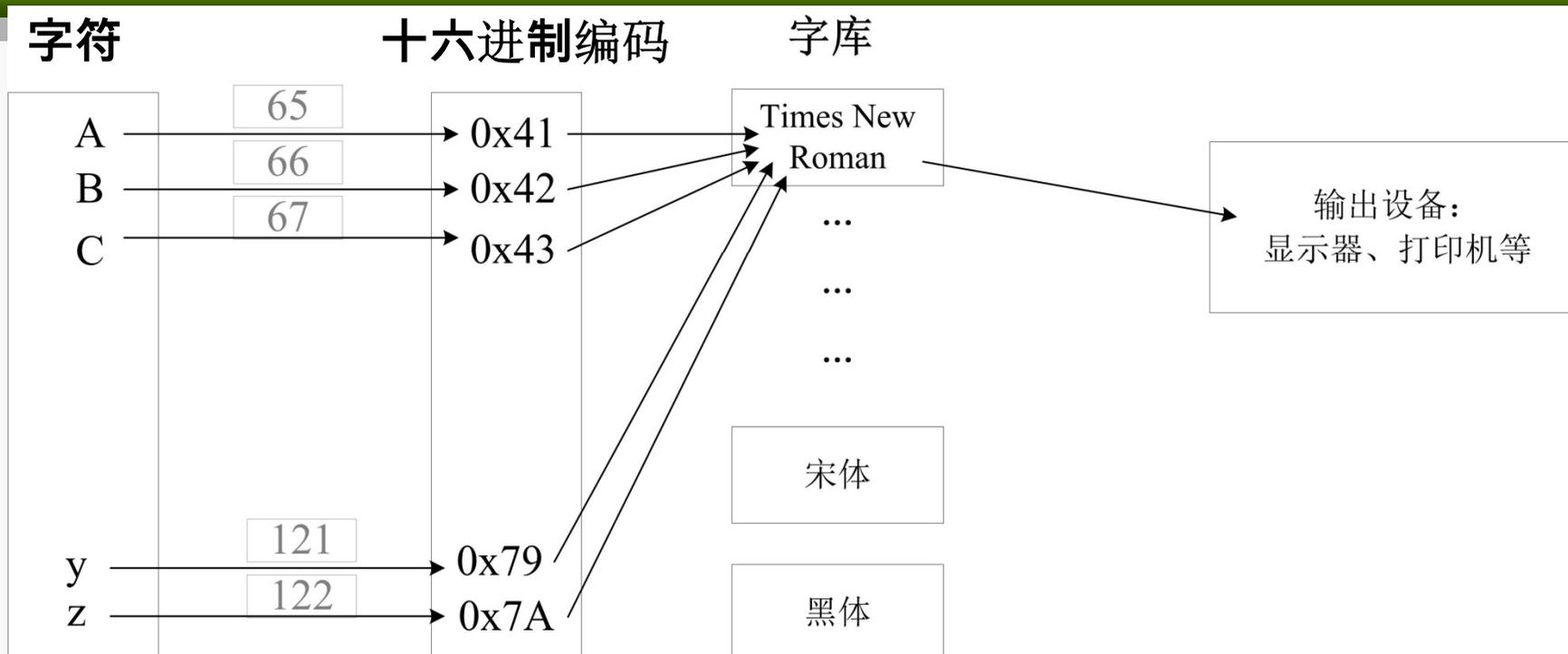


计算机现实字符的原理



点阵字库：也叫位图字体，其中每个字形都以一组二维像素信息表示。DOS时代被普遍采用。

矢量字库：保存的是对每一个汉字的描述信息，比如一个笔划的起始、终止坐标，半径、弧度等等。要经过一系列的数学运算才能输出结果，理论上可以被无限地放大，现在普遍采用，打印时使用的字库均为此类字库。



ASCII 码表

ASCII值	控制字符	ASCII值	控制字符	ASCII值	控制字符	ASCII值	控制字符
0	NUL	32	(space)	64	@	96	`
1	SOH	33	!	65	A	97	a
2	STX	34	"	66	B	98	b
3	ETX	35	#	67	C	99	c
4	EOT	36	\$	68	D	100	d

.....

25	EM	57	9	89	Y	121	y
26	SUB	58	:	90	Z	122	z
27	ESC	59	;	91	[123	{
28	FS	60	<	92	/	124	
29	GS	61	=	93]	125	}
30	RS	62	>	94	^	126	`
31	US	63	?	95	_	127	DEL

GB2312-80 (1981年)

1. 一个汉字的机内码目前常用两个字节来表示：第一字节是区位码的区号加 $(160)_{10}$ ；第二字节是区位码的位号加 $(160)_{10}$ 。已知：汉字“却”的机内码是 $(11001000\ 10110100)_2$ ，则根据下图所示的汉字区位码表，汉字“确”的机内码用十六进制表示，应为(D)
- A. C7B4 B. C7B6
C. C8B4 D. C8B7

区\位	01	...	19	20	21	22	23	...	94
01
...
16	啊	...	吧	笆	八	疤	巴	...	剥
17	薄	...	鄙	笔	彼	碧	蓖	...	炳
...	...								
40	取	...	瘸	却	鹊	榷	确	...	叁
...	...								
94	...								



ASCII：美国信息交换标准代码，到目前为止共定义了128个字符
GB2312-80: 1980年制定的中国汉字编码国家标准。共收录7445个字符，其中汉字6763个。GB2312兼容标准ASCII码，采用扩展ASCII码的编码空间进行编码，一个汉字占用两个字节，每个字节的最高位为1。

GBK《汉字内码扩展规范》(GBK)于1995年制定，兼容GB2312、GB13000-1、BIG5编码中的所有汉字，使用双字节编码，编码空间为0x8140~0xFEFE，共有23940个码位，其中GBK1区和GBK2区也是GB2312的编码范围。收录了21003个汉字、

GB18030-2005《信息技术中文编码字符集》是我国制订的以汉字为主并包含多种我国少数民族文字（如藏、蒙古、傣、彝、朝鲜、维吾尔文等）的超大型中文编码字符集强制性标准，其中收入汉字70000余个



Unicode、UTF-8、UTF-16、UTF-32相互编码和解码。**Unicode** 是容纳世界所有文字符号的国际标准编码，使用四个字节为每个字符编码。

UTF 是英文 Unicode Transformation Format 的缩写，意为把 Unicode 字符转换为某种格式。目前，这些衍变方案中 UTF-8 被广泛使用，而 UTF-16 和 UTF-32 则很少被使用。

UTF-8 使用一至四个字节为每个字符编码，其中大部分汉字采用三个字节编码，少量不常用汉字采用四个字节编码。因为 UTF-8 是可变长度的编码方式，相对于 Unicode 编码可以减少存储占用的空间，所以被广泛使用。

UTF-16 使用二或四个字节为每个字符编码，其中大部分汉字采用两个字节编码，少量不常用汉字采用四个字节编码。

UTF-32 使用四个字节为每个字符编码，使得 UTF-32 占用空间通常会其它编码的二到四倍。

中日韩统一汉字

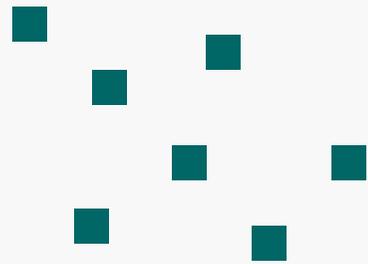
- CJK方案

东亚汉字统一，包括GBK、BIG5、韩文、日文、越南语用到的汉字。

- 编码范围： U+0800~U+FFFF

实际只用到U+4E00~U+9FA5

- Unihan.txt



Thank you

还有什么疑问可以到www.mrbccd.com提出
也可以发送邮件到 mingrisoft@mingrisoft.com

