国家经济信息系统设计与应用

标准化规范

（六）

中文信息处理

国家信息中心　　编著

航空工业出版社

国家经济信息系统设计与应用

标 准 化 规 范

（六）

中文信息处理

国家信息中心 编著

航空工业出版社

1993

（京）新登字161号

国家经济信息系统设计与应用

标 准 化 规 范

（六）

中 文 信 息 处 理

国家信息中心 编著

航空工业出版社出版发行
（北京市安定门外小关东里14号）
邮政编码：100029
全国各地新华书店经售
北京市通县向阳印刷厂印刷

1993年6月第1版 1993年6月第1次印刷
开本：787×1092 1/16 印张：8.5
印数：1—3000 字数：209千字
ISBN 7-80046-541-1/Z·092
定价：8.00元

序

早在1986年初国家经济信息系统建设初期，国家信息中心就注意到及早制定出符合我国信息系统实际情况的一整套标准与规范的重要意义。十分重视系统标准规范的制定工作。为此国家信息中心和原国家标准局共同组织编制了《国家经济信息系统设计与应用标准化规范》（下称红皮书）并于1986年7月由国家计委、原国家标准局联合批准，在系统内试行，成为我国经济信息系统建设的一个重要的指导性文件。

3年来，红皮书在指导我国各级信息系统建设中发挥了十分重要的作用，成为各级信息中心及其他信息机构从事系统设计与应用开发工作必不可少的依据。现已发行近1万册，需求还在增加。

与此同时，我们广泛征求了系统内外各级用户的意见，得到了大家的热情支持，收集到各使用单位的许多宝贵意见，感到书中确有一些不足之处；同时这几年系统建设和标准化理论发展很快，红皮书的内容也需要进一步修改完善。为此，国家信息中心在1988年初决定对红皮书进行重新修订。经过了近两年的努力，现在终于定稿。

这次公开发行的规范文本，是国家信息中心根据几年来信息系统建设实践中总结的经验，组织了国内著名的专家学者在该书原有基础上进行修改、充实、细化而成。修订后的规范文本覆盖了信息系统建设可行性研究、系统设计、应用开发、使用维护的全过程。具有技术先进、实用性强等特点，是各级信息系统建设的必备文件。

新的规范文本由八个分册组成。它们是：《概述》、《信息分类编码·通用文件格式》、《信息记录格式》、《数据库开发工程实施细则》、《软件工程规范》、《数据通信、计算机网络及系统互连》、《中文信息处理》、《信息安全与保密·系统建设必备文件》。

国家经济信息系统的建设是一项庞大而复杂的系统工程，标准化工作在这一工程的建设中占有十分重要的地位。这次红皮书的修订是对这一面向未来的课题的又一次积极尝试，同时也是系统标准化建设的一个新的开端。目前国内信息系统的标准化工作仍然处在初创阶段，要做的事还很多。但只要我们积极进取，勇于开拓，就能够引导这一工作向预期的目标发展。

编　者

1990年3月

出 版 说 明

《国家经济信息系统设计与应用标准化规范》经国家计划委员会和原国家标准批准作为指导国家信息系统建设与开发的纲领性文件，并于1987年7月正式颁布试行。3年来本规范在指导全国范围内的国家信息系统建设，保证各种信息系统在设计与应用开发中，按系统工程规律健康发展等方面发挥了巨大的作用。

此次公开发行的规范文本是国家信息中心根据几年来国家信息系统各级设计与开发部门对《规范》提出的意见和要求，组织国内著名专家、学者，经过一年多的努力在原有基础上进行修改、充实、细化而成的。修改后的《规范》覆盖了信息系统设计、开发、应用的全过程，具有技术上的先进性，使用上的实用性和权威性等突出特点，将成为各级政府经济信息系统建设的必备文件，对其他信息系统的建设开发、使用和管理也具有一定的指导意义。

全《规范》由8个分册组成。包括《概述》《信息分类编码·通用文件格式》、《信息记录格式》、《软件工程规范》、《数据通信、计算机网络及系统互连》、《中文信息处理》、《数据库开发实施细则》、《信息安全与保密·系统建设必备文件》将陆续出版发行。

本分册为第6分册。这一分册结合国家经济信息系统设计与应用开发的需要，提出中文信息处理需要使用的基础性技术标准和规范，对标准的主要内容作了扼要的说明，它们包括：七位和八位编码字符集、代码扩充技术、汉字交换码和内部码、少数民族文字编码、控制功能、输入和输出方式、汉字属性以及语词集等。这些内容中大部分已颁发有国家标准，但有一部分属于正在制定或修订的国家标准，还有一些属于将要制定的标准。本书的目的在于帮助国家经济信息系统设计者和广大用户了解这些标准的主要内容，帮助他们正确贯彻执行已有的标准，提前考虑将要制定的标准，有利于系统的互联。本分册由机械电子工业部十五研究所王之燿、国家技术监督局姚世全、程女范、刘碧松等同志编写。

目　录

1 编码字符集……（1）
1.1 信息处理交换用七位编码字符集及其代码扩充技术……（1）
1.1.1 信息处理交换用七位编码符集……（1）
1.1.2 八位编码字符集……（2）
1.1.2.1 GB11383规定的八位代码的结构和编码规则……（5）
1.1.2.2 EBDIC 码……（11）
1.1.3 七位和八位代码扩充技术……（11）
1.1.3.1 七位代码环境中的代码扩充技术……（12）
1.1.3.1.1 用移位功能进行图形字符集的扩充……（13）
1.1.3.1.2 用转义序列进行代码扩充……（14）
1.1.3.1.3 二字符转义序列……（15）
1.1.3.1.4 三字符转义序列……（15）
1.1.3.1.5 四字符及四字符以上的转义序列……（17）
1.1.3.2 八位代码环境中的代码扩充技术……（19）
1.1.3.3 代码扩充机制的宣布……（24）
1.1.3.4 转义序列汇总……（24）
1.1.4 电报用五单位电码及其与七位代码之间的转换……（25）
1.1.5 文本通信用编码字符集……（25）
1.1.5.1 文本通信用图形字符集……（25）
1.1.5.2 按页成象格式用控制功能……（39）
1.1.6 有关图形字符集……（41）
1.2 信息处理交换用汉字编码的字符集……（42）
1.2.1 信息交换用汉字编码字符集　基本集和第一辅助集……（42）
1.2.2 信息交换用汉字编码字符集　第二辅助集和第三辅助集……（43）
1.2.3 信息交换用汉字编码字符集　第四辅助集和第五辅助集……（43）
1.2.4 信息系统内部用的汉字代码……（43）
1.3 通用编码字符集……（45）
1.3.1 通用编码字符集的总结构……（45）
1.3.2 图形字符子集……（47）
1.3.3 GB13000……（47）
1.4 七位和八位编码字符集用的控制功能……（48）
1.4.1 控制功能组成……（48）
1.4.2 控制序列表示的控制功能……（49）
1.4.3 控制串……（50）
1.4.4 模式……（50）

1.4.5 控制功能的含义……（54）
1.5 少数民族文字编码……（80）
1.5.1 信息处理交换用蒙古文七位和八位编码图形字符集……（80）
1.5.2 信息处理交换用维吾尔文编码图形字符集……（81）
1.5.3 信息交换用朝鲜文字编码字符集……（82）
1.5.4 信息交换用彝文编码字符集……（83）
1.5.5 其他少数民族文字编码字符集……（84）
2 输入方式……（98）
2.1 键盘输入……（98）
2.1.1 汉字键盘输入方法(通用标准小键盘汉字输入方法)……（98）
2.1.2 汉字键盘输入方法评测……（99）
2.1.2.1 总则……（99）
2.1.2.2 术语……（99）
2.1.2.3 申请参测必备的基本条件……（100）
2.1.2.4 静态测试……（101）
2.1.2.5 定性评价……（101）
2.1.2.6 动态测试……（102）
2.1.2.6.1 试题文本与统选操作员……（102）
2.1.2.6.2 基本项目测试……（102）
2.1.2.6.3 专门项目测试……（103）
2.1.2.7 数据处理……（103）
2.1.2.8 综合评定……（103）
2.1.3 汉字整字键盘输入……（104）
2.1.3.1 术语……（104）
2.1.3.2 汉字整字键盘输入……（105）
2.2 文字识别输入……（106）
2.2.1 联机汉字识别……（106）
2.2.2 脱机印刷体识别……（107）
2.2.3 脱机手写体识别……（108）
2.3 语音识别输入……（109）
3 输出方式……（111）
3.1 汉字点阵输出和检测……（111）
3.1.1 汉字点阵输出……（111）
3.1.2 汉字点阵字模数据的检测……（116）
3.2 汉字字形压缩与还原输出……（118）
3.3 少数民族文字点阵输出……（118）
3.3.1 适用范围……（118）
3.3.2 少数民族文字字形点阵系列标准……（118）
3.3.3 主要有关标准……（119）

3.4 语音输出…… (119)
4 现代汉语属性与词语集…… (120)
4.1 汉字属性…… (120)
4.1.1 汉字基本属性…… (120)
4.1.2 汉字排序…… (120)
4.2 汉字词语切分规则…… (121)
4.2.1 词类的划分…… (121)
4.2.2 切分原则…… (121)
4.2.3 说明…… (124)
4.3 汉语词语库…… (124)
4.3.1 汉字键盘输入用通用词语库…… (124)
4.3.2 专业词语库…… (125)
4.4 汉字部首顺序表…… (126)

1 编码字符集

国家经济信息系统是一个全国性的综合性大型信息系统，它是一个多层次、跨部门、地理上分布的信息系统，因此在系统设计和应用开发过程中必须采用统一的编码字符集，以利于信息直接交换。

1.1 信息处理交换用七位编码字符集及其代码扩充技术

1.1.1 信息处理交换用七位编码字符集

本系统使用的七位编码字符集采用国家标准"信息处理交换用七位编码字符集"(GB1988—89)，它是在1980年制定的，1989年参照国际标准ISO646—1983进行了修订。修订版与第一版是完全一致的，它增加了一些名词术语，更明确地规定了七位编码字符集的版本和某些图形符号的名称。这个编码字符集在本系统内已广泛采用。它与目前个别设备上使用的美国国家标准ASCII七位编码字符集是不完全相同的，GB1988规定的七位代编码表在国内使用时位置2/4的图形符号是¥，在国际上交换信息时可以使用$符号，而ASCII规定的代码表中位置2/4的图形符号是$。因而标识这两个字符集用的转义序列也是不同的，前者为ESC 2/8 5/4，而后者是ESC 2/8 4/1。

GB1988规定一个由128个字符组成的字符集及其编码表示，这个编码字符集中的每个字符用七个二进制位表示，七位代码中，每个位组的各位用b_7、b_6、b_5、b_4、b_3、b_2、b_1表示，这里b_7是最高位，b_1是最低位。在二进制记数法中利用下列权值对各位进行加权，则可以把位组看成是代表0至127范围内的数。

位	b_7	b_6	b_5	b_4	b_3	b_2	b_1
权值	64	32	16	8	4	2	1

位组也可用它在七位代码表中位置的列号和行号 (X/Y) 的形式表示，X是b_7、b_6和b_5表示的数，赋予b_7、b_6和b_5的权值分别为4、2和1，则X表示成0至7范围内的数。Y是b_4、b_3、b_2和b_1表示的数，赋予b_4、b_3、b_2和b_1的权值分别为8、4、2和1，则y表示成0至15范围内的数。

七位编码字符集的结构如图1所示。这个字符集中各位组的分配如下：

(1) 位组0/0至1/15表示的32个控制字符组成的控制字符集。

(2) 位组2/0表示的间隔字符，这个字符既可解释成控制字符，它起到格式控制符的作用，使操作位置正向移动一个字符位置，也可解释成图形字符，表示这个图形字符是一个没有可视图形的图形符号。

(3) 位组2/1至7/14表示的94个图形字符组成的图形字符集。

(4) 位组7/15表示的控制字符抹掉。

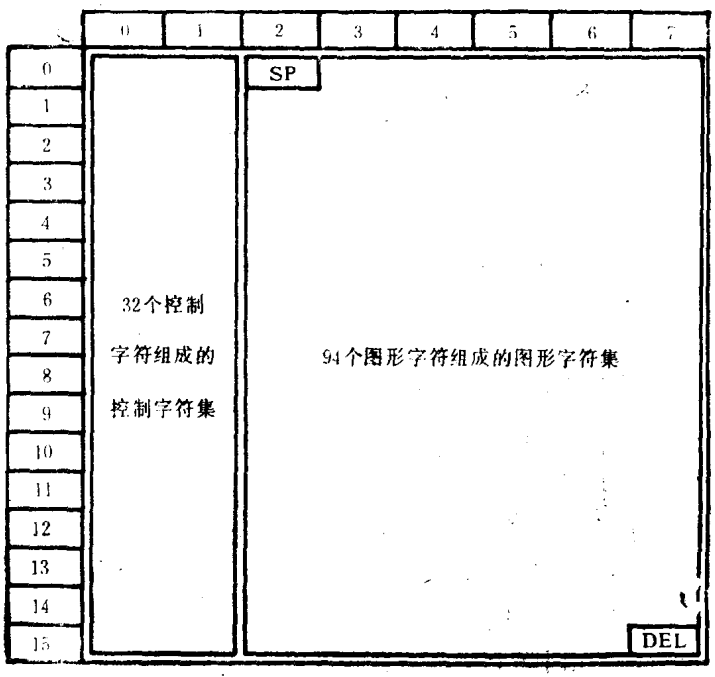

图1 七位编码字符集的结构

由32个控制字符组成的控制字符集用作C0集，由转义序列ESC 2/1 4/0指明和调用。由94个图形字符组成的图形字符集可以用作G0、G1、G2或G3集，但一般都作G0集用，由转义序列ESC 2/8 5/4指明，由控制字符移入(SI)调用。当用作G1、G2或G3集时，则它们分别由转义序列ESC 2/9 5/4，ESC 2/10 5/4和ESC 2/11 5/4指明，并分别由SO、LS2和LS3调用整个字符集或由SS2和SS3调用G2集和G3集中的单个字符。图2为GB 1988规定的七位代码表。

七位代码表中控制字符的名称和定义将在1.4.5中给出。它的图形字符的名称和编码表示在表1中给出。

GB 1988规定的图形字符都是使操作位置正向移动的进格字符。为了在同一字符位置上产生两个或多个图形字符合成的图形符号，则应使用控制字符退格或回车。例如斜线(2/15)和等于符号(3/13)可以合成"不等"的图形(≠)。这时在斜线和等于符号之间必须有一个控制字符退格(BS)。

1.1.2 八位编码字符集

本系统内除使用1.1.1中规定的七位代码外，还可使用八位代码。系统内使用的八位代码有两种形式。第一种形式是在GB 1988规定的七位代码基础上增加一位构成八位，它的表示形式为b_8至b_1，其中b_8为最高位，b_1为最低位，也可用其代码表中位置的列号和行号xx/yy表示，其中xx为列号，对应于b_8至b_5，yy为行号，对应于b_4至b_1。在这种代码系中，小于9的列号，其前面的"0"不能省掉。在这种八位代码系中第八位为"0"时构成的128个位组

b_7	0	0	0	0	1	1	1	1
b_6	0	0	1	1	0	0	1	1
b_5	0	1	0	1	0	1	0	1

b_4	b_3	b_2	b_1		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	1	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	IS4	,	<	L	\	l	|
1	1	0	1	13	CR	IS3	-	=	M	〕	m	}
1	1	1	0	14	SO	IS2	.	>	N	∧	n	—
1	1	1	1	15	SI	IS1	/	?	O	_	o	DEL

*在国际上进行信息交换时，位置2/4的图形符号可以使用$符号。

图2　GB 1988规定的七位代码表

表1　图形字符的名称和编码表示

图　形	名　　称	编码表示	图　形	名　　称	编码表示
!	感叹号	2/1	P	大写拉丁字母P	5/0
"	双引号	2/2	Q	大写拉丁字母Q	5/1
#	数码记号	2/3	R	大写拉丁字母R	5/2
¥	元货币符	2/4	S	大写拉丁字母S	5/3
%	百分比	2/5	T	大写拉丁字母T	5/4
&	和	2/6	U	大写拉丁字母U	5/5
'	右单引号	2/7	V	大写拉丁字母V	5/6
(	左圆括号	2/8	W	大写拉丁字母W	5/7
)	右圆括号	2/9	X	大写拉丁字母X	5/8
*	星号	2/10	Y	大写拉丁字母Y	5/9
+	正号	2/11	Z	大写拉丁字母Z	5/10
,	逗号	2/12	[	左方括号	5/11
-	连字符、负号	2/13	\	反斜线	5/12
.	句号	2/14	]	右方括号	5/13
/	斜线	2/15	^	向上箭头	5/14
0	数字 0	3/0	_	下横线	5/15
1	数字一	3/1	`	左单引号	6/0
2	数字二	3/2	a	小写拉丁字母a	6/1
3	数字三	3/3	b	小写拉丁字母b	6/2
4	数字四	3/4	c	小写拉丁字母c	6/3
5	数字五	3/5	d	小写拉丁字母d	6/4
6	数字六	3/6	e	小写拉丁字母e	6/5
7	数字七	3/7	f	小写拉丁字母f	6/6
8	数字八	3/8	g	小写拉丁字母g	6/7
9	数字九	3/9	h	小写拉丁字母h	6/8
:	冒号	3/10	i	小写拉丁字母i	6/9
;	分号	3/11	j	小写拉丁字母j	6/10
<	小于符号	3/12	k	小写拉丁字母k	6/11
=	等于符号	3/13	l	小写拉丁字母l	6/12
>	大于符号	3/14	m	小写拉丁字母m	6/13
?	问号	3/15	n	小写拉丁字母n	6/14
@	商业用单价符号	4/0	o	小写拉丁字母o	6/15
A	大写拉丁字母A	4/1	p	小写拉丁字母p	7/0
B	大写拉丁字母B	4/2	q	小写拉丁字母q	7/1
C	大写拉丁字母C	4/3	r	小写拉丁字母r	7/2
D	大写拉丁字母D	4/4	s	小写拉丁字母s	7/3
E	大写拉丁字母E	4/5	t	小写拉丁字母t	7/4
F	大写拉丁字母F	4/6	u	小写拉丁字母u	7/5
G	大写拉丁字母G	4/7	v	小写拉丁字母v	7/6
H	大写拉丁字母H	4/8	w	小写拉丁字母w	7/7
I	大写拉丁字母I	4/9	x	小写拉丁字母x	7/8
J	大写拉丁字母J	4/10	y	小写拉丁字母y	7/9
K	大写拉丁字母K	4/11	z	小写拉丁字母z	7/10
L	大写拉丁字母L	4/12	{	左花括号	7/11
M	大写拉丁字母M	4/13	|	竖线	7/12
N	大写拉丁字母N	4/14	}	右花括号	7/13
O	大写拉丁字母O	4/15	~	上横线	7/14

与GB 1988中规定的128个字符(位组00/14和00/15除外)完全相同。第八位为"1"时构成另外128个位组，其中列08和列09的32个位组构成另外一个控制字符集，称它为C_1集，列10至列15构成另外一个图形字符集。它的结构如图3所示。

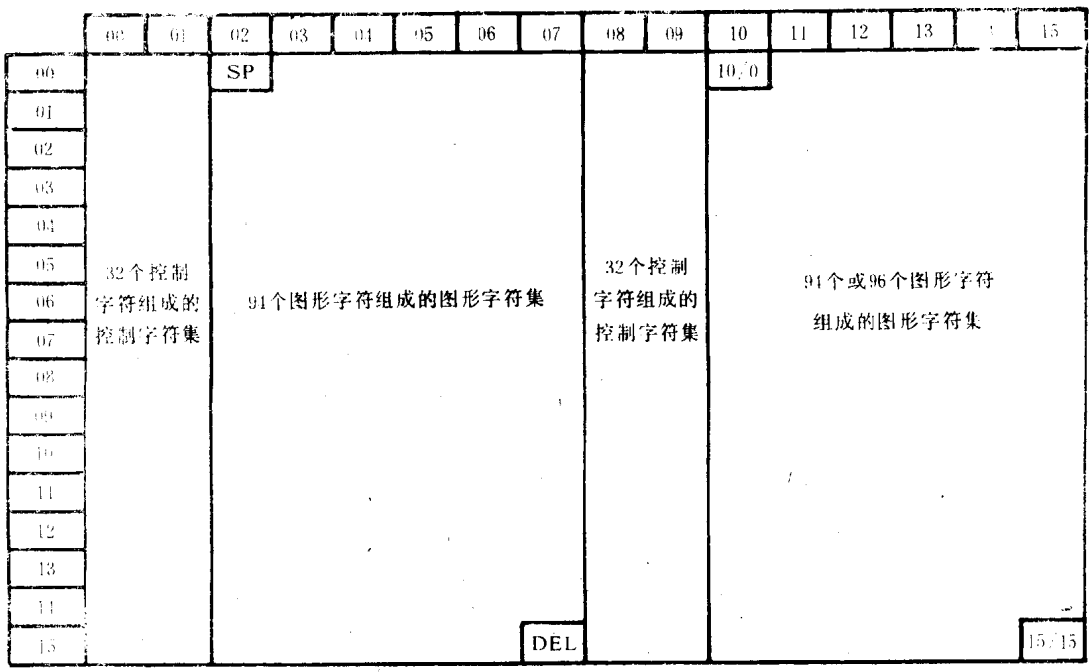

图3 八位代码的结构

本系统内使用的另一种八位代码形式是EBCDIC码，这种代码体系与GB 1988规定的代码体系是不同的，两者是不一致的。这种八位代码只限于系统内部使用，不应扩大其使用范围。这两种代码之间应按统一的转换规则进行转换。

1.1.2.1 GB11383规定的八位代码的结构和编码规则

GB 11383规定的八位代码是在GB 1988规定的七位代码基础上增加一位构成。它由下列几部分组成：

a. C_0 集

由30个控制字符组成的控制字符集，这些字符由位组00/00至01/15表示，其中不应使用位组00/14和00/15。

b. 间隔字符

由位组02/00表示的字符，可以把它解释为控制字符或图形字符，也可以把它解释为既是控制字符又是图形字符。

c. G_0 集

由位组02/01至07/14表示的最多为94个图形字符组成的图形字符集。

d. 抹掉字符

由位组07/15表示的控制字符。

e. C1集

由位组08/00至09/15表示的最多为32个控制字符组成的控制字符集。

f. G1集

由位组10/00至15/15表示的最多为96个图形字符组成的图形字符集。

g. G2集

由最多为96个图形字符组成的图形字符集。

h. G3集

由最多为96个图形字符组成的图形字符集。

这些组成部分由相应的转义序列和移位功能来指明和调用。

C0集应安排在列00和01上，它应由转义序列ESC 02/01 F指明和调用。

间隔字符安排在列02的位置02/00上，它不必明确地指明和调用。

G0集应安排在列02至列07的位置02/01至07/14上，它应由转义序列ESC 02/08 F指明。

抹掉字符应安排在列07的位置07/15上，它不必明确地指明和调用。

C1集应安排在列08和列09上，它应由转义序列ESC 02/02 F指明和调用。

G1集应由转义序列ESC 02/09 F或ESC 02/13 F指明，调用时它应安排在列10至列15上，若G1集是由94个字符组成的字符集，则G1集的字符由位组10/01至15/14表示，若G1集是由96个字符组成的字符集，则G1集的字符由位组10/00至15/15表示。

G2集应由转义序列ESC 02/10 F或ESC 02/14 F指明，它可以由锁移功能LS2R将整个字符集调入列10至列15，或者由单移功能SS2调用这个字符集中的单个字符，当它由LS2R调用时，若G2集是由94个字符组成的字符集，则G2集的字符由位组10/01至15/14表示。若G2集是由96个字符组成的字符集，则G2集的字符由位组10/00至15/15表示。当由SS2调用时，若G2集是由94个字符组成的字符集，则紧跟着的那个位组是02/01至07/14范围内的位组，若G2集是由96个字符组成的字符集，则紧跟着的那个位组是02/00至07/15范围内的位组。

G3集应由转义序列ESC 02/11 F或ESC 02/15 F指明，它可以由锁移功能LS3R将整个字符集调入列10至列15，或者由单移功能SS3调用这个字符集中的单个字符。当它由LS3R调用时，若G3集是由94个字符组成的字符集，则G3集的字符由位组10/01至15/14表示，若G3集是由96个字符组成的字符集，则G3集的字符由位组10/00至15/15表示。当由SS3调用时，若G3集是由94个字符组成的字符集，则紧跟着的那个位组是02/01至07/14范围内的位组，若G3集是由96个字符组成的字符集，则紧跟着的那个位组是02/00至02/15范围内的位组。图4列出了八位代码的组成及其调用和位置安排。

为了简化信息交换，在这种八位代码中，规定了三种嵌套的等级，每种等级可由宣布序列标识。

ESC 02/00 04/12标识等级一

ESC 02/00 04/13标识等级二

ESC 02/00 04/14标识等级三

三种嵌套的等级是：

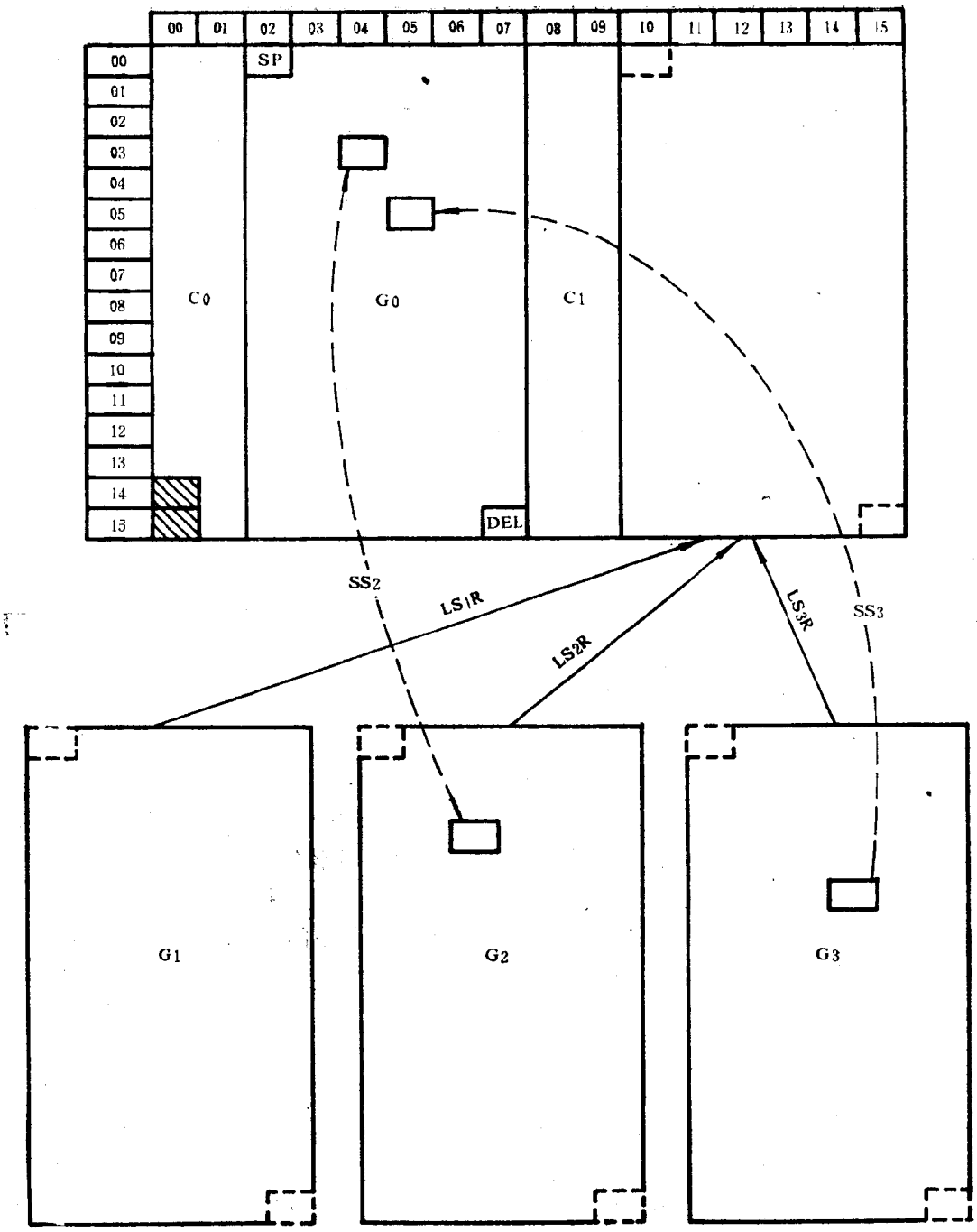

图4　八位代码的组成及其调用和位置安排

等级一(见图5)包括：

a. C0集，由ESC 02/01 F指明和调用；

b. 在位置02/00上的间隔字符；

c. G0集，由ESC 02/08 5/4指明；

d. 在位置07/15上的抹掉字符；

e. C_1集，由ESC 03/02 F指明和调用；

f. G_1集，由ESC 02/09 F或ESC 02/13 F指明。

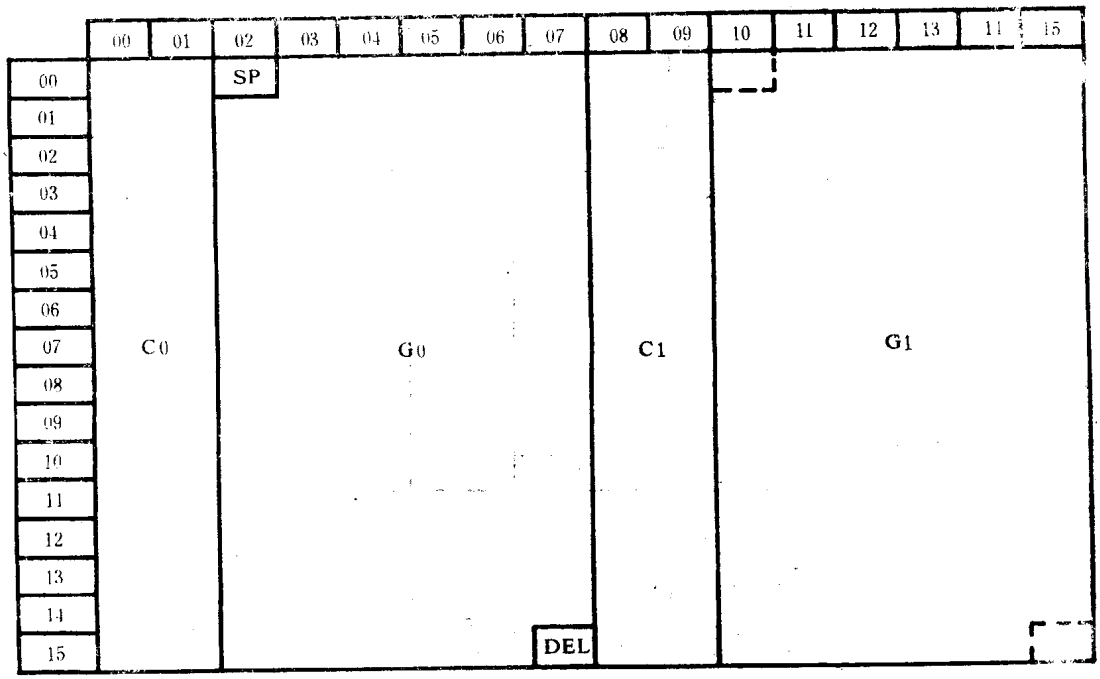

图5 等级一

等级一中不应使用移位功能，且假定将G0集和G1集永久地分别调入列02至列07和列10至列15，所以指明G0集和G1集的转义序列也暗示着调用了该字符集。

在等级一中，如果除C0集提供的控制字符外，不需要其它控制字符；除G0集提供的图形字符外，不需要其它图形字符，则C_1集和G_1集可以是空的。空的C_1集应由ESC 02/02 07/14指明和调用，空的G_1集应由ESC 02/09 07/14或ESC 02/13 07/14指明和调用。在等级一不应使用G2集和G3集。

等级二(见图6)包括等级一，并增加：

a. G2集，由ESC 02/10 F或ESC 02/14 F指明，其字符应由SS2单个地调用；

b. G3集，由ESC 02/11 F或ESC 02/15 F指明，其字符应由SS3单个地调用。

在等级二中不应使用其它移位功能。

G_1集不应是空的，G2集和G3集不应都是空的。若在G0、G_1和G3集所提供的图形字符以外，不需要其它图形字符，则G2集可以是空的，在这种情况下，G2集应由ESC 02/10 07/14或ESC 02/14 07/14指明。若在G0集、G_1集和G2集所提供的图形字符以外，不需要其它图形字符，则G3集可以是空的，在这种情况下，G3集应由ESC 02/11 07/14或ESC 02/15 07/14指明。

C_1集不应是空的，至少应在位置08/14和08/15处分别设置单移功能SS_2和SS_3。仅在位置08/14和08/15处设置单移功能SS_2和SS_3的C_1集，由转义序列ESC 02/02 04/07指明和调用。

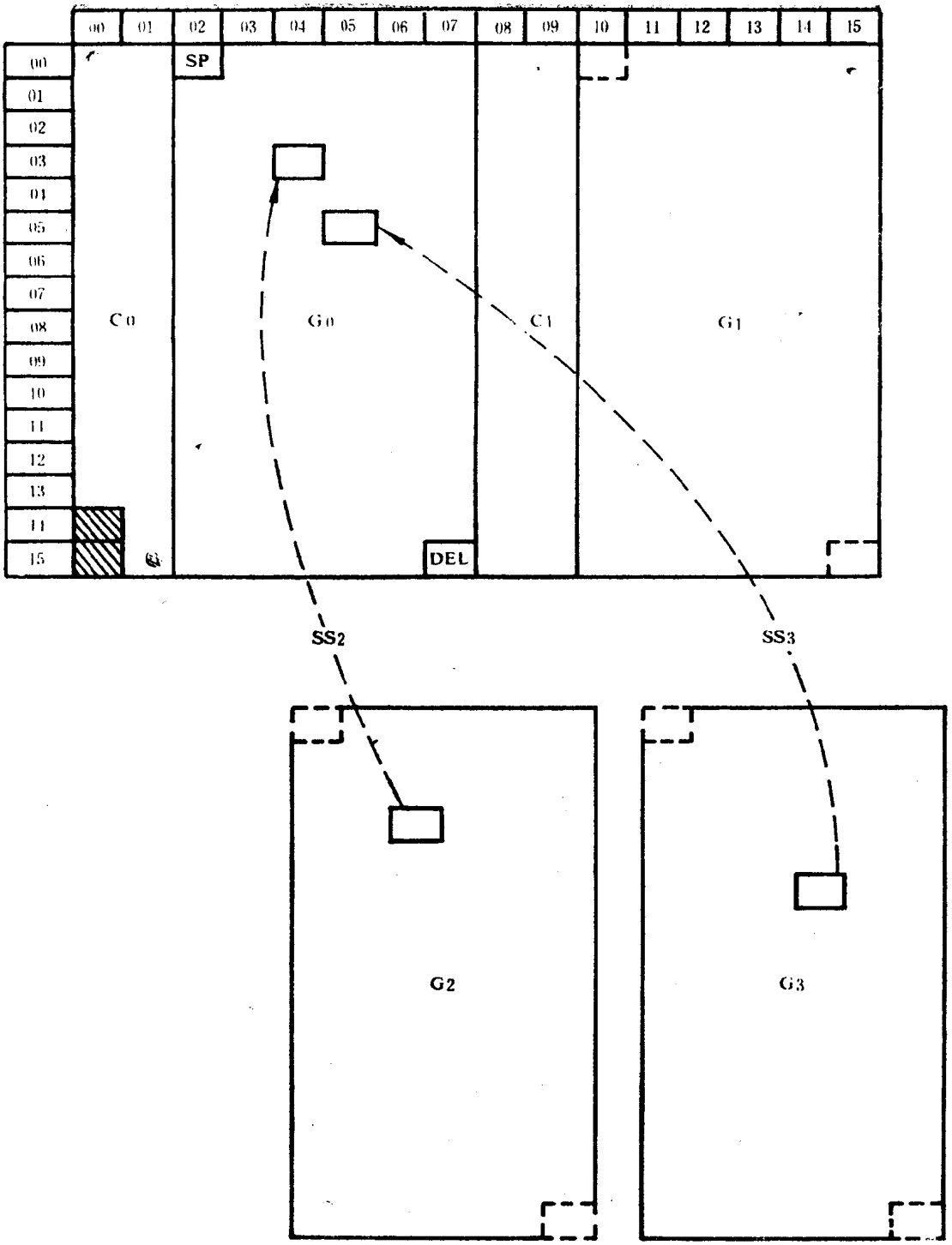

图6　等级二

等级三(见图7)包括等级二，并增加下列三个移位功能：

a. LS_1R