



中山大學 软件工程学院
SUN YAT-SEN UNIVERSITY SCHOOL OF SOFTWARE ENGINEERING

计算机组成原理

授课老师：吴炜滨

➤ 非数值数据的表示

- 字符表示
- 汉字编码

➤ 数据信息的校验

- 码距与校验
- 奇偶校验
- 海明校验



➤ 非数值数据的表示

- 字符表示

■ 非数值数据

- 没有数值大小之分，如字符和汉字等

■ 字符表示

- 如何使用二进制代码进行字符编码？
 - ASCII-American Standard Code for Information Interchange（美国信息交换标准码）

■ ASCII码

- 使用7bit表示128个字符 (2^7)：从000 0000 到 111 1111
- 计算机中数据存储以字节为单位，故字节最高位（Most Significant Bit, MSB）为0
- 单字节编码

ASCII码



b ₄ b ₃ b ₂ b ₁	b ₇ b ₆ b ₅							
	000	001	010	011	100	101	110	111
0000	NUL	DLE	(Space)	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

- 英文字母
- 十进制数码
- 专用符号
- 控制字符



➤ 非数值数据的表示

- 汉字编码

■ 对汉字信息进行处理

- 汉字输入 → 汉字输入码
- 汉字交换 → 汉字交换码
- 汉字存储、处理 → 汉字机内码
- 汉字输出 → 汉字字形码

■ 汉字输入码

- 使用英文键盘**输入汉字时**所使用的编码
 - 输入法：将从键盘输入的汉字输入码转化为汉字的机器内码
- 流水码：用数字组成的等长编码，如**国标码**、区位码
- 音码：根据汉字读音组成的编码，如**拼音码**，常见的有全拼、简拼等
- 形码：根据汉字的形状、结构特征组成的编码，如**五笔字型码**

金金钅勺 夕夕夕力人备 鸟鸟鸟儿儿弟 35 Q	人八 亻亻 几几 34 W	月月心 用用力力象毛 衣衣衣白白 33 E	白白厂 斤丘气 义扌手 32 R	禾禾 欠欠 彳彳 31 T	言言讠 文言言 方藏圭 41 Y	立立 辛羊羊 六门門丹丹 42 U	水水 水水水水水 浦小小小 43 I	火火 业业业 广广广 44 O	之之 之之 之之 45 P
工工 戈戈戈戈戈 廿廿廿廿廿 15 A	木木 西西甫 14 S	犬犬 其其 石厂厂 13 D	土土 十十 甘雨雨 12 F	王王 土土 夫夫牛牛 11 G	目目 上上 止止少少 21 H	日日 リリリ 早早 22 J	口口 川川 川川 23 K	田田 甲甲車車 四四四四 24 L	
幺幺 五五 弓弓 55 X	又又 巴巴牛馬 34 C	女女 刀刀 艮艮 53 V	子子 耳耳 也也乃口 52 B	己己 吕吕 心心小小 51 N	山山 由由 贝贝 25 M				



■ 汉字交换码

- 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**所用的代码标准：国标码

■ 区位码

- 为检索方便，采用 $94 \times 94 = 8836$ 的二维矩阵对汉字、特殊符号、数字、英文字符、制表符等进行编码
- 每一行称为“区”，每一列称为“位”，编号从1开始
- 区号和位号的组合（4位10进制）构成该字的区位码
- 中（区位码）：5448
- 双字节编码
- 同一英文字母、数字和符号
 - 在区位码中以两个字节表示，称之为全角字符
 - 在ASCII码中以一个字节表示，称之为半角字符，在屏幕上的显示宽度为全角字符的一半
 - A（区位码）：0333；A（ASCII码）：01000001

■ 国标码

- 为了复用ASCII码中的控制码（0~31）、空格字符（32）且不发生冲突
- 将区位码的区号和位号分别加上32（20H）得到国标码：国标码=区位码+2020H
- 中（国标码）：5650H=3630H（区位码）+2020H

■ 汉字机内码

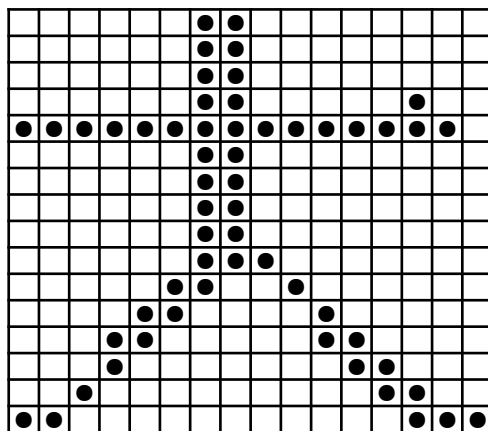
- 计算机内部存储、处理汉字时所用的统一编码
- 国标码前后字节的最高位为0，与ASCII码发生冲突
 - 中：5650H误认为V（56H）和P（50H）
- ASCII码MSB为0，汉字机内码MSB为1
- 汉字机内码=汉字国标码+8080H
- 中（机内码）：D6D0H= 5650H（国标码） +8080H

■ 汉字字形码（汉字字模）

- 表示**汉字字形信息**（结构、形状、笔画等）的编码，以实现计算机对汉字的输出（显示、打印）
- 最常用的表示方式：点阵形式和矢量形式

■ 点阵形式

- 将字符的**字形分解成若干“点”**组成的点阵，有字形笔画的点用黑色，反之用白色
- 在存储时，用1表示黑色点，0表示没有笔画的白色点，顺序存储，即构成汉字字形码



汉字编码



■ 点阵形式

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	十六进制码			
0							●	●									0	3	0	0
1							●	●									0	3	0	0
2							●	●									0	3	0	0
3							●	●						●			0	3	0	4
4	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●		F	F	F	E
5							●	●									0	3	0	0
6							●	●									0	3	0	0
7							●	●									0	3	0	0
8							●	●									0	3	0	0
9							●	●	●								0	3	8	0
10						●	●			●							0	6	4	0
11					●	●					●						0	C	2	0
12				●	●						●	●					1	8	3	0
13				●								●	●				1	0	1	8
14			●										●	●			2	0	0	C
15	●	●												●	●	●	C	0	0	7

■ 点阵形式

- 占用存储空间大，以 32×32 为例：
 - 每个汉字要占用128个字节（1字节=8位）
- 只用来构成汉字库，不用于机内存储，需要时才到字库中检索汉字并输出
- 不同字体（如黑体、微软雅黑等）对应不同的汉字库

黑体

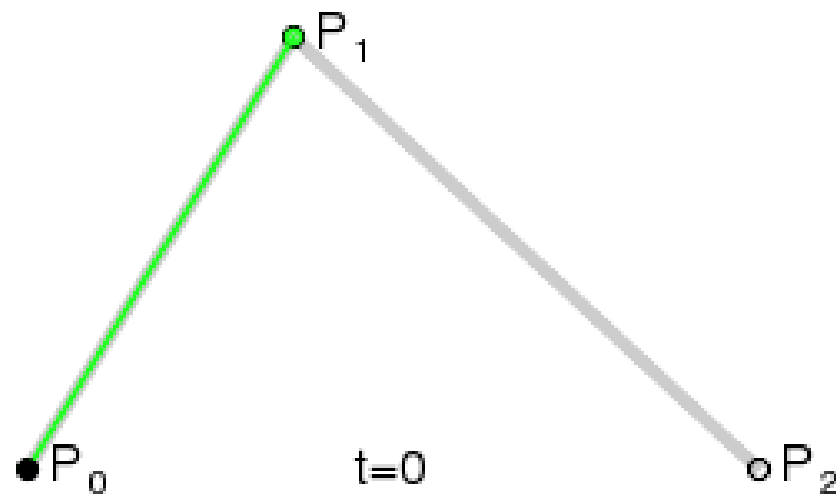
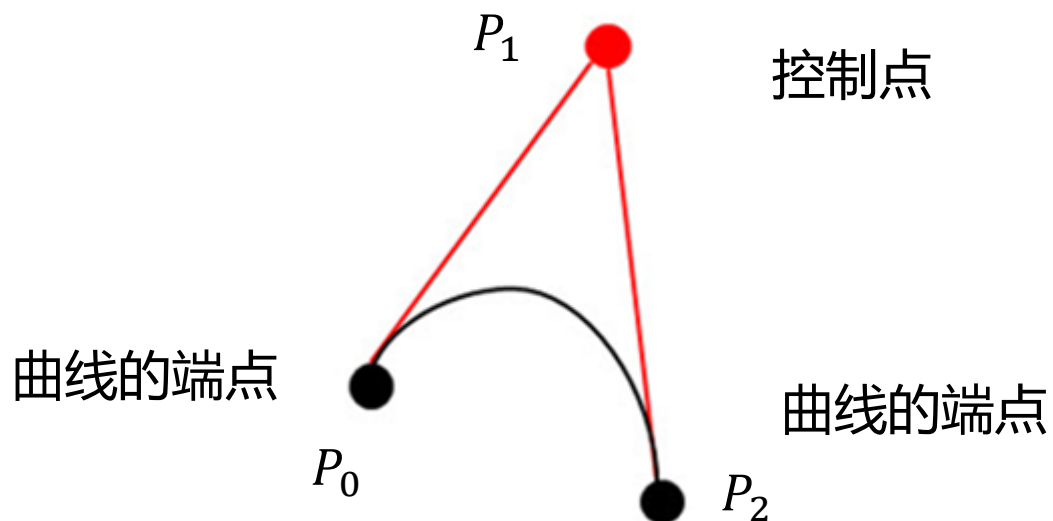


微软雅黑



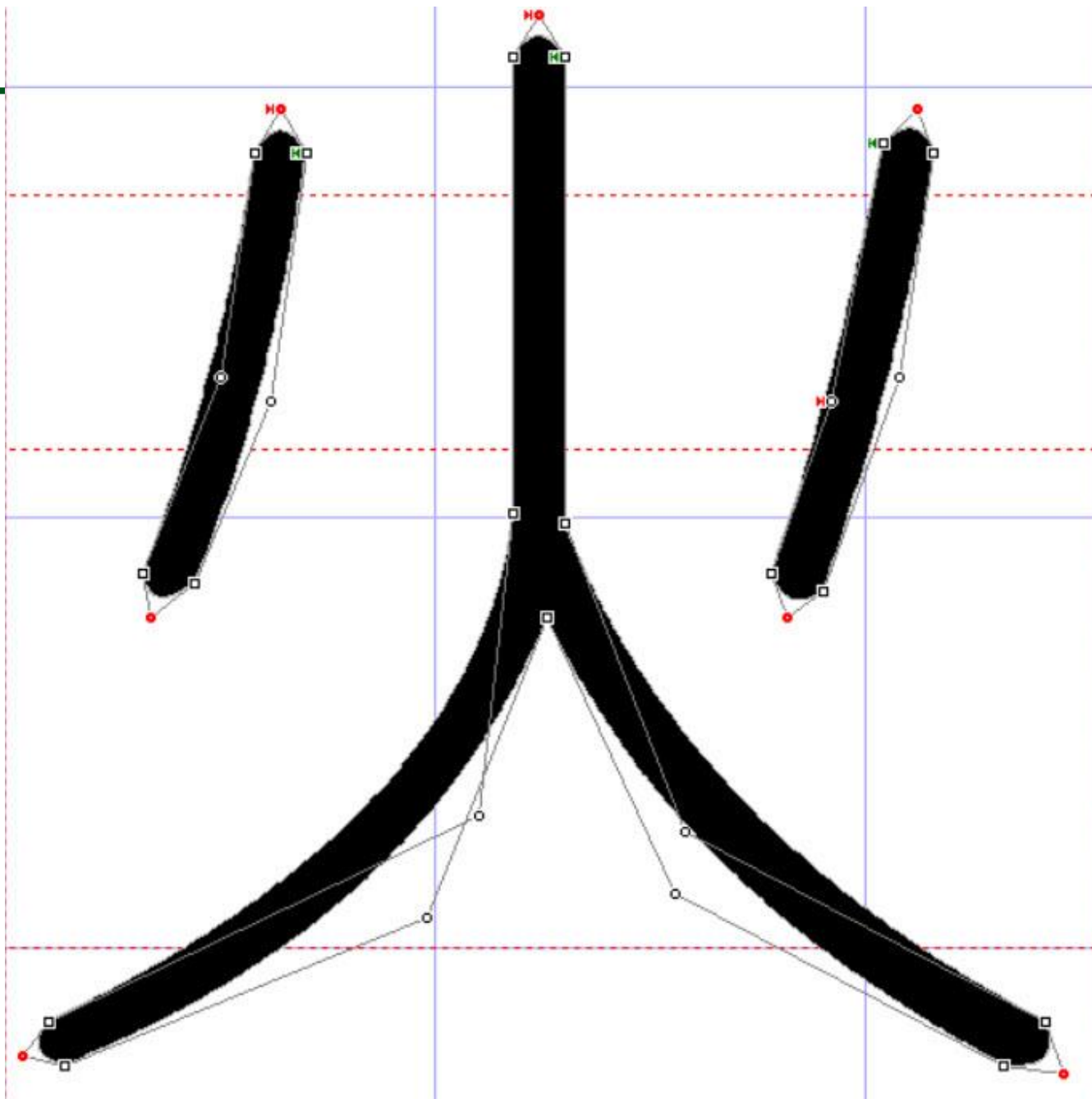
■ 矢量形式

- 通过直线、曲线来描述每一个汉字字形
 - 存储这些线的关键点
 - 利用这些点来绘制曲线或直线，描绘出字体的轮廓，最后进行黑色填充
- 二次贝塞尔曲线



汉字编码

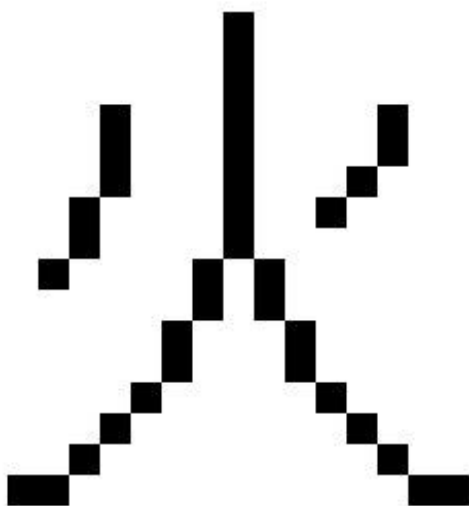
- 矢量形式
- 火



■ 矢量形式

- 点阵形式输出的字体放大后会变形
- 矢量形式描述的字形与最终文字显示的大小、分辨率无关，放大或缩小字体的时候，只需要按比例缩放改变端点值的相对位置，故可以输出高质量的字形

点阵字体



矢量字体





➤ 数据信息的校验

- 码距与校验

■ 数据校验的必要性

- 受元器件的质量、电路故障或噪音干扰等因素的影响，**数据**在被处理、传输、存储的过程中**可能出现错误**

■ 校验码

- 具有发现错误或纠正错误能力的数据编码
- 在**被校验数据（原始数据）**中引入部分**冗余信息（校验数据）**，使最终的**校验码（原始数据+校验数据）**符合某种编码规则
- 当校验码中某些位发生错误时，会破坏预定**规则**，使错误可以被检测，甚至被纠正





■ 码距

- 两个编码对应二进制位不同的个数
- 10101 vs. 00110
 - 码距为3

■ 最小码距

- 同一编码系统中，任意两个合法编码之间**码距的最小值**

想象码字是地图上的城市，码距是俩城市间的最短路（不同位=弯路）。最小码距小=城市挤，出门容易走错；大=分散，导航稳。

■ 现有两种编码体系，分别分析它们各自的最小码距

- 合法编码集合为：{000, 001, 010, 011, 100, 101, 110, 111}
 - 解：最小码距为1。任何一个合法编码发生一位错误时，就会变成另外一个合法编码，不具备检测错误的能力
- 合法编码集合为：{000, 011, 101, 110}
 - 解：最小码距为2。任何一位发生改变，如000变成100，就从有效编码变成了无效编码，可以检测一位错误
 - 但发生两位错误时，可能会变成另一个合法编码，如000变成011，无法识别错误

■ 增大码距能把一个不能检错的编码变成能检错的编码

■ 校验码的工作原理

- 编码中引入一定冗余，增加最小码距，使编码符合某种规则，当编码出现一个或多个错误时变成非法代码（不符合规则）

■ 码距越大，抗干扰能力越强，检错、纠错能力越强

- 数据冗余越大，编码效率低，编码电路也相对复杂
- 选择码距应综合考虑信息出错概率和系统容错率等因素



➤ 数据信息的校验

- 奇偶校验

■ 奇偶校验的基本原理

- 增加冗余码 (1位 校验位 P)



■ 奇校验

- 编码规则：校验码 (原始数据 + 校验位) 中1的个数为奇数

■ 偶校验

- 编码规则：校验码 (原始数据 + 校验位) 中1的个数为偶数

奇偶校验



■ 奇偶校验

- 最小码距为2

原始数据 (7位)	奇校验码 (8位)	偶校验码 (8位)
000 0000	0000 000 1	0000 000 0
111 1111	1111 111 0	1111 111 1

■ 如何使用逻辑电路自动生成奇偶校验位

- 设原始数据 $D = D_1D_2D_3 \cdots D_n$, 校验位为 P
- 奇偶校验编码电路的逻辑表达式

$$\text{偶校验: } P = D_1 \oplus D_2 \oplus D_3 \cdots \oplus D_n$$

$$\text{奇校验: } P = \overline{D_1 \oplus D_2 \oplus D_3 \cdots \oplus D_n}$$

■ 奇偶校验过程

- 发送方生成 校验码 ($D_1 D_2 \cdots D_n P$) 并发送
- 接收方收到发送方传输的校验码 $D_1' D_2' \cdots D_n' P'$ 后, 生成如下检错码 G

$$\text{偶校验: } G = P' \oplus D_1' \oplus D_2' \cdots \oplus D_n'$$

$$\text{奇校验: } G = \overline{P' \oplus D_1' \oplus D_2' \cdots \oplus D_n'}$$

- $G=1$
 - 数据**一定**出错
- $G=0$
 - **较大概率**正常

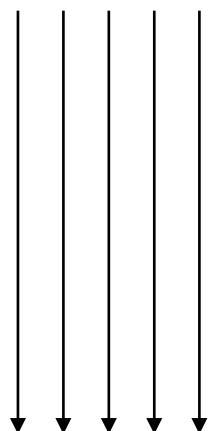
奇偶校验



■ 奇偶校验性能

- 假设数据采用偶校验

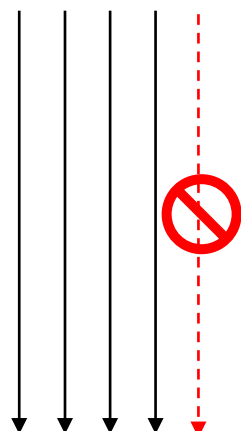
00011



00011

正确传输

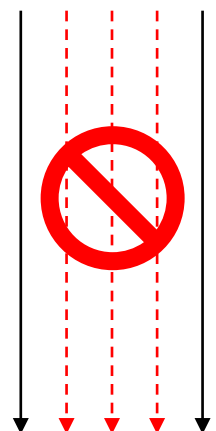
00011



00010

正常检错

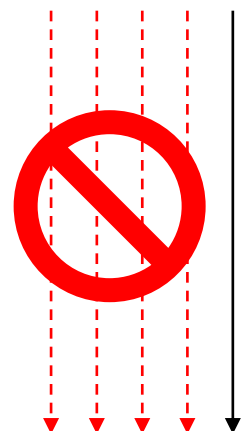
00011



01101

正常检错

00011



11101

不能检错

只能**检测奇数错**，不能纠错，不保证正确，实现简单，编码效率高



➤ 数据信息的校验

- 海明校验

海明校验



■ 海明码具有一位纠错能力

■ 海明码采用奇偶校验

- 偶校验

- 00100011 → 001000111



校验位

无法定位出错位置

■ 海明码采用分组校验

- 分组校验: 00100011 → 0010100110 (偶校验)



校验位

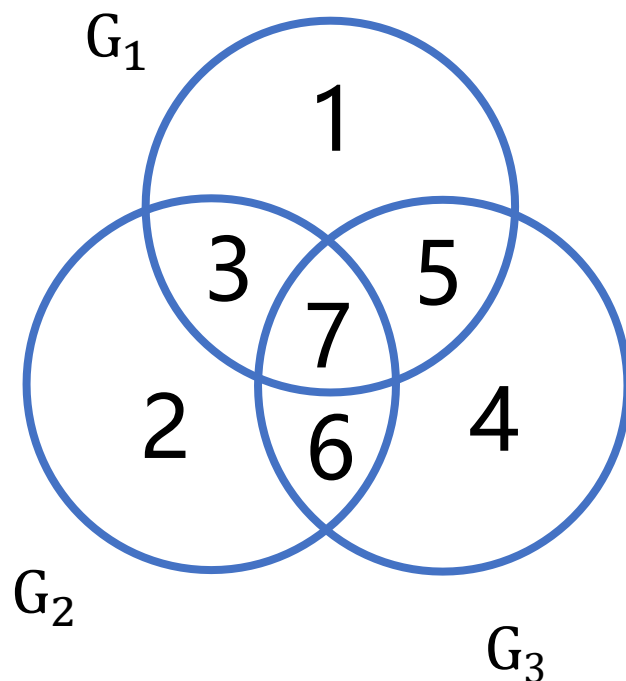
■ 海明码的分组是一种非划分方式

- 数据分组之间是有交叉的: 有些位是属于多个组

1	2	3	4	5	6	7
---	---	---	---	---	---	---

■ 海明码如何分组？

- 分成3组，每组有1位校验位，共包括4位数据位



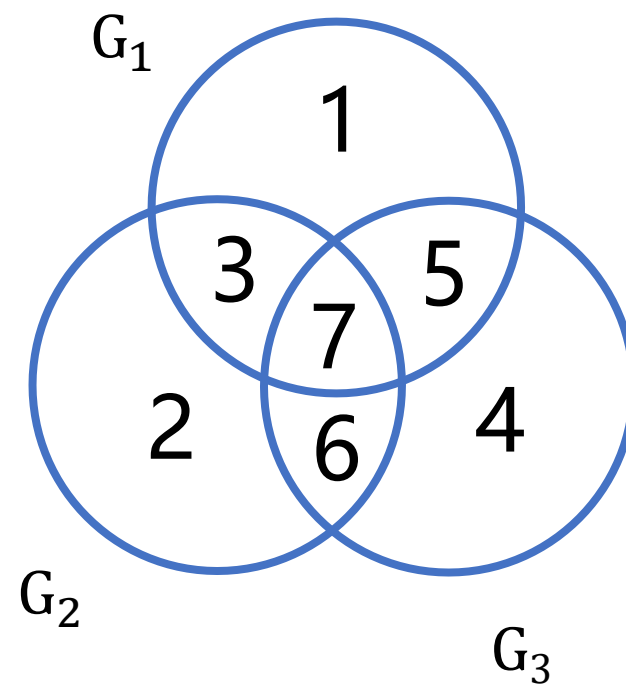
■ 假定最多只有一位错，检错时生成 $G_3G_2G_1$ （检错码）

$G_3G_2G_1$	出错位
0 0 0	无差错
0 0 1	1
1 0 1	5
1 1 0	6
1 1 1	7

好处：检错码指出了出错的位置

■ 海明码如何编码？

- 应该增添多少位校验位？
- 校验位应该放在哪些位置？
- 数据位应该如何分组？
- 校验位的取值？



■ 海明码编码规则

- 原始数据信息被分成若干个校验组，每组设置一个奇/偶校验位
- 每个校验组在检错时会得到一个检错位
- 所有校验组的检错位的值组成检错码
 - 检错码值为0时，大概率无错误
 - 检错码值不为0时，指出一位出错的位置

■ 海明码需增添多少位校验位？

- 设海明码 $H_1H_2 \cdots H_n$ 共 n 位，原始数据 $D_1D_2 \cdots D_k$ 共 k 位，校验位 $P_1P_2 \cdots P_r$ 共 r 位
 - $n = k + r$
 - 包含 r 个校验组，即 r 个检错位，来指出 $k + r + 1$ 种状态

$$2^r \geq r + k + 1$$



■ 海明码校验位的位置？

- 设海明码 $H_1H_2 \cdots H_n$ 共 n 位，原始数据 $D_1D_2 \cdots D_k$ 共 k 位，校验位 $P_1P_2 \cdots P_r$ 共 r 位
- 检错码 $G_r \cdots G_2G_1$ 共 r 位
- 校验位只对数据位进行校验 → 校验位只位于一个校验组
 - 出错时，其检错码只有一位为1
 - $001, 010, 100 \rightarrow 2^i \ (i = 0, 1, 2, \cdots)$

海明码	H_1	H_2	H_3	H_4	H_5	H_6	H_7
检错码/出错位置	001	010	011	100	101	110	111
映射关系	P_1	P_2	D_1	P_3	D_2	D_3	D_4
G_1 校验组	√						
G_2 校验组		√					
G_3 校验组				√			



- 海明码数据位如何分组？
 - 设海明码 $H_1H_2 \cdots H_n$ 共 n 位，原始数据 $D_1D_2 \cdots D_k$ 共 k 位，校验位 $P_1P_2 \cdots P_r$ 共 r 位
 - 检错码 $G_r \cdots G_2G_1$ 共 r 位
 - H_3 (D_1) 出错：
 - 检错码为011，应参与 G_1 ， G_2 校验组
 - H_5 (D_2) 出错：
 - 检错码为101，应参与 G_1 ， G_3 校验组
 - 以此类推

海明码	H_1	H_2	H_3	H_4	H_5	H_6	H_7
检错码/出错位置	001	010	011	100	101	110	111
映射关系	P_1	P_2	D_1	P_3	D_2	D_3	D_4
G_1 校验组	√		√		√		√
G_2 校验组		√	√			√	√
G_3 校验组				√	√	√	√



■ 海明码数据位如何分组？

- H_j 位的 **数据** 被编号小于 j 的若干个海明位号之和等于 j 的 **校验位** 所校验

海明码	H_1	H_2	H_3	H_4	H_5	H_6	H_7
检错码/出错位置	001	010	011	100	101	110	111
映射关系	P_1	P_2	D_1	P_3	D_2	D_3	D_4
G_1 校验组	√		√		√		√
G_2 校验组		√	√			√	√
G_3 校验组				√	√	√	√

■ 校验位的取值?

- 校验位的取值与该位所在的校验组中承担的奇/偶校验任务有关
- 常用偶校验

$$\text{偶校验: } P_1 = H_3 \oplus H_5 \oplus H_7 \cdots$$

$$\text{奇校验: } P_1 = \overline{H_3 \oplus H_5 \oplus H_7 \cdots}$$



■ 按配偶原则配置 0011 的海明码

解： $\because k = 4$

根据 $2^r \geq r + k + 1$

得 $r = 3$

海明码	H_1	H_2	H_3	H_4	H_5	H_6	H_7
检错码/出错位置	001	010	011	100	101	110	111
映射关系							
值							

练习



■ 按配偶原则配置 0011 的海明码

解：

海明码	H_1	H_2	H_3	H_4	H_5	H_6	H_7
检错码/出错位置	001	010	011	100	101	110	111
映射关系	P_1	P_2	D_1	P_3	D_2	D_3	D_4
值	1	0	0	0	0	1	1

$$P_1 = H_3 \oplus H_5 \oplus H_7 = 1$$

$$P_2 = H_3 \oplus H_6 \oplus H_7 = 0$$

$$P_3 = H_5 \oplus H_6 \oplus H_7 = 0$$

\therefore 0011 的海明码为 1000011

练习



■ 求 0101 按 “偶校验” 配置的海明码

解： $\because k = 4$

根据 $2^r \geq r + k + 1$

得 $r = 3$

海明码排序如下：

海明码	H_1	H_2	H_3	H_4	H_5	H_6	H_7
检错码/出错位置	001	010	011	100	101	110	111
映射关系	P_1	P_2	D_1	P_3	D_2	D_3	D_4
校验值	0	1	0	0	1	0	1

\therefore 0101 的海明码为 0100101

海明码的检错、纠错过程



■ 约定好编码规则

- 一位纠错海明码，奇/偶检验

■ 对每一组进行校验生成检错码

- 位数等于校验位位数，如校验位 $P_1P_2 \cdots P_r$ 共 r 位，检错码 $G_r \cdots G_2G_1$ 也共 r 位
- 发送方发送数据位为 $D_1D_2 \cdots D_k$ ，校验位为 $P_1P_2 \cdots P_r$
- 接收方收到数据位为 $D_1'D_2' \cdots D_k'$ ，校验位为 $P_1'P_2' \cdots P_r'$
- 以 $k + r = 7$ 为例

位置	001	010	011	100	101	110	111
海明码	H_1'	H_2'	H_3'	H_4'	H_5'	H_6'	H_7'
映射关系	P_1'	P_2'	D_1'	P_3'	D_2'	D_3'	D_4'

海明码的检错、纠错过程



位置	001	010	011	100	101	110	111
海明码	H_1'	H_2'	H_3'	H_4'	H_5'	H_6'	H_7'
映射关系	P_1'	P_2'	D_1'	P_3'	D_2'	D_3'	D_4'

■ 对每一组进行校验生成检错码

- 对于按“偶校验”配置的海明码, G_i 的取值为

$$G_1 = H_1' \oplus H_3' \oplus H_5' \oplus H_7'$$

$$G_2 = H_2' \oplus H_3' \oplus H_6' \oplus H_7'$$

$$G_3 = H_4' \oplus H_5' \oplus H_6' \oplus H_7'$$

- 检错码值为0时, 大概率无错误
- 检错码值不为0时, 指出一位出错的位置, 可进行纠错

举例



- 已知接收到的海明码为 0100111（按配偶原则配置），试问要求传送的信息是什么？

解：

位置	001	010	011	100	101	110	111
海明码	H_1'	H_2'	H_3'	H_4'	H_5'	H_6'	H_7'
映射关系	P_1'	P_2'	D_1'	P_3'	D_2'	D_3'	D_4'
值	0	1	0	0	1	1	1

$$G_1 = H_1' \oplus H_3' \oplus H_5' \oplus H_7' = 0 \quad \text{无错}$$

$$G_2 = H_2' \oplus H_3' \oplus H_6' \oplus H_7' = 1 \quad \text{有错}$$

$$G_3 = H_4' \oplus H_5' \oplus H_6' \oplus H_7' = 1 \quad \text{有错}$$

$$\therefore G_3 G_2 G_1 = 110$$

第 6 位出错，可纠正为 0100101，故要求传送的信息为 **0101**

练习



■ 写出按偶校验配置的海明码0101101 的纠错过程

解：

位置	001	010	011	100	101	110	111
海明码	H_1'	H_2'	H_3'	H_4'	H_5'	H_6'	H_7'
映射关系	P_1'	P_2'	D_1'	P_3'	D_2'	D_3'	D_4'
值	0	1	0	1	1	0	1

$$G_1 = H_1' \oplus H_3' \oplus H_5' \oplus H_7' = 0$$

$$G_2 = H_2' \oplus H_3' \oplus H_6' \oplus H_7' = 0$$

$$G_3 = H_4' \oplus H_5' \oplus H_6' \oplus H_7' = 1$$

$$\therefore G_3 G_2 G_1 = 100$$

第 4 位出错，为校验位，不参与运算，故一般情况下可以不纠正



■ 按配奇原则配置 0011 的海明码

解： $\because k = 4$

根据 $2^r \geq r + k + 1$

得 $r = 3$

海明码	H_1	H_2	H_3	H_4	H_5	H_6	H_7
检错码/出错位置	001	010	011	100	101	110	111
映射关系							
值							

练习



■ 按配奇原则配置 0011 的海明码

解：

海明码	H_1	H_2	H_3	H_4	H_5	H_6	H_7
检错码/出错位置	001	010	011	100	101	110	111
映射关系	P_1	P_2	D_1	P_3	D_2	D_3	D_4
值	0	1	0	1	0	1	1

$$P_1 = \overline{H_3 \oplus H_5 \oplus H_7} = 0$$

$$P_2 = \overline{H_3 \oplus H_6 \oplus H_7} = 1$$

$$P_3 = \overline{H_5 \oplus H_6 \oplus H_7} = 1$$

纠错过程同样采用奇校验产生检错码

∴ 0011 的海明码为 0101011



谢谢！