"十二五"普通高等教育本科国家级规划教材

国家精品在线开放课程配套教材

国家级优秀教学团队教学成果

计算机组成原理

（第5版）

◎纪禄平　罗克露　刘辉　张建　编著

中国工信出版集团　電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

“十二五”普通高等教育本科国家级规划教材
国家精品在线开放课程配套教材
国家级优秀教学团队教学成果

计算机组成原理

（第5版）

纪禄平 罗克露 刘 辉 张 建 编著

電子工業出版社
Publishing House of Electronics Industry
北京 · BEIJING

内 容 简 介

本书为“十二五”普通高等教育本科国家级规划教材。

本书以当前主流微型计算机技术为背景，以建立系统级的整机概念为目的，深入介绍计算机各功能子系统的逻辑组成和工作机制。全书共 6 章，第 1 章概述计算机的基本概念、发展历程和系统的硬件、软件组织及计算机相关的性能指标；第 2 章介绍数据信息的表示、运算和校验方法；第 3 章介绍 CPU 的一般模型、指令系统和 x86 架构、MIPS32 架构简易 CPU 的设计；第 4 章介绍存储子系统的存储原理、主存设计和计算机三级存储体系等；第 5 章介绍总线与 I/O 子系统，包括接口、总线以及中断、DMA 和 IOP、PPU 等 I/O 传输控制模式；第 6 章介绍键盘原理、显示器件和打印机等外围设备。

本书可作为高等院校计算机及相关专业“计算机组成原理”及相关课程的配套教材，也可作为从事计算机专业考研和工程技术人员的参考书。

图书在版编目(CIP)数据

计算机组成原理/纪禄平等编著. —5 版. —北京：电子工业出版社，2020.1

ISBN 978-7-121-37684-9

Ⅰ. ① 计…　Ⅱ. ① 纪…　Ⅲ. ① 计算机组成原理－高等学校－教材　Ⅳ. ① TP301

中国版本图书馆 CIP 数据核字（2019）第 246996 号

责任编辑：章海涛

印　　刷：三河市鑫金马印装有限公司

装　　订：三河市鑫金马印装有限公司

出版发行：电子工业出版社

北京市海淀区万寿路 173 信箱　邮编　100036

开　　本：787×1092　1/16　印张：27.25　字数：764 千字

版　　次：2004 年 9 月第 1 版

2020 年 1 月第 5 版

印　　次：2020 年 3 月第 2 次印刷

定　　价：58.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888，88258888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

本书咨询联系方式：192910558（QQ 群）。

前　言

本书源自电子科技大学《计算机组成原理》教材40余年的传承。

本书的第1版（2004年）和第2版（2010年）均由罗克露教授主编，电子工业出版社出版发行，2014年入选**“十二五”普通高等教育本科国家级规划教材**。在这两个版次基础之上，纪禄平继续进行了修订和改编，并由电子工业出版社分别出版了第3版（2014年）和第4版（2017年）。遵照电子科技大学的本科教学改革精神，纪禄平于2019年再次主持了针对本书的修订工作，删减了一部分陈旧的教学内容，增补了若干前沿知识点，同时对组织结构进行了进一步优化。

全书共6章，主要按计算机硬件功能子系统来安排章节内容。第1章概述计算机的概念、发展历程、硬软件组织，以及性能评价指标。第2章介绍数据信息的表示、运算和校验方法。第3章先介绍CPU模型和指令系统，然后展示了x86架构CISC型处理器和MIPS32架构RISC型处理器的典型设计实例。第4章介绍存储子系统，主要讨论存储的基本原理、主存储器设计和三级存储体系架构，以及高性能存储系统的介绍。第5章介绍总线与输入/输出子系统，包括设备接口、总线以及程序中断方式、DMA和IOP等经典的I/O控制模式。第6章对典型输入/输出设备，如键盘、显示器和打印机进行了介绍。

全书由纪禄平博士主持编写和统稿。第1章和第3章由纪禄平编写，第2章和第4章由罗克露编写，第5章由刘辉编写，第6章由张建编写。作为计算机组成原理教学和教材编写的资深前辈，罗克露教授对本书的编写给予了悉心指导，提出了许多宝贵建议，为本书审阅付出大量心血，谨此向罗克露教授致以崇高敬意！

本书的修订编写工作还得到了电子科技大学的专门立项支持，电子工业出版社也提供了积极的协助，一并向他们致以诚挚谢意！

本书还有配套的辅导书——《计算机组成原理（第5版）学习指导与实验》，其中包含与本书相关的实验等内容。

本书还为任课教师提供了配套的教学资源（包含电子教案），需要者可**登录华信教育资源网站**（http://www.hxedu.com.cn），注册之后进行免费**下载**。

由于编者水平有限，书中内容难免会有疏漏和错误，恳请读者批评指正，全体作者将不胜感激。如有任何问题，请直接通过邮件与作者联系：jiluping@uestc.edu.cn。

作　者

目 录

第 1 章　概　论

“计算机组成原理”课程的主要内容是以单机系统为对象，阐述计算机系统的硬件组成，其核心是建立一个计算机系统的整机概念。这里提到的“整机概念”包括两层面，即 CPU 级的整机和硬件系统级的整机，且每个层面都涉及硬件的逻辑组成及其工作原理机制。本书将从这两层面逐步建立前述的整机概念。为此，本章首先阐明三个重要的基本概念：信息的数字化表示、存储程序工作方式和计算机系统的层次结构，并将这些概念作为了解计算机的逻辑组成结构和硬件系统工作原理的基本出发点。

1.1　计算机的基本概念

计算机是 20 世纪人类最伟大的发明之一，是一种能够存储程序，通过执行程序指令，自动、高速、精确地对各种信息进行复杂运算处理，并输出运算结果的一种高科技电子设备。计算机是人类进入信息化时代的里程碑式标志之一，也是最基本、最常用的信息处理工具。

计算机系统通常由硬件和软件两大部分组成。硬件是指看得见、摸得着且物理存在的设备实体，如运算器、控制器、存储器和鼠标、键盘等输入、输出设备，如图 1-1 所示。

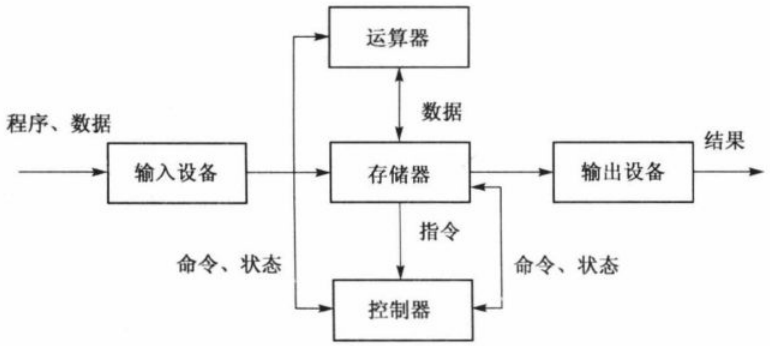

图 1-1　计算机组成结构示意

软件是指不能直接触摸但确实在逻辑上存在的可感知对象，如程序和文档等。设计计算机硬件系统的基本原则是功能部件的模块化，即用逻辑电路构造各种功能独立部件，如用门电路、触发器等来构造运算器和存储器等。在硬件基础上，再根据应用需要配置各种软件，如操作系统、编程语言及各种支撑软件等。硬件和软件按层次逻辑组成一个复杂的计算机系统。

通过软硬件协同，计算机能够灵活地执行面向特定任务而编写的程序，兼具高复杂性、自动化和智能化等特性，这些都是计算机与其他普通电子设备的显著区别。

不管是进行复杂的数学计算，还是对大量的数据进行处理，或者对一个过程进行自动控制，用户都应先按照处理的步骤，用编程语言编写程序，然后通过输入设备（如键盘），将程序和需处理的数据送入计算机并存放在存储器中。用户编写的程序称为源程序，是不能被计算机直接执行的。计算机只能执行机器指令，即要求计算机完成某种操作的命令，如执行加法操作的加法指令、执行乘法操作的乘法指令、执行传送操作的传送指令等。因此，计算机在运行程序前，必须将源程序编译转换为计算机可识别的机器指令，并将这些指令按一定顺序存放在存储器的若干存储单元中。每

个单元对应一个称为地址的固定编号，只要给出确定的地址，就能访问相应的存储单元，对该单元进行读/写操作，从中读出指令，并将执行结果写回到存储器。

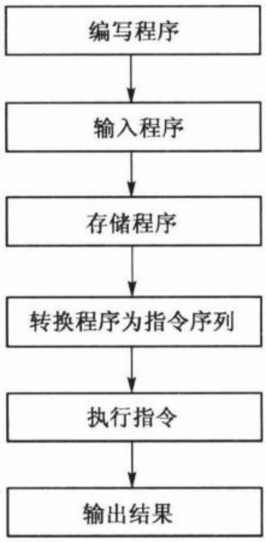

图 1-2　计算机工作流程

当计算机启动运行后，控制器将某个地址送往存储器，从该地址单元取回一条指令。控制器根据这条指令的含义，发出相应的操作命令，控制该指令的执行。比如，执行一条加法指令，先要从存储单元或寄存器取出操作数，送入运算器，再将两个操作数相加，并将运算处理的结果送回存储单元或寄存器存放。如果用户要了解处理结果，则计算机可通过输出设备（如显示器、打印机等），将结果显示在屏幕或打印在纸上。图 1-2 给出了计算机工作流程。

从以上描述可看出，计算机作为一个能够自动处理信息的智能工具，它的设计过程必须考虑很多因素。首先，信息如何表示才能方便计算机进行识别和处理。其次，计算机硬件系统应该由哪些部件组成，每部分的相互关系是怎样的，以及如何控制它们协同工作。再次，采用什么样的工作方式才能使计算机能够自动地对信息进行处理。最后，应该提供怎样的人机交互接口才能方便操作和使用计算机。

纵观计算机的诞生和发展历程可以看出，上述几个核心问题自始至终贯穿于计算机发展的各阶段中。不同时代对计算机有特定的应用需求，而基于这些需求提出的解决思路完美地回答了这些问题，从而促进了计算机由弱到强、从低级到高级发展。

1.1.1　信息的数字化表示

计算机是通过执行程序（指令序列）来实现对数据的加工处理的。因此，计算机中的信息可以分为两大类：控制信息和数据信息。控制信息用来控制计算机的工作。计算机执行指令时，用指令产生的控制命令（称为微命令）控制有关操作，所以指令序列和微命令序列属于控制信息类。数据信息是计算机加工处理的对象。计算机根据指令要求取出的操作数以及对操作数处理的结果等，都属于数据信息类。数据信息又分为数值型数据和非数值型数据两类。数值型数据有数值大小与正负之分，如 6、−15 等。非数值型数据则无数值大小，也不分正负，如字符、文字、图像、声音等人们能够识别的信息，以及条件、状态、命令等用于判定的逻辑信息。那么，在计算机中如何表示这些信息呢？

前面讲过，计算机的主要部件是由逻辑电路即电子电路构成的，所以，在电子数字计算机中传输和处理的信息都采用数字化表示方法。信息的数字化表示包含了两层含义：① 用数字代码表示各种信息；② 用数字信号表示数字代码。信息表示数字化的概念是我们理解计算机工作原理的一个基本出发点。下面通过几个例子对这两层含义加以说明。

1．在计算机中用数字代码表示各类信息

数字代码是指一组数字的集合，这里的数字代码通常指二进制数字代码。我们可以根据需要描述的信息（某类控制信息或某类数据信息），用一组约定了含义的数字代码来表示它。

【例 1-1】 用数字代码表示数值型数据。

6 和−7 是两个数值型数据。可以约定，用一位二进制代码表示每个数的符号，如用 0 表示正数，用 1 表示负数；再用 4 位二进制代码表示每个数的大小。这样，代码 00110 表示 6（左边第一个 0 代表正号），代码 10111 表示−7（左边第一个 1 代表负号）。

当然，也可以用 8 位二进制代码表示一个数的大小。代码位数增多，数的表示范围将扩大。例

如，用 4 位二进制代码不能表示 200 这个数，但用 8 位二进制代码可以表示它。

【例 1-2】 用数字代码表示字符。

字符本身没有大小和正负之分，但仍然可以用数字代码来表示它。计算机中常约定用 7 位代码表示一个西文字符，如用 1000001 表示字符 A，用 1000010 表示字符 B；或用 7 位代码表示一个控制字符，如用 0001100 表示换页（FF），用 0001101 表示回车（CR）。字符的这种编码称为 ASCII，是国际上广泛采用的一种字符表示方法。另外，可以约定用两组 8 位二进制代码表示一个中文字符，如用 01010110 01010000 表示“中”，用 00111001 01111010 表示“国”等。总之，用数字代码可以表示各种字符，以字符为基础又可以表示范围广泛的各种文字。

【例 1-3】 用数字代码表示图像。

字符的种类总是有限的，因而可以用若干位编码来表示。图像则不然，其变化是无穷无尽的，那么，如何用数字代码来表示这些随机分布的图像信息呢？实际上，一幅图像可以被细分为若干个点，这些点称为像素。也就是说，可以用像素的组合来逼近真实的图像。图像划分得越细，像素越多，组成的图像也就越真实。按照信息表示数字化的思想，可以用数字代码表示像素。例如，用一位代码表示一个像素，若像素是亮的，则用代码 1 表示；若像素是暗的，则用代码 0 表示，再将表示一幅图像所有像素的代码按照像素在图像中的位置进行组织，就可实现用数字代码来表示图像。

【例 1-4】 用数字代码表示声音。

为了对声音信息数字化，首先将声波转换为电流波，再按一定频率对电流波进行采样，即在长度相同的时间间隔内分别对电流波的幅值进行测量，每次测到的电流幅值都用一个数字量来表示。只要采样频率足够高，得到的数字信息就能逼真地保持声波信息，还原后，真实地再现原来的声音。

【例 1-5】 用数字代码表示指令。

指令属于控制信息。通常，一条指令需提供要求计算机做什么操作及如何获取操作数等信息。因此可以用一段数字代码表示操作类型，这段代码称为操作码；用另一段代码表示获取操作数的途径，这段代码称为地址码。将操作码和地址码组合在一起，就形成了机器指令代码。例如，操作码取 4 位，可以约定，0000 表示传送操作，0001 表示加法操作，0010 表示减法操作……地址码取 12 位，用 6 位表示一个操作数的来源，用其余 6 位表示另一个操作数的来源。例如，约定 000000 表示操作数来自 0#寄存器，000001 表示另一操作数来自 1#寄存器。这样，16 位代码 0001000000000001 表示一条加法指令，其含义是将 0#寄存器的内容与 1#寄存器的内容相加，结果存放在 1#寄存器中。

【例 1-6】 用数字代码表示设备状态。

计算机在工作时往往需要了解外部设备的状态，根据外设状态决定做什么操作。不同的外部设备可能有不同的工作状态，如打印机将字符打印在纸上，显示器则将字符显示在屏幕上。这些设备的状态可以归纳为三种：空闲（设备没有工作）、忙（设备正在工作）、完成（设备做完一次操作）。相应地，可以用约定的数字代码表示这三种状态，如用 00 表示空闲，01 表示忙，10 表示完成。

2．在物理机制上用数字信号表示数字代码

为什么能用数字代码来表示各种信息呢？这就涉及计算机的物理机制。计算机是一种复杂的电子线路，传送和处理的实际对象是电信号。电信号又分为模拟信号和数字信号两种。

模拟信号是一种随时间连续变化的电信号，如电流信号、电压信号等。我们可以用电流或电压的幅值来模拟数值或物理量的大小，如模拟温度的高低、压力的大小等。处理模拟信号的计算机称为模拟计算机，只应用在极特殊的领域中。用模拟信号表示数据的大小有许多缺点，如表示的精度低、表示的范围小、抗干扰能力弱、不便于存储等。如果用数字信号表示信息，则可克服以上缺点。

数字信号是一种在时间上或空间上断续变化的电信号，如电平信号和脉冲信号。单个电信号一

般只取两种状态，如电平的高或低、脉冲的有或无，这样可以用这两种状态分别表示数字代码 1 和 0，称为二值逻辑。比如，用高电平状态表示 1，低电平状态表示 0；或者用有脉冲的状态表示 1，无脉冲的状态表示 0。用一位数字信号表示 1 位数字代码，用多位数字信号的组合可以表示多位数字代码。处理数字信号的计算机称为数字计算机，电平信号和脉冲信号是数字计算机中最基本的电信号形式。用数字信号可以表示数字代码，用数字代码又可以表示各种信息，因而数字计算机能用于各行各业，处理广泛的信息。下面通过例子说明如何用多位电信号的组合来表示多位数字代码。

【例 1-7】 用一组电平信号表示 4 位数字代码。

电平信号利用信号电平的高、低状态表示不同的代码，所以电平信号通常需要一段有效维持时间。可以用 4 根信号线分别输出 4 个电平信号，每个电平信号表示 1 位代码。我们约定，+5 V 为高电平，表示 1；0 V 为低电平，表示 0。如图 1-3 所示，4 位电平信号表示 4 位数字代码 1011，它们可能表示一个 4 位的二进制数，也可能表示一个命令或一种状态的编码。

每个信号各占用一根信号线，因而这一组电平信号在空间上的分布是离散的。在计算机中，常用电平信号表示并行传送的信息，如用若干信号线同时传输的数据、地址或其他信息的编码。

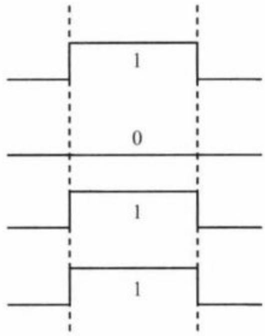

图 1-3　用一组电平信号表示多位数字代码

【例 1-8】 用一串脉冲信号表示 4 位数字代码。

与电平信号不同，脉冲信号的电平维持时间很短，如信号电平从 0 V 向+5 V（或−5 V）跳变，维持极短时间后再回到原来的 0 V 状态。因此，信号出现时，其电平为+5 V（或−5 V），信号未出现时，其电平为 0 V，如图 1-4 所示。由于脉冲信号在时间上的分布是离散的，因此可以用一根信号线发出一串脉冲信号，在约定的时间内有脉冲表示 1，无脉冲表示 0。图 1-4 中的脉冲串表示 4 位数字代码 1011。

可以用脉冲信号的上升边沿或下降边沿表示某一时刻，对某些操作定时。例如，在计算机中，在脉冲上升边沿将数据送入某个寄存器中，常用脉冲信号表示串行传输的数据。

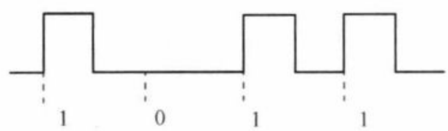

图 1-4　用一串脉冲信号来表示多位数字代码

3．用数字化方法表示信息的主要优点

（1）在物理上容易实现信息的表示与存储

每个信号只取两种可能的状态表示 1 或 0，因此在物理上可以用多种方法来实现，如开关的接通或断开、晶体管的导通或截止、电容上有电荷或无电荷、磁性材料的正向磁化或反向磁化、磁化状态的变化或不变等。凡是具有两种稳定状态的物理介质均可用来存储信息，如用双稳态触发器存储信息，利用电容上存储的电荷来存储信息，还可以用磁性材料记录信息，或者用激光照射过的介质记录信息。

（2）抗干扰能力强，可靠性高

由于单个数字信号的两种状态（高电平与低电平，或者有脉冲与无脉冲）差别较大，即使信号受到一定程度的干扰，仍能比较可靠地鉴别出电平的高低或信号的有无。例如，高电平+5 V 表示 1，低电平 0 V 表示 0，假设信号处于 0 状态，如果出现了 2 V 的干扰信号，也不会将原来信号的 0 状态改变到 1 状态。

（3）数值的表示范围大，表示精度高

一位数字信号的表示范围很窄，用多位数字信号的组合表示一个数时，可以获得很大的表示范围和很高的精度。例如，用 4 位电平信号表示一个 4 位的二进制整数时（不考虑符号），能够表示

的最大数值是 15。若要表示一个 4 位的二进制小数，同样不考虑符号，则数的精度为 2^{-4}。位数越多，数的表示范围越大，或者数的表示精度越高。从理论上讲，位数的增加是没有限制的，但位数增多，花费的硬件开销也相应增大。

（4）可表示的信息类型极其广泛

各种非电量类型的信息可以先转换为电信号，模拟电信号又可以转换为数字电信号，因此表示的信息类型和范围几乎没有限制。

（5）能用数字逻辑技术进行信息处理

根据处理功能逻辑化的思想，计算机的所有操作最终是用数字逻辑电路来实现的。因此，用逻辑代数对信息进行处理就形成了计算机硬件设计的基础，可以用非常有限的几种逻辑单元（如与门、或门、非门等）构造出变化无穷的计算机系统和其他数字系统。

从事计算机技术工作的重要基础是善于用约定的数字代码表示各种需要描述的信息。这里再次强调“信息数字化”的概念：① 计算机中的各种信息都用数字代码表示，这些信息包括数值型的数字、非数值型的字符、图像、声音以及逻辑型的命令、状态等；② 数字代码中的每一位都用数字信号来表示，数字信号可以是电平信号或脉冲信号。

1.1.2 存储程序工作方式

存储程序是计算机的核心内容，表明了计算机的工作方式，包含三个要点：事先编写程序，存储程序，自动、连续地执行程序。这三点体现了用计算机求解问题的过程，下面分别加以说明。

（1）根据求解问题事先编写程序

计算机处理任何复杂的问题都是通过执行程序来实现的。因此，在求解某一问题时，用户要根据解决这个问题所采用的算法事先编写程序，规定计算机需要做哪些事情，按什么步骤去做。程序中还应提供需要处理的数据，或者规定计算机在什么时候、什么情况下从输入设备取得数据或向输出设备输出数据。

（2）事先将程序存入计算机中

如前所述，用户用某种编程语言编写的程序称为源程序，它是由字符组成的，计算机不能识别。因此，需要通过编译器将源程序转换为二进制代码，保存在存储器中。这时的程序还不是指令代码，不能被计算机执行，需进一步被转换为机器指令序列。所以，事先编写的程序最终将变为指令序列和原始数据，并被保存在存储器中，提供给计算机执行。

（3）计算机自动、连续地执行程序

程序已经存储在计算机内部，计算机被启动后，不需要人工干预，就能自动、连续地从存储器中逐条读取指令，按指令要求完成相应操作，直到整个程序执行完。当然，在某些采用人机对话方式工作的场合，也允许用户以外部请求方式干预程序的运行。

指令和数据都是以二进制代码的形式存放在存储器中的，那么计算机如何区分它们？又如何自动地从存储器中读取指令呢？首先，将指令和数据分开存放。由于在多数情况下，程序是顺序执行的，因此大多数指令需要依次相邻存放，而将数据放在该程序区中不同的区间。其次，可以设置一个程序计数器（Program Counter，PC），用它存放当前指令所在的存储单元的地址。如果程序顺序执行，则在读取当前指令后将 PC 的内容加 1（当前指令只占用一个存储单元），指示下一条指令的地址。如果程序要进行转移，则将转移目标地址送入 PC，以便按照转移地址读取后续指令。所以，依靠 PC 的指示，计算机就能自动地从存储器中读取指令，再根据指令提供的操作数地址读取数据。

对于传统的冯•诺依曼机而言，存储程序工作方式是一种控制流驱动方式，即按照指令的执行序列依次读取指令，再根据指令所含的控制信息调用数据进行处理。这里的控制流也称为指令流，是指在程序执行过程中，各条指令逐步发出的控制信息，它们始终驱动计算机工作。而依次被处理的数据信息称为数据流，它们是被驱动的对象。

1.1.3 计算机的分类

计算机的种类有多种多样，从不同的角度出发，可以将它们分成不同的种类。

首先，按计算机中处理信息的制式，计算机可以分为数字计算机和模拟计算机。

数字计算机是通过信号的两种不同状态来表示数字信息（1 和 0）的，可以方便地对数字信号进行算术和逻辑运算，具有速度快、精度高、便于存储等优点。目前，通常讲的计算机一般是指数字计算机。

模拟计算机一般只能处理模拟信号，如连续变化的电压、电流和温度等，主要由模拟信号运算器件组成，适合求解微分方程等。这种计算机在模拟计算和实时控制系统中有应用，但通用性不强，信息不易表示和存储，精度也不高，因此它的应用范围很窄。

其次，如果按照计算机通用性的差异，计算机可以分为专用计算机和通用计算机。

专用计算机，即专门为解决某一特定问题而设计制造的计算机。它一般只具有为实现某一个特定任务而定制的最小软硬件结构，也拥有相对固定的存储程序。专用计算机具有执行速度快、可靠性高、结构简单和价格便宜等优点，但它的功能单一，对应用环境的适应性很差。例如，控制轧钢过程的轧钢控制计算机、计算导弹弹道的专用计算机等，都属于专用计算机的范畴。

通用计算机指各行业和各种工作环境都能使用的计算机，一般具有较高的运算速度、较大的存储容量、配备了比较齐全的外部设备及软件。与专用计算机相比，为了实现其通用性，通用计算机的结构比专用计算机更复杂、价格也更昂贵。

通用计算机的适应能力很强，其应用范围也很广。它的运行效率、速度和经济性等指标在不同的应用场景下表现会有很大差别。我们平常使用的个人台式计算机等都属于通用计算机。

如果按照计算机系统的规模和处理能力等技术指标，从小到大、从弱到强地进行区分，典型的通用计算机可以分为如下几类。

（1）微型机

微型机，俗称微机，是计算机领域中目前发展得最快、应用也最广泛的一种计算机。第一台典型意义上的微机 MCS-4，是由美国的 Intel 公司在 1971 年以自主研发的 4 位微处理器（Intel 4004）为基础，扩展并增加存储系统和输入/输出接口等后形成。从那以后，Intel、MITS 和 IBM 等公司又陆续推出了更新型的微机产品。

常见的微型机有个人台式计算机、笔记本电脑、一体机和工作站等。

微型计算机和与其他类型计算机的主要区别在于，微型计算机广泛采用了集成度相当高的电子元件和独特的总线（Bus）结构。除此之外，微型计算机具有轻便、小巧、价格低、操作和使用方便等特点，其应用范围最广，发展普及也最快，已经成为大众化的信息处理和数字娱乐工具。

（2）小型机

小型机是相对于大型机而言的，小型机的软件、硬件系统规模比较小，但价格低、可靠性高、便于维护和使用。

小型机最初是在 20 世纪 70 年代由美国 DEC 公司首先开发的一种高性能计算产品，曾经风行一时。小型机也曾用来表示一种多用户、采用主机/终端模式的计算机，它的规模和性能介于大型

机和微型机之间。目前，主流小型机的内部一般集成了几十或上百个 CPU，且采用不同版本的 UNIX 操作系统，常作为中高端的专业服务器。国内的服务器领域还习惯性地将各类 UNIX 服务器简称为小型机。

小型机采用的是主机/终端模式，并且各家厂商均有各自的体系结构，如处理器架构、I/O 通道和操作系统软件等都是特别设计的，一般彼此之间互不兼容。与普通服务器相比，小型机还具有高 RAS（Reliability，Availability，Serviceability）特性，即：① 高可靠性，计算机可以 7×24 持续工作永不停机；② 高可用性，重要资源都有备份，能检测到潜在异常，能转移任务到其他资源以减少停机时间保持持续运行，且具备实时在线维护和延迟性维护等功能；③ 高服务性，能够实时在线诊断，精确定位发生的故障，并能做到准确无误的快速修复。

（3）大型机

大型计算机简称大型机，一般作为大型的高性能商业服务器，因其具有较大的体积（通常占地面积几十平方米）而得名。

大型机通常使用专用的处理器指令集（如 IBM 公司的 Z/Architecture 架构 CISC 指令集）、专用的操作系统（如 Z/OS）和专用的应用软件（如 IBM DB2 数据库系统），通常具有较高的运算速度，一般为每秒数千万亿次级别，还具有较大的存储容量，具备较好的通用性，功能也比较完备，能支持大量用户同时使用计算机数据和程序，具有强大的数据处理能力，但大型机的价格相对昂贵。

大型机除了像小型机那样要求高 RAS 性能外，还特别强调了 I/O 数据的吞吐率和处理器指令架构的兼容性。大型机一般通过专用处理器来控制通道进行 I/O 处理，一个 I/O 通道能同时处理多个 I/O 操作、控制上千个 I/O 设备，因此能同时处理上千个数据流，还能保证每个数据流高速运转。大型机的高 RAS、分区和负载能力、I/O 性能优势是其他类型服务器不能匹敌的。大型机处理复杂的多任务时能表现出超强的处理能力，其宕机时间远远低于其他类型的服务器。大型机 I/O 能力强，擅长超大型数据库的访问，采取动态分区管理，根据不同应用负载量的大小，灵活地分配系统资源；从底层防止入侵的设计策略使大型机安全性提高。

目前，能生产大型机的企业主要有美国的 IBM 和 UNISYS 等公司，其中 IMB 公司生产的大型机系列产品几乎占据了全球 90%以上的市场份额，曾于 2005 年与电子科技大学签署合作协议，并提供一台 IBM eServer Z900 大型机用于教学和培训。大型机通常应用在银行、证券和航空等大型企业中，对大数据处理能力和系统的安全性、稳定性等都有极为苛刻要求的应用场合。我国仅有中科院计算所、国防科学技术大学和浪潮等单位能设计和生产大型机。

（4）超级计算机

超级计算机，早期叫巨型机，现在常简称为“超算”。与大型机相比，超级计算机通常由成千上万个计算节点和服务节点组成，具有强大的计算和处理数据的能力，主要特点表现为超高的计算速度和超大的存储容量，并配有多种外部和外围设备及功能丰富的软件系统。

超级计算机是计算机中功能最强、运算速度最快、存储容量最大的一类计算机，多用于国家高科技领域和尖端技术研究，是一个国家科研实力的体现，它对国家安全，经济和社会发展具有举足轻重的意义，是国家科技发展水平和综合国力的重要标志。

超级计算机与大型机的主要区别如下。

① 大型机使用专用指令系统和操作系统，而超级计算机使用通用处理器及 UNIX 或类 UNIX 操作系统，如 Linux 等。

② 大型机擅长非数值计算（数据处理），而超级计算机擅长数值计算（科学计算）。

③ 大型机主要用于商业领域，如银行和电信等，而超级计算机常用于尖端科学领域，特别是

国防和天气预报等领域。

④ 大型机大量使用冗余等技术确保其安全性及稳定性，所以内部结构通常会有备份。而超级计算机使用大量的处理器，通常由数十个机柜组成，体积比大型机更大。

超级计算机主要由高速运算部件和大容量主存等部件构成，一般还需半导体快速扩充存储器和海量（磁盘）存储子系统的支持。超级计算机的主机一般不直接管理低速 I/O 设备，而是通过 I/O 接口连接前端机，一般是小型机，由前端机处理 I/O 任务。此外，I/O 的另一种途径是通过网络，联网用户借助其终端机（微型、小型或大型机）与超级计算机交互，I/O 均由用户终端机来完成，这种方式可大大提高超级计算机的利用率。

为提高系统性能，现代的超级计算机都在系统结构、硬件、软件、工艺和电路等方面采取各种支持并行处理的技术。例如，一般都采用多处理机结构，且处理器除了支持传统的标量数据外，还增加了向量或数组类型数据；硬件方面大多采用流水线、多功能部件、阵列结构或多处理机、向量寄存器、标量运算、并行存储器等多种先进技术。

我国的超级计算机研制开始于 20 世纪 60 年代，由国防科学技术大学慈云桂教授主持研发的国内首台超级计算机“银河-I”于 1983 年 12 月 22 日诞生，使我国继美国、日本之后成为第三个高性能计算机研制生产国。中国现阶段超级计算机的拥有量超过 60 台，代表性的有国防科学技术大学近年研制成功的“天河”系列、中科院研制的“曙光”系列超级计算机以及国家并行计算机工程技术研究中心研制的“神威·太湖之光”超级计算机。无论是保有量还是计算能力，目前我国在超级计算机领域都处于世界领先水平。

注意，计算机领域中对于微型机、小型机、大型机和超级计算机的划分都是一个相对的概念，其划分标准也会随着技术的不断发展而发生动态变化。正因为如此，目前的微型机，其计算性能或许相当于数十年前的小型机，又或者现在的小型机，若干年后，说不定它还赶不上那时微型机的综合性能指标。

1.2 计算机的诞生与发展

1.2.1 冯·诺依曼体系

计算机系统作为一个能够自动地处理信息的智能化工具，必须解决好两个最基本的问题：第一，信息如何表示才能方便地让计算机识别和处理；第二，采用什么工作方式才能使计算机能自动地对信息进行处理。

图 1-5 冯 • 诺依曼

对上述两个问题的解决方案做出杰出贡献并产生深远影响的是一位美籍匈牙利科学家约翰 • 冯 • 诺依曼（John von Neumann，1903—1957），如图 1-5 所示。冯 • 诺依曼在 1944 年加入了美国军方 ENIAC（Electronic Numerical Integrator And Computer，电子数字积分器与计算器）计算机的研制项目。在此研制基础上，他在 1945 年提出并发表了一个全新的“存储程序通用电子计算机”方案——EDVAC（Electronic Discrete Variable Automatic Computer）。在这份方案中，冯 • 诺依曼具体介绍了制造电子计算机和程序设计的新思想，这份方案是计算机发展史上一个划时代的文献，它向世界宣告了电子计算机时代的来临。

在 EDVAC 技术方案中，冯 • 诺依曼明确提出新型计算机的结构应由五大部分组成：运算器、控制器、存储器、输入设备和输出设备，并且阐述了这五大部分的功能和相互之间的逻辑关系。此外，他还对 EDVAC 的两大设计思想做了进一步的论证，为新型电子计算机的设计指明了方向、树

立了一座里程碑，具有划时代的意义。

设计思想之一是信息采用二进制来表示。他根据电子元件的双稳态工作特点，建议在电子计算机中采用二进制来表示信息，他还分析了二进制的优点，并预言二进制的采用将大大简化计算机硬件系统的逻辑结构。

设计思想之二是程序存储的思想。把运算程序存在计算机的存储器中，程序设计员只需在存储器中寻找运算指令，计算机就能自行计算，这样就不必对每个问题重新编程，从而大大加快了运算的进程。程序存储的思想标志着计算机期望的自动运算的实现，同时标志着电子计算机设计思想的成熟，且已成为电子计算机设计的基本原则。

总体上，冯·诺依曼体系的主要思想可概括为：① 计算机硬件系统由五大部件（存储器、运算器、控制器、输入设备和输出设备）组成；② 计算机中采用二进制形式表示信息（数据、指令）；③ 采用存储程序的工作方式，这也是冯·诺依曼体系最为核心的思想。

冯·诺依曼针对 EDVAC 提出的新型电子计算机体系结构及设计思想，奠定了现代电子计算机设计的理论基础，并开创了程序设计的新时代。因此，他提出的这些计算机设计思想也被称为冯·诺依曼思想，相应的计算机体系结构也被称为冯·诺依曼体系结构，也正是因为对计算机的这种划时代贡献，他被后世尊称为“计算机之父”。

采用冯·诺依曼思想体制的计算机就称为冯·诺依曼计算机。几十年来，尽管计算机的体系结构已经发生了许多演变，但它们总体上仍然没有脱离冯·诺依曼体制的核心思想原则，现在的绝大多数计算机仍然属于冯·诺依曼机体系结构的计算机。当然，从本质上讲，传统的冯·诺依曼机采用的是串行处理的工作机制、逐条执行指令序列。要想提高计算机的性能，其根本方向之一是采取并行处理机制，如用多个处理部件形成流水线，依靠时间上的重叠来提高处理效率，或者用多个冯·诺依曼机组成多机系统，以支持并行算法等。

1.2.2 计算机发展历程

1946 年 2 月 14 日，在美国的宾夕法尼亚大学，世界上第一台严格意义上的电子数字计算机 ENIAC 诞生了，它标志着电子计算机时代的到来。这台计算机由美国军方定制，专门为了计算导弹的弹道和射击特性而研制。承担开发任务的“莫尔小组”主要由四位科学家和工程师埃克特、莫克利、戈尔斯坦、博克斯组成，后来冯·诺依曼作为技术顾问也加入了该小组。ENIAC 的主要元器件采用的是电子管，共使用了 1500 个继电器，18800 个电子管，占地约 170 m^2，重达 30 多吨，耗电 150 kW，造价高达 48 万美元，每秒能完成 5000 次加法运算，400 次乘法运算，比当时最快的计算工具快 300 倍，是继电器计算机的 1000 倍、手工计算的 20 万倍。由于研制工期和技术原因，ENIAC 并没遵照冯诺依曼体系进行设计，比如无存储器，仍然采用打孔卡记录信息。

第一台电子计算机于 1946 年诞生后，发展到现在已经历了几十年，也经历了不同的技术发展阶段，且每个阶段各具特色。

1．第一代，电子管计算机（1946—1957 年）

这一阶段计算机的主要特征是采用电子管元件作基本器件，一般使用光屏管或汞延时电路作为存储器，输入或者输出主要采用穿孔卡片或纸带等，通常体积比较大、功耗较高、计算速度较慢、存储容量较小、可靠性比较差、维护很困难而且价格昂贵。在程序设计上，通常使用机器语言或者汇编语言来编写程序，这一代计算机主要用于军事科学计算。

电子计算机在我国的研制起步较晚。1957 年 4 月，我国购买了 M-3 计算机的技术资料，在这些技术资料的基础上开始研制。以中科院计算所研究员张梓昌和莫根生为主，组织了 M-3（代号为