① GB2312 及区位码、国标码、机内码、外码、字形码

GB2312 标准共收录 6763 个汉字，其中一级汉字 3755 个，二级汉字 3008 个；同时，GB 2312 收录了包括拉丁字母、希腊字母、日文平假名及片假名字母、俄语西里尔字母在内的 682 个全角字符。整个字符集分成 94 个区，每区有 94 个位。

- GB2312，又称为 GB0，由中国国家标准总局发布，1981 年 5 月 1 日实施
- GB2312 标准共收录 6763 个汉字，其中一级汉字 3755 个，二级汉字 3008 个
- GB2312 是一种区位码。分为 94 个区(01-94)，每区 94 个字符(01-94)
- 01-09 区为特殊符号
- 10-15 区没有编码
- 16-55 区为一级汉字，按拼音排序，共 3755 个
- 56-87 区为二级汉字，按部首 / 笔画排序，共 3008 个
- 88-94 区没有编码
- GB2312 只是编码表，在计算机中通常都是用"EUC-CN"表示法，即在每个区位加上 0xA0 来表示。区和位分别占用一个字节。

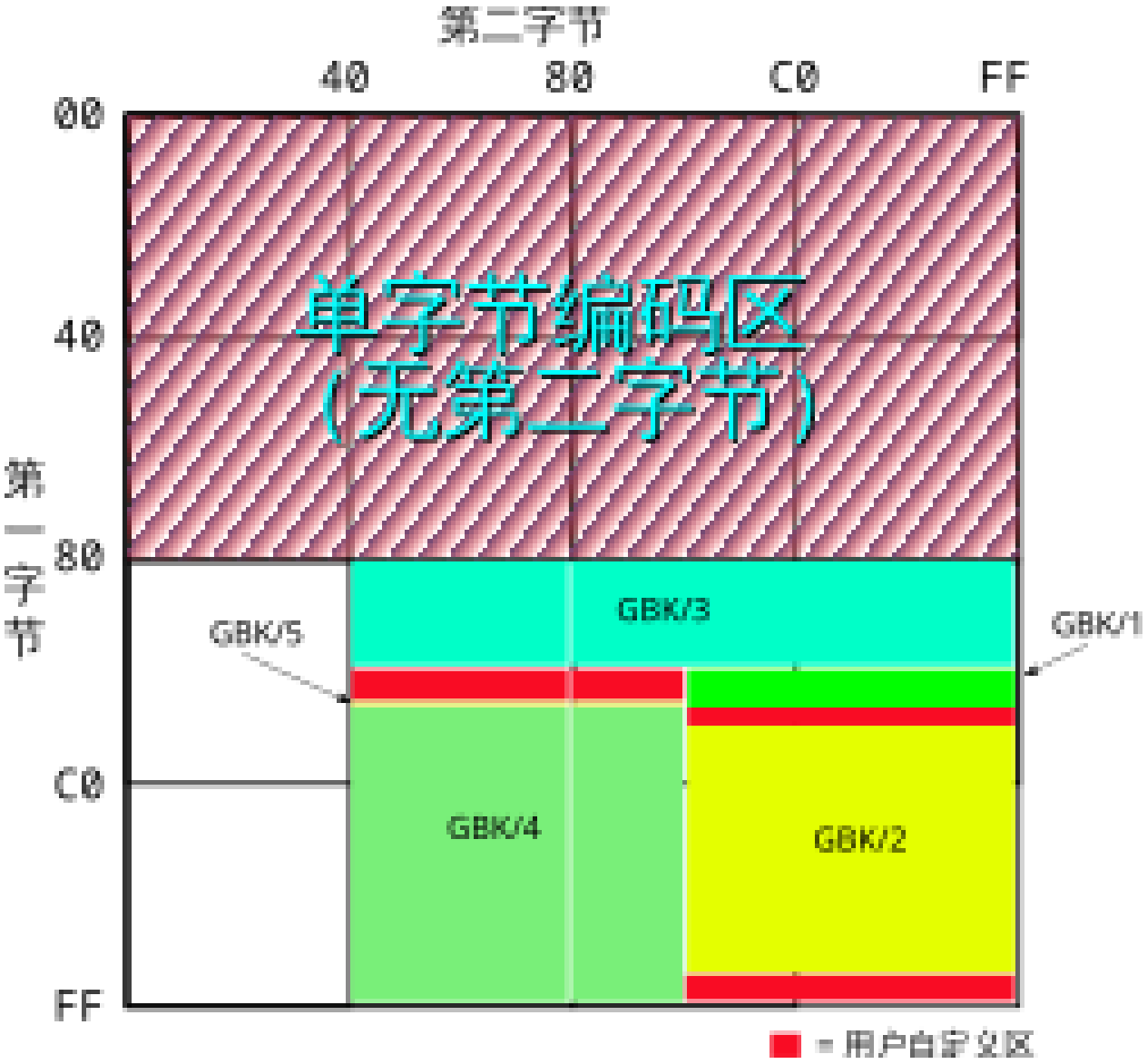
第 01 区	+0	+1	+2	+3	+4	+5	+6	+7	+8	+9
A1A0			、	。	·	ˉ	ˇ	¨	〃	々
A1B0	―	‖	…	‘	’	“	”	〔	〕	〈
A1C0	±	×	÷	∶	∧	∨	∑	∏	∪	∩
A1D0	∈	∷	√	⊥	∥	∠	⌒	⊙	∫	∮
A1E0	∴	♂	♀	°	′	″	℃	$	¤	¢
A1F0	‰	●	◎	◇	◆	□	■	△	▲	※

后面略

高字节 0xA1-0xFE　　FE-A1 = 5D　　共 94

低字节 0xA1-0xFE

其中汉字的编码范围为 B0A1-F7FE，第一字节 0xB0-0xF7（对应区号：16－87），第二个字节 0xA1-0xFE（对应位号：01－94）。



GB2312 汉字区即 GBK/2：B0A1-F7FE 收录汉字 6763 个。

为了避开 ASCII 字符中的不可显示字符，十六进制为 0 ~ 1F，十进制为 0 ~ 31。及空格字符 0010 0000(十六进制为 20，十进制为 32)，国标码(又称为交换码)规定表示汉字的范围，十六进制为(21，21) ~ (7E，7E)，十进制为(33，33) ~ (126，126)。

可以算出“万”字的国标码十进制为：(45+32，82+32) = (77，114)，十六进制为：(4D，72)。

“万”字国标码中的高位字节 77 与 ASCII 的“M”冲突，低位字节 114 与 ASCII 的“r”冲突。因此，为避免与 ASCII 码冲突，规定国标码中的每个字节的最高位都从 0 换成 1，即相当于每个字节都再加上 128(十六进制为 80，即 80H；二进制为 1000 0000)，从而得到国标码的“机内码”表示，简称“内码”。

从区位码(国家标准定义) ---> 区码和位码分别+32(即+20H)得到国标码 ---> 再分别+128(即+80H)得到机内码(与 ACSII 码不再冲突)。

因此，区位码的区和位分别+160(即+A0H，32+128=160)可直接得到内码。用十六进制表示就是：

区位码(区码, 位码) + (20H, 20H) + (80H, 80H)

= 区位码(区码, 位码) + (A0H, A0H)

= 内码(高字节, 低字节)。

外码也叫输入码，目前常用的汉字外码，按照编码形式上的不同，大致上可分为以下几类：

1）数字编码，比如区位码；

2）拼音编码，比如全拼、双拼、自然码等；

3）字形编码，比如五笔、表形码、郑码等。

为了将汉字在显示器或打印机上输出，把汉字按图形符号设计成点阵图，就得到了相应的点阵代码(字形码)。

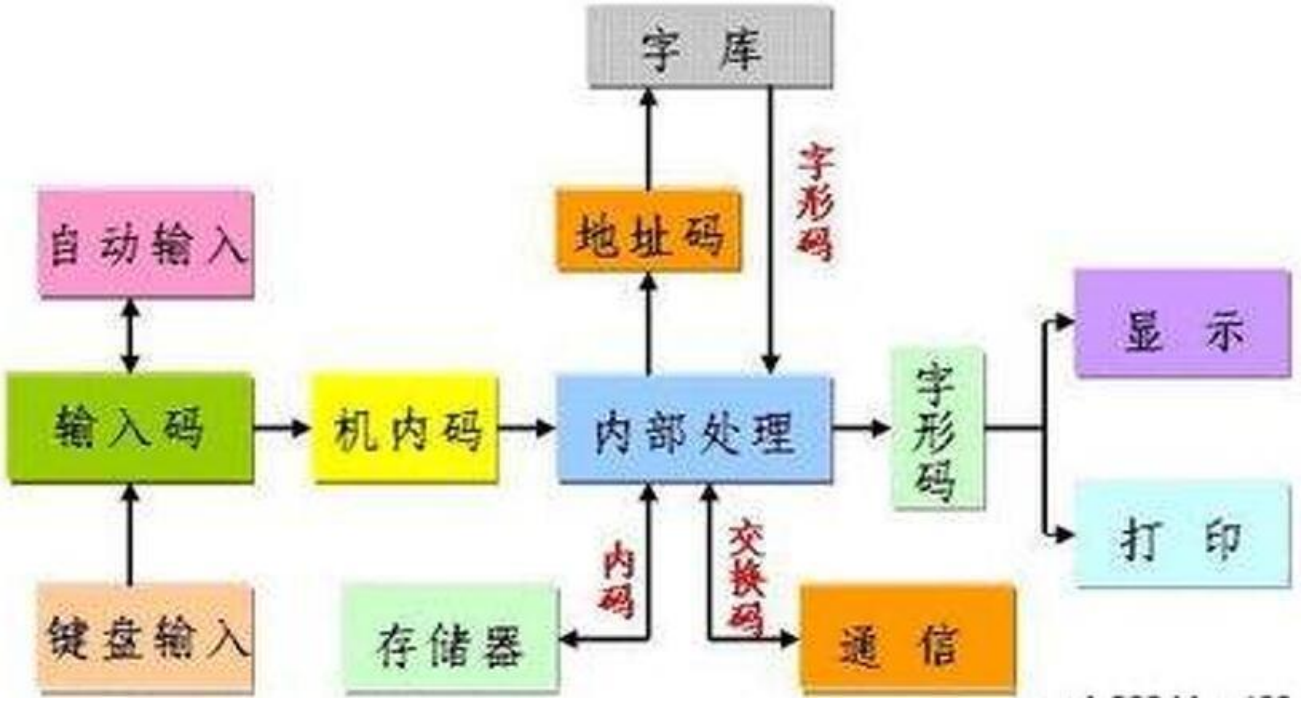
也就是用 0、1 表示汉字的字形，将汉字放入 n 行*n 列的正方形(即点阵)内，该正方形共有 n^2 个小方格，每个小方格用一位二进制数表示，凡是笔划经过的方格其值为 1，未经过的方格其值为 0。

为了将汉字的字形显示输出或打印输出，汉字信息处理系统还需要配有汉字字形库，也称字模库，简称字库，它集中存储了汉字的字形信息。

字库按输出方式可分为显示字库和打印字库。用于显示输出的字库叫显示字库，工作时需调入内存。用于打印输出的字库叫打印字库，工作时无需调入内存。

字库按存储方式也可分为软字库和硬字库。软字库以字体文件(即字形文件)的形式存放在硬盘上，现在多用这种方式。硬字库则是将字库固化在一个单独的存储芯片中，再和其它必要的元器件组成接口卡，插接在计算机上，通常称为汉卡，比如当年的巨人汉卡、联想汉卡。不过这种方式现已淘汰。

- 为在计算机内表示汉字而采取统一的编码方式所形成的汉字编码叫内码；
- 为方便汉字输入而形成的汉字编码为外码，也叫输入码；
- 为显示输出和打印输出汉字而形成的汉字编码为字形码，也称为字模码、输出码。



② **GBK 汉字及符号大致范围 0x8140-0xFEFE。** *GBK 的编码框架(Code Scheme)：其中 GBK/1 收录除 GB2312 字符外的其他增补字符，GBK/2 收录 GB2312 字符，GBK/3 收录 CJK 字符，GBK/4 收录 CJK 字符和增补字符，GBK/5 为非中文字符，UDC 为用户自定义字符。*

虽然 GBK 跟 GB2312 一样是双字节编码，但 GBK 只要求第一个字节即高字节大于 127 就固定表示这是一个汉字的开始(即 GBK 编码高字节的首位必须是 1；0~127 当然表示的还是 ASCII 字符)，不再像 GB2312 一样要求第二个字节即低字节也必须大于 127(即 GBK 编码低字节首位既可以是 0，也可以是 1)。

BIG5 **汉字及符号大致范围 0x8140-0xFEFE。**大五码是双字节字符集，以十六进制表示，使用双八码存储方法，以两字节安放一字。第一字节称为“高位字节”，第二字节称为“低位字节”。“高位字节”使用了 0x81 至 0xFE，“低位字节”使用了 0x40 至 0x7E，及 0xA1 至 0xFE。

UNICODE 汉字范围 0x4E00-0x9FA5

GBK BIG5 是内码。

Unicode 是一种字符编码标准，它为世界上几乎所有的字符分配了一个唯一的数字代码，旨在解决不同字符编码系统之间的兼容性问题。

③ **输入法**如何处理不同编码的字符，解析了从键盘输入到屏幕显示的全过程，包括 ASCII 码转换、编码到字体的匹配及文件存储的编码转换。

事实上我用输入法对着什么程序输入都不会出问题。那么这是怎么做到的呢？

我猜有两种情况：

1、程序会告诉输入法它要哪种编码的字符。

2、输入法输出的字符在输入程序的时候被程序转换了。

当软件或者没有给出任何编码指示的时候，都推脱给操作系统了，操作系统是中文的，默认就 GB2312 了，如果操作系统是其他默认编码方式，你的输入也会变成相应的其他的方式。

"直接用输入法打出来的字" -- 这里含好几个过程。

（1）当你用键盘打字时，从键盘进入计算机的是 ASCII 码序列。

（2）“输入法”把 ASCII 码序列 转换成 输入法 自己规定的 码。

（3）你在一个窗上看到的东西，例如 notepad 的文本编辑窗，wordpad 的文本编辑窗，或 DOS 黑窗，那是 把 “输入法自己规定的码”显示出来，这里有一个编码到字体（font)到 bitmap 点阵图形的转换。只有当它们匹配时才能显示出有意义的字的形状。否则看上去是“乱码”。