) 是为了方便用户操作、使用计算机而提供的人机交互接口,也是计算系统与用户之间进行交互和通信的通路。目前,常用操作系统的用户接口可分为 3 种:命令接口 (Command Interface)、程序接口 (Program Interface) 和图形接口 (Graphic Interface)。

(1) **命令接口**是为便于用户直接或间接控制自己的程序而设置的接口,用户通过该接口向计算机发出命令,执行功能操作。命令接口又分为联机用户接口和脱机用户接口。

① **联机用户接口**由一组操作命令及其对应的命令解释程序所组成,当用户输入一条命令后,系统便立即对该命令进行解释和执行。例如,MS-DOS 和 UNIX 操作系统提供的就是联机用户接口。

② **脱机用户接口**是为批处理作业的用户提供的接口,由一组作业控制语言组成。所谓作业,是指用户在运行程序和进行数据处理过程中要求计算机所做工作的集合。批处理作业的用户不能直接与自己的作业交互,只能委托系统代替用户对作业进行控制和干预,其实现就是由脱机用户接口来完成的。

(2) **程序接口**是为用户程序访问系统资源而设置的接口,是用户使用操作系统提供的系统服务功能的唯一途径,使用户能在程序中调用操作系统提供的子功能模块。各类操作系统提供了不同程序设计语言的接口,例如,MS-DOS 以系统功能调用的方式提供程序接口,这种调用是一种特殊的过程调用,通常是由特殊的机器指令实现的,可以在编写汇编语言程序时直接调用功能函数。Windows 操作系统以应用程序接口的方式提供各种功能函数,直接调用这些功能函数,就能编写出各种界面友好、功能强大的应用程序。

【提示】可视化编程环境 (VC++、Java、Delphi 等) 提供了大量的类库和各种控件,如微软基础类 (Microsoft Foundation Classes, MFC), 这些类库和控件提供了方便的调用方法,从而极大地简化了 Windows 应用程序的开发。

(3) **图形接口**是联机用户接口的另一种方式,是为克服命令方式存在的不足而采用的接口方

式。通过图形界面提供的“对话框”“图标”“菜单”，能将系统的各项功能、各种应用程序和文件直观地表示出来。例如，在 Windows 操作系统中，通过鼠标、菜单和对话框来完成对各种应用程序和文件的操作。此时用户不必像使用命令接口那样记住命令名及格式，只要轻点鼠标就能实现很多操作，从而使用户从烦琐且单调的操作中解放出来，因此 Windows 操作系统能够被很多非专业人员使用。

3.1.4 软件与硬件的关系

一台能操作、使用的计算机必须具有硬件和软件，两者相辅相成，缺一不可，从而构成一个不可分割的整体。硬件与软件既相互支持，又相互制约，只有在取得“共识”的前提下，“相互支撑、齐心协力”地努力工作，才能高效完成用户给定的工作任务。计算机的运行就是硬件和软件彼此配合、共同作用的结果。计算机硬件和软件之间的“相互支撑”主要体现在以下 3 个方面。

1. 层次结构关系

一个完整的计算系统，如果按照系统的层次结构抽象，可以抽象为硬件系统层、系统软件层、支撑软件层和应用软件层 4 层，如图 3-3 所示。

(1) **硬件系统层**是计算系统的物理实现，位于计算系统的底层。计算机指令系统是硬件系统与系统软件交互的工作界面，其信息交互的语言是由 0 和 1 组成的二进制代码。

(2) **系统软件层**能够直接与硬件系统层进行交互，其中最重要的是操作系统，用来帮助用户摆脱硬件的束缚，对硬件进行扩充和改造，为利用硬件资源提供使用环境，是实现管理和使用的核心。操作系统是用户操作、使用计算机的界面，其交互信息是 ASCII 码或图标。

(3) **支撑软件层**利用操作系统提供的接口及系统调用来使用计算系统中的各类系统资源，并为应用软件提供支撑。最典型的支撑软件是测试工具软件、数据库管理系统和其他系统程序。

(4) **应用软件层**建立在支撑软件层之上，为用户提供应用系统界面，使用户能够方便地利用计算机来解决具体问题。在应用系统界面上，用户与计算机之间进行信息交换的“公用”语言是程序设计语言，利用程序设计语言，把要解决的问题以命令的形式进行有序描述。



图 3-3 计算系统的层次结构

【提示】计算系统中的每层都具有特定的功能，并提供了相应的接口界面，接口屏蔽了层内的实现细节，并对外层提供了使用约定。

2. 相互依赖关系

现代计算系统中的硬件和软件相互依赖和支持。有了软件的支持，硬件才能正常运转和高效工作。如果把硬件比作计算系统的躯体，那么软件就是计算系统的灵魂。

(1) **硬件是计算系统的躯体**，是构成计算机的物理装置，也是计算机的物理实现。硬件为软件提供物理支撑，任何软件都是建立在硬件基础之上的，如果脱离了硬件，软件则无法栖身。

(2) **软件是计算系统的灵魂**，是为运行、管理和维护计算机而编制的各种程序的总和。软件为硬件提供使用环境，对于一个完整的计算系统，如果没有软件的支持，硬件将变得毫无意义。

3. 功能等价关系

计算机的硬件和软件在逻辑功能上是等价的，即计算系统的许多功能既可用硬件实现，也可用软件实现。例如，在计算机设计中，如果硬件成本高，软件成本低，为了取得较高的性能价格比，可用软件来实现更高一级的性能，这种做法称为硬件软化。随着集成电路技术的发展，硬件

价格逐渐降低、可靠性逐渐提高,因而可以用硬件替代软件来实现较强的功能,这种做法称为软件硬化。一般情况下,用硬件实现功能往往可提高速度和简化程序,但将使系统结构复杂,造价提高;用软件实现功能可降低系统造价,但使程序变得复杂,运行速度降低。例如,计算机处理汉字时,既可使用硬字库,也可使用软字库。前者造价高,但运行速度快;后者造价低,但运行速度慢。

正是由于软件和硬件在功能上的等价关系,促进了软件、硬件技术的协同发展。一方面,许多生产厂家为实现某一功能或达到某一技术指标,分别用软件或硬件的办法来实现,并各自评价其优缺点。另一方面,硬件技术的发展及性能的改善,为软件的应用提供了广阔的前景,并为新软件的诞生奠定了基础。同时,软件技术的发展给硬件技术提出了新的要求,从而促进新的硬件产生与发展。正是这种彼此促进的关系和激烈竞争,形成了推动软件、硬件技术不断向前发展的强大动力。

下面按照计算思维的思想方式,介绍各硬件系统的基本原理以及操作系统如何实现对各硬件系统的管理。

§ 3.2 处理器系统——“控制执行”的基本思维

现代计算机是一个高度自动化的计算系统,有人把计算机称为“电脑”,是因为它具有类似于人脑,按照人的意愿,指挥和控制计算机程序执行各种操作的功能,这个指挥和控制机构就是由 CPU 和指令系统构成的处理器系统,它是计算机硬件系统的核心,由操作系统实现调配管理。

3.2.1 CPU 的结构组成

随着超大规模集成电路的高速发展,电子器件的体积越来越小,为了使主体器件具有一致性能,将运算器和控制器集成在一块芯片上,称为中央处理器或 CPU,在个人计算机中称为微处理器 (Microprocessor)。

1. CPU 的基本组成

CPU 是计算机的核心部件,用来实现运算和控制,并由运算器、控制器、控制线路等组成。

(1) **运算器**是用来完成算术运算和逻辑代数运算的部件,其功能包括:快速地对数据进行加、减、乘、除(包括变更数据符号)等算术运算;完成“与”“或”“非”等逻辑代数运算,逻辑左移、逻辑右移、算术左移、算术右移等移位操作;及时存放算术运算和逻辑代数运算的中间结果(由通用寄存器组实现),实现选择参与运算的操作数并执行运算的功能,并且把运算结果送到存储器中。运算器的逻辑结构如图 3-4 所示。

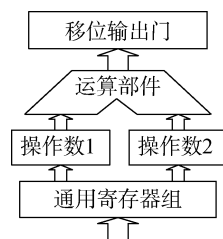


图 3-4 运算器的逻辑结构

运算器由多功能算术逻辑部件 (Arithmetic and Logical Unit, ALU)、通用寄存器组[包括累加器 (ACcumulator, AC)、缓冲寄存器、状态寄存器 (Flag Register, FR)] 及其控制线路组成。通用寄存器组用来存放运算器的数据和运算中间结果,减少访问存储器的次数,以提高运算器的运算速度。整个运算过程在控制器的统一指挥下,对取自 RAM 的数据按程序编排顺序进行算术运算或逻辑代数运算,然后将运算结果送到 RAM。

(2) **控制器**是计算系统发布操作指令的部件,其功能是根据指令提供的信息实现对系统各部

件的操作和控制。计算机程序和原始数据的输入、CPU 内部的信息处理、处理结果的输出、外部设备与主机之间的信息交换等都是是在控制器的控制下实现的。

2. CPU 的主要部件

CPU 是计算机硬件系统的指挥中心，有人将它形容为人脑的神经中枢。CPU 的指挥控制功能由指令控制、操作控制、时间控制、数据加工等部件实现。CPU 的基本组成如图 3-5 所示。

(1) **数据缓冲寄存器** (Data Buffer Register, DBR) 用来存储从 RAM 中取出的指令或数据字。

(2) **指令寄存器** (Instruction Register, IR) 用来存储从 RAM 中取出的将要执行的指令。

(3) **指令译码器** (Instruction Decoder, ID) 用于在执行 IR 中的指令时，对指令的操作码进行检测，以便识别要实现的操作。

(4) **地址寄存器** (Address Register, AR) 用来存储当前 CPU 所访问的内存单元的地址。

(5) **程序计数器** (Program Counter, PC) 用来存储下一条将要执行的指令的地址码 (Address Code)。

(6) **累加器** 用来在 ALU 执行算术或逻辑代数运算时，为 ALU 提供工作区。

(7) **状态寄存器** 用来存储为算术运算和逻辑代数运算或测试的结果建立的条件码内容，如运算进位标志、运算结果溢出标志、运算结果为零标志、运算结果为负标志，等等。

(8) **微操作控制单元和时序部件** (Micro-operation Control Unit and Timing Unit) 根据指令操作码和时序信号产生微操作控制信号，对各种操作实施时间上的控制。

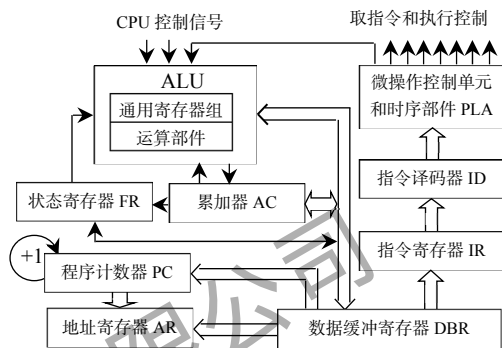


图 3-5 CPU 的基本组成

3.2.2 计算机指令系统

计算机之所以能自动工作，是因为它能顺序地执行存放在存储器中的程序（一系列指令）。我们把指示计算机执行各种基本操作的命令称为指令，计算机中所有指令的集合（Instruction Set）称为该机的指令系统，它是衡量 CPU 性能的重要指标之一。

1. 指令的基本格式

计算机指令由一串二进制代码（即机器语言）表示的操作码和地址码组成，如图 3-6 所示。

(1) **操作码** 用来表示进行何种操作，如进行加法、减法、乘法、除法等操作。在计算机中，机器语言是计算机唯一能识别的语言，所以操作码和地址码都用二进制代码来表示。由于机器语言不便于记忆，而且容易出错，给编程带来了很大困难，因此常采用类似英文单词的符号来表示指令操作码，并称之为助记符，其应用实例如表 3-1 所示。



图 3-6 指令的基本格式

表 3-1 助记符的应用实例

操作码	000	001	010	011	100	101	110	111
助记符	ADD (加法)	SUB (减法)	MUL (乘法)	DIV (除法)	AND (逻辑与)	LD (取数)	MOV (数据传送)	STOP (停机)

表 3-1 中的操作码为 3 位，可执行 8 种操作。由此可见，操作码的位数越长，计算机的功能

越强。一个 n 位操作码最多能够表示 2^n 种操作。

(2) **地址码**表示从哪个地址中取出操作数及操作的结果存放到哪个地址中去。由于操作运算有操作数、被操作数和操作结果,因此分为三地址指令、二地址指令、一地址指令等。不同操作运算的指令格式如图 3-7 所示。

三地址指令	操作码	A_1 (第一操作数)	A_2 (第二操作数)	A_3 (第三操作数)
二地址指令	操作码	A_1 (第一操作数)		A_2 (第二操作数)
一地址指令	操作码	A (操作数)		

图 3-7 不同操作运算的指令格式

对于一地址指令而言,通常是以运算器中的累加器 AC 中的数据为隐含操作数,指令字的地址码字段中的数为操作数,运算结果又被放回累加器 AC 中。例如,00111110 00001001 这条指令中的操作码是 00111110,表示向累加器 AC 传送操作数,操作数是 00001001 (十进制数 9),这条指令的含义是把 9 送到累加器 AC 中。

2. 指令的基本类型

不同的指令系统,指令的数目和种类有所不同。按照功能,指令可以分为如下 4 类。

(1) **数据处理指令**用于对数据进行算术运算、逻辑代数运算、移位和比较等,包括算术运算指令、逻辑代数运算指令、移位指令、比较指令、其他专用指令等。

(2) **数据传送指令**用于把数据从计算机的某一部件传送到另一部件,但数据内容不变。这类指令包括存储器传送指令、内部传送指令、输入/输出传送指令、堆栈指令等。

(3) **程序控制指令**用于改变 PC 的内容、改变程序执行的顺序,包括无条件转移指令、条件转移指令、跳转子程序指令、暂停指令、空操作指令等。

(4) **状态管理指令**用于改变计算机中表示其工作状态的状态字或标志,但不改变程序执行的顺序。它不执行数据处理,只进行状态管理,如允许中断指令、屏蔽中断指令等。

3. 指令的控制执行

计算机指令的执行是由控制器来实现的,为了使计算机能自动工作,在控制器中设置了一个 PC,用来存放程序地址,并自动递增,使得自动取出和自动执行地址指令。控制器的主要任务是取指令、分析指令、执行指令、控制程序和数据的输入与结果输出、对异常情况 (Unexpected Exception) 和某些请求进行处理。指令执行流程如图 3-8 所示。

(1) **取指令**。以 PC 中的内容作为指令地址,并且每取出一条指令后,PC 中的内容自动加 1 ($PC=PC+1$),并将其作为下一条指令的地址。

(2) **分析指令**。对取出的指令进行译码,分析其功能。

(3) **执行指令**。以指令的地址码为内存地址 (指向数据区),取出操作数进行运算;如果是转移指令,将用指令的地址码部分取代 PC 的内容,从而实现按地址转移。

上述 3 个步骤周而复始,直至停机,这就实现了自动、连续、灵活地执行程序。通常把其中的一个循环 (取指令、分析指令、执行指令) 称为计算机的一个指令周期。这样可以把程序对计算机的控制归结为每个指令周期中指令对计算机的控制。

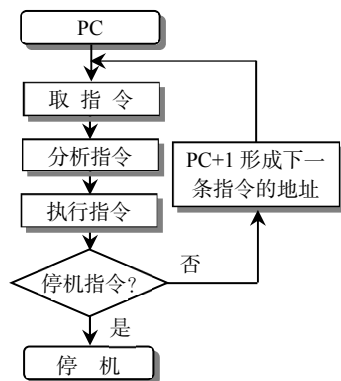


图 3-8 指令执行流程

这种以存储过程控制为核心,以 PC 为关键部件而自动连续执行的存储程序原理决定了 CPU 的体系结构,这就是沿用至今的著名的冯·诺依曼体系结构。

3.2.3 操作系统对 CPU 的管理

CPU 是计算系统中最为重要的硬件资源,如何对 CPU 进行资源分配和控制管理,最大限度地提高 CPU 的利用率,减少其空闲时间等,是操作系统的重要功能任务之一,可以把实现这一功能任务的管理称为 CPU 管理 (CPU Management)。为了实现对 CPU 的高效管理,现代操作系统对 CPU 的分配和运行是以进程为基本单位的,因而对 CPU 的管理主要可归结为进程管理,进程管理包括作业过程 (Operation Process)、进程控制 (Process Control)、进程同步 (Process Synchronization)、进程通信 (Process Communication)、CPU 调度 (CPU Scheduling)、线程控制等。

1. 作业过程

作业过程可被理解为程序执行工作任务的作业方式。计算机中的作业是指用户提交给计算机的工作任务的基本单位,主要用来描述批处理过程,由完成某一独立任务的程序及其所需数据组成。在早期的计算机中,只允许有一道(单道)用户程序在内存中运行,程序运行时完全控制了系统中的所有资源。在现代计算机中,允许有多道程序在内存中运行,从而使 CPU 的运行效率得到大幅度提高。多道程序的作业过程包含 3 个核心概念:进程、程序和执行。

(1) **进程**是为了描述程序执行时的特征而引入的概念,是从程序运行的执行方式到状态转换的过程抽象,是分配资源的基本单位,也是驻留在 RAM 中的作业。只要作业被装入 RAM 就成为进程,每个进程都是作业,而每个作业未必都是进程。我们可将其描述为

进程=程序+执行

(2) **程序**是问题求解步骤的命令集合,一个作业中至少包括一个业务处理程序。CPU 管理的实质是实现在多道程序环境下对 CPU 的分配和调度,使一个 CPU 能为多道程序交替服务,最大限度地提高 CPU 的利用率。

(3) **执行**是计算系统的一种工作状态,在程序运行时可分为两种执行方式。

① **顺序执行** (Sequential Execution) 是指在单道程序处理过程中每次只能有一道程序运行,并以顺序方式执行,且运行时独占全部资源。例如,对于有 3 条语句的程序段:

```
S1: x=a+b;  
S2: y=x+c;  
S3: z=y+d;
```

必须按照 S₁、S₂、S₃ 的先后顺序执行,即 S₂ 必须在 x 被赋值后才能执行, S₃ 必须在 y 被赋值后才能执行。显然,如果计算系统中任何时刻只能运行一道程序,则系统的处理能力无法提高。

如果用节点 I 代表输入操作——对应输入设备,用节点 C 代表计算操作——对应 CPU,用节点 P 代表输出操作——对应输出设备;用箭头表示前后操作顺序,那么,程序顺序执行的运行状态如图 3-9 所示,它表示两个或两个以上操作在计算系统中的某个部件按先后顺序进行。

② **并发执行** (Concurrent Execution) 是指在多道程序处理过程中,让多道程序或程序段同时运行。当某道程序因为某种原因无法继续运行而需要等待时,可让 CPU 去执行另一道可以运行的程序。程序并发执行的运行状态如图 3-10 所示,它表示两个或两个以上操作在计算系统中的不同部件同时进行,在执行时间上是重叠的,即一道程序或程序段的执行尚未结束,另一道程序或程序段的执行已经开始。

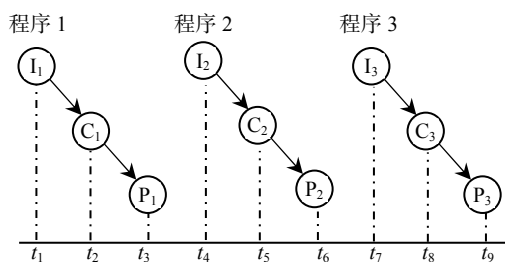


图 3-9 程序顺序执行的运行状态

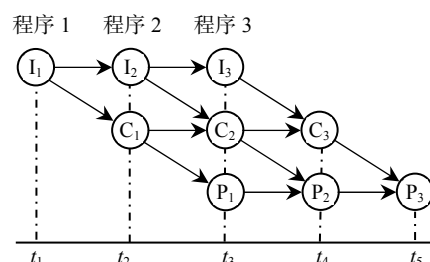


图 3-10 程序并发执行的运行状态

【提示】计算机中并发不同于并行。并发是指两个或多个事件在同一时间段内发生，并行是指两个或多个事件在同一时刻发生。此外，进程和程序有如下关联和区别。

- ◆ 进程是动态的，程序是静态的。进程是程序的一次执行过程，程序是一组代码的集合。
- ◆ 进程是暂时的，程序是永久的。进程有一个状态变化的过程，程序是计算机执行的步骤。
- ◆ 进程可以演化，程序不能演化。一个进程可以创建其他进程，一个程序不能形成新程序。
- ◆ 进程包含程序，一个程序可对应多个进程。一个进程的组成包括程序、数据和进程控制块。

2. 进程控制

进程控制是为解决程序并发问题提出来的。在多道程序环境下，在程序运行前必须为它创建一个或几个进程，并为之分配必要的资源。当进程运行结束时，立即撤销该进程，以便及时回收该进程所占用的各类资源，这一工作是由进程控制来实现的。进程控制的主要任务是为程序创建进程、撤销已结束进程、控制进程在运行过程中的状态转换。

【提示】一个程序被加载到内存中，系统就创建了一个进程，它是一个程序在其自身的地址空间中的一次执行活动。程序执行完毕，该进程也就结束了。当一个程序同时被执行多次时，系统就创建多个进程。一个程序可以被多个进程执行，一个进程也可以同时执行一个或几个进程。在 Windows 操作系统中，按下 [Ctrl] + [Alt] + [Delete] 组合键，即可看到机器当前执行的进程。

3. 进程同步

进程同步是为了保证相互有关的进程能正常运行而设置的机制，即对存在制约关系的多个进程的运行进行协调，并且分为同步方式（Synchronous Mode）和互斥方式（Mutually Exclusive Mode）。

（1）**同步方式**是指两个并发执行的进程为共同完成一个任务而相互配合的进程间通信。例如，设有一个计算进程 P_1 和一个打印进程 P_2 ，为使计算结果打印输出，需设置一个输出缓冲区。当进程 P_1 计算完后，立即将结果送入输出缓冲区，此时进程 P_2 便可以从该缓冲区中取出结果，并进行打印。进程 P_1 和 P_2 的这种协同动作，保证了“计算—打印”工作的顺序进行。

（2）**互斥方式**是指两个并发执行的进程在同一时刻要求共享同一资源而相互排斥。例如，设有两个进程 P_1 和 P_2 ，在某一时刻同时要求使用一台打印机，打印机一次只能让一个进程使用。此时，进程 P_1 和 P_2 是互斥的，操作系统只能让其中一个进程（如 P_1 ）先使用打印机，让另一个进程（ P_2 ）处于“挂起”状态，一旦进程 P_1 结束打印，释放该资源，就立即“唤醒”进程 P_2 ，使进程 P_2 使用该打印机。这种一次只能允许一个进程使用的资源称为临界资源（Critical Resource），而把每个进程中访问临界资源的那段程序称为临界区（Critical Section）。

【提示】在多道程序环境下，进程同步问题十分重要，因此引发了不少学者对它进行研究。例如，生产者—消费者问题、哲学家共餐问题、读—写问题（Reader-Writer Problem）等都是典型的进程同步问题。

4. 进程通信

进程通信用来实现相互合作的进程之间的信息交换。在多道程序环境下可由系统为一个应用程序建立多个进程，这些进程相互合作，共同完成一个任务，并且在活动过程中彼此间存在相互依赖或者相互制约的关系，因此存在进程的通信问题。例如，在输入进程、计算进程和打印进程的活动中，输入进程将输入数据传送给计算进程；计算进程对输入数据进行计算，并把计算结果传送给打印进程；打印进程把计算结果打印出来。

5. CPU 调度

CPU 调度用来为并发执行的多个进程分配 CPU 资源。换句话说，在多道程序系统中，一个作业提交后必须经 CPU 调度后才能执行。通常情况下，CPU 调度可分为三级。

(1) **作业调度** (Job Scheduling) 也称为高级调度，其基本任务是从存储在外存中的后备作业队列中，按照一定的算法选择若干作业调入内存并准备执行。例如，用户要求计算机把编好的程序进行编译、连接并执行就是一个作业，作业的运行是以进程方式实现的。作业调度是多道批处理操作系统的重要功能，现代操作系统管理下的多道程序是并发执行的，因此不再需要进行作业调度。

(2) **交换调度** (Swapping Scheduling) 也称为中级调度，其基本任务是根据进程的当前状态决定外存和内存的进程交换。当内存容量不足时，将暂时不执行的进程从内存调至外存中（虚拟存储），而将需要执行的进程调入内存。显然，这种方式能极大地提高内存资源的利用率。

(3) **进程调度** (Process Scheduling) 也称为低级调度，其基本任务是从进程的就绪队列中按照一定的算法选出一个进程，然后把 CPU 分配给它，并为其设置运行环境，使该进程进入运行状态。因此，进程调度是操作系统 CPU 管理中最重要、最频繁的调度。

§ 3.3 存储器系统——“存储优化组合”的基本思维

现代计算机是以存储器为中心的，所有数据和程序都存储在存储器中，所以人们总是希望存储器的存储容量越大越好，存储速度越快越好，存储时间越久越好，存储器的价格越低越好。然而，没有哪种存储器能够同时满足存储容量大、存储速度快、存储时间长、价格低的要求。为此，可以将内存和外存虚拟结合，以扩大存储容量和降低成本，将不同性能的存储器进行优化组合，构成一个高性能的存储器系统，以满足应用需求，且均由操作系统实现高效管理。

3.3.1 内存

内存是与 CPU 直接相连并协同工作的存储器，包括 ROM 和 RAM。RAM 与 CPU 是计算机中最宝贵的硬件资源，是决定计算机性能的重要因素。

1. ROM

ROM 是一种只能读出存储器内的信息，而不能写入信息的存储器，用来存放系统的引导、检测、诊断、设置等程序。ROM 中的程序内容是在生产制作 ROM 时预先设定好的，是通过一次性固化得到的，即使停电或关机，ROM 中的信息也不会丢失。

计算机启动后，自动加载 Flash ROM 内的 BIOS (Basic Input/Output System, 基本输入输出系统) 固件程序；该程序执行上电自检 (Power-On Self-Test, POST)，检测内存容量并校验各类硬件设备的状态。当检测到存在硬件故障时，通过蜂鸣器发出对应的报警提示。用户可按需进入 BIOS 界面修改硬件配置参数，该步骤并非开机必须执行的流程。自检完成后，BIOS 读取磁盘引

导程序, 转交系统控制权。系统加载完成后, ROM 中的 BIOS 并未彻底失效, 计算机运行底层硬件操作时仍会调用 BIOS 中内置的中断服务。

2. RAM

RAM 是由半导体器件组成的可以随时写入和读出其内容的存储器, 用来存放当前使用的操作系统、应用软件、计算程序、运行数据、交换信息等。为了实现高效存取, 通常按地址写入和读出数据。RAM 的基本组成如图 3-11 所示。

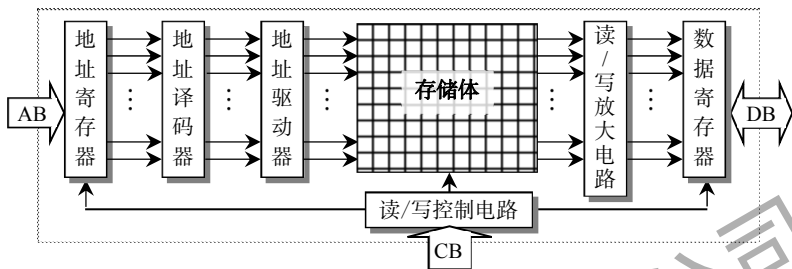


图 3-11 RAM 的基本组成

其中, 存储体 (Memory Bank) 由存储单元 (Memory Location) 组成, 每个存储单元包括若干存储元 (Memory Cell), 每个存储元可存储 1 位二进制数, 全部存储单元的总位数 (或总字节数) 为 RAM 存储容量。RAM 接收来自地址总线 (Address Bus, AB) 的信息, 经地址寄存器、地址译码器、地址驱动器进入存储体; 然后经读/写放大电路、数据寄存器与数据总线 (Data Bus, DB) 相连; 读/写控制电路与控制总线 (Control Bus, CB) 相连。

与 ROM 相比, RAM 的读/写有三个特点: 一是可以随机读/写, 且读/写操作所需时间相同; 二是读出时原有内容不会被破坏, 写入时被写存储元中的原有内容被所写内容替代; 三是只能用来临时存储信息, 断电后 RAM 中保存的信息会立即丢失。这三个特点是 RAM 与 ROM 的重要区别。

【提示】虽然 RAM 和 ROM 都属于内存, 但对用户而言, 只有 RAM 才是可用的存储空间, 所以常把 RAM 称为主存。通常所说的内存容量实际上是指 RAM 容量。在计算机运行过程中, 所有参与处理的数据信息和控制器执行各种操作的指令信息都来自 RAM。RAM 容量越大, 从外存中一次读入 RAM 的数据量越大, 计算机的运行速度就越快。

3. 存储单位

存储单位 (Storage Unit) 用来表示存储容量的大小。计算机中的数据信息是以二进制数的形式存储的, 所以存储单位是指数据存放时占用的二进制位数, 常用的存储单位有位 (bit)、字节 (Byte) 和字 (Word)。

(1) **位**。位是计算机中存储数据的最小单位, 用来存放一位二进制数 (0 或 1)。一个二进制位只能表示 $2^1=2$ 种状态, 若要表示更多的信息, 就需要组合多个二进制位。

(2) **字节**。字节是计算机中的一个存储单元, ASCII 码中的英文字母、阿拉伯数字、特殊符号和专用符号有 128 个, 刚好可以用 8 个二进制位 (1 字节) 表示。计算机中表示存储容量时通常用 KB、MB、GB、TB、PB、EB 等计量单位, 换算关系如下:

$$\begin{aligned} 1 \text{ KB} &= 1024 \text{ B} = 2^{10} \text{ B} & 1 \text{ MB} &= 1024 \text{ KB} = 2^{20} \text{ B} & 1 \text{ GB} &= 1024 \text{ MB} = 2^{30} \text{ B} \\ 1 \text{ TB} &= 1024 \text{ GB} = 2^{40} \text{ B} & 1 \text{ PB} &= 1024 \text{ TB} = 2^{50} \text{ B} & 1 \text{ EB} &= 1024 \text{ PB} = 2^{60} \text{ B} \end{aligned}$$

(3) **字**。计算机在存储、传送或操作时, 作为一个数据单位的一组二进制位称为一个计算机字, 简称“字”, 每个字所包含的位数称为字长。一个字由若干字节组成, 而字节是计算机进行数据处理和数据存储的基本单位, 所以字长通常是字节的整数倍。

3.3.2 外存

外存是相对主存而言的，所以又称为辅助存储器（Auxiliary Storage），用来存储当前不参与运行的程序和数据。与 RAM 相比，外存的存储容量较大、价格较低、速度较慢，并且不能直接与 CPU 相连。由于它是一种磁质存储器，因而能永久保存磁盘中的信息。目前常用的外存有硬盘、光盘、U 盘、移动硬盘（Mobile Hard Disk）等。在大型计算机和视频设备中，还有磁带，有利于大数据的存储、复制和保存。

1. 硬盘

硬盘又称为硬盘存储器，是最主要的辅助存储媒介，是目前存储大量程序和数据的理想设备，因而计算系统中都配置了硬盘，它是所有数据文件的“栖身”之地。

硬盘是相对于软盘（Soft Disk，又称为软盘存储器）而言的。在 U 盘和移动硬盘出现之前，软盘曾经是应用最广的外存设备。现在虽已被淘汰，但其概念是硬盘的基础。磁盘中的盘片是用类似薄膜唱片的柔性材料制成的，盘片外形如图 3-12 所示。

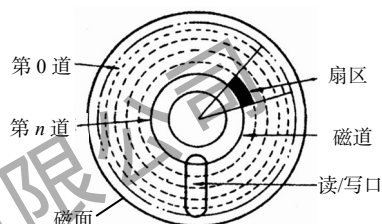


图 3-12 盘片外形

（1）**磁道（Track）**是指以盘片中心为圆心的一组同心圆，每一圆周为一个磁道。磁道的编号从 0 开始，若磁道数为 80，则数据在磁道中的存放顺序为 0→79。

（2）**扇区（Sector）**是指磁盘划分的区域，是存储数据的基本单位。计算机进行数据读/写时，无论数据多少，总是读/写一个完整的扇区或几个扇区。因此，一个扇区又称为一个记录。

（3）**柱面（Cylinder）**是指直线沿着一条定曲线平行移动所形成的曲面，由磁面上相同位置的磁道构成。一个磁盘通常有多个盘片，每个盘片中有多个磁道。

（4）**磁头（Head）**是指通过磁性原理读/写磁盘数据的部件，可以采用 CHS（Cylinder-Head-Sector）作为寻址方式，或者对扇区采用逻辑块寻址（Logical Block Addressing, LBA）方式。

（5）**容量（Capacity）**是指磁盘能存储的数据字节总数，是在对磁盘进行格式化时形成的。所谓格式化，就是对磁盘按磁道数和扇区数进行划分，并写入地址码、识别码等信息。向磁盘中写入数据或从磁盘中读取数据时，就是按照区域信息表（文件表）进行操作的，这样有利于快速存取。

硬盘和软盘都属于磁表面存储器，具有同样的存储特性，我们可把硬盘看作由盘片叠加而成，如图 3-13 所示。

硬盘通常采用温切斯特技术（Winchester Technology），故称为温氏硬盘，由一个或多个不可更换的盘片作为存储介质组成。这些盘片外覆盖有铁磁性材料，硬盘盘片与其驱动器合二为一。绝大多数硬盘都是固定硬盘，被永久性地密封固定在硬盘驱动器中，所以硬盘机被称为硬盘驱动器（Hard Disk Drive, HDD）或固定磁盘（Fixed Disk）。除此之外，还有可更换盘片的硬盘，但个人计算机中使用的是固定硬盘。

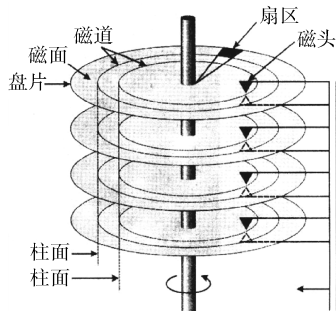


图 3-13 硬盘的组成结构

2. 光盘

光盘是继记录存储器之后出现的一种存储设备，由于它利用激光在磁性介质上存储信息，因此又称为光存储器。光盘是一种记录密度高、存储容量大的外存设备，广泛用于存储文字、图形、图像、语音等多元信息。目前，绝大多数光盘只能读不能

写。随着更优存储器的出现,光盘将慢慢退出“历史舞台”。

3. U 盘

U 盘是使用 USB 接口而无须物理驱动器的微型大容量移动存储设备,通过 USB 接口与个人计算机连接,可实现即插即用 (Plug and Play, PnP),因而得到了广泛应用。与硬盘和光盘相比,U 盘不仅体积小、质量轻、便于携带,而且具有防潮防磁、耐高低温、抗振性能强(无任何机械装置)等特性。近年来生产的 U 盘容量越来越大,并具有多种适用类型,已成为计算机工作者随身携带的普适工具。

4. 移动硬盘

移动硬盘具有存储容量大、便于携带、使用方便等优点。使用移动硬盘时,只需使用一根 USB 接口线便可将其连接在 USB 接口上,并且即插即用。因此,移动硬盘具有许多与 U 盘相似的优良特性。

3.3.3 存储体系

为了使计算机能自动、高速运行,除采用存储程序控制外,还需要采取相应的技术措施:一是协调 RAM 与 CPU 之间的运行速度;二是动态地组织存储空间;三是充分利用各种存储器的性能特点并进行优化组合,构成一个存储体系,通过彼此协调工作,提高计算系统的整体性能。

1. 高速缓冲存储器

高速缓冲存储器 (Cache) 是一种在 RAM 与 CPU 间起缓冲作用的存储器。CPU 在执行程序时,总是按指令地址或操作数地址访问主存。理论上,主存的读/写速度与 CPU 的工作速度不仅越快越好,两者的速度还应该相匹配。然而,现代计算系统对 CPU 性能的要求越来越高,要求主存容量越来越大。如果用同 CPU 一样性能的材料制作主存,则会使计算机的成本大幅度提高。因此,一般通用计算机的主存速度较 CPU 速度低,CPU 的时钟频率远远超过了主存的响应速度,故使得 CPU 的执行速度受主存的限制而不能充分发挥自身的快速作用,从而降低了计算机整体的运行速度。为了协调主存与 CPU 在速度上的差异,目前解决这个问题的最有效办法是采用高速缓冲技术,即在 CPU 与主存两者之间增加一级在速度上与 CPU 相等,在功能上与主存相同的高速缓冲存储器,在两个工作速度不同的部件之间交换信息的过程中起缓冲(协调)作用。高速缓冲存储器的引入既提高了系统性能,又保证了造价低廉。现在个人计算机内部一般都含有高速缓冲存储器,否则难以高速运行。

高速缓冲存储器的容量通常为 512 KB 左右,并嵌入 CPU。计算机开始运行后,将当前正要执行的一部分程序批量地从主存复制到高速缓冲存储器,CPU 读取指令时,先到高速缓冲存储器中查找。若在高速缓冲存储器中找到相关程序,则直接从高速缓冲存储器中读取(称为命中),否则从主存中读取。

2. 虚拟存储器

虚拟存储器 (Virtual Memory) 不是一个实际的物理存储器,而是建立在主-辅层次结构上,由主存、外存和操作系统存储管理软件组成的存储体系。高速缓冲存储器的引入相当于提高了主存的速度。对整体而言,不仅要求主存速度快,还要求存储容量大,但主存容量毕竟是有限的。为了解决这一供需矛盾,现代操作系统中普遍采用虚拟存储技术。虚拟存储的基本思想是在装入程序时,只将当前需要执行的内容装入内存,暂时不用的其余部分保留在外存中;在程序执行过程中,如果需要用到的数据不在内存中,则由操作系统从外存上将其调入内存,从而使用户可以

使用一个比实际内存容量大得多的“虚拟存储”空间。

虚拟存储器的实现有三个基本要素：一是有一定的内存容量，能够存放基本程序和数据；二是有足够的外存空间，能够存放多个用户程序；三是有地址变换机构。

【提示】虚拟存储器的最大容量与 CPU 的寻址能力有关，例如，CPU 的地址线是 20 位的，则虚拟存储器的存储容量最多是 1 MB。Pentium 芯片的地址线是 32 位的，所以虚拟存储器的存储容量可达 4 GB。硬盘通常采用温切斯特（Winchester）技术，因此也被称作温氏硬盘，其以一片或多片不可由用户自行更换的盘片作为存储载体，盘片表面涂覆铁磁性存储介质，盘片与读/写驱动机构一体化封装。这类一体化封装设备统称为硬盘驱动器或固定磁盘。历史上曾出现过盘片可拆卸的磁盘设备，但个人计算机配套的硬盘均为固定磁盘结构。进行虚拟存储时，动态实现存储过程中逻辑地址到物理地址的变换，也称为地址映射。

3. 存储体系结构

计算机中的存储器包括内存（主存 RAM 和 ROM）、外存和高速缓冲存储器，它们各有功能特点：主存容量较大，运行速度较快，用来在计算机运行时存放操作系统和其他程序代码；外存是相对主存而言的，主要指硬盘，其存储容量最大，速度慢，用来存储各种程序和数据；高速缓冲存储器的存储容量最小，运行速度最快，用来协调主存与 CPU 的速度。为了充分利用三种存储器各自的特点，“存储体系”应运而生，采用“高速缓冲存储器+主存+硬盘”的三级存储体系来解决存储容量和运行速度之间的矛盾。三级存储体系的逻辑结构如图 3-14 所示，其对应的层次结构如图 3-15 所示，三级存储器的基本特性如表 3-2 所示。

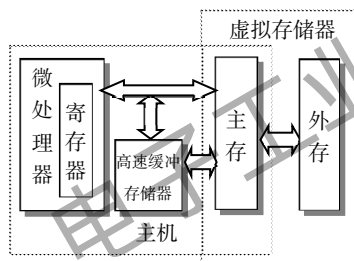


图 3-14 三级存储体系的逻辑结构

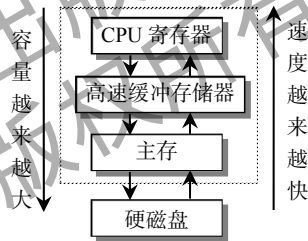


图 3-15 三级存储体系的层次结构

表 3-2 三级存储器的基本特性

性能	类别		
	高速缓冲存储器	主存	硬盘
容量	小	中	大
速度	最快	中等	最慢
价格	最高	中等	最低

三级存储体系既能满足速度、容量要求，又具有良好的性能价格比，因此其已成为现代计算系统中普遍采用的存储体系结构模式。三级存储体系由操作系统实施调配和协调。

3.3.4 操作系统对存储器的管理

计算机运行时，凡是经过 CPU 处理的所有程序和数据都需要先存储在内存（RAM）中。然而，内存的容量是极为有限的，因而这便成为制约计算机高效、快速、可靠运行的技术瓶颈，如何突破这个瓶颈是操作系统的一项重要任务。操作系统对内存的管理可概括为以下 4 个方面。

1. 内存分配管理

内存分配管理（Memory Allocation Management）的主要任务：一是为了提高内存的利用率，最大限度地减少不可用的内存空间；二是当有存储申请时，根据需要选定分配区域；三是当用户作业不再使用某个区域时，及时将其收回。根据存储形式，内存分配管理可分为 3 种基本方式。

（1）连续存储管理（Continuous Storage Management）又称为单一连续分配管理，是指将内存分为两个连续存储区域，其中一个存储区域固定分配给操作系统使用，另一个存储区域分配给用

户作业使用。通常,用户作业只占用所分配存储区域的一部分,剩下的一部分存储区域实际上被浪费了。例如,在容量为 256 KB 的内存中,操作系统占用 32 KB,剩下的 224 KB 全部分配给用户作业,如果用户作业仅需要 64 KB,就有 160 KB 的存储空间被浪费。连续存储管理如图 3-16 所示。

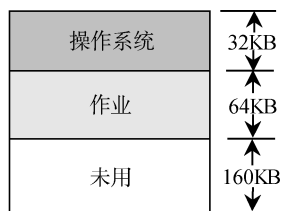


图 3-16 连续存储管理

连续存储管理是静态分配方式,只有等到作业结束后才能释放内存,并且同一时刻只能运行一道作业,因而不能使资源得到充分利用。

(2) **分区存储管理** (Partition Storage Management) 是满足多道程序设计需要的一种最简单的存储管理方式,它将内存划分为若干分区。分区的大小可以相等,也可以不等,除操作系统占用一个分区之外,其余的每个分区可以容纳一个用户作业。按照分区数目的变化,可以进一步划分为固定分区分配 (Fixed Partition Allocation) 和动态分区分配 (Dynamic Partition Allocation)。

① **固定分区分配**是指将内存的用户区划分成若干大小不等的区域,每个进程占据一个或多个区域,从而实现多道程序设计环境下各并发进程对内存空间的共享。为了便于分配管理,需要建立分区表,记录分区号、分区大小、起始地址和状态信息。当某个用户程序被装入内存时,需要为它分配存储空间,操作系统根据分区表中的信息,找出一个足够大的尚未分配的分区分配给它,然后将该进程调入内存。

如图 3-17 所示的固定分区表将内存划分为 4 个固定分区,并给出了各分区的基本信息。图 3-18 给出了当前 3 个进程 A、B、C 所占内存空间的分配情况。由于作业的大小并不一定与某个分区大小相等,因此绝大多数已分配的分区中都有一部分存储空间被浪费,即固定分区分配方式下的内存仍然不能得到充分利用。

② **动态分区分配**是指根据程序的实际需要分配内存空间,使分区大小正好满足作业需要。为了实现动态分区,系统中需要配置一张空闲分区表,以记录每个空闲分区的情况,包括分区号、起始地址、分区大小及状态等信息,如图 3-19 所示。

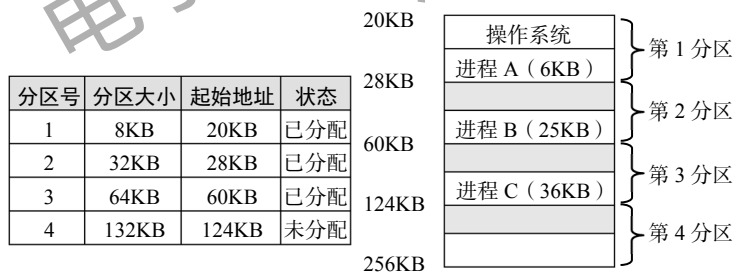


图 3-17 固定分区表

图 3-18 固定分区内存状态

分区号	分区大小	起始地址	状态
1	32KB	352KB	空闲
2	...	...	空表目
3	520KB	505KB	空闲
4	...	...	空表目

图 3-19 空闲分区表

为了实现对空闲分区的分配和链接,在每个分区的起始部分设置用于控制分区分布的信息以及用于链接各分区的前向指针,将内存的空闲分区链接起来,构成如图 3-20 所示的空闲分区链。

分配内存时,从分区中按请求的大小划分一块内存空间,余下的部分留在空闲分区链中,然后将分配区的首地址返回给调用者。当进程运行完毕释放内存时,系统根据回收区的首地址,从空闲分区链中找到相应的插入点,将回收区插入适当的位置。

分区存储管理克服了连续存储管理的不足,但将作业存放在一片连续的存储区域中会产生内存碎片。为此,可将一个作业存放到多个不相邻的存储区域中。

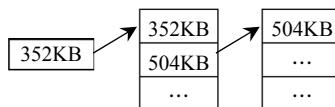


图 3-20 空闲分区链

(3) **分页存储管理** (Paging Storage Management) 是指将一个作业 (进程) 直接分配到许多

互不相邻的分区，因而又称为离散分配方式。根据分配单位不同，分页存储管理可分为以下 3 种方式。

① **页式存储管理** (Page Storage Management) 是指把用户程序的地址空间划分成若干固定大小的区域，称为“页”，页的典型大小为 1 KB；相应地，也将物理存储空间分成若干物理块，页和块的大小相等。这样，用户程序的任意一页可以放在内存的任意一块中，实现离散分配。

② **段式存储管理** (Segmentation Storage Management) 是指把用户程序的地址空间分成若干大小不等的段，每个段定义一组相对完整的逻辑信息，在内存中可以不相邻，以实现离散分配。UNIX 操作系统采用的就是段式存储管理。

③ **段页式存储管理** (Segment-Page Storage Management) 是指先把程序的地址空间划分成“段”，再将段划分成“页”，每页包含物理空间中的一个或者多个物理块，段内的页可以存储在不相邻的物理块中。这样既提高了存储器的利用率，又能满足用户要求，是目前广泛使用的一种存储管理方式。

2. 内存保护管理

内存保护管理 (Memory Protection Management) 是为确保每道程序都只能在自己的内存空间中运行而建立的内存保护机制。现代操作系统可实现多道程序管理，在多道程序设计环境下，内存中的许多用户程序或系统程序和数据段可供不同的用户进程共享，因而极大地提高了内存的利用率。内存保护的主要任务是绝不允许用户程序访问操作系统的程序和数据，也不允许转移到非共享的其他用户程序中去执行，各进程不能对其他进程的程序和数据段产生干扰和破坏。

3. 内存扩充管理

内存扩充管理 (Memory Expansion Management) 是解决内存容量不足的技术措施，如虚拟存储。操作系统把暂时不用的程序和数据保留在外存中，同时将内存中暂时不用的数据信息调到外存中去，从而腾出空间存放将要调入内存的数据信息。这种调入和调出，是由操作系统自动完成的，它使用户能够在较小的内存中执行较大的用户程序，从而能够在内存中容纳更多的程序并发执行。虚拟存储技术通常采用分页虚拟存储、分段虚拟存储和段页式虚拟存储。

4. 地址映射管理

地址映射管理 (Address Mapping Management) 也称为地址重定位。一个源程序编译后通常会形成若干目标程序，它们经过连接形成可执行程序。这些程序的起始地址从“0”开始，程序中的其他地址都是相对于起始地址计算的，称为“逻辑地址”或“相对地址”，把内存中的一系列存储单元所限定的地址范围称为“内存空间”，其中的地址称为“物理地址”。在多道程序设计环境下，每道程序不可能都从“0”地址装入内存，这就致使地址空间中的逻辑地址和内存空间中的物理地址不一致。为了使程序能正确运行，必须将地址空间中的逻辑地址变换为内存空间中与之对应的物理地址，从而实现地址映射。这个映射是由软件和硬件共同实现的，并且映射过程较为复杂。

3.3.5 典型案例——计算机的工作过程

现代计算机的基本工作原理仍然遵循冯·诺依曼提出的“存储程序控制”理论，根据存储在存储器中的程序，由控制器发出控制指令，指挥各有关部件有条不紊地执行各项操作。

1. 存储地址与指令

下面我们以计算 $y=5+9$ 且将结果放在累加器中为例，描述计算机的工作过程。为了便于理解，我们用专用助记符和操作代码来描述计算 $y=5+9$ 的工作过程。例如，MOV 表示数据传送指

令, ADD 表示加法指令, HALT 表示停机指令等。对于一个具体问题, 各种计算机都有自己的指令表。计算机计算 $y=5+9$ 的汇编程序, 可用以下三条语句来实现。

MOV	A, 05H	; 将立即数 05H 送至累加器 A 中
ADD	A, 09H	; 将立即数 09H 加到累加器 A 中
HALT		; 停机

该汇编语言程序通过汇编 (编译) 形成由 0 和 1 组成的机器语言程序, 并且在编译过程中形成地址空间及操作指令。上述程序的三条操作指令及其在 RAM 中存储地址的对应关系如图 3-21 所示。

存储地址	RAM 中内容	指令助记符
40H	0011 1110	MOV A,05H
41H	0000 0101	
42H	1100 0110	ADD A,09H
43H	0000 1001	
44H	1111 0100	HALT

图 3-21 上述程序的三条操作指令及其在 RAM 中存储地址的对应关系

计算机工作时, 把上述经过编译形成的程序由外存调入 RAM, 并以二进制数的形式按区域连续存储, 这种连续存储的单元编号称为存储地址 (Memory Address)。RAM 中存储的内容为操作指令, 操作指令和操作数各占一个存储单元 (字节), 上述三条指令占据 5 个存储单元。假如把它们存储在以 40H 地址开始的 5 个连续单元中, 则相应的存储地址为 40H、41H、42H、43H、44H。

2. 指令的执行顺序

程序输入计算机后, 只要告诉计算机程序的起始地址 (这里是 40H), 并发出一个启动指令, 计算机就会执行这段程序, 执行的过程是反复进行取出指令 (取指) 和执行指令这两个基本操作。

第一条指令取指阶段的基本操作如图 3-22 所示, 其他阶段指令操作按照表 3-3 所示的步骤执行。

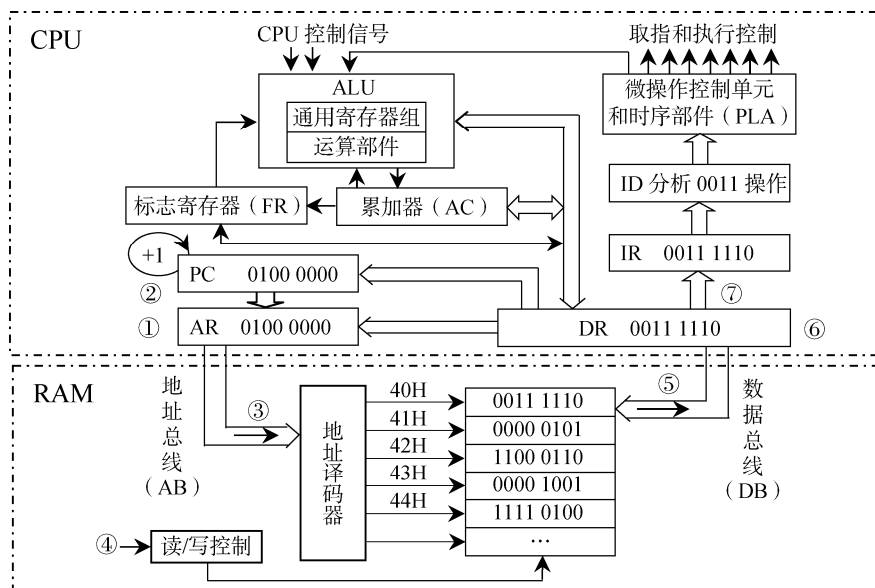


图 3-22 第一条指令取指阶段的基本操作

表 3-3 指令的执行步骤

步骤	第一条指令取指阶段	第一条指令执行阶段	第二条指令取指阶段	第二条指令执行阶段
①	将 PC=40H 送入 AR, 使 AR=40H	把 PC 内容 41H 送入 AR	把 PC 内容 42H 送入 AR	把 PC 内容 43H 送入 AR
②	PC 加 1 变为 41H	PC 自动加 1, 变为 42H	PC 自动加 1, 变为 43H	PC 自动加 1, 变为 44H
③	AR 把地址 40H 通过 AB 送入存储器, 经地址译码器译码选中 40H 号单元	AR 把 41H 通过 AB 送入存储器, 经地址译码器译码选中 41H 号单元	AR 把 42H 通过 AB 送入存储器, 经地址译码器译码选中 42H 号单元	AR 把 43H 通过 AB 送入存储器, 经地址译码器译码选中 43H 号单元
④	CPU 发出读命令	CPU 发出读命令	CPU 发出读命令	CPU 发出读命令
⑤	将 40H 号单元内容 3EH 读入 DB	将 41H 号单元内容 05H 读入 DB	将 42H 号单元内容 C6H 读入 DB	将 43H 号单元内容 09H 读入 DB
⑥	读出的内容经过 DB 送入 DBR	读出的内容经过 DB 送入 DBR	读出的内容经过 DB 送入 DBR	读出的内容经过 DB 送入 DBR
⑦	DBR 把指令操作码送入 IR, 经过 ID 和 PLA 发出微操作命令, 见图 3-22	05H 由 DBR 通过 DB 送入 AC, 按图 3-22 改变相关值	取指阶段读出指令, 与第一条指令取指相同	09H 由 DBR 通过 DB 送入 ALU 的另一个输入端

第二条指令执行阶段, AC 中内容与 ALU 中的内容相加, 由 ALU 输出到 AC 中。至此, 第二条指令执行完毕, 转入第三条指令取指阶段。按照上述过程取出第三条指令并经过译码后, 控制器停止产生控制信号而停机。程序执行完毕, $y=5+9$ 的计算任务就此完成, 累加器 AC 中存储运算结果。

3. 虚拟存储的使用

前面介绍了运行一个简单程序时 CPU 与 RAM 的基本工作情况。当运行复杂程序时, 操作系统可采用分页、分段或段页式虚拟存储调度机制, 仅将程序当前所需部分调入内存, 其余部分暂存在外存。例如, 物理内存容量为 10MB, 同时运行 10 个程序, 每个程序逻辑占用 3MB, 全部程序的总逻辑地址需求为 30MB。任意时刻仅有合计 10MB 的程序内容驻留在内存中, 剩余未使用的程序页/段存放于磁盘, 系统通过页、段调度按需换入/换出。该机制即为虚拟存储技术, 现代操作系统普遍采用该技术, 以此突破物理内存容量的限制。

§ 3.4 文件管理系统——“长治久安”的基本思维

3.3.4 节中介绍了操作系统对 RAM 的管理, 由于 RAM 的容量有限, 且不能长久保存, 因而需要将程序和数据以文件形式存放在外存上, 计算机工作时随时将它们调入内存。为了管理外存上的文件, 操作系统中设置了管理文件的功能模块——文件管理系统, 简称文件系统, 它负责管理文件, 并为用户提供对文件进行存取、共享及保护的手段, 这不仅方便了用户, 保证了文件的安全性, 还可以有效地提高系统资源利用率, 实现对磁盘文件的“长治久安”。

3.4.1 文件与文件系统

操作系统对文件进行管理的实质是对存放文件的磁盘存储空间进行管理。具体地说, 就是当建立一个新的文件时, 文件系统要为其分配相应的存储空间; 删除一个文件时, 要及时收回其所占用的空间。

1. 文件

文件是指逻辑上具有完整意义的信息集合,计算机中所有的程序和数据都是以文件的形式进行存储和管理的。文件通常由若干记录组成,而记录是一些相关数据项的集合,其数据项是数据组织中可以命名的最小逻辑单位。例如,一位教师的信息记录由姓名、性别、籍贯、出生日期、学历、职称等数据项组成,而一个单位职工信息记录的集合就组成了一个文件。文件、记录、数据项的关系如图 3-23 所示。

在计算系统中,一个完整的程序或一组完整的数据记录都是一个文件,它是操作系统进行信息管理的基本单位。文件中记录的符号元素不仅可以是文字、公式、表格、插图信息,还可以是视频信息和音频信息等,而承载这些信息的媒介是外存(各类磁盘),磁盘中的文件称为磁盘文件。

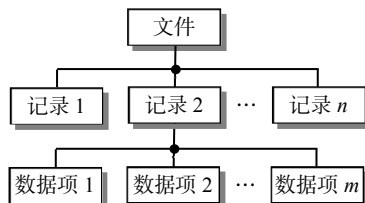


图 3-23 文件、记录、数据项的关系

2. 文件系统

文件系统是操作系统中与文件管理有关的软件和数据结合,是操作系统中负责存取和管理信息的模块,是对文件的存储空间进行组织和分配,负责文件的存储,并对存入文件进行保护和检索的系统。它用统一的方式管理用户和系统信息的存储、检索、更新、共享和保护,并为用户提供一套高效的文件使用方法,是对文件进行组织和管理的层次抽象。

文件系统由 3 部分组成:与文件管理有关的软件;被管理的文件;实施管理所需的数据结构。其层次结构如图 3-24 所示。

(1) **逻辑文件系统层**又称为输入/输出文件系统底层,用来处理文件及记录的相关操作,允许用户利用磁盘文件名访问文件及文件中的记录,实现对文件及记录的保护、目录操作等。

(2) **文件组织模块层**又称为基本输入/输出管理程序层,用来完成与磁盘输入/输出有关的工作,包括选择文件所在设备,进行文件逻辑块号到物理块号的转换,文件空闲存储空间管理,指定输入/输出缓冲区。

(3) **基本文件系统层**又称为物理输入/输出层,用来处理主存和外存之间的数据块交换,只关心数据块在外存和主存缓冲区中的位置,无须了解所传送数据块的内容或文件结构。

(4) **设备驱动程序层**又称为输入/输出控制层,由磁盘驱动程序组成,用来启动设备输入/输出操作及对设备发来的中断信号进行处理。

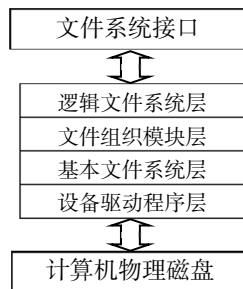


图 3-24 文件系统的层次结构

3.4.2 文件组织与结构

文件组织 (File Organization) 是指根据用户和系统管理的需要,采用不同的存储方式来组织文件。将一个文件存储到外存中,涉及文件逻辑结构 (File Logical Structure) 和文件物理结构 (File Physical Structure)。

1. 文件逻辑结构

文件逻辑结构是指用户能够观察到的文件组织形式,也是用户可以直接处理的数据及其结构,独立于文件的物理特性。文件逻辑结构可分为流式文件 (Streaming File) 和记录式文件 (Record File)。

(1) **流式文件**。用户对流式文件内的信息不再划分信息单位, 整个文件由顺序的一串信息组成, 如 Word 文件、图片文件等。

(2) **记录式文件**。用户对记录式文件内的信息可以按逻辑上独立的含义再划分信息单位, 每个单位为一个逻辑记录, 每个记录可以独立存取, 如班级成绩单文件, 其中每个学生的各科成绩就是一条记录。

2. 文件物理结构

文件物理结构是指文件在磁盘上的组织形式, 又称为文件的存储结构。常用的文件物理结构有顺序文件 (Sequential File)、链接文件 (Link File)、索引文件 (Index File) 和哈希文件 (Hash File)。

(1) **顺序文件**是指按其逻辑记录顺序依次存入存储介质而建立的文件组织形式, 顺序文件的读取从第一个记录开始。顺序文件的存储结构如图 3-25 所示。

顺序文件的优点是管理简单、存取速度快; 缺点是增删困难, 磁盘存储空间的利用率不高, 因此只适用于长度不变的只读文件。

(2) **链接文件**是指把各个逻辑记录任意存储到一些磁盘块中, 用指针将各个块按逻辑记录的顺序 (物理块号、链接指针和逻辑块号) 链接起来, 在文件目录中只记录第一个块和最后一个块的地址。链接文件的存储结构如图 3-26 所示。

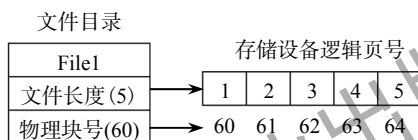


图 3-25 顺序文件的存储结构

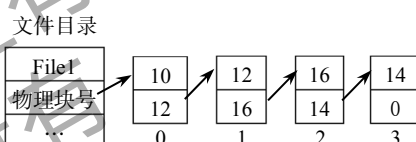


图 3-26 链接文件的存储结构

链接文件的优点是文件长度可以动态增长, 容易插入和删除逻辑记录; 缺点是只能按块队列中的串联指针顺序搜索, 只适用于逻辑上连续的文件。因此, 链接文件存储结构的应用存在较大的局限性。

(3) **索引文件**是指为每个文件建立一个索引表, 在每个索引表中指出文件的逻辑块号与物理块号的对应关系的组织形式, 其存储结构如图 3-27 所示。

索引文件的优点是既可以满足文件动态增长的要求, 又能实现随机存取, 适用于对信息处理的及时性要求较高的场合, 如飞机订票系统; 缺点是访问索引的过程降低了存取速度。

(4) **哈希文件**为了实现文件存储空间的动态分配, 利用 Hash 函数将记录关键字转换为相应记录的地址。Hash 函数所求得的并非相应记录的地址, 而是指向目录表的指针, 该目录表的内容指向相应记录所在的物理块, 如图 3-28 所示。

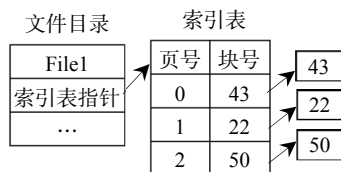


图 3-27 索引文件的存储结构

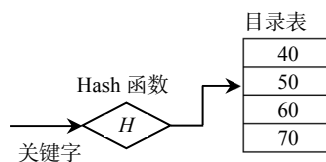


图 3-28 哈希文件的存储结构

令 Key 为记录关键字, 用 A 作为通过 Hash 函数 H 的转换所形成的该记录在目录表中对应表项的位置, 则有关系 $A=H(\text{Key})$ 。在索引文件中, 必须将文件的索引保存在磁盘上, 在需要处理数据文件时, 先把索引导入内存, 搜索索引找到数据记录的地址, 然后访问数据文件存取记录。

而在哈希文件中,用函数来寻找地址不需要索引和伴随的所有开销,因而它是目前最常用的一种存储结构。

3.4.3 文件目录与目录结构

计算机中最常用的文件载体是磁盘。就其存储容量而言,一个磁盘可以存储多个文件。但如果简单地将文件堆放到磁盘中,不仅难以搜索,而且即使有足够的存储空间存储许多个文件,一个区域允许存储的文件个数也是有限的。因此,磁盘文件管理采用了“逐级分化”的计算思维,即建立便于文件存储和查找的文件目录 (File Directory) 及其目录结构 (Directory Structure)。

1. 文件目录

文件目录是文件的排列顺序,如同书本的“章”“节”目录。文件是信息的集合,文件与信息的关系如同书本的“节”与节中的符号元素,“节名”便是文件名,节中的文字、公式、表格、插图等信息便是文件的内容。“章”“节”的有序排列,便构成了书本的目录。

【提示】在 Windows 操作系统中,文件目录通常称为文件夹,文件夹名就是目录名。为了便于文件归类存储和查找,通常将磁盘划分为多个“区”(将磁盘划分为 C、D、E、F 等区),然后在各个区中建立多个文件目录,每个文件目录中可以存储多个不同名称的文件。

2. 目录结构

目录结构是文件管理的一种规范模式。Windows 操作系统采用树形目录结构,目录和文件的隶属关系像一棵倒置的树,树枝称为子目录,树叶称为文件,如图 3-29 所示。

树形目录结构有利于在磁盘上组织、管理文件,并具有明确的层次关系。树形目录结构中包括根目录和子目录,根目录是在磁盘格式化时建立的,也称为系统目录。一个根目录下可以有若干子目录或文件,在每一个子目录下又可以再下挂若干子目录(如 D1、D2...)或文件(如 F1、F2...)。

(1) 子目录名 (Subdirectory Name)。根目录下的每个目录项既可以是文件名,也可以是另一个目录名,这些目录名称称为子目录名。它是用专用命令建立起来的,例如, D1、D01、D2、D3 都是子目录名。

(2) 当前目录 (Current Directory)。为了方便表示或描述树形目录结构中文件所在的位置,把任何时刻当前正在使用的那个目录称为当前目录。例如,机器刚启动时,根目录就是当前目录,在 DOS 状态下的显示标识符为“C:\>”。

(3) 目录路径 (Directory Path)。目录路径是指搜索文件所经过的子目录名称的顺序,即目录之间的通路 (Path),用“\”标识,也称为“路径”。例如,在根目录下运行 F33 文件,其目录路径为 C:\D1\D2\D3\F33。在 DOS 中,用目录命令“CD”改变当前目录;在 Windows 操作系统中,通过鼠标操作改变当前目录。

【提示】在一个目录下只能存储不同名称的文件,而在不同的目录下,可以存储相同名称的文件。例如,在目录 D1 和 D2 中,允许存储同名的 F1 文件,这也是目录结构的重要意义所在。

3. 文件管理

文件管理是操作系统的四大职能之一,以实现操作系统对磁盘文件的高效管理,也是建立文件目录结构的目的所在,为计算机的应用、管理和维护提供了极大便利。

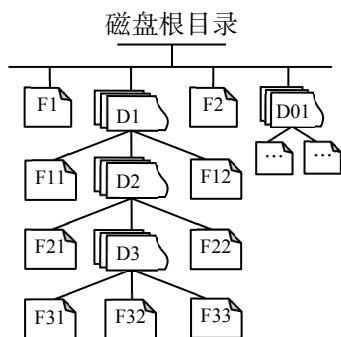


图 3-29 树形目录结构

(1) **文件分类**。为了便于管理和控制,可将文件分为多种类型。按用途划分,可分为系统文件、库文件和用户文件;按数据形式划分,可分为源文件、目标文件和可执行文件。

(2) **文件命名**。为了便于操作使用,每个文件必须有唯一的标记,即文件名。文件名由主文件名和扩展名两部分组成:主文件名可以由多个连续的字符组成,类似一个人的姓名;扩展名用来标识文件的性质(如系统文件、库文件、可执行文件等,类似于标明某人的性别),跟在主文件名之后,主文件名与扩展名之间用“.”符号分开。传统(如DOS)的主文件名可以由1~8个连续的字符组成,扩展名由1~3个连续的字符组成,称为“8.3命名规则”,其基本格式为

□□□□□□□□.□□□

Windows 7 操作系统突破了DOS中8.3命名规则的限制,文件名的长度可达255个字符(包括盘符和路径在内)。文件命名的原则是便于归类 and 记忆(望文生义),其字符可以是英文、汉字、数字,但不能用特殊字符。

(3) **文件属性**。操作系统为不同类型的文件定义了独特性质,使每个文件能选择不同的属性。

① **隐含**(Hidden, H)。具有H属性的文件不会在文件列表中显示,这类文件不能被删除、重命名和复制。

② **系统**(System, S)。具有S属性的文件是系统文件,这类文件由系统建立,不能被删除、修改和复制。

③ **只读**(Read only, R)。具有R属性的文件只能被读或被执行,这类文件不能被删除或修改。

④ **档案**(Archive, A)。具有A属性的文件是用户建立的文件,这类文件可以被修改但不能被执行。

(4) **文件操作**。为了便于实现文件管理,操作系统提供了许多文件操作功能,如创建文件、复制文件、修改文件、删除文件等。不同的操作系统具有不同的操作指令及其操作方式。

§3.5 输入/输出系统与总线系统——“组织与协调”的基本思维

计算机硬件系统由功能不同、性能各异的部件组成。为了使系统中的各功能部件能够“彼此适应、齐头并进”地高效工作,必须进行有效“组织与协调”,而担负起该工作的是输入/输出设备、输入/输出接口、总线系统等,它们在操作系统的管理和控制下,有条不紊地高效工作。

3.5.1 输入/输出设备

输入/输出设备是计算机硬件系统的功能部件,通过输入设备,把程序、数据、图形、图像甚至语音等信息输入计算机。经过计算机运算或处理后,由输出设备输出运算或处理结果。

1. 输入设备

输入设备是指向计算机输入各种信息的设备,目前常用的输入设备有键盘、鼠标、光电笔(Photoelectric Pen)、触摸屏(Touch Screen)、扫描仪(Scanner)、数码相机、数字化仪、字符阅读器及智能输入设备等。

(1) **键盘**是计算机中实现人机交互的主要设备,也是在信息录入过程中实现数码转换的重要设备。键盘由一组排列成矩阵形式的按键组成,每按下一个按键,相当于接通了相应的开关电路,产生一个相应的字符代码(每个按键的位置码),然后将它转换成ASCII码或其他编码送到主机。

键盘的按键结构如图 3-30 所示, 4×4 键盘的按键矩阵开关电路如图 3-31 所示。

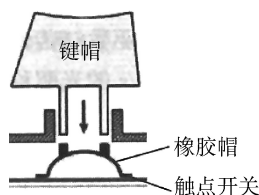


图 3-30 键盘的按键结构

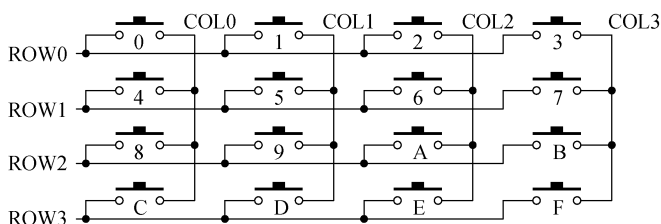


图 3-31 4×4 键盘的按键矩阵开关电路

(2) **鼠标**因其“长相”像一只长尾巴的老鼠而得名,它是随着 Windows 操作系统的出现而诞生的一种输入设备。在软件的支持下,通过快速地移动和控制光标,点击鼠标按键来完成某些特定的功能(菜单选择或绘图)。鼠标可分为光学鼠标和机械鼠标两大类。由于光学鼠标没有活动部件,所以可靠性和精度都比较高。机械鼠标价格便宜,但准确性较差,现已淘汰。

(3) **光电笔**是利用光信号来完成信息输入的标准输入设备,它能实现选择显示屏上的清单,在屏幕上作图、改图,图形放大、移位及旋转等多种功能,因而它既能指定显示屏上图形的位置,又能向计算机输入新的数据信息。

(4) **触摸屏**是一种基于光电笔功能和原理的屏幕触摸输入设备,它利用红外线对屏幕进行监测,屏幕上的项目(标题)相当于“伪按键”,用户只要用手指点一下屏幕上的项目便完成信息输入。触摸屏已广泛应用在各领域的信息控制、登记、查询等系统中。

(5) **扫描仪**是用于扫描条码信息的输入设备。条码是一种用线条和间隔,按编码规则表示数据的图形标记,可分为一维码和二维码。

一维码是指在水平方向上表示信息的条码,用来识别商品的基本信息,目前我国推行使用的一维码是 EAN-128 码,如图 3-32 所示。

二维码是指在水平和垂直方向的二维空间存储信息的条码,目前常用的二维码有 PDF417 码和 QR (Quick Response) 码,QR 码如图 3-33 所示。与一维码相比,二维码具有信息容量大、可靠性高、可表示汉字和图像等多种信息、保密防伪性强、扫码反应时间短等特点,因而在许多业务中得到了广泛应用。



图 3-32 EAN-128 码



图 3-33 QR 码

用来获取信息的相关设备还有数码相机、数字化仪、字符阅读器、智能输入设备等,现在智能输入(声音、人脸、指纹、手写等)已广泛应用于信息认证和文字输入。随着计算机科学和人工智能技术的高速发展,输入设备将越来越智能化,越来越适应各类信息输入的需求。

2. 输出设备

输出设备是指输出计算机处理结果(数字、文字、表格、图形、图像、语音等)的设备。目前常用的输出设备有显示器、打印机、绘图仪等。利用输出设备,可将处理结果呈现给用户。由于计算机已将所输入的各类信息转换成数字信息,所以输出设备相对简单些。

3.5.2 输入/输出接口

计算机主机是集成电路芯片的组合,而外部设备是由机械和电子结合的装置,所以主机与外

部设备在数据形式、传送方式、传递速率等方面因存在巨大差异而不能直接相连和进行信息交换，为了解决这一适配问题，人们采用了输入/输出接口，简称 I/O 接口。

1. 接口功能

接口是对输入/输出设备与数据的一种抽象，是主机与外部设备之间实现信息交换的界面，称为交互界面（Interactive Interface）。接口功能（Interface Function）是实现各类信息交互的“协调”机制，以解决主机与外部设备在信息传递速度、时序、格式、类型等方面的不匹配问题。

（1）**速度不匹配**。外部设备的工作速度比主机慢得多，而且各种外部设备之间的速度差异很大，有每秒钟能传输几兆位的硬盘驱动器，也有每秒钟只能传输几百个字符的打印机。

（2）**时序不匹配**。因为外部设备与主机的工作速度不同，所以一般外部设备都有自己的时序控制电路（时钟），以自己规定的速率传输数据，它无法与 CPU 的时序取得同步。

（3）**信息格式不匹配**。不同的外部设备，信息的存储和处理格式不同，既有并行和串行之分，也有二进制位、ASCII 码、BCD 码之分等。

（4）**信息类型不匹配**。不同的外部设备，其信号类型有可能不同（可能是数字电压，也可能是连续电流或其他模拟电信号），而且信号量的幅值也可能不同。

2. 接口电路

接口电路（Interface Circuit）用来解决主机与外部设备不匹配的问题，尽管不同输入/输出接口的具体任务有所不同，但基本功能是大致相同的。接口电路由 5 个功能部件组成，如图 3-34 所示。

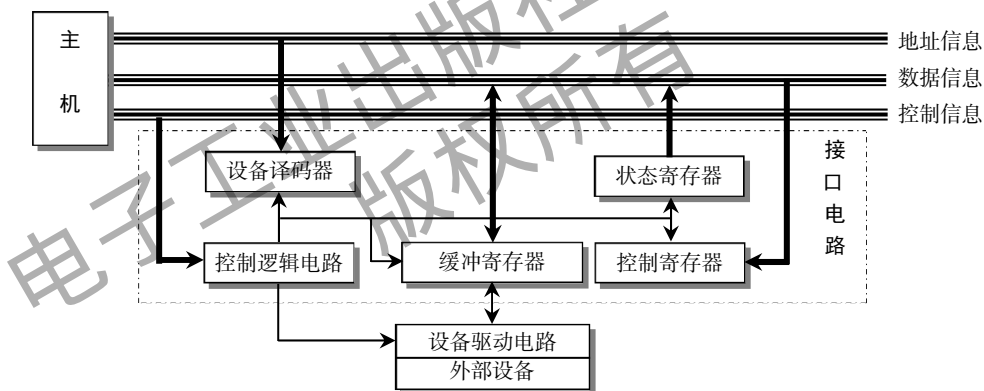


图 3-34 接口电路的组成

（1）**设备译码器（Device Decoder）**。为了便于数据交换，需要对外部设备或端口进行编码，设备译码器用来判别主机是否选中该接口及其连接的外部设备，确定主机与哪一台外部设备交换数据。

（2）**控制逻辑电路（Control Logic Circuit）**。为了使主机与外部设备之间的数据传输高速有效，需要通过控制逻辑电路，实现主机与外部设备之间的通信控制，包括同步控制、中断控制等。

（3）**状态寄存器（Status Register）**。为了提高计算机的作业效率，需要记录外部设备工作状态，如“空闲”“繁忙”“结束”等，以便及时调整作业进程，这一进程是由状态寄存器来完成的。

（4）**缓冲寄存器（Buffer Register）**。为了确保数据传输的可靠性，在主机与外部设备交换数据时，需要先将数据缓存在缓冲寄存器中，避免因外部设备运行缓慢而丢失数据。

（5）**控制寄存器（Control Register）**。为了使外部设备按照主机命令有条不紊地执行各种操作，需要寄存主机发来的控制命令信息，以免外部设备因运行速度跟不上主机下发命令的速度而失去控制。

【提示】将计算机中的二进制代码转换成不同的输出信息，是由输出设备接口编码实现的。个人计算机中输入/输出系统的物理连接是通过主板（Main Board）来实现的，如 CPU、控制芯片组、内存、输入/输出接口芯片组、时钟等，均插在主板的插座或扩展槽中。

3.5.3 总线系统

计算机硬件系统中的主机、输入/输出设备、外存及其他设备，可通过一组导线按照某种连接方式组织起来，构成一个完整的硬件系统。这组导线称为总线，它是各部件之间的数据通道。

1. 总线类型

总线类型（Bus Type）是指为了适应各种设备的连接而划分的不同总线的类型。如果按照通信方式划分，总线可分为串行总线（一次传输一个二进制位的数据，如鼠标、键盘、调制解调器）和并行总线（一次传输多个二进制位的数据，如打印机）。如果按照功能作用划分，总线可分为 3 种类型。

（1）**内部总线**（Internal Bus）是 CPU 芯片内部、寄存器与寄存器之间、寄存器与 ALU 之间的连接总线，因而又称为片内总线，从外部是看不见的。

（2）**外部总线**（External Bus）是实现中低速输入/输出设备、外存之间互连的总线，是接口电路与外部设备之间进行信息传送的公共通道，用来实现计算机主机与外部设备之间的互连。

（3）**系统总线**（System Bus）是 CPU、RAM 及各接口之间的连线，是各部件之间进行信息传送的公共通道，是整个计算系统的“中枢神经”，所有的地址、数据和控制信息都经由这组总线传输。

2. 总线结构

总线结构（Bus Structure）是指连接各部件的“组织”形式，是 CPU 与内存、外存、接口电路、输入/输出设备之间的连接方式，通常分为单总线结构、双总线结构、三总线结构等。

（1）**单总线结构**。单总线结构是将计算机中所有的部件、设备都连接到一组导线上的最简单连接方式。在单 CPU 计算机中，常采用单总线结构，它包括三根总线，如图 3-35 所示。

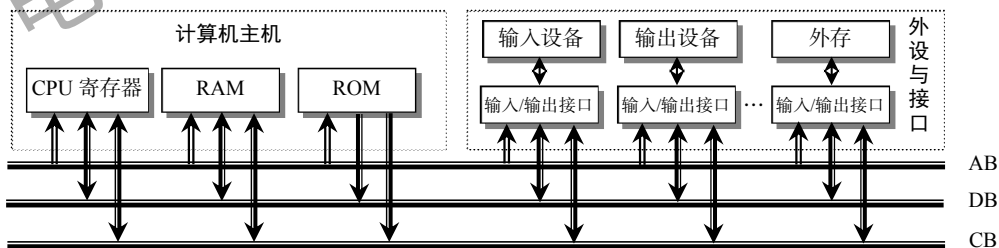


图 3-35 单总线结构

① AB 是 CPU 向存储器或接口传送地址信息的单向通路，用来传送存储单元或输入/输出接口的地址信息。AB 决定了计算机内存空间的范围。

② DB 是 CPU 与内存或接口传送数据信息的双向公共通道，用于在 CPU 与内存或输入/输出设备之间传送数据。DB 的宽度用二进制位来衡量，如 32 位或 64 位。

③ CB 是 CPU 向存储器或接口传送命令及 CPU 接收信息的通道，用来传送各种控制信号。控制信号可分为两类：一类是由 CPU 向内存或外设发送的控制信号；另一类是由外设或有关接口向 CPU 送回的信号，包括内存的应答信号。

由于单总线结构中的所有部件都挂在系统总线（AB、DB、CB）上，只能分时传送数据信息，即某一时间只允许一对部件传送数据信息，因此计算机中各部件间的数据传送受到了极大限制。

(2) **双总线结构**。为了克服单总线结构分时传送数据的缺点,双总线结构增加了主存总线,如图 3-36 所示,其中输入/输出总线为系统总线。双总线结构既具有单总线结构简单、易于扩充的优点,又可以使 CPU 与主存交换信息,从而减轻了系统总线的负担,使系统的吞吐量及速度得到提高。

(3) **三总线结构**。三总线结构在双总线结构的基础上增加了直接内存访问(Direct Memory Access, DMA)总线,使外设与主存之间直接交换信息,以进一步提高 CPU 的工作效率,如图 3-37 所示。

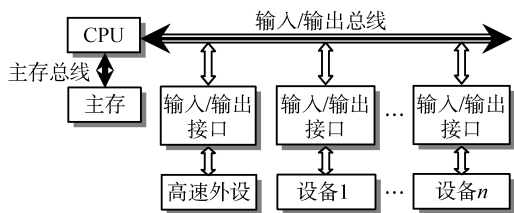


图 3-36 双总线结构

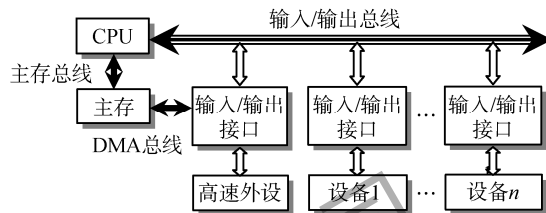


图 3-37 三总线结构

三总线结构中任意时刻只能执行一种总线传输,主存总线与 DMA 总线不能同时对主存进行存储,输入/输出总线只在 CPU 执行输入/输出指令时才会用到。

3. 总线标准

总线标准(Bus Standard)是指不同厂家生产同一功能的设备时必须遵守的技术规范。由于不同厂家生产同一功能设备时的实现方法不同,因而其技术性能会存在一定差异。若要实现各厂家生产的部件可以替换使用,就需要使其遵守相同的标准——硬件协议和软件协议,包括时钟频率、型号定义、总线结构、仲裁机制、电气规范、物理规范,以及实施总线协议的驱动程序等。

3.5.4 操作系统对输入/输出的管理

现代计算系统中存在着大量的输入/输出设备,输入/输出设备的种类越丰富,计算机的功能越强大。为了方便用户连接和使用各种输入/输出设备,操作系统提供了输入/输出管理功能,包括输入/输出设备管理和输入/输出控制管理。

1. 输入/输出设备管理

输入/输出设备管理的目的是有效地分配和使用外设、协调 CPU 与外设之间的时间差异、提高系统总体性能。输入/输出设备管理包括设备分配、设备驱动、设备无关性、缓冲区管理和即插即用等。

(1) **设备分配**。当进程向系统发出输入/输出请求后,由设备分配程序按照一定的策略把所要求的设备分配给该进程。为确保 CPU 与设备之间能进行通信,还应分配相应的控制器和通道。

(2) **设备驱动**。为了实现对外设的操作控制,必须有一个实现 CPU 与外设通信的程序,称为设备驱动程序,操作系统通过设备驱动程序指挥外设执行操作,完成具体的输入/输出任务。

(3) **设备无关性**。为了使用户编写的程序与实际使用的物理设备无关,由操作系统把用户程序中使用的逻辑设备映射到物理设备,以实现逻辑设备与物理设备的对接。

(4) **缓冲区管理**。计算机中 CPU 的运行速度快,而外设的运行速度慢,这降低了 CPU 的使用效率,甚至会中断 CPU 的运行。为了解决 CPU 与外设之间速度不匹配的矛盾,常采用存储缓冲技术,实行缓冲区管理,以提高外设与 CPU 之间的并行性,从而提高整个系统的性能。

(5) **即插即用**。当把设备连接到计算机后,操作系统能随即自动检测到新接入的设备并自动

安装驱动程序，而无须进行任何配置，故而称为“即插即用”。目前，绝大多数操作系统支持即插即用技术，如在计算机上使用 U 盘便是典型的即插即用。

2. 输入/输出控制管理

输入/输出控制管理是指操作系统对 CPU 与输入/输出设备之间数据传送的控制方式，要求速度快、开销小、充分发挥硬件资源的能力。随着计算机技术的发展，控制方式经历了由低级到高级、由简单到复杂、由集中到分散的发展过程。按发展的先后次序和主机与外设并行工作的程度，输入/输出控制管理可分为 3 种方式。

(1) **程序中中断控制** (Program Interrupt Control, PIC) 是指计算机在执行程序的过程中，当出现异常情况或特殊请求时，立即中断现行程序的运行而转向对异常情况或特殊请求的处理，处理结束后返回程序的中断处继续执行。

(2) **DMA** 是指在外设和 RAM 之间开辟直接的数据交换通路进行数据传送，即在计算机内增设 DMA 控制器，由它直接控制主机与外设之间的数据交换。

(3) **输入/输出通道控制** (I/O Channel Control) 是 DMA 方式的进一步发展，具有存储传输数据的存储器、类似 DMA 的通道程序、解释和执行通道指令的控制部件，使用通道来控制 RAM 和外设之间的数据传送。它与 DMA 方式类似，以 RAM 为中心与设备直接交换数据，只要 CPU 发出启动指令，指出要求通道执行的操作和使用的输入/输出设备，就可以启动该通道，使该通道从 RAM 中调出相应的通道程序并执行。

本章小结

1. 计算系统是对计算机的抽象，一个完整的计算系统包括硬件系统和软件系统，硬件系统是计算系统的物理支撑，软件系统是计算系统的灵魂。它们相互支撑和配合，齐心协力地高效工作。

2. 要使计算系统高效可靠地工作，需要从系统科学的观点出发，以计算思维分析系统的结构组成，寻找提高系统性能的思想策略。

3. 操作系统是软件系统的核心，用来管理和控制计算系统中的硬件及软件资源。操作系统对计算机的管理包括进程管理、存储管理、设备管理、文件管理。

4. 存储管理的主要任务是内存分配管理、内存保护管理、内存扩充管理、地址映射管理。其中，内存扩充管理与虚拟存储实际上是同一个概念，通过对内存中程序和数据的调度，使用户能够在较小的内存中执行较大的用户程序，并且可以在内存中容纳更多的程序并发执行。

5. 为了管理外存上的文件，操作系统中设置了管理文件的功能模块——文件系统，它负责管理文件，为用户提供对文件进行存取、共享及保护的手段，以提高系统资源的利用率。

6. 系统总线是计算机中各部件之间的数据通道，也是计算机内部设备与外设间的枢纽。基于计算机主机与外设之间性能的差异性，采用接口技术来解决各类部件之间的不匹配问题。

7. 系统总线和接口电路解决的是部件间的数据传输和匹配问题，但如何实现对设备的管理，如何高效利用数据通道，充分发挥硬件资源的能力，则是计算机操作系统的重要职能。

习题 3

一、选择题

1. 计算机硬件系统的五大部件是指运算器、控制器、输入设备、输出设备和 ()。

- A. 存储器 B. 接口电路 C. 外部设备 D. 指令系统
2. 一个完整的计算系统，可分成硬件系统、支撑软件、应用软件和（ ）4个层次。
A. 软件系统 B. 程序设计语言
C. 操作系统 D. 指令系统
3. 计算机操作系统具有并发性、共享性、虚拟性和（ ）4项基本特征。
A. 同步性 B. 可靠性 C. 异步性 D. 可控性
4. 20世纪90年代以来的主流操作系统有DOS、Windows、UNIX和（ ）。
A. 网络操作系统 B. 手机操作系统
C. 智能操作系统 D. Linux
5. 计算机中指令执行的基本流程可分为（ ）、取指、分析指令和执行指令。
A. PC+1 B. 编写指令 C. 存储指令 D. 操作指令
6. 操作系统对CPU的管理主要包括进程控制、进程通信、进程同步和（ ）等。
A. 作业管理 B. CPU调度 C. 并发控制 D. 交通控制
7. 目前，个人计算机中常用的外存主要有硬盘、（ ）和移动硬盘。
A. 软盘 B. 光盘 C. U盘 D. 磁带
8. 输入/输出接口电路用来解决速度不匹配、时序不匹配、格式不匹配和（ ）不匹配的问题。
A. 操作方式 B. 存储格式 C. 电器特性 D. 信息类型
9. 输入/输出的控制方式可分为程序中断方式、DMA方式和（ ）。
A. 程序查询方式 B. 作业管理方式
C. 输入/输出通道控制方式 D. 通道管理方式
10. 操作系统对文件的管理主要包括存储空间管理、文件目录管理、文件读/写管理和（ ）。
A. 文件操作管理 B. 文件安全保护
C. 文件删除管理 D. 文件建立管理

二、问答题

1. 什么是计算系统？
2. 如何定义计算机硬件与硬件系统？如何定义计算机软件与软件系统？
3. 什么是操作系统？
4. 计算机硬件与软件之间存在哪些关系？
5. 什么是作业？
6. 为什么在操作系统中引入进程和线程？
7. 并发与并行有何区别？
8. 什么是存储管理？
9. 什么是输入/输出控制？
10. 什么是文件和文件系统？

三、讨论题

1. 本章介绍了3种典型的操作系统，为什么要研发多种类型的操作系统？
2. 如果计算机不使用操作系统，你认为计算机的硬件设计应该解决哪些问题？