

内部讲义

计算机中文信息处理

第四册

赵珀璋编著

空军第三研究所

北京信息控制研究所

中国科学院研究生院

一九八四年七月

第四分册目录

第十章 中文信息通信	3
§ 10-1 数据通信基础知识	7
10-1-1 数据传输线路分类	7
10-1-2 数据交换分类	10
§ 10-2 传输控制	13
§ 10-3 汉字通信有关问题	18
§ 10-4 通信规程转换器	21
§ 10-5 国家标准 GB34 53-82 “数据通信基本型控制规程”简介	25
10-5-1 字符结构	26
10-5-2 传输控制字符	27
10-5-3 文电格式	30
10-5-4 数据通信阶段	35
10-5-5 差错控制	35
10-5-6 恢复规程	36
10-5-7 会话型传输控制规程	37
10-5-8 编码独立的信息传输控制规程	37
10-5-9 多站选择	37
10-5-10 全双向传输控制规程	38
§ 10-6 国家标准 GB34 54-82 “数据终端设备 (DTE) 和数据电路终接设备 (DCE) 之间的接口电路定义表”介绍	38

10-6-1 适用范围	38
10-6-2 分界线	40
10-6-3 接口电路的定义	40
10-6-4 操作要求	58
参考文献	65

第十一章 汉字和汉语语音识别	67
§ 11-1 汉字的特征与表示	70
§ 11-2 汉字识别的预处理	74
11-2-1 汉字的平滑处理	78
11-2-2 汉字图形的细化处理	81
§ 11-3 汉字笔画有向特征抽取方法	87
§ 11-4 印刷体汉字识别	95
11-4-1 轮廓投影及其幅质谱法	95
11-4-2 分级图形匹配法	96
11-4-3 模板匹配法	97
11-4-4 几种判决函数方法	97
§ 11-5 印刷体汉字识别的粗分类方法	98
11-5-1 汉字R变换码粗分类方法	98
11-5-2 汉字外轮廓四角八码粗分类方法	100
11-5-3 汉字内轮廓四角八码粗分类方法	103
11-5-4 汉字脉冲码粗分类方法	105
11-5-5 汉字长短码粗分类方法	108
§ 11-6 轮廓投影识别法	111
§ 11-7 手写体汉字识别	114
11-7-1 联机识别系统	114

11-7-2	脱机识别系统	127
§ 11-8	模糊集合论简介	134
11-8-1	模糊集合的定义	134
11-8-2	模糊集合的运算	135
11-8-3	模糊关系	137
11-8-4	模糊矩阵	138
11-8-5	模糊逻辑和模糊算子	140
11-8-6	模糊算子	142
§ 11-9	汉字模糊笔画特征抽取方法	143
11-9-1	汉字笔画特征	143
11-9-2	特征判别方法	145
§ 11-10	汉字模糊笔画笔顺分析	148
11-10-1	汉字的模糊结构	148
11-10-2	汉字的 2 维构形向一维笔序的变换	153
§ 11-11	语音信号处理基础	156
11-11-1	语音信号	156
11-11-2	语音信号处理	157
11-11-3	数字信号处理	157
11-11-4	数字语音处理	158
11-11-5	语音处理的时域方法	160
11-11-6	语音波形的数字表示	164
11-11-7	短时付里叶分析	165
11-11-8	同态语音处理	166
11-11-9	语音的线性预测编码	167
11-11-10	人机语音通信的数字语音处理	167

§ 11-12 汉语的语音特点	176
§ 11-13 微型机语音识别接口	178
11-13-1 设计原理	178
11-13-2 软件结构及调用	192
参考文献	195
第十二章 汉语自然语言处理	197
§ 12-1 语言工程	200
§ 12-2 汉语语言理解系统	204
§ 12-3 汉语拼音处理系统	209
12-3-1 汉语拼音输入系统 HP-2	210
12-3-2 自动切分系统	212
12-3-3 同音的自动处理	215
§ 12-4 机器翻译概述	216
§ 12-5 机器翻译的两个主要方面	222
12-5-1 机器翻译中汉语有关问题	222
12-5-2 机器翻译专用软件包	229
§ 12-6 机器翻译系统设计的几种方法	233
12-6-1 机器翻译的分类法	233
12-6-2 机器翻译程序设计方法	233
§ 12-7 机器翻译中目标语的生成问题	238

12-7-1 短语(成分)词组的转换与 生成	238
12-7-2 以动词为核心的句法结构转 换与生成	242
§ 12-8 自然对话智能机	245
12-8-1 自然对话基本方法	246
12-8-2 语义框架及其在语言理解过 程中的操作	247
§ 12-9 中文信息五维模型	250
参考文献	259
第十三章 几个国家标准	261
§ 13-1 国家标准“文字和符号成形设备 的增补控制功能”编制说明	261
13-1-1 原则	262
13-1-2 几个具体问题	262
§ 13-2 国家标准“信息交换用汉字编码 字符集辅助集”设计思想	276
§ 13-3 汉字编码字符集的构成	283
附录 I 国家标准 GB 1988-80 “信息 处理交换用的七位编码字符集”	292

附录 II	国家标准 GB 2311-80 “信息处理交 换用七位编码字符集的扩充方法”	303
附录 III	国家标准 GB 2312-80 “信息交换用 汉字编码字符集基本集” (摘录)	419

计算机诞生至今已快40年了，它经历了四代更新，即电子管时代，晶体管时代，集成电路时代和大规模集成电路时代，第五代计算机即超大规模集成电路及智能化时代正处在孕育阶段，不久将会诞生。伴随着计算机本身的发展，同样计算机的应用发展也经历了四个阶段：

第一阶段：点上应用——“0维”应用。

早期，计算机作为一台单机，放在机房里。由操作员进机房使用计算机，而用户不与计算机直接打交道。从几何来看，计算机房是一个点，即“0”维应用。如图10-1所示。

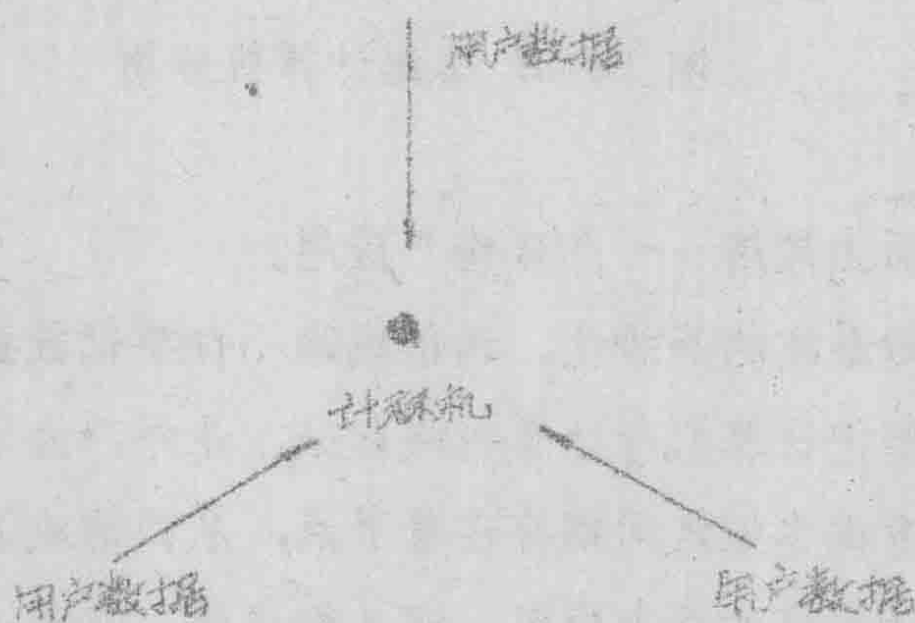


图10-1 “0”维计算机应用

第二阶段，线上应用——“1维”应用。

随着计算机实时系统的发展，远离计算机的用户，通过数据通信线路，把其数据终端与计算机联接起来，可直接与计算机打交道，而不必由人将数据送到计算机房交操作员处理。显然，这就是联机实时

系统，可称为计算机的线上应用，即 1 维应用。它提高了计算机的使用效率，方便了用户。如图 10-2 所示。

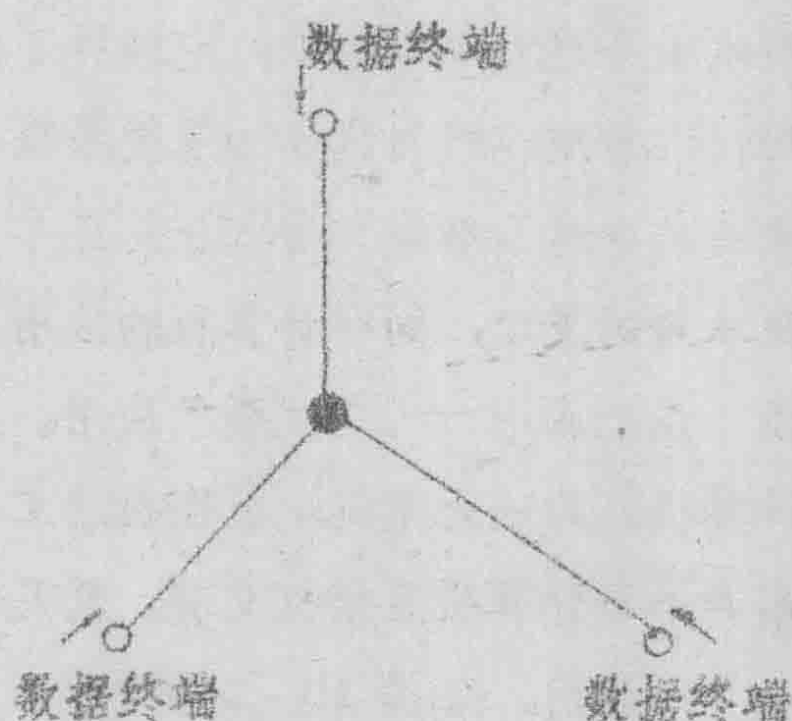


图 10-2 1 维计算机应用

第三阶段：面上应用——“2 维”应用。

随着联机实时系统的高级化，通信线路不仅要把数据终端与计算机相联，而且还要把计算机与计算机联起来。此外，由于计算机相互之间、计算机与数据终端之间都存在着节点，从而构成了复杂的计算机网络。这就是计算机的面上应用，即 2 维应用。它进一步提高了资源共享，促进了分布式处理系统、网络通信和数据库系统的发展。其结构示意图如图 10-3 所示。

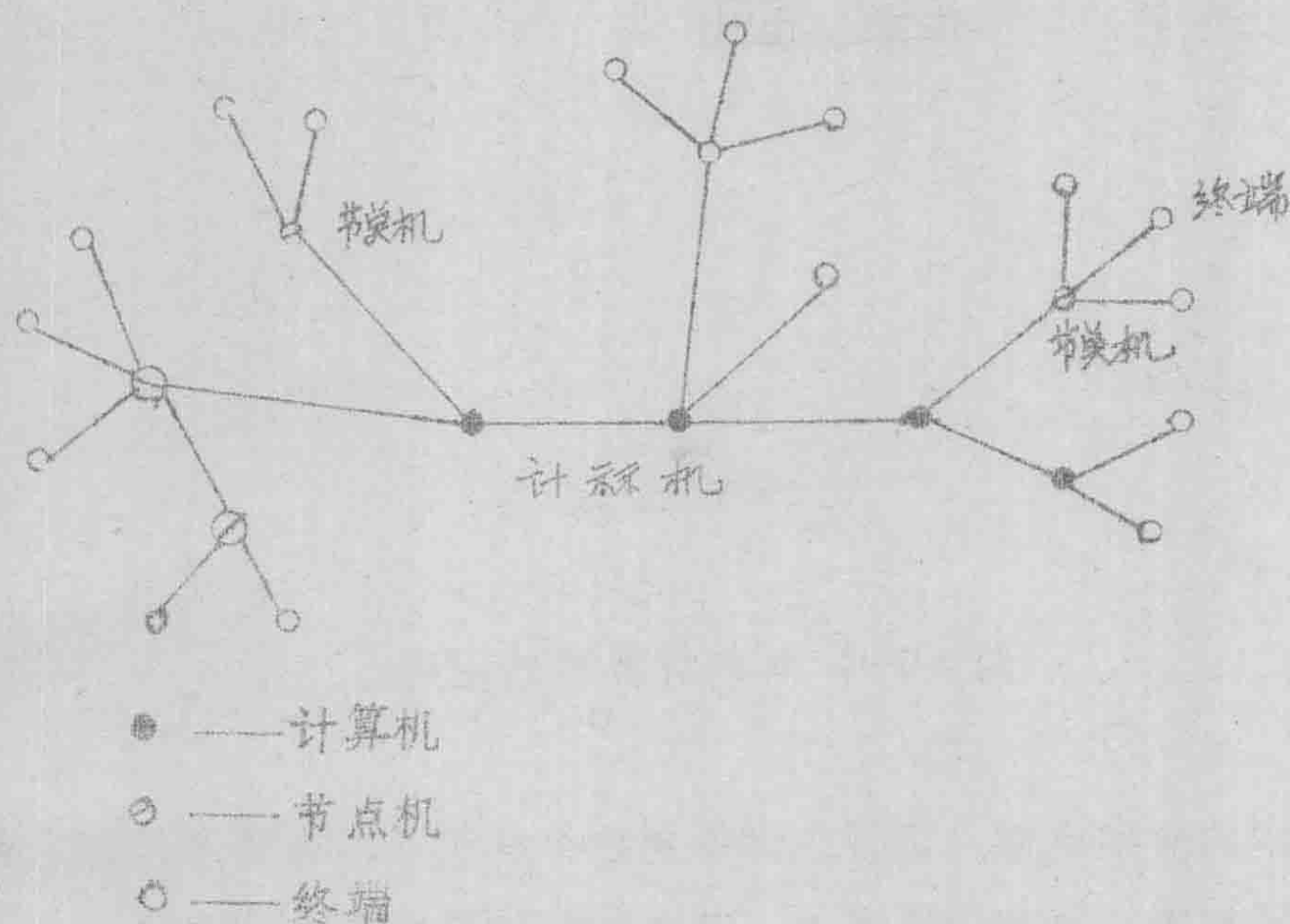


图 10-3 2 维计算机应用

第四阶段 体上应用——“3 维”应用

随着航天技术的发展，以通信卫星为介质的数据通信网也得到了飞速发展。各个分散的计算机网络，可以通过通信卫星进一步联起来，这就是计算机的空间应用，即“3 维”应用。它实现了整个人类的数据共享。促进了国际性数据库的发展。图 10-4 示出了计算机空间应用的示意图。

通过计算机应用的四个发展阶段可以看出，计算的应用与计算机的自身发展是紧密相关的，同时也与数据通信紧密相连。换句话说，计算机与数据通信这两个原来相互独立的分支，现在已发展成彼此紧密相关。进而还将发展使之合二为一成为一体。

事实上，数据处理和数据传输是数据系统中不可分割的一体，也是数据系统设计的基本。文献〔10-3〕指出，在数据系统设计中

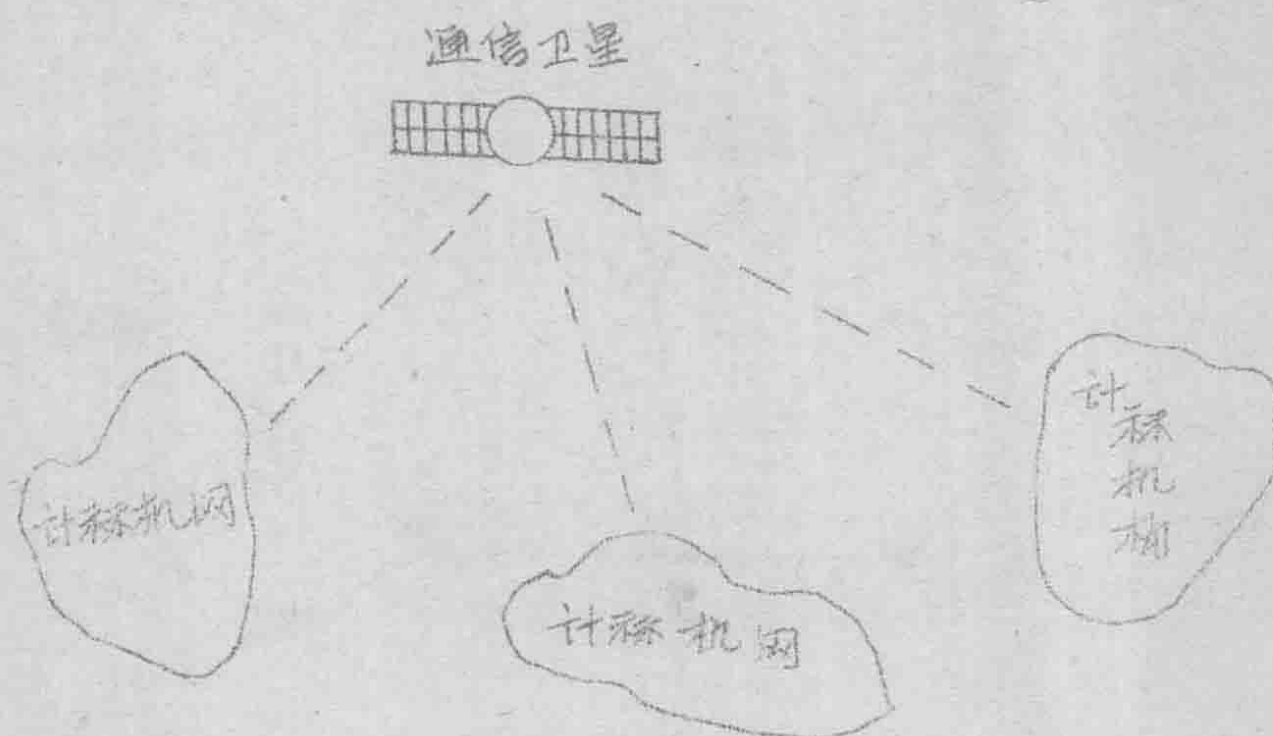


图 10—4 3 维计算机的应用

(从基本的数字电路、部件、设备到整个系统)，都要注意到，数据在处理过程中，要有传输能力，而在传输过程中要有处理能力。今天从计算机应用的四个阶段来看，已充分证明了这个论断。

从电信角度来看，通信与计算机相结合也是必然的结果。通信内容由文字符号发展到图形图象，由静态发展到动态，进而发展到实时通信，这些都需要计算机去处理、控制。

中文信息处理和中文信息通信，事实上都是以计算机为基础，也是计算机应用的一个领域。

正如前面各章介绍的那样，从计算机处理角度来看，中西文区别仅在字符代码集。西文以一个字节为单位，而中文以两个字节为单位。同样从通信角度来看，也是如此。

实际上，由于通信系统的透明传输方式，因而能将中西文的差别减至更小。

和中西文信息处理系统一样，中西文信息通信时需要特征控制，即如何区别中文和西文，这在中文通信时必须解决之。

为便于讨论中文信息通信，为此先介绍一点数据通信基础知识。

§ 10-1 数据通信基础知识

电气通信按其业务内容可分为电报、电话、传真、电视和数据通信等方面。而按其传递的信号形式又可分为模拟通信和数字通信。如果从传输介质来分，可分为有线通信和无线通信。

10-1-1 数据传输线路分类

要通信就得有两个或两个以上的站点，也就是说数据通过通信线路从一个地方传到另一个地方去。通常线路连接方式有下面几种：

1. 点一点方式

用一条线路连接两个通信站进行通信，其中一个称为主站，另一个称为从站。采用通信线路通常有二种。

(1) 专用线路

在终端设备和主机之间构成“一对一”连接线路。这种线路适用于连接速率较高，数据较长，连续传输的终端设备。这类线路连接的终端数目可较多些，但控制简单。如图 10-5 所示。



图 10-5 专用线路

(2) 交换线路

在终端数目较多但使用次数较少的系统中，终端设备通过交换机和主机连接。这比专用线路要经济的多。如图 10-6 所示。



图 10-6 交换线路

2. 多点方式

用一条通信线路连接两个以上通信站进行通信。其中一个为控制站，余下为从属站。在控制站控制下，可以进行探测、选择、传输和广播等操作。其特点是线路可被多个站共享。通信线路通常有下面几种：

① 分支线路

分支线路能从主机引出一条或多条多点线路。每条多点线路接有多个终端，但一次只能有一个终端和主机进行通信。每个终端在多点线路上拥有自己的地址，并能识别线路上传输的地址信息。对于使用次数频繁，数据输入输出速率较慢的终端设备，但每次要处理的信息长度较短，运算操作较多比较合适。如图 10-7 所示。

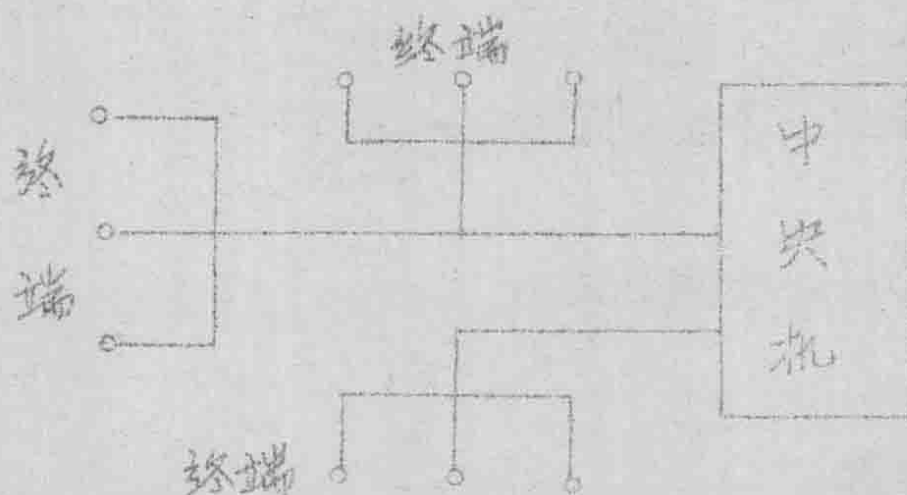


图 10-7 分支线路

② 环形线路

环形线路是多点线路的特殊形式。它适合于在局部区域终端设备密集的系统。在线路上的每个报文，包括终端地址和数据，将环形地同步传输。如图 10-8 所示。

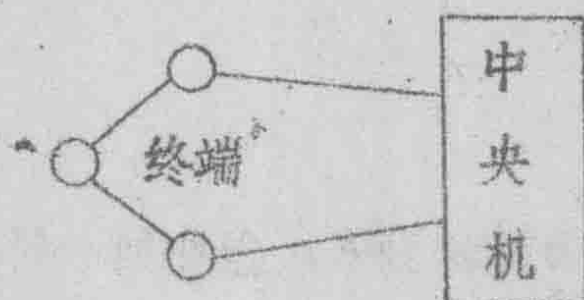


图 10-8 环形线路

③ 集中线路

此方式是当远地终端与中心机联接时，在终端群近旁设置线路汇集站，用线路集中器把终端线路（即连接终端与集中器的线路）集中起来，再通过中继线（即连接集中器和中心的线路）与中心相连接。如图 10-9 所示。

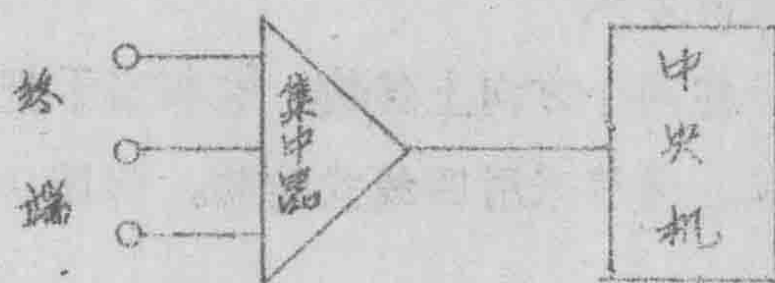


图 10-9 集中线路

根据通信线路的通信方式，数据传输线路可分为下面三种：

1. 单工通信方式

此方式下数据仅能按一固定方向作单向传输。如图 10-10 所示。图中数据只能由 A 传送到 B，而不能由 B 传送到 A，即 B 点只能接收数据，但 B 却能把监视信号传送到 A。

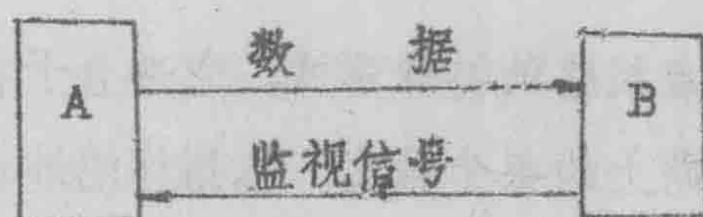


图 10-10 单工通信方式

2. 半双工通信方式

数据从 A 传送到 B 或由 B 传送到 A，但在同一时间内只能作单向传输。同样监视信号也不可同时按两方向传送。如图 10-11 所示。此方式适用于终端之间的会话式通信。通常采用二线式结构。

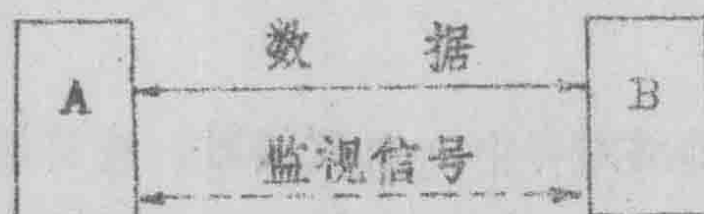


图 10-11 半双工通信方式

3. 全双工通信方式

数据可以同时两个方向上传输。它相当于把两个相反的单工通信方式组合在一起。通常采用四线式结构。如图 10-12 所示。

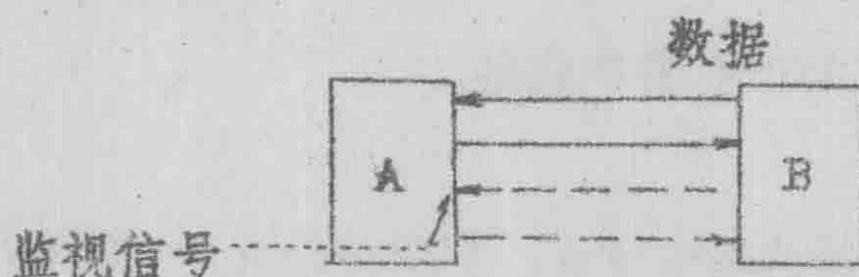
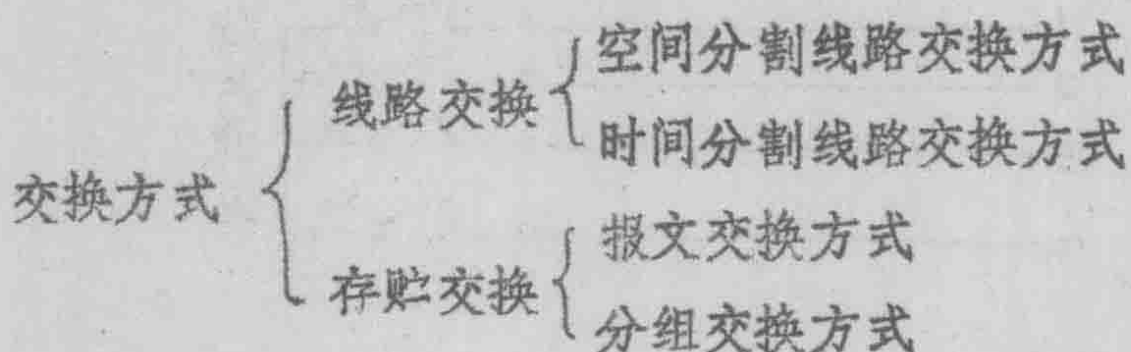


图 10-12 全双工通信方式

10-1-2 数据交换分类

通信各方需要交换信息。数据交换方式通常按下述方式分类：



1、线路交换方式

通过直接切换通信线路来实现交换的方式称为线路交换方式。根据线路通道构成方式，交换可以分为空间分割线路交换和时间分割线路交换两种。

空间分割线路交换是指每一条通道都是由一单独的导线构成的。在空间分割的交换机接续网络中，每一用户各占有一定的空间位置，有自己独用的导线。整个接续网很象一个坐标网，用户之间的接续均在相应空间位置的一个交换结点上进行，如图 10-13 所示。

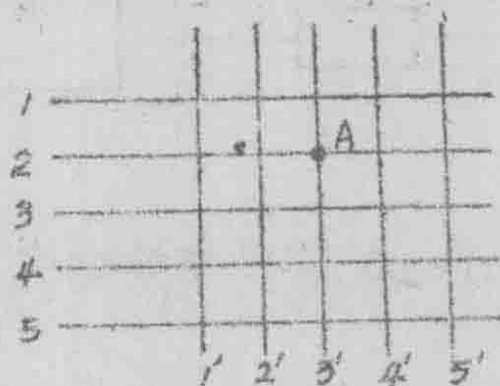


图 10-13 空间分割的接线器

时间分割线路交换是指许多话路合用一个公共通道，只是时间上分成先后占用它。按时间分割构成话路时，每一话路在公共通道的循环周期中只占用其中一个特定的时间片。在时间分割电子交换机中，许多用户共用一根公共导线（公共通道），每一用户通过一个电子开关与公共通道相连。如图 10-14 所示。

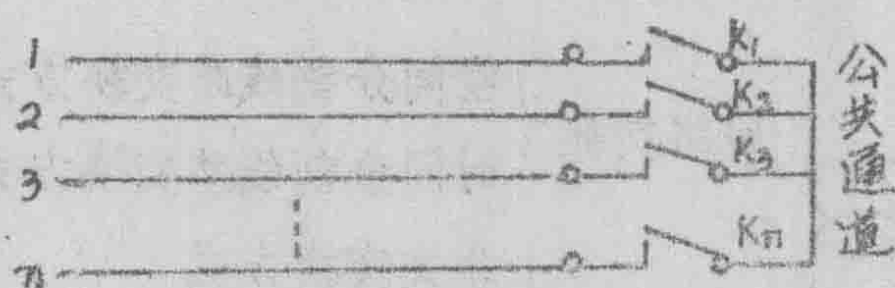


图 10-14 时间分割的接续设备

2. 存贮交换方式

存贮交换方式是指将信息暂时存入输入缓冲存贮器，然后对存贮好的信息作适当的处理，并等待输出口空闲。当输出口一空便送出去。如图 10-15 所示。

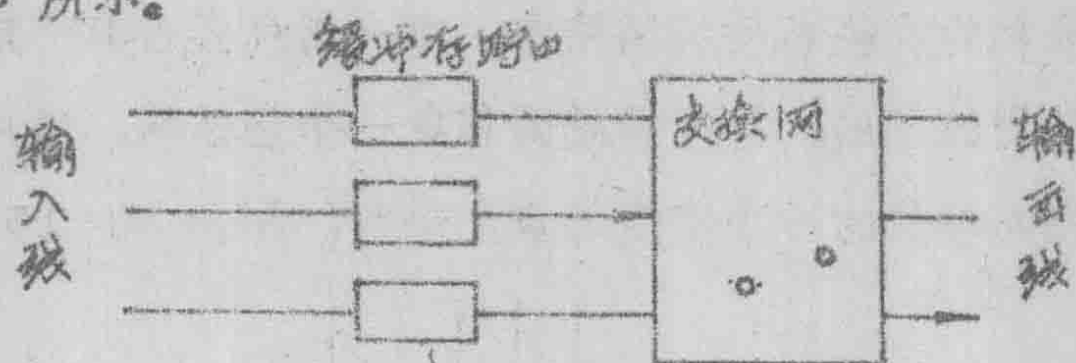


图 10-15 存贮交换方式

3. 分组交换方式

分组交换方式是把数据报文分割成若干个长度较短（最长不超过 255 个字符）的分组进行传输。每个分组内，除数据信息外还包括有地址标识，分组编号和校验信息等。报文是在发送端进行分割的。每个分组可经过不同的路径传输，途中历经多个结点时，象报文交换那样，也采用存贮转发方式传输，并进行检查、重发，回送应答等工作；最后当接收端正确收到全部分组时，再按分组编号重新编集成数