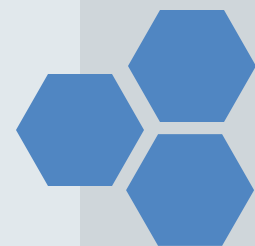


计算机组成原理与系统结构

第三章 信息编码与数据表示

<http://jpkc.hdu.edu.cn/computer/zcyl/dzkjdx/>





第三章 信息编码与数据表示

3.1 数值数据的表示

3.2 数据格式

3.3 定点机器数的表示方法

3.4 浮点机器数的表示方法

3.5 非数值数据的表示

3.6 校验码

3.7 现代计算机系统的数据表示

本章小结

BACK



3.5 非数值数据的表示

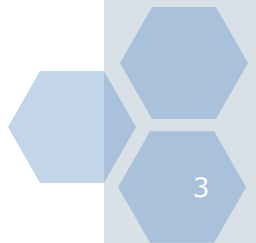
- ❖ 非数值数据：文字和符号（**字符**）、图像、声音等
- ❖ 非数值数据的表示：对其进行二进制编码



字符编码



汉字编码





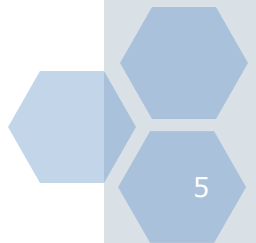
一、字符编码

1. **字符的表示**：采用字符编码，即用规定的二进制数表示文字和符号的方法。
2. **ASCII码**：美国标准信息交换码，为国际标准，在全世界通用。
 - ❖ 常用的**7位ASCII码**的每个字符都由7个二进制位b6~b0表示，有128个编码，最多可表示128种字符；其中包括：
 - 10个数字 ‘0’ ~ ‘9’ : 30H~39H，顺序排列
 - 26个小写字母 ‘a’ ~ ‘z’ : 61H~7AH，顺序排列
 - 26个大写字母 ‘A’ ~ ‘Z’ : 41H~5AH，顺序排列
 - 各种运算符号和标点符号等。



ASCII码分类

- ❖ **95个可打印或显示的字符**：称为图形字符，可在打印机和显示器等输出设备上输出；可在计算机键盘上找到相应的键。
- ❖ **33个控制字符**：不可打印或显示，分成5类：
 - ① 10个传输类控制字符：用于数据传输控制； ■
 - ② 6个格式类控制字符，用于控制数据的位置 ■
 - ③ 4个设备类控制字符，用于控制辅助设备； ■
 - ④ 4个信息分隔类控制字符，用于分隔或限定数据 ■
 - ⑤ 9个其他控制字符、空格字符和删除字符。





ASCII 码编码表

	000	001	010	011	100	101	110	111
0000	NUL	DLE	SP	0		P	,	p
0001	SOH	DC1	!	1	A	Q	a	q
0010	STX	DC2	“	2	B	R	b	r
0011	ETX	DC3	#	3	C	S	c	s
0100	EOT	DC4	¥	4	D	T	d	t
0101	ENQ	NAK	%	5	E	U	e	u
0110	ACK	SYN	&	6	F	V	f	v
0111	BEL	ETB	'	7	G	W	g	w
1000	BS	CAN	(8	H	X	h	x
1001	HT	EM)	9	I	Y	i	y
1010	LF	SUB	*	:	J	Z	j	z
1011	VT	ESC	+	;	K	[k	{
1100	FF	FS	,	<	L	,	l	
1101	CR	GS	-	=	M]	m	}
1110	SO	RS	.	>	N	^	n	~
1111	SI	US	/	?	O	_	o	DEL





基于IBM ProPrinter打印机的扩展ASCII码

$B_7B_6B_5B_4$ $B_3B_2B_1B_0$	0000	0001	1000	1001	1010	1011	1100	1101	1110	1111
0000		▶	Ç	É	á	▒	Ⓕ	Ⓖ	α	≡
0001	☺	◀	ü	æ	í	▒	Ⓖ	Ⓙ	β	±
0010	●	↕	é	Æ	ó	▒	Ⓙ	Ⓟ	Γ	≥
0011	♥	!!	â	ô	ú		Ⓡ	Ⓖ	π	≤
0100	♦	¶	ä	ö	ñ	Ⓡ	—	Ô	Σ	∫
0101	♣	§	à	ò	Ñ	Ⓡ	Ⓡ	Ⓕ	σ	∫
0110	♠	■	å	û	ª	Ⓡ	Ⓡ	Ⓟ	μ	÷
0111	●	↕	ç	ù	º	Ⓟ	Ⓡ	Ⓡ	τ	≈
1000	◻	↑	ê	ÿ	¿	Ⓡ	Ⓖ	≠	Φ	≈
1001	○	↓	ë	Ö	⌈	Ⓡ	Ⓕ	Ⓡ	Θ	·
1010	■	→	Ⓡ	Ü	⌋		Ⓖ	Ⓡ	Ω	·
1011	♂	←	ï	ç	½	Ⓡ	Ⓙ	■	δ	√
1100	♀	└	î	£	¼	Ⓡ	Ⓡ	■	∞	n
1101	♪	↔	ì	¥	¡	Ⓡ	=	■	φ	²
1110	♫	▲	Ä	Pts	«	Ⓡ	Ⓡ	■	ε	■
1111	⚙	▼	Å	f	»	Ⓡ	Ⓖ	■	∩	





二、汉字编码

对于汉字，计算机的处理技术必须解决三个问题：汉字输入、汉字储存与交换、汉字输出，它们分别对应着**汉字输入码、交换码、内码、字形码**的概念。

1. 汉字输入码

汉字输入码也称**外码**，是为了将汉字输入计算机而编制的代码，**是代表某一汉字的一串键盘符号**。

- 汉字输入码种类：
 - **数字编码**：如区位码、国标码、电报码等。
 - **拼音编码**：如全拼码、双拼码、简拼码等。
 - **字形编码**：如王码五笔、郑码、大众码等。
 - **音形编码**：如表形码、钱码、智能ABC等。



二、汉字编码

2、汉字交换码：指不同的具有汉字处理功能的计算机系统之间在**交换汉字信息**时所使用的代码标准。

- **目前国内标准信息处理交换码：**基于1980年制定的国家标准《信息交换用汉字编码字符集·基本集》（GB2312-80）修订的**国标码**。
- 共收录了**6763个汉字和682个图形符号**。6763个汉字分为一级常用汉字3755个，二级次常用汉字3008个。其中一级汉字按拼音字母顺序排列，二级汉字按偏旁部首排列。
- 采用**两个字节对每个汉字进行编码**，每个字节各取七位，可对 **$128 \times 128 = 16384$** 个字符进行编码。



二、汉字编码

两种典型的数字编码作为交换码：

- ① **区位码**：是将国家标准局公布的6763个两级汉字分为**94个区**，**每个区分94位**，实际上把汉字表示成二维数组，**每个汉字在数组中的下标就是区位码**。例如“中”字位于54区48位，“中”字的区位码即为“5448”。
- ② **国标码**：将**区位码加2020H**，占用两个字节。例如“中”字的国标码为区位码5448的区码和位码转化为16进制，为3630H，再加2020H得国标码5650H。



二、汉字编码

3、汉字内码

- 汉字内码是用于**汉字信息的存储、交换、检索**等操作的机内代码，一般采用**两个字节**表示。
- 汉字可以通过不同的输入法输入，但其内码在计算机中是**唯一**的。
- **英文字符**：七位的ASCII码，字节的**最高位为“0”**。
- **汉字机内代码**：2个字节的**最高位均为“1”**。
- **汉字机内码=汉字国标码+8080H**。例如“中”字的机内码为D6D0H。
- 文本文件中储存的是汉字内码。



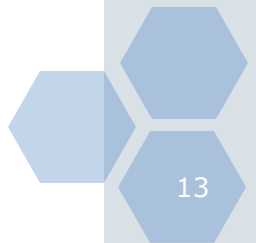
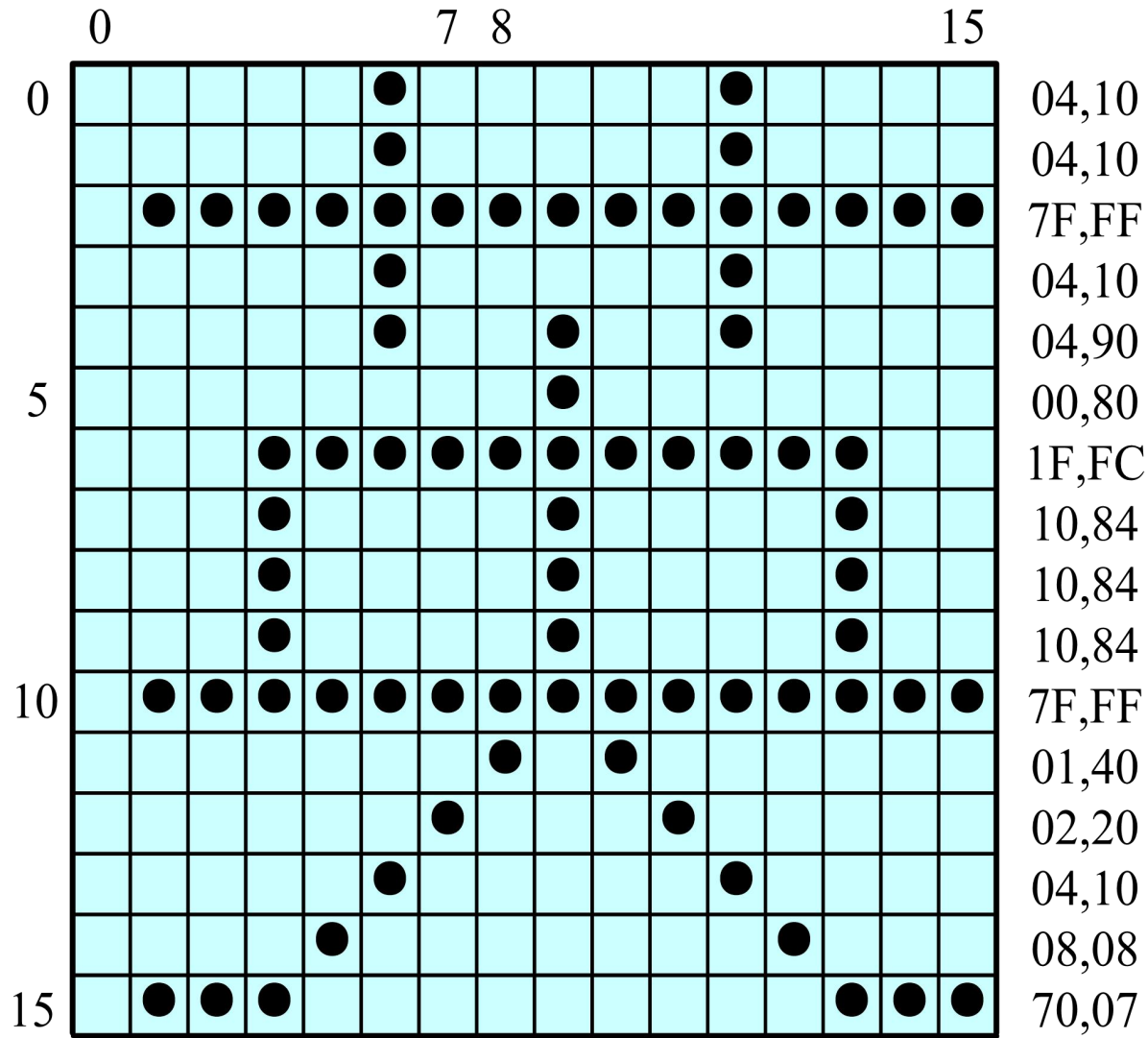
二、汉字编码

4、汉字字形码

- 汉字字形码是将汉字字形经过点阵数字化后形成的一串二进制数，用于汉字的**显示和打印**。
- 根据汉字输出的要求不同，点阵有以下几种：
 - **简易型汉字**：16×16， 32字节/汉字
 - **普通型汉字**：24×24， 72字节/汉字
 - **提高型汉字**：32×32， 128字节/汉字。
- **汉字字库**：将所有汉字的**字模点阵代码**按内码顺序集中起来，构成了汉字库。

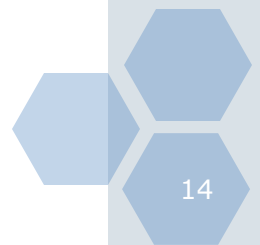
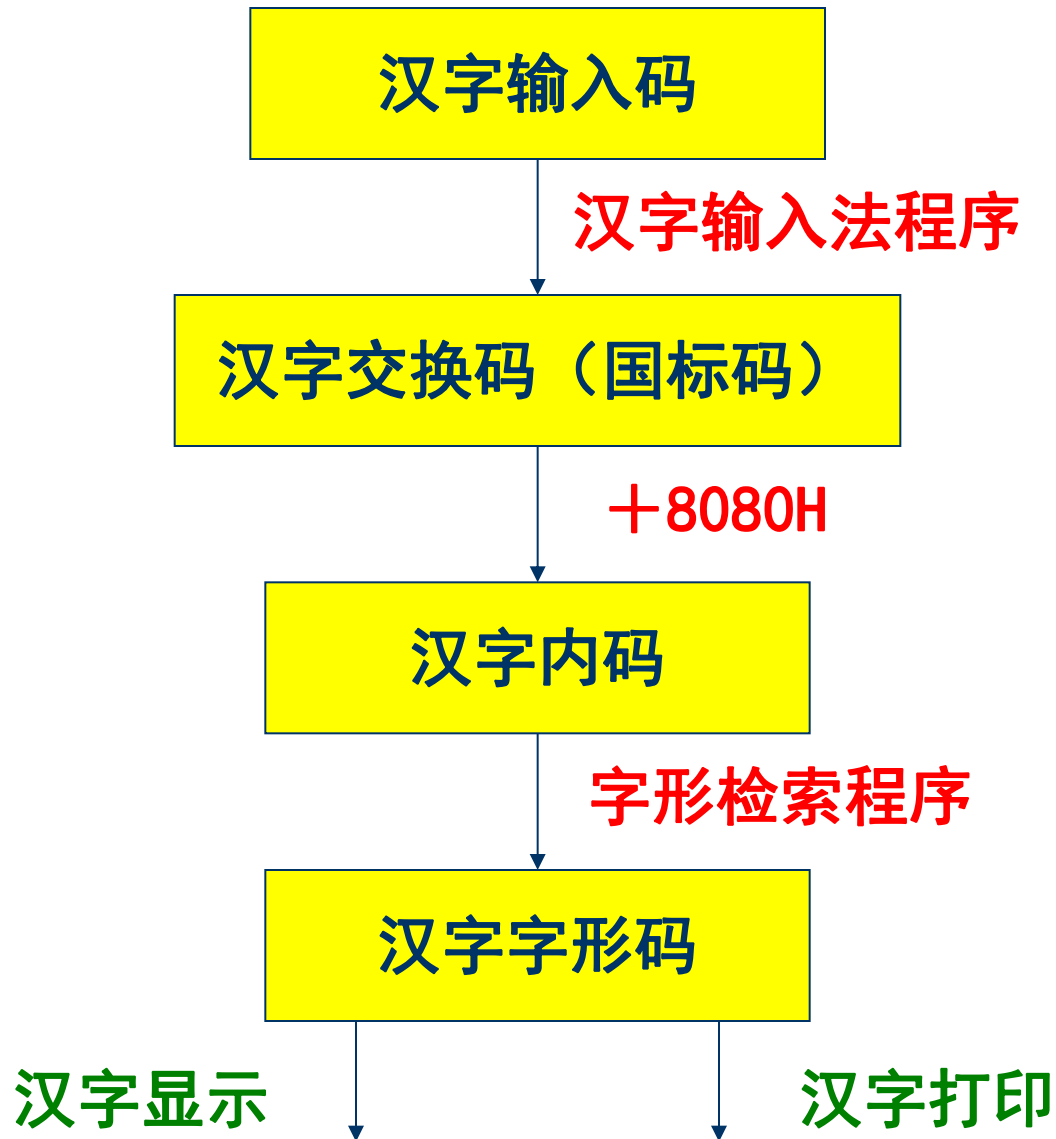


字模码：16×16点阵，需要16×16b=32B/汉字





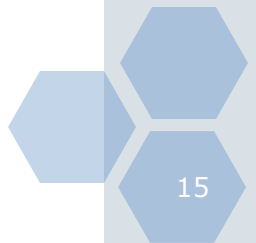
汉字编码之间的关系





Unicode

- ❖ Unicode是国际组织制定的可以**容纳世界上所有文字和符号**的字符编码方案，又称统一码、万国码、单一码，是一种在可在计算机上使用的字符编码。
- ❖ 它为每种语言中的每个字符设定了统一并且唯一的二进制编码，以满足**跨语言、跨平台**进行文本转换、处理的要求。
- ❖ 目前普遍采用的是**UCS-2，即Unicode 16**，它用**2个字节来编码一个字符**；包含了GB18030里面的所有汉字（27484个字）。“中”的Unicode 16编码是4E2DH。
- ❖ UCS-4（Unicode 32bit）：预备纳入康熙字典的所有汉字。





The End!

