02

OPEN ORIENTED

凹 凸 实 验 室

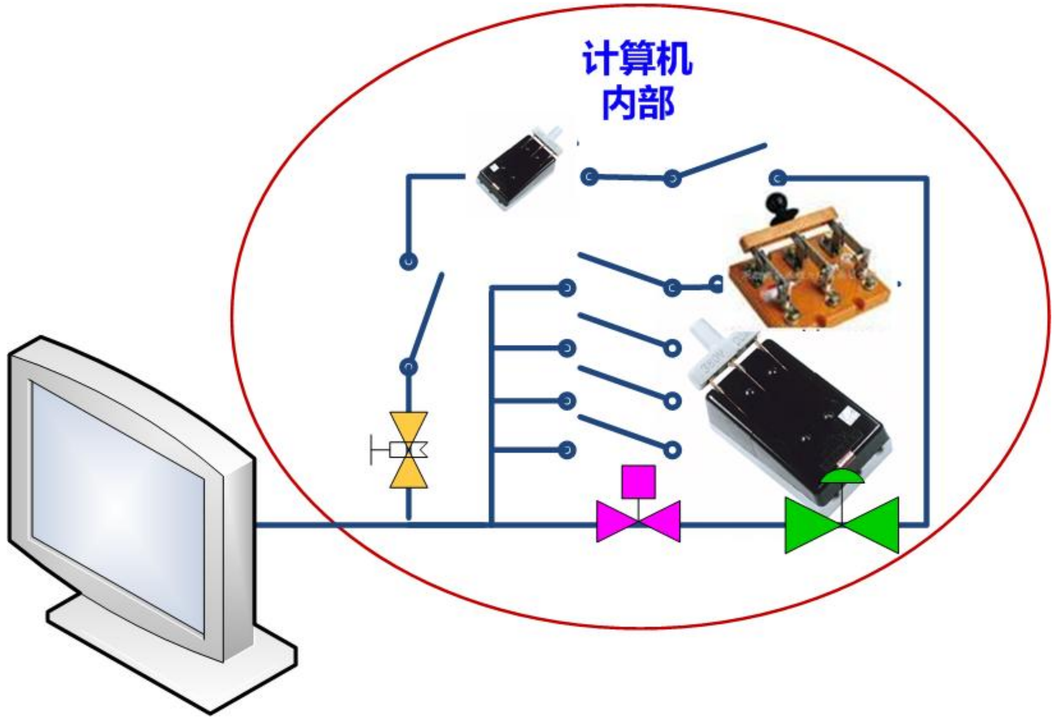
漫谈字符编码

有山

1. 字符编码的由来
2. 各种字符编码
3. 大一统的创想Unicode
4. 字符编码和存储编码

字符编码的由来

字符编码的由来



计算机底层硬件通过电路开和闭形成两种状态
分别用0和1的二进制表示，用于数学计算

字符编码的由来

计算机是用来解决数学计算问题的

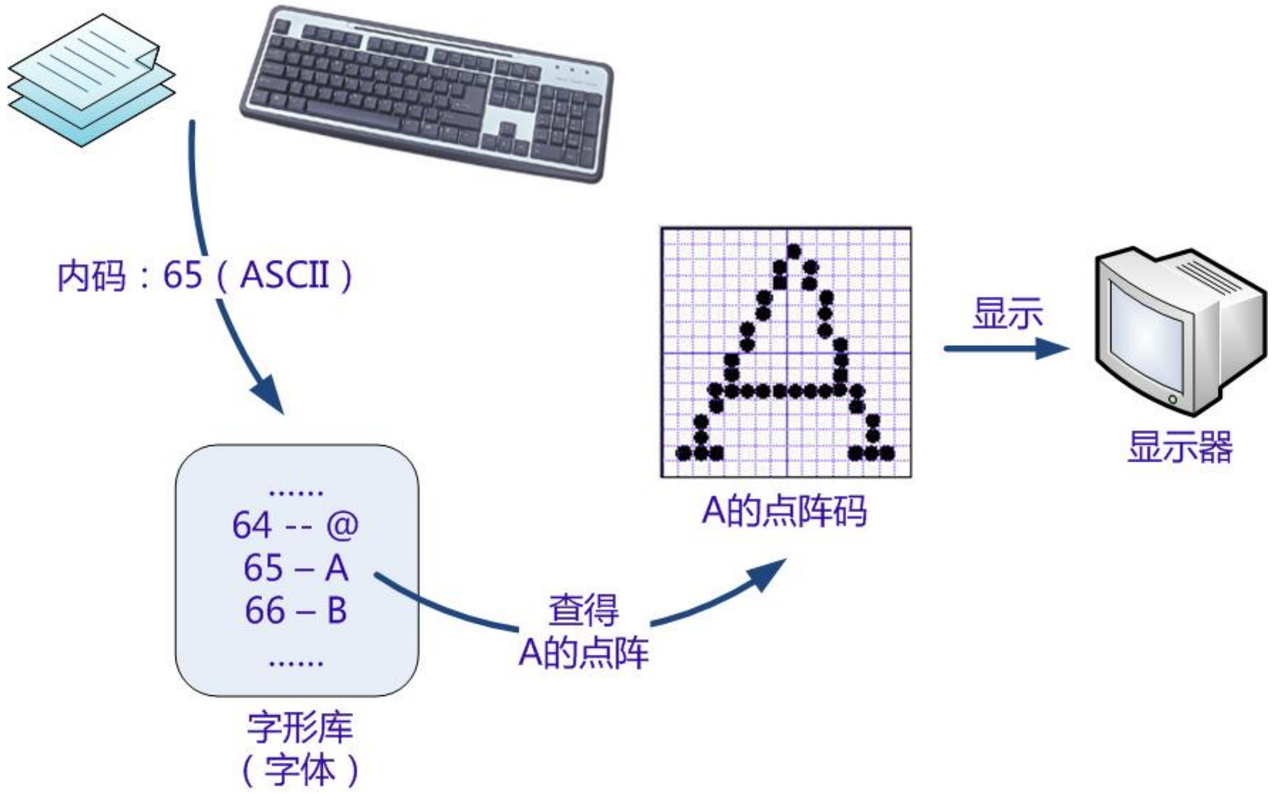
人们希望计算机做更多的事情，比如文本处理

但是计算机很单纯，它只认识01011100...这样的二进制数字

为了在计算机上能表示、存储和处理文本、符号等字符
就必须将字符转成二进制

那么计算机应该如何转换呢?

计算机处理文本的过程



各种字符编码

EBCDIC

EBCDIC CP037																
	x0	x1	x2	x3	x4	x5	x6	x7	x8	x9	xA	xB	xC	xD	xE	xF
0x	NUL	SOH	STX	ETX	ST	HT	SSA	DEL	EPA	RI	SS2	VT	FF	CR	SO	SI
1x	DLE	DC1	DC2	DC3	OSC	NEL	BS	ESA	CAN	EM	PU2	SS3	FS	GS	RS	US
2x	PAD	HOP	BPH	NBH	IND	LF	ETB	ESC	HTS	HTJ	VTS	PLD	PLU	ENQ	ACK	BEL
3x	DCS	PU1	SYN	STS	CCH	MW	SPA	EOT	SOS	SGCI	SCI	CSI	DC4	NAK	PM	SUB
4x	SP	NBSP	â	ä	à	á	ã	å	ç	ñ	¢	.	<	(	+	|
5x	&	é	ê	ë	è	í	î	ï	ì	ß	!	$	*	)	;	¬
6x	-	/	Â	Ä	À	Á	Ã	Å	Ç	Ñ	¦	,	%	_	>	?
7x	ø	É	Ê	Ë	È	Í	Î	Ï	Ì	`	:	#	@	'	=	"
8x	Ø	a	b	c	d	e	f	g	h	i	«	»	ð	ý	þ	±
9x	°	j	k	l	m	n	o	p	q	r	ª	º	æ	¸	Æ	¤
Ax	µ	~	s	t	u	v	w	x	y	z	¡	¿	Ð	Ý	Þ	®
Bx	^	£	¥	·	©	§	¶	¼	½	¾	[	]	¯	¨	´	×
Cx	'{'	A	B	C	D	E	F	G	H	I	SHY	ô	ö	ò	ó	õ
Dx	'}'	J	K	L	M	N	O	P	Q	R	¹	û	ü	ù	ú	ÿ
Ex	\	÷	S	T	U	V	W	X	Y	Z	²	Ô	Ö	Ò	Ó	Õ
Fx	0	1	2	3	4	5	6	7	8	9	³	Û	Ü	Ù	Ú	APC

EBCDIC

Ebcdic(Extended Binary Coded Decimal Interchange Code)
扩展二进制编码的十进制交换码

于1964年IBM开发设计的

EBCDIC码中英文字母不是连续的，在处理上带来麻烦

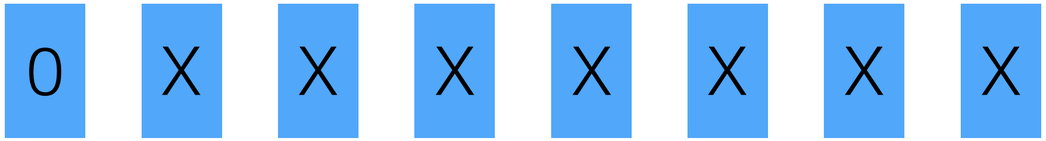
因此，在后来IBM的个人计算机和工作站操作系统中
并没有采用EBCDIC码，而是采用了晚于EBCDIC码推出，
且后来成为了英文字符编码工业标准的ASCII编码方案。

ASCII

ASCII码(American Standard Code for Information Interchange)
美国信息交换标准码

由美国国家标准学会ANSI(American National Standard Institute)
于1968年正式制定

ASCII



通常最高位为0

共有2^7=128个字符

ASCII

0～31：控制字符或通讯专用字符

32～126：可显示可打印字符

48～57为0-9的阿拉伯数字

65～90为26个大写英文字母

97～122为26个小写英文字母

其余的是一些标点符号、运算符号等。

127：控制字符DELETE(删除符号)

ASCII码

USASCII code chart

b_7 b_6 b_5 →	0 0 0	0 0 1	0 1 0	0 1 1	1 0 0	1 0 1	1 1 0	1 1 1
Bits ↓ / Column → / Row ↓	0	1	2	3	4	5	6	7
0 0 0 0	0	NUL	DLE	SP	0	@	P	`	p
0 0 0 1	1	SOH	DC1	!	1	A	Q	a	q
0 0 1 0	2	STX	DC2	"	2	B	R	b	r
0 0 1 1	3	ETX	DC3	#	3	C	S	c	s
0 1 0 0	4	EOT	DC4	$	4	D	T	d	t
0 1 0 1	5	ENQ	NAK	%	5	E	U	e	u
0 1 1 0	6	ACK	SYN	&	6	F	V	f	v
0 1 1 1	7	BEL	ETB	'	7	G	W	g	w
1 0 0 0	8	BS	CAN	(	8	H	X	h	x
1 0 0 1	9	HT	EM	)	9	I	Y	i	y
1 0 1 0	10	LF	SUB	*	:	J	Z	j	z
1 0 1 1	11	VT	ESC	+	;	K	[	k	{
1 1 0 0	12	FF	FS	,	<	L	\	l	|
1 1 0 1	13	CR	GS	-	=	M	]	m	}
1 1 1 0	14	SO	RS	.	>	N	^	n	~
1 1 1 1	15	SI	US	/	?	O	_	o	DEL

GB2312

天朝专家规定：一个小于127的字符的意义与原来相同，但两个大于127的字符连在一起时，就表示一个汉字，前面的一个字节（他称之为高字节）从0xA1用到 0xF7，后面一个字节（低字节）从0xA1到0xFE，这样我们就可以组合出大约7000多个简体汉字了。在这些编码里，还把数学符号、罗马希腊的 字母、日文的假名们都编进去了，连在ASCII里本来就有的数字、标点、字母都统统重新编了两个字节长的编码

GBK

GBK编码能够用来同时表示繁体字和简体字，而GB2312只能表示简体字，GBK是兼容GB2312编码的。GBK工作小组于1995年10月，同年12月完成GBK规范。该编码标准兼容GB2312，共收录汉字 21003个、符号883个，并提供1894个造字码位，简、繁体字融于一库。

BIG5

是目前台湾、香港地区普遍使用的一种繁体汉字的编码标准，包括440个符号，一级汉字5401个、二级汉字7652个，共计13060个汉字。BIG5又 称大五码或五大码，1984年由台湾财团法人信息工业策进会和五间软件公司宏碁(Acer)、神通(MiTAC)、佳佳、零壹(Zero One)、大众(FIC)创立，故称大五码。Big5码的产生，是因为当时台湾不同厂商各自推出不同的编码，如倚天码、IBM PS55、王安码等，彼此不能兼容；另一方面，台湾政府当时尚未推出官方的汉字编码，而中国大陆的GB2312编码亦未有收录繁体中文字。

混乱的字符编码


0
127
256
65535
ASMO
阿拉伯文
ISO8859-1
拉丁文
GB2312
GBK
简体中文
ASCII
Big5
繁体中文
Shift-JIS
日文
其它编码 …… ……
7位
8位
16位

大一统的Unicode

Unicode码

Unicode字符集（简称为UCS）,国际标准组织于1984年4月成立ISO/IEC JTC1/SC2/WG2工作组，针对各国文字、符号进行统一性编码。1991年美国跨国公司成立Unicode Consortium，并于1991年10月与WG2达成协议，采用同一编码字集。

Unicode码

Unicode一开始是采用16位编码体系，内容包含符号6811个，汉字20902个，韩文拼音11172个，造字区6400个，保留20249个，共计65534个。Unicode编码后的大小是一样的.

例如一个英文字母 "a" 和　一个汉字 "好"，编码后都是占用的空间大小是一样的，都是两个字节

Unicode码

随着中文，日文和韩文引入，原有的Unicode定义的字符集已经不能满足。Unicode定义的字符集已经超过16位所能表达的范围，把所有这些CodePoint分成17个代码平面（Code Plane）：

U+0000 ～ U+FFFF划入基本多语言平面（Basic MultilingualPlane，简记为BMP）

其余划入16个辅助平面（Supplementary Plane），代码点范围U+10000 ～ U+10FFFF。

字符编码和存储编码

字符编码和存储编码

字符编码：是抽象的概念，是一个纯粹的字符集合。
集合中每个字符有一个ID号用于识别字符。

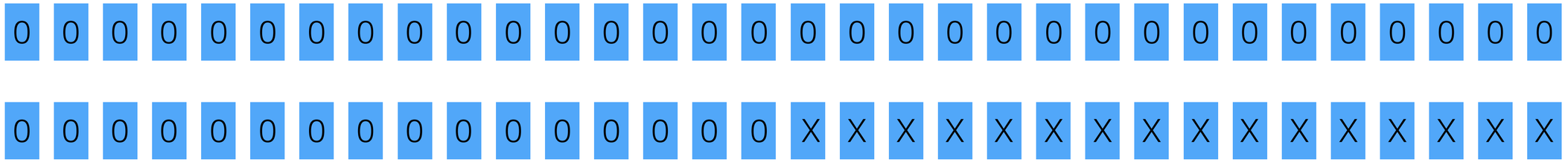
存储编码：解决字符的实际存储问题
即字符在存储设备（内存、硬盘）上如何存储

Unicode存储方案-UTF

UTF：Unicode Transformation Format

目前流行的存储方案有：UTF-32 UTF-16 UTF-8

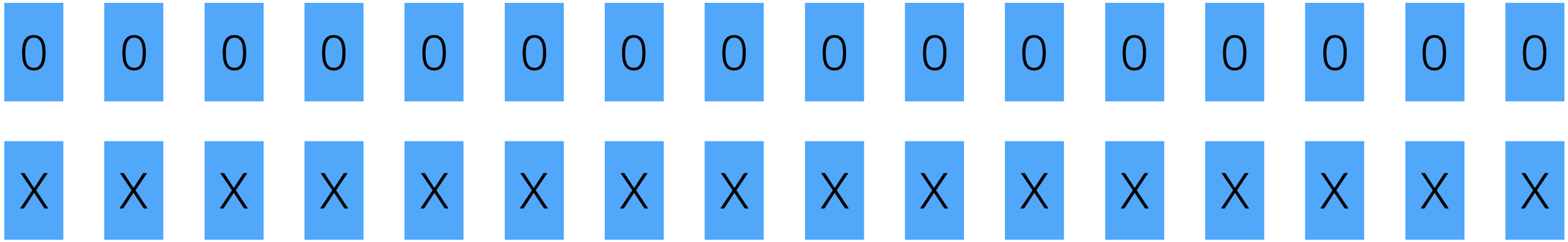
UTF-32



UTF-32：用32位存储一个Unicode字符

存储每个字符需要占用4字节，造成存储空间的浪费

UTF-16



UTF-16：以16位为单位，用1~2个16位来存储一个字符

UTF-8

Code point	UTF-8字节流
U+00000*00* – U+0000007*F*	**0**xxxxxxx
U+00000080 – U+000007FF	**110**xxxxx ***10***xxxxxx
U+00000800 – U+0000FFFF	**1110**xxxx ***10***xxxxxx ***10***xxxxxx
U+00010000 – U+001FFFFF	**11110**xxx ***10***xxxxxx ***10***xxxxxx ***10***xxxxxx

UTF-8 （8-bit Unicode Transformation Format）是一种针对Unicode的可变长度字符编码，使用1~4个字节为每个字符编码

THANKS

FOR YOUR WATCHING

02

OPEN ORIENTED

凹 凸 实 验 室