



第二章 汉字代码体系



内容提要

- ✓ **ASCII码及其扩展**
- ✓ 中文信息在计算机内的表示
- ✓ **ISO/IEC 2022**
- ✓ 汉字编码字符集
- ✓ **GB2312-80**
- ✓ **BIG-5**
- ✓ **Unicode和ISO10646**
- ✓ **GBK和GB18030**



1.1 ASCII码及其扩展

- ✓ ASCII码
- ✓ 扩展ASCII
- ✓ CJK-Roman



1.1.1 ASCII码





1.1.1 ASCII码

✓ ASCII

- **American Standard Code for Information Interchange**
- 表示英文、数字及其常用符号
- 和现有的英文键盘相对应



✓ 1991年ISO定义为ISO/IEC 646:1991

- 信息交换用7-位编码字符集 (**ISO 7-bit coded character set for information interchange**)



1. 1. 2 ASCII码内容

✓ 7位二进制数，定义128个字符：

➢ 94个图形字符（可显示字符）

➢ ‘0’-‘9’: 30H-39H

➢ ‘A’-‘Z’: 41H-5AH

➢ ‘a’-‘z’: 61H-7AH

➢ 30个控制字符

✓ 00-19H

➢ 1个空格字符

✓ 20H

➢ 1个Del（删除）符

✓ 7FH

| ASCII码 | | 字符 | 控制字符 | 意义 | ASCII码 | | 字符 | 控制字符 | 意义 |
|--------|------|----|------|-----------------|--------|------|----|------|--------|
| 十进制 | 十六进制 | | | | 十进制 | 十六进制 | | | |
| 000 | 00 | | NULL | | 016 | 10 | ► | DLE | |
| 001 | 01 | ☺ | SOH | | 017 | 11 | ◄ | DC1 | |
| 002 | 02 | ☹ | STX | | 018 | 12 | ↕ | DC2 | |
| 003 | 03 | ♥ | ETX | | 019 | 13 | ❗ | DC3 | |
| 004 | 04 | ♦ | EOT | | 020 | 14 | ¶ | DC4 | |
| 005 | 05 | ♣ | ENQ | | 021 | 15 | § | NAK | |
| 006 | 06 | ♠ | ACK | | 022 | 16 | — | SYN | |
| 007 | 07 | • | BELL | 振铃 | 023 | 17 | ↑ | ETB | |
| 008 | 08 | ■ | BS | 退格键 | 024 | 18 | ↑ | CAN | |
| 009 | 09 | | HT | 定位键 | 025 | 19 | ↓ | EM | |
| 010 | 0A | | LF | line feed | 026 | 1A | → | SUB | 档案结束 |
| 011 | 0B | ♂ | VT | home | 027 | 1B | ← | ESC | escape |
| 012 | 0C | ♀ | FF | form feed | 028 | 1C | L | FS | 向右键 |
| 013 | 0D | | CR | carriage return | 029 | 1D | ↔ | GS | 向左键 |
| 014 | 0E | ♪ | SO | | 030 | 1E | ▲ | RS | 向上键 |
| 015 | 0F | ☼ | SI | | 031 | 1F | ▼ | US | 向下键 |

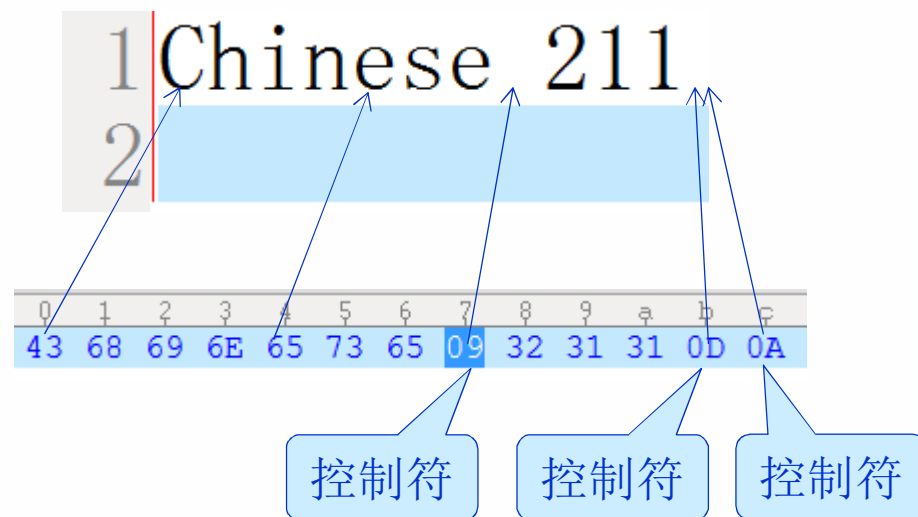


1. 1. 3 ASCII码一布局

| | 00 | 01 | 02 | 03 | 04 | 05 | 06 | 07 |
|----|----|----|----|----|----|----|----|----|
| 00 | | | | | | | | |
| 01 | | | | | | | | |
| 02 | | | | | | | | |
| 03 | | | | | | | | |
| 04 | | | | | | | | |
| 05 | | | | | | | | |
| 06 | | | | | | | | |
| 07 | | | | | | | | |
| 08 | | | | | | | | |
| 09 | | | | | | | | |
| A | | | | | | | | |
| B | | | | | | | | |
| C | | | | | | | | |
| D | | | | | | | | |
| E | | | | | | | | |
| F | | | | | | | | |

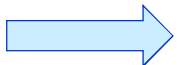
控制字符区

图形字符区





1.2 扩展ASCII

- ✓ 8位表示扩展
 - 128  256
- ✓ 扩展的字符集有16个定义：从ISO 8859-1到ISO 8859-16，分别定义了相应国家的文字和符号。



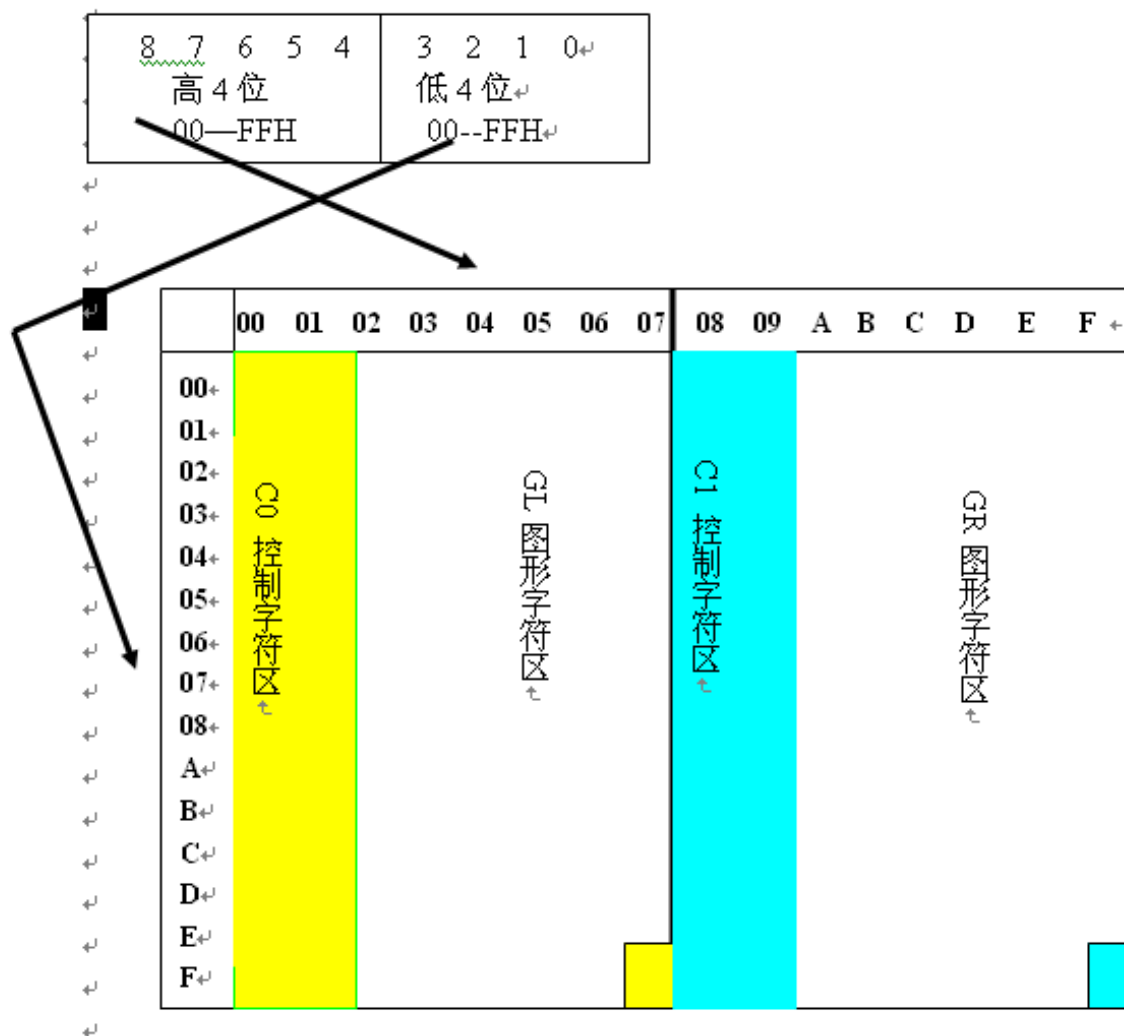
1.2.1 ISO 8859内容

√ ISO 8859

- 第1至第16部分（Information Processing—8 Bit Single-Byte Coded Graphic Character Set）。
- 定义新增的128个码元。
- 每个部分分别定义ASCII码和其扩展的字符集（针对不同拉丁语言）。



1.2.2 ISO 8859代码空间图





1.2.3 ISO/IEC 8859 举例

| ISO/IEC 8859-1 | | | | | | | | | | | | | | | | | | | | | | | | |
|----------------|------|----|----|----|----|----|----------------|------|----|----|----|----|----|----|----|----|----|----|----|-----|----|----|----|---|
| | x0 | x1 | x2 | x3 | x4 | x5 | ISO/IEC 8859-7 | | | | | | | | | | | | | | | | | |
| 0x | | | | | | | | x0 | x1 | x2 | x3 | x4 | x5 | x6 | x7 | x8 | x9 | xA | xB | xC | xD | xE | xF | |
| 1x | | | | | | | 0x | | | | | | | | | | | | | | | | | |
| 2x | SP | ! | " | # | \$ | % | 1x | | | | | | | | | | | | | | | | | |
| 3x | 0 | 1 | 2 | 3 | 4 | 5 | 2x | SP | ! | " | # | \$ | % | & | ' | (|) | * | + | , | - | . | / | |
| 4x | @ | A | B | C | D | E | 3x | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | : | ; | < | = | > | ? | |
| 5x | P | Q | R | S | T | U | 4x | @ | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | |
| 6x | ` | a | b | c | d | e | 5x | P | Q | R | S | T | U | V | W | X | Y | Z | [| \ |] | ^ | _ | |
| 7x | p | q | r | s | t | u | 6x | ` | a | b | c | d | e | f | g | h | i | j | k | l | m | n | o | |
| 8x | | | | | | | 7x | p | q | r | s | t | u | v | w | x | y | z | { | | } | ~ | | |
| 9x | | | | | | | 8x | | | | | | | | | | | | | | | | | |
| Ax | NBSP | ı | ç | £ | ¤ | ¥ | 9x | | | | | | | | | | | | | | | | | |
| Bx | ° | ± | ² | ³ | ´ | µ | Ax | NBSP | ı | £ | € | ¸ | ı | § | ¨ | © | ª | « | ¬ | SHY | | — | | |
| Cx | À | Á | Â | Ã | Ä | Å | Bx | ° | ± | ² | ³ | ´ | µ | À | Á | Â | Ã | Ä | Å | » | Ö | ½ | Υ | Ω |
| Dx | Đ | Ñ | Ò | Ó | Ô | Õ | Cx | İ | A | B | Γ | Δ | E | Z | H | Θ | I | K | Λ | M | N | Ξ | O | |
| Ex | à | á | â | ã | ä | å | Dx | Π | P | | Σ | T | Υ | Φ | X | Ψ | Ω | İ | Ÿ | á | é | ñ | í | |
| Fx | ð | ñ | ò | ó | ô | õ | Ex | Ü | α | β | γ | δ | ε | ζ | η | θ | ι | κ | λ | μ | ν | ξ | ο | |
| | | | | | | | Fx | π | ρ | ς | σ | τ | υ | φ | χ | ψ | ω | ϊ | ϋ | ό | ύ | ώ | | |

ISO/IEC 8859-1

符集显示。

语字母换走，加入土耳其语字母。
来代替Latin-4。
集演化而来。

芬兰语字母和大写法语重音字母，以及
E语使用，并加入欧元符号。

ISO/IEC 8859-1



1.3 CJK-Roman

- ✓ ASCII码一样，7位二进制数编码。
- ✓ 收录字符基本与ASCII码一样,个别字符作了调整。
- ✓ 符合本国使用需要
 - 货币单位 （\$（美国）----- ¥（中国））
- ✓ 中、日、韩字符编码标准：
 - GB-Roman(中国ASCII码字符集ASCII字符编码标准，代号为GB 1988-89)；
 - CNS-Roman（台湾ASCII码标准，代号为CNS 5205-1989）；
 - JIS-Roman（日本ASCII码标准，代号为JIS X 0201-1997）；
 - KS-Roman（韩国ASCII码标准，代号为KS X 1003:1993）。



CJK-Roman—特殊字符

| 码元值 | ASCII码 | GB-Roman | CNS-Roman | JIS-Roman | KS-Roman |
|------|--------|----------|-----------|-----------|----------|
| 0x24 | \$(美元) | ¥（人民币） | \$ | \$ | \$ |
| 0x5C | \(反斜杠) | \(反斜杠) | \(反斜杠) | ¥（日圆） | ₩（韩圆） |
| 0x7E | ~（波浪线） | —（顶线） | —（顶线） | —（顶线） | —（顶线） |



内容

- ✓ ASCII码及其扩展
- ✓ 中文信息在计算机内的表示
- ✓ ISO/IEC 2022
- ✓ 汉字编码字符集
- ✓ GB2312-80
- ✓ BIG-5
- ✓ Unicode和ISO10646
- ✓ GBK和GB18030



2.1 概述

- ✓ 最早在计算机内表示中文信息：
 - IBM、富士通、日立等计算机生产厂家。
 - 采用的编码形式互不兼容。
- ✓ 为了通用性，**ISO、IEEE**以及各个使用汉字的国家和地区，都制定了各种各样的汉字编码字符集。
- ✓ 汉字代码：**汉字在计算机内表示。**
- ✓ 通过扩充**ASCII**码编码长度实现
 - **ASCII**码（扩展）最多**256**个码位
 - 汉字数量成千上万
 - 如何放？



2.1.1 汉字代码

- ✓ 汉字代码是真实世界的汉字信息在计算机系统中的最基本表示。
- ✓ 根据在计算机内部使用的目的和存储的方式，汉字代码有各种不同的形式和称谓：
 - 交换码
 - 机内码
 - 输入码
 - 字形码
 -



2.1.2 汉字交换码

- ✓ 用于信息交换的汉字代码。
- ✓ 双字节、3字节和4字节。
- ✓ 一般不能直接用于信息处理
 - 例如，在GB2312中，“码”字的交换码为十六进制的42H/6BH。无法与ASCII码的“Bk”相区别。
- ✓ 在实际使用中，**交换码必须转换为机内码。**
- ✓ 例外：
 - ISO/IEC 10646和Unicode中，交换码与机内码一致
 - ASCII码也采用双字节表示



2.1.2 汉字机内码

- ✓ 用于信息处理的汉字代码，也称：
 - ✓ 汉字处理码
 - ✓ 处理码
 - ✓ 机内码
 - ✓ 内码
- ✓ 汉字内码长度**2-4**字节，通常是双字节。
- ✓ 单字节操作系统内核，汉字代码为了与**ASCII**码相区分，往往把内码的两字节（至少把第一个字节）的最高位（**Bit 7**）置为**1**。



2.1.3 相互关系

√ GB2312

中

> 56 50 (交换码)

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|

> D6 D0 (机内码)

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|

√ Unicode

中

> 4E2DH (交换码)

> 4E2DH (机内码)



内容

- ✓ ASCII码及其扩展
- ✓ 中文信息在计算机内的表示
- ✓ **ISO/IEC 2022**
- ✓ 汉字编码字符集
- ✓ GB2312-80
- ✓ BIG-5
- ✓ Unicode和ISO10646
- ✓ GBK和GB18030

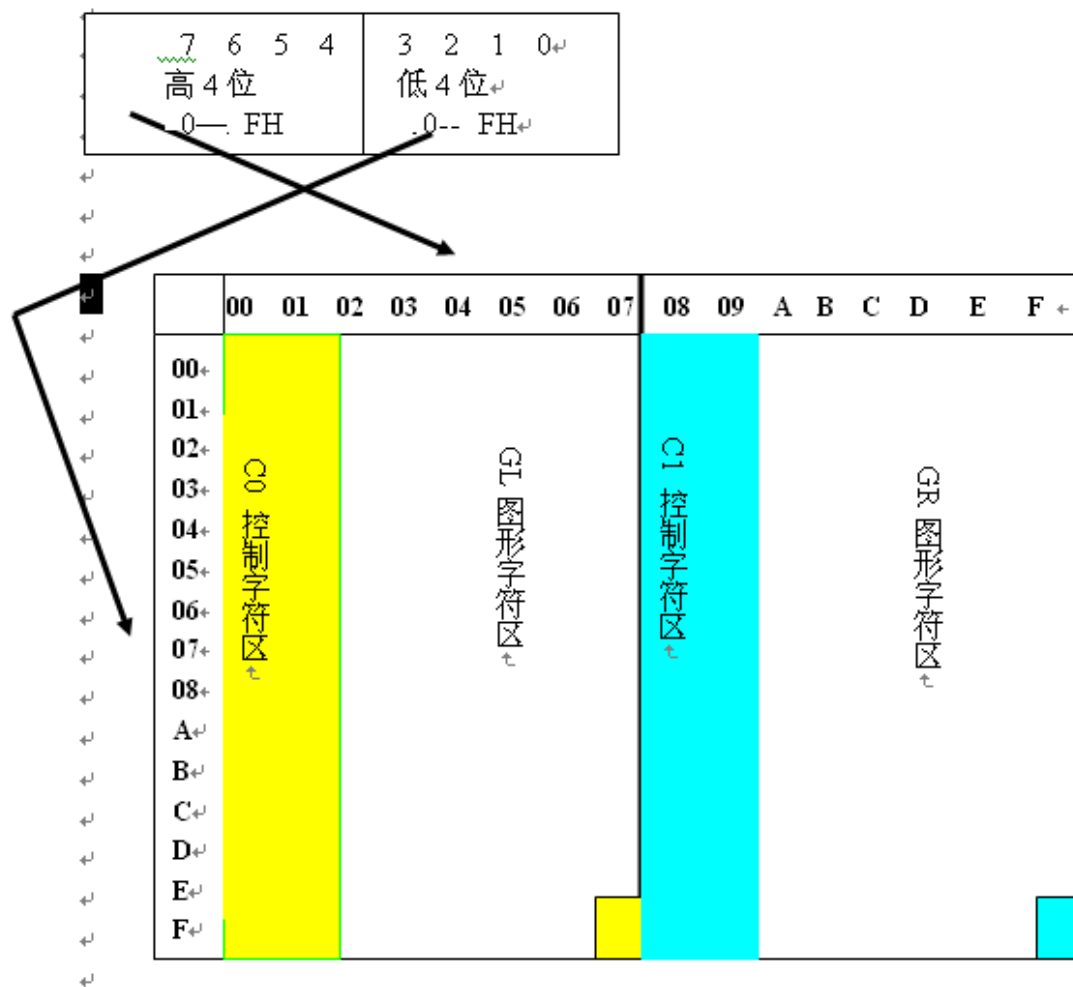


3. ISO 2022标准

- ✓ 国际标准化组织于1976年制订了ISO 2022国际标准，ISO 2022定义了七位代码和八位代码的空间及其代码空间扩充的技术。
- ✓ 多数计算机系统所采用的字符集都是以ISO 2022为基础。
- ✓ 我国根据ISO 2022制订了国家标准GB 2311。



3.1 单八位代码空间图





3.2 单八位代码空间布局

- ✓ **00-31** (00H-1FH)
 - › 第一个控制字符集C0编码区域
- ✓ **32** (20H) : Space
- ✓ **127** (十六进制为7FH) : DELETE
- ✓ **128-160** (80H-A0H)
 - › 第二个控制字符集C1编码区域
- ✓ **33-126** (GL) 和 **161-254** (GR)
 - › 两个图形字符编码区域

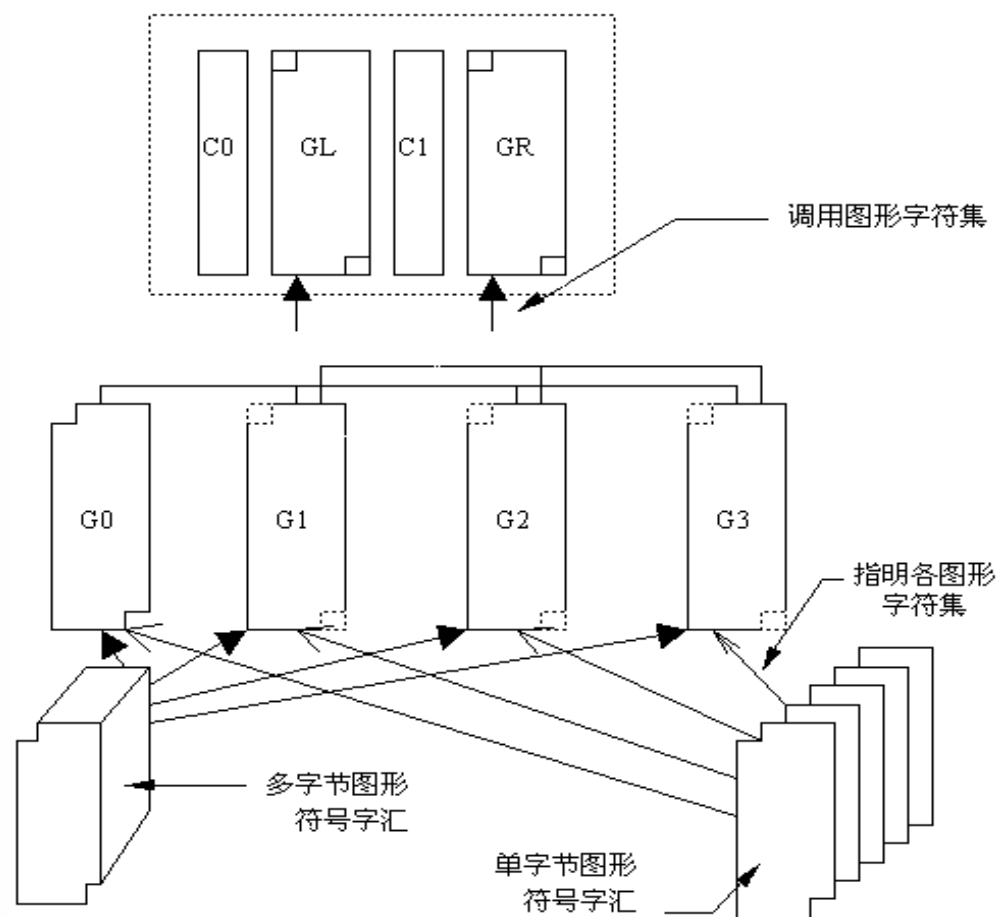


3.3 扩充编码空间的方法

- ✓ ISO 2022扩充编码空间的方法，可以用多个7位单元或8位对字符进行编码，但是必须跳过控制字符区（即C0和C1的区域）。
- ✓ 采用该标准扩充的编码空间为 94^n ， n 为编码单元的个数，若 $n=2$ ，则可以获得8836个编码，若 $n=3$ ，则可以获得830584个编码。

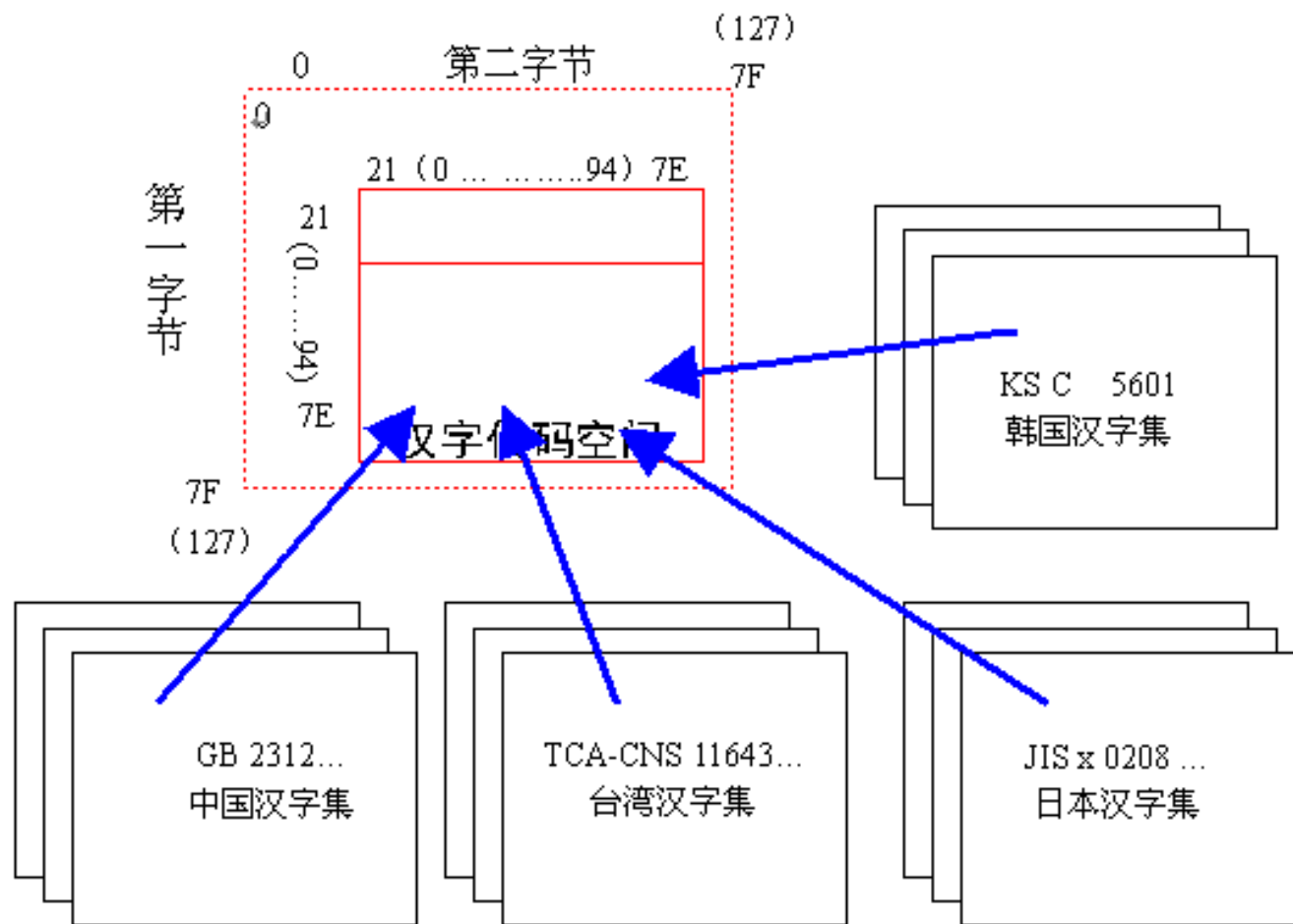


3.4 字符代码空间





3.5 汉字位置





3.5.1 体系结构特点

- ✓ 代码空间狭小
 - C0, C1回避不用
 - 没有利用80 H以上的空间
- ✓ 按国家/地区分别编码。
- ✓ 需要一整套复的控制功能来区分不同代码空间。



3.5.2 问题

✓ 字符集判别问题

学校简介



苏州大学坐落于素有“人间天堂”之称的古城苏州，是国家“211工程”重点建设高校、“2011计划”首批认定高校，是江苏省属重点综合性大学，其主要前身为创建于1900年的东吴大学。一个多世纪以来，一代代苏大人始终秉承“养天地正气，法古今完人”的校训精神；坚守“学术至上，学以致用，培养模范公民”的办学理念；传承和弘扬“自由开放，包容并蓄，追求卓越”的优良校风和“博学笃行，止于至



3.5.3 ISO 2022字符集

- ISO-2022-JP - 日语文字
 - ✓ ISO-2022-JP-1 - 加上一组转义字符串
 - ✓ ISO-2022-JP-2 ESC \$ (D 转为JIS X 0212-1990 多语言支援
 - ✓ ISO-2022-JP-3 - 加上两组转义字符串
 - ✓ ISO-2022-JP-2004 - 加上一组转义字符串
- ISO-2022-KR - 朝鲜文
- ISO-2022-CN - 中文
 - ✓ ISO-2022-CN-EXT - 加上六组转义字符串



内容

- ✓ ASCII码及其扩展
- ✓ 中文信息在计算机内的表示
- ✓ ISO/IEC 2022
- ✓ 汉字编码字符集
- ✓ GB2312-80
- ✓ BIG-5
- ✓ Unicode和ISO10646
- ✓ GBK和GB18030



4. 汉字编码字符集

- ✓ 按照一组无歧义的规则而定义的汉字字汇的有序集合。
 - 每一个汉字与它的代码表示之间具有一一对应关系
- ✓ 在信息处理技术中用于汉字信息的表示、交换、传输、处理、存储、输入及显示
- ✓ ISO定义中：
 - “无歧义的规则”很重要，确保编码的唯一性，避免重码



4.1 常用汉字编码字符集

- ✓ GB2312-80
- ✓ BIG-5
- ✓ ISO10646/Unicode
- ✓ GB13000
- ✓ GBK
- ✓ GB18030-2000



4.2 代码页

✓ 代码页

- 可用于信息
- 支持多文

✓ IBM称呼电 的名称

- EBCDIC

✓ Microsoft在

- 每个具体的
页ID”

874 (泰语)

932 (日语Shift-JIS)

936 (简体中文GBK)

949 (韩文)

950 (繁体中文Big5)

1258 (越南语)

支持的字符集编码

ws使用代码页

个代号，称为“代码



内容

- ✓ ASCII码及其扩展
- ✓ 中文信息在计算机内的表示
- ✓ ISO/IEC 2022
- ✓ 汉字编码字符集
- ✓ **GB2312-80**
- ✓ BIG-5
- ✓ Unicode和ISO10646
- ✓ GBK和GB18030



5. GB2312-80

- ✓ 信息交换用汉字编码字符集（基本集）
- ✓ 双字节内码
- ✓ 每个字节使用低7位
 - “0000, 0001” --- “0101, 1110”
 - 1-0x5E (1-94)
- ✓ 内码的空间: $94 \times 94 = 8836$
- ✓ 收录汉字6763个, 符号682个
- ✓ 简体字符集



5.1 国标码和区位码

- ✓ 高位字节（1-94）：94个区
- ✓ 低位字节（1-94）：94个位
- ✓ 国标码：16进制
- ✓ 区位码：10进制
- ✓ 如汉字“啊”，在第16区中的第1位，则
 - 国标码：1001（H）
 - 区位码：1601



5.2 符号区

✓ 1-9区，682个符号

- 一般符号（间隔、标点、运算、制表）202个
- 序号60个
- 数字22个
- 希腊字母48个
- 俄文字母66个
- 汉语拼音26个
- 拉丁字母52个
- 日文假名169个
- 汉语注音37个

✓ 2-9区有空位164个



5.2 符号区

国标第 01 区⁺

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|-------------------|----|---|-----|---|---|----|----|---|---|
| 0 | <u> </u> | 、 | 。 | . | - | ˘ | ” | ” | 々 | ↓ |
| 1 | — | ~ | | ... | ‘ | ’ | “ | ” | { | } |
| 2 | < | > | 《 | 》 | 「 | 」 | 『 | 』 | [|] |
| 3 | 【 | 】 | ± | × | ÷ | : | ∧ | ∨ | Σ | Π |
| 4 | ∪ | ∩ | ∈ | :: | √ | ⊥ | // | ∠ | ∩ | ⊙ |
| 5 | ∫ | ℳ | ≡ | ≡ | ≈ | ∞ | ≠ | ≠ | ≠ | ↓ |
| 6 | ≤ | ≥ | ∞ | ∴ | ∴ | ♂ | ♀ | ° | ' | ” |
| 7 | ℃ | \$ | ⊙ | ⊙ | £ | % | § | No | ☆ | ★ |
| 8 | ○ | ● | ◎ | ◇ | ◆ | □ | ■ | △ | ▲ | ※ |
| 9 | → | ← | ↑ | ↓ | = | | | | | |



5.2 符号区

国标第 02 区⁺

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|------|------|------|------|------|------|------|------|------|------|
| 0 | | i | ii | iii | iv | v | vi | vii | viii | ix |
| 1 | x | | | | | | | 1. | 2. | 3. |
| 2 | 4. | 5. | 6. | 7. | 8. | 9. | 10. | 11. | 12. | 13. |
| 3 | 14. | 15. | 16. | 17. | 18. | 19. | 20. | (1) | (2) | (3) |
| 4 | (4) | (5) | (6) | (7) | (8) | (9) | (10) | (11) | (12) | (13) |
| 5 | (14) | (15) | (16) | (17) | (18) | (19) | (20) | ① | ② | ③ |
| 6 | ④ | ⑤ | ⑥ | ⑦ | ⑧ | ⑨ | ⑩ | | | (一) |
| 7 | (二) | (三) | (四) | (五) | (六) | (七) | (八) | (九) | (十) | |
| 8 | | I | II | III | IV | V | VI | VII | VIII | IX |
| 9 | X | XI | XII | | | | | | | |



5.2 符号区

国标第 09 区⁺

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|------|------|---|---|---|---|-----|-----|
| 0 | | | | | — | — | | | --- | --- |
| 1 | ⋮ | ⋮ | ---- | ---- | ⋮ | ⋮ | ┐ | ┐ | ┐ | ┐ |
| 2 | ┐ | ┐ | ┐ | ┐ | └ | └ | └ | └ | └ | └ |
| 3 | └ | └ | └ | └ | └ | └ | └ | └ | └ | └ |
| 4 | └ | └ | └ | └ | └ | └ | └ | └ | └ | └ |
| 5 | └ | └ | └ | └ | └ | └ | └ | └ | └ | └ |
| 6 | └ | └ | └ | └ | └ | └ | └ | └ | └ | └ |
| 7 | └ | └ | └ | └ | └ | └ | └ | └ | └ | └ |
| 8 | | | | | | | | | | |
| 9 | | | | | | | | | | |



5.3 汉字区

- ✓ 10-15区：空
- ✓ 88-94区：空
- ✓ 16-87区：6763个汉字
 - 16-55区：一级汉字3755个
 - 55区有5个空位，从89-94
 - 56-87区：二级汉字3008个
 - 一级汉字按照音、笔形排列
 - 二级汉字按照部首排列



5.3 汉字区

国标第 16 区⁺

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|----------|----------|----------|---|---|----------|---|----------|----------|----------|
| 0 | | <u>啊</u> | 阿 | 埃 | 挨 | 哎 | 唉 | 哀 | <u>皑</u> | 癌 |
| 1 | <u>蔼</u> | 矮 | 艾 | 碍 | 爱 | <u>隘</u> | 鞍 | 氨 | 安 | 俺 |
| 2 | 按 | 暗 | 岸 | 胺 | 案 | <u>肮</u> | 昂 | <u>盎</u> | <u>凹</u> | <u>敖</u> |
| 3 | 熬 | <u>翱</u> | 袄 | 傲 | 奥 | <u>懊</u> | 澳 | <u>芭</u> | 捌 | 扒 |
| 4 | 叭 | 吧 | <u>笆</u> | 八 | 疤 | 巴 | 拔 | 跋 | 靶 | 把 |
| 5 | 耙 | 坝 | 霸 | 罢 | 爸 | 白 | 柏 | 百 | 摆 | 佰 |
| 6 | 败 | 拜 | <u>裨</u> | 斑 | 班 | 搬 | 扳 | 般 | 颁 | 板 |
| 7 | 版 | 扮 | 拌 | 伴 | 瓣 | 半 | 办 | 绊 | <u>邦</u> | 帮 |
| 8 | 梆 | 榜 | <u>膀</u> | 绑 | 棒 | 磅 | 蚌 | 镑 | 傍 | <u>谤</u> |
| 9 | <u>苞</u> | 胞 | 包 | 褒 | 剥 | | | | | |



5.3 汉字区

国标第 86 区[↓]

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | | 觥 | 觥 | 觥 | 觥 | 觥 | 觥 | 霁 | 雳 | 雯 |
| 1 | 霆 | 霁 | 霁 | 霁 | 霁 | 霁 | 霁 | 霁 | 霁 | 霁 |
| 2 | 龋 | 龋 | 龋 | 龋 | 龋 | 龋 | 龋 | 龋 | 龋 | 龋 |
| 3 | 龋 | 佳 | 隼 | 隼 | 隼 | 隼 | 隼 | 隼 | 隼 | 隼 |
| 4 | 鎔 | 鎔 | 鎔 | 鎔 | 鎔 | 鎔 | 鎔 | 鎔 | 鎔 | 鎔 |
| 5 | 鲚 | 鲚 | 鲚 | 鲚 | 鲚 | 鲚 | 鲚 | 鲚 | 鲚 | 鲚 |
| 6 | 鲚 | 鲚 | 鲚 | 鲚 | 鲚 | 鲚 | 鲚 | 鲚 | 鲚 | 鲚 |
| 7 | 鲚 | 鲚 | 鲚 | 鲚 | 鲚 | 鲚 | 鲚 | 鲚 | 鲚 | 鲚 |
| 8 | 鲚 | 鲚 | 鲚 | 鲚 | 鲚 | 鲚 | 鲚 | 鲚 | 鲚 | 鲚 |
| 9 | 鳅 | 鳅 | 鳅 | 鳅 | 鳅 | 鳅 | 鳅 | 鳅 | 鳅 | 鳅 |



5.4 区位码

- ✓ 汉字交换码的另一种形式
- ✓ 在**GB2312**中，交换码方阵为**94×94**
- ✓ 区位码
 - 纵向定义为区号（取值范围为十进制数的**0-94**）
 - 横向定义为位号（取值范围为十进制数的**0-94**）
 - 两个坐标明确了一个汉字的位置
 - 区号和位号的编号：**1-94**
- ✓ 例如，在**GB2312-80**中
 - “码”字所在的区号为“**34**”，位号为“**75**”，故其区位码为“**3475**”



5.5 交换码/区位码/内码关系

- ✓ 存在着简单的转化关系
 - ✓ 假如：
 - 交换码为JH（J为高位，H为低位，为十六进数）
 - 区位码为QW（Q为区号，W为位号，为十进制数）
 - 处理码为CL（C为高位，L为低位，为十六进制数）
- 则：
- $J=Q+32$ --à 再转换为十六进制数
 - $H=W+32$ --à 再转换为十六进制数
 - $C=J+80H$
 - $L=H+80H$



5.6 转换例子

“心”

√ 区位码:

➢ 48 36 -> 30H 24H

√ 交换码: 50H 44H

➢ $30H + 20H = 50H$

➢ $24H + 20H = 44H$

√ 机内码: D0H C4H

➢ $50H + 80H = D0H$

➢ $44H + 80H = C4H$

国标第 48 区

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | | 小 | 孝 | 校 | 肖 | 啸 | 笑 | 效 | 楔 | 些 |
| 1 | 歇 | 蝎 | 鞋 | 协 | 挟 | 携 | 邪 | 斜 | 胁 | 谐 |
| 2 | 写 | 械 | 卸 | 蟹 | 懈 | 泄 | 泻 | 谢 | 屑 | 薪 |
| 3 | 芯 | 锌 | 欣 | 辛 | 新 | 忻 | 心 | 信 | 衅 | 星 |
| 4 | 腥 | 猩 | 惺 | 兴 | 刑 | 型 | 形 | 邢 | 行 | 醒 |
| 5 | 幸 | 杏 | 性 | 姓 | 兄 | 凶 | 胸 | 匈 | 汹 | 雄 |
| 6 | 熊 | 休 | 修 | 羞 | 朽 | 嗅 | 锈 | 秀 | 袖 | 绣 |
| 7 | 墟 | 戌 | 需 | 虚 | 嘘 | 须 | 徐 | 许 | 蓄 | 酗 |
| 8 | 叙 | 旭 | 序 | 畜 | 恤 | 絮 | 婿 | 绪 | 续 | 轩 |
| 9 | 喧 | 宣 | 悬 | 旋 | 玄 | | | | | |



内容

- ✓ ASCII码及其扩展
- ✓ 中文信息在计算机内的表示
- ✓ ISO/IEC 2022
- ✓ 汉字编码字符集
- ✓ GB2312-80
- ✓ **BIG-5**
- ✓ Unicode和ISO10646
- ✓ GBK和GB18030



6. BIG-5

- ✓ 繁体用汉字编码字符集
- ✓ 交换码和内码一致
- ✓ 台湾、香港、澳门等地使用
- ✓ 取码范围：
 - 高位：0x81-0xfe 94
 - 低位：0x40-0x70, 0xa1-0xfe 157

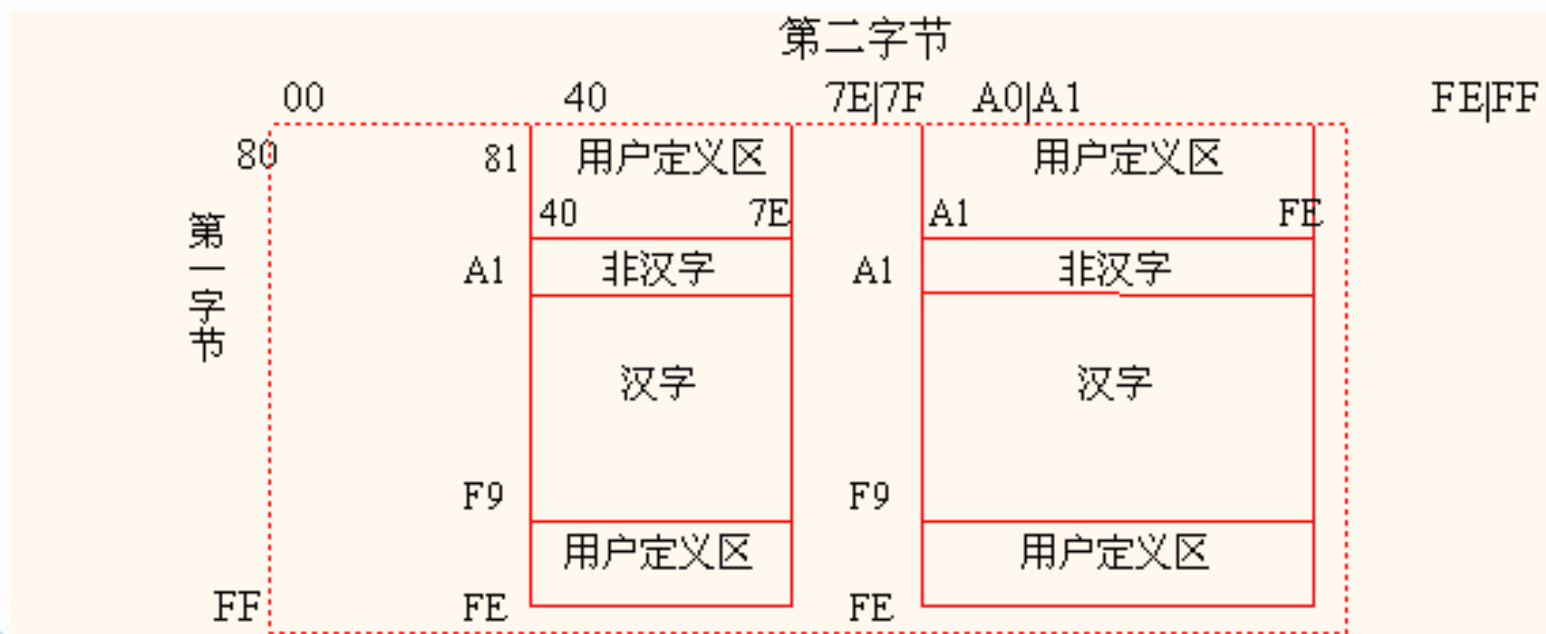


6.1 BIG-5布局

- ✓ 94*157的矩阵
 - 94区，每区157位
 - 最多14758个码位
- ✓ 收录了13494个字符
 - 13053个汉字
 - 441个非汉字图形字符



6.2 BIG-5 代码空间图





6.3 代码分布举例

| C9 | 00 | 01 | 02 | 03 | 04 | 05 | 06 | 07 | 08 | 09 | 0A | 0B | 0C | 0D | 0E | 0F |
|----|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
| 40 | 乂 4E42 | 乚 4E5C | 凵 51F5 | 匸 531A | 厂 5382 | 万 4E07 | 丌 4E0C | 乇 4E47 | 亅 4E8D | 口 56D7 | 兀 FA0C | 中 5C6E | 彳 5F73 | 𠂇 4E0F | 有 5187 | 与 4E0E |
| 50 | 𠂇 4E2E | 𠂇 4E93 | 𠂇 4EC2 | 𠂇 4EC9 | 𠂇 4EC8 | 𠂇 5198 | 𠂇 52FC | 𠂇 536C | 𠂇 53B9 | 𠂇 5720 | 𠂇 5903 | 𠂇 592C | 𠂇 5C10 | 𠂇 5DFF | 𠂇 65E1 | 𠂇 68B3 |
| 60 | 𠂇 6BCC | 𠂇 6C14 | 𠂇 723F | 𠂇 4E31 | 𠂇 4E3C | 𠂇 4EE8 | 𠂇 4EDC | 𠂇 4EE9 | 𠂇 4EE1 | 𠂇 4EDD | 𠂇 4EDA | 𠂇 520C | 𠂇 531C | 𠂇 534C | 𠂇 5722 | 𠂇 5723 |
| 70 | 𠂇 5917 | 𠂇 592F | 𠂇 5B81 | 𠂇 5B84 | 𠂇 5C12 | 𠂇 5C3B | 𠂇 5C74 | 𠂇 5C73 | 𠂇 5E04 | 𠂇 5E80 | 𠂇 5E82 | 𠂇 5FC9 | 𠂇 6209 | 𠂇 6250 | 𠂇 6C15 | |
| 80 | | | | | | | | | | | | | | | | |
| 90 | | | | | | | | | | | | | | | | |
| A0 | | 𠂇 6C36 | 𠂇 6C43 | 𠂇 6C3F | 𠂇 6C3B | 𠂇 72AE | 𠂇 72B0 | 𠂇 738A | 𠂇 79B8 | 𠂇 808A | 𠂇 961E | 𠂇 4F0E | 𠂇 4F18 | 𠂇 4F2C | 𠂇 4EF5 | 𠂇 4F14 |
| B0 | 𠂇 4EF1 | 𠂇 4F00 | 𠂇 4EF7 | 𠂇 4F08 | 𠂇 4F1D | 𠂇 4F02 | 𠂇 4F05 | 𠂇 4F22 | 𠂇 4F13 | 𠂇 4F04 | 𠂇 4EF4 | 𠂇 4F12 | 𠂇 51B1 | 𠂇 5213 | 𠂇 5209 | 𠂇 5210 |
| C0 | 𠂇 52A6 | 𠂇 5322 | 𠂇 531F | 𠂇 534D | 𠂇 538A | 𠂇 5407 | 𠂇 56E1 | 𠂇 56DF | 𠂇 572E | 𠂇 572A | 𠂇 5734 | 𠂇 593C | 𠂇 5980 | 𠂇 597C | 𠂇 5985 | 𠂇 597B |
| D0 | 𠂇 597E | 𠂇 5977 | 𠂇 597F | 𠂇 5B56 | 𠂇 5C15 | 𠂇 5C25 | 𠂇 5C7C | 𠂇 5C7A | 𠂇 5C7B | 𠂇 5C7E | 𠂇 5DDF | 𠂇 5E75 | 𠂇 5E84 | 𠂇 5F02 | 𠂇 5F1A | 𠂇 5F74 |
| E0 | 𠂇 5FD5 | 𠂇 5FD4 | 𠂇 5FCF | 𠂇 625C | 𠂇 625E | 𠂇 6264 | 𠂇 6261 | 𠂇 6266 | 𠂇 6262 | 𠂇 6259 | 𠂇 6260 | 𠂇 625A | 𠂇 6265 | 𠂇 65EF | 𠂇 65EE | 𠂇 673E |
| F0 | 𠂇 6739 | 𠂇 6738 | 𠂇 673B | 𠂇 673A | 𠂇 673F | 𠂇 673C | 𠂇 6733 | 𠂇 6C18 | 𠂇 6C46 | 𠂇 6C52 | 𠂇 6C5C | 𠂇 6C4F | 𠂇 6C4A | 𠂇 6C54 | 𠂇 6C4B | |



6.4 两岸文字的不一致性

- √ 苏 州 大 学
- √ CB D5 D6 DD B4 F3 D1 A7 苏州大学 GB2312
- √ CC 4B D6 DD B4 F3 8C 57 蘇州大學 GBK
- √ C4 AC A6 7B A4 6A BE C7 默 既 BIG-5



内容

- ✓ ASCII码及其扩展
- ✓ 中文信息在计算机内的表示
- ✓ ISO/IEC 2022
- ✓ 汉字编码字符集
- ✓ GB2312-80
- ✓ BIG-5
- ✓ **Unicode和ISO10646**
- ✓ GBK和GB18030



7. Unicode和 ISO10646

√ 本地化编码问题



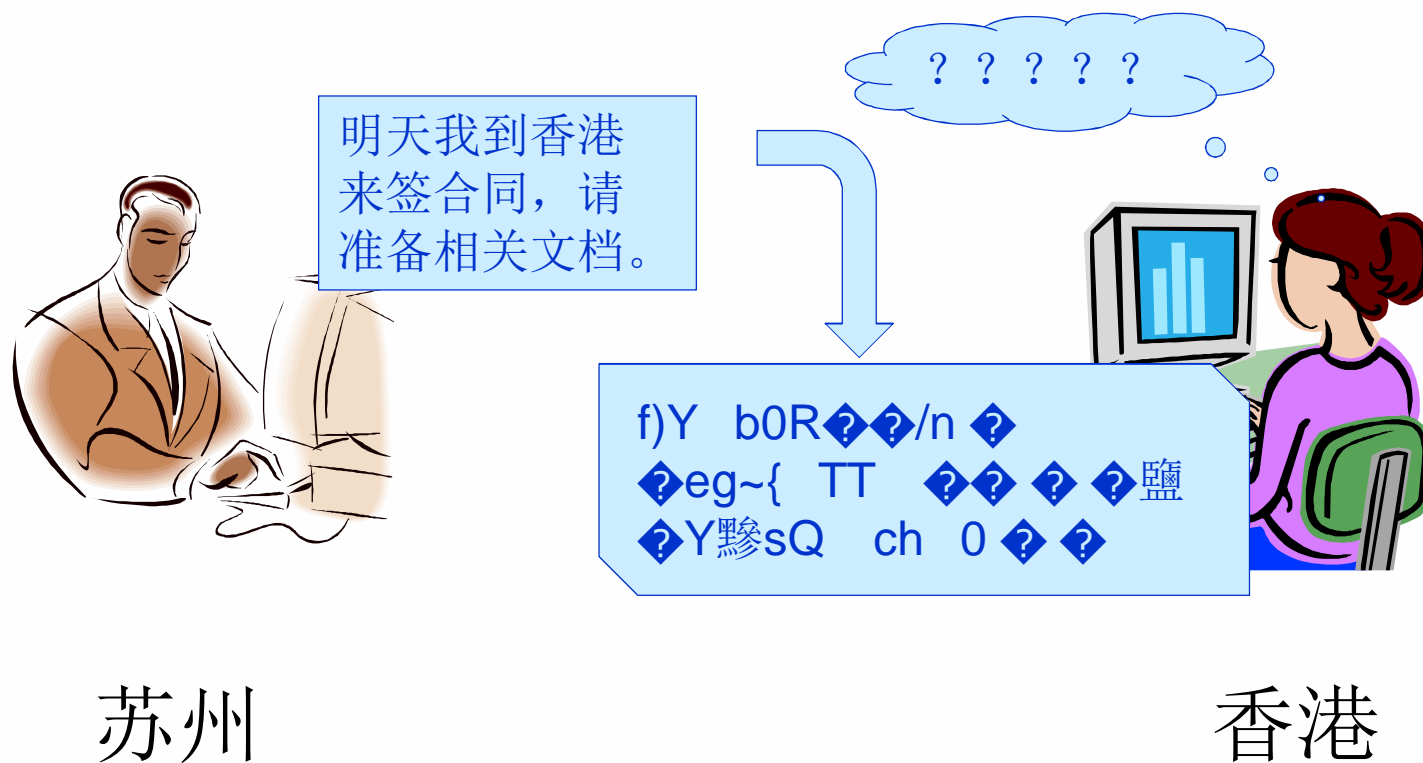
7.1.1 本地化编码的问题

√ 现状

- 世界各国采用了不同的编码标准。
- 例如:香港及台湾使用繁体字, 通常采用「BIG-5」
 - √ 中国内地使用简体字, 通常采用「GB2312」
- 各种不同的编码标准互不兼容。
- 一个编码在不同的编码标准内可能代表不同的字符。



7.1.2 例子





7.1.3 统一文字编码

- ✓ 提供一套统一的字符编码标准
 - 包含世界上所有文字
 - 使通讯及资料交换不需转码
 - 在一个电脑上处理多种语言文本
- ✓ 采用该标准后
 - 不同的电脑系统之间能更准确地储存、处理、传递及显示各种文字信息
 - 加强各地间文字信息的流通
 - 推动电子交易



7.2 ISO 10646

- ✓ 1984年发起制定新的编码字符集国际标准
- ✓ WG2负责，命名为UCS（Universal Character Set）
- ✓ 字符码长为4个八位的字节(Octet)
- ✓ 编码仍坚持遵循ISO 2022
- ✓ 字符编码区必须要避开C0和C1控制区
- ✓ 编号为ISO 10646



7.3 Unicode

- ✓ 一些著名的IT公司认为:
 - ISO 2022避开C0、C1区，降低编码效率
 - **主张采用统一、连续编码**
- ✓ 1988年初，施乐Joe Becker倡议以新编码标准:
 - 字符集编码的基本单位由7位或者8位扩充为16位
 - 充分利用65536个编码位置
 - 容纳全世界各种语言的字符和常用符号
 - 新标准被命名为Unicode
- ✓ 1991年1月，IBM、DEC、Sun、Microsoft、Xerox、Apple、Novell等成立Unicode技术委员会



7.3.1 Unicode的含义

- ✓ Unicode委员会负责Unicode字元搜集、整理、编码等
- ✓ Unicode的含义和目标是“3Uni”:
 - Unique(唯一)
 - Unified(统一)
 - Universal(通用)
- ✓ 所有文字均采用16位代码
- ✓ 任何代码没有二义性



7.3.2 ISO 10646和Unicode

- √ 由于
 - Unicode技术委员会成员的实力和影响力
 - Unicode方案的科学性
 - Unicode技术委员会对WG2持续的游说和施压
- √ WG2改用Unicode的编码方式：
 - 所有字符的码长均等同
 - 进行连续编码
 - 不再避开C0和C1区
- √ WG2在1991年10月达成了协议
 - 将Unicode并入ISO10646，成为ISO 10646的第0字面



7.3.4 UCS-4

- ✓ ISO10646的正规形式为32位
 - 4个八位字节，称为UCS-4
 - ✓ 组 (Group) : 128组 (组号为00~7Fh)
 - ✓ 面 (Plane) : 256面 (面号为00~FFh)
 - ✓ 行 (Row) : 256行 (行号为00~FFh)
 - ✓ 位 (Cell) : 256位 (位号为00~FFh)
 - 编码的Bit31 (即首字节最高位) 必须为0



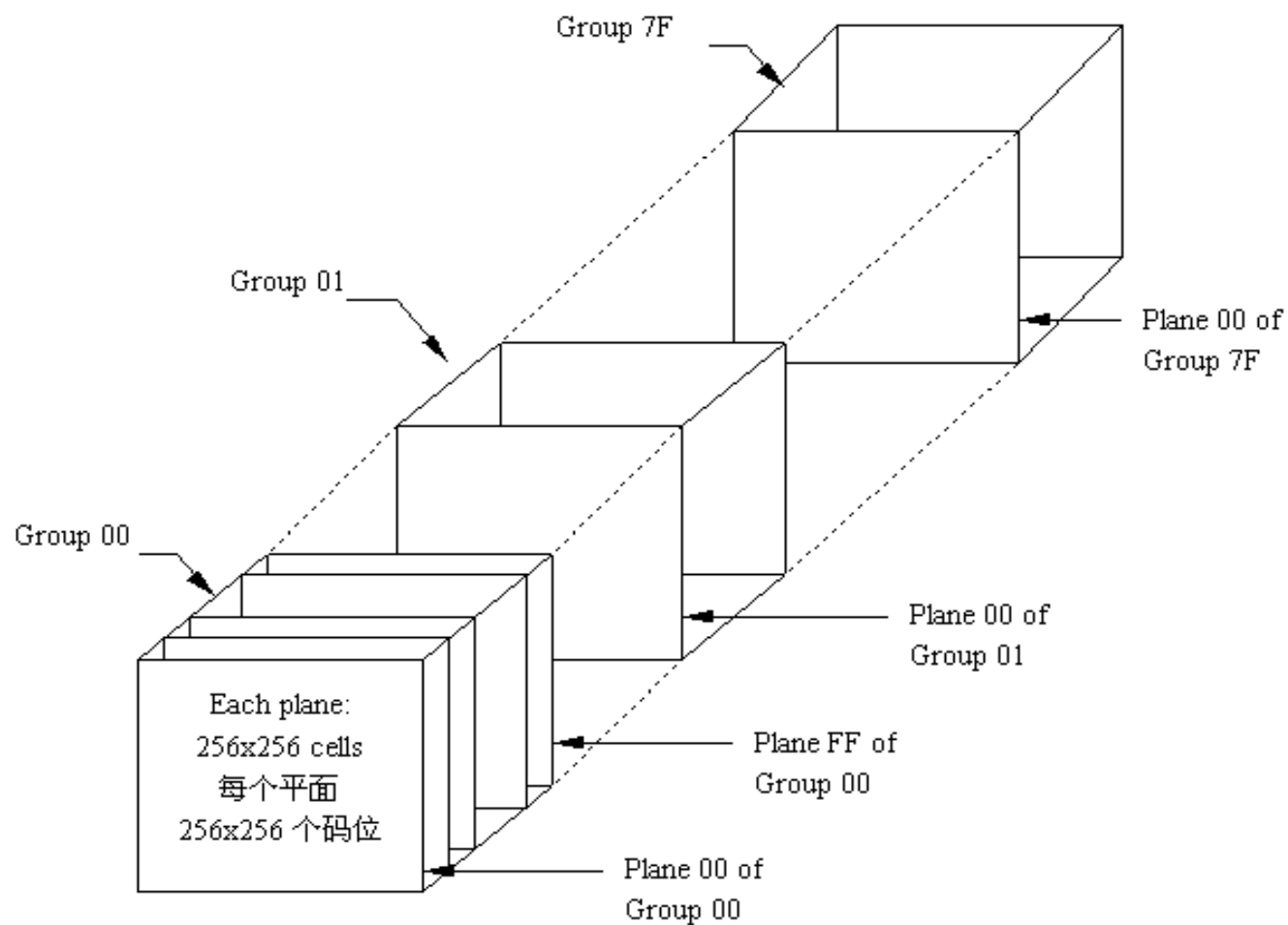
7.3.5 ISO 10646的编码空间

√ ISO10646编码空间总共为:

- $256 \times 128 = 32768$ 个字面
- 每个字面为 $256 \times 256 - 2 = 65534$ 个编码位置
- 合计 $65534 \times 32768 = 2147418112$ 个编码位置
- ISO10646规定，每个字面的最后两个编码位置FFFEh和FFFFh保留不用



7.3.6 编码结构示意图





7.3.7 基本多文种字面

BMP Chart

| | | | | | |
|----|---|-------------------|------------------|-----------------------------|---------|
| 00 | A-Zone 拼音文字 | | | | |
| 33 | See Detailed Chart | | | | |
| 34 | | | | | |
| 4D | CJK Unified Ideographs Extension A (新) | | | | |
| 4E | 中日韩统一汉字扩充集 A | | | | |
| : | CJK Unified Ideographs | | | | |
| : | | | | | |
| 9F | 中日韩统一汉字 | | | | |
| A0 | Yi Syllables (A000-A48F) & Yi Radicals (A490-A4BF) 彝文音节字和字根 (新) | | | | |
| A5 | | | | | |
| AB | O-Zone | | | | |
| AC | | | | | |
| : | Hangul Extended 韩文 (新) | | | | |
| D8 | | | | | |
| DF | S-Zone (For Use in UTF-16 only) (新) | | | | |
| E0 | | | | | |
| F8 | Private Use Area | | | | |
| F9 | | | | | |
| FA | CJK Compatibility Ideographs | | | | |
| FB | | | | | |
| FC | Alphabetic Presentation Forms | | | | |
| FD | Arabic Presentation Forms-A | | | | |
| FE | | | | | |
| FF | Comb. Half M'ks | CJK Compat. Forms | Small Form Vars. | Arabic Presentation Forms-B | |
| | Halfwidth And Fullwidth Forms | | | | Special |

基本多文种字面 (Basic Multi-lingual Plane, BMP) :

- ISO10646的第0组第0字面（组和面的值都为00h）
- 编码字元与Unicode相同。

UCS-2: 只用BMP, 每个字符只用16位编码



BMP(Unicode)编码

✓ 0000~007Fh: 基本拉丁平面 BMP Chart

- 0000~001Fh为C0控制码
- 0020h为空格 (space)
- 0021~007Eh为ASCII码
- 007Fh为控制码DEL
- 把前8位去掉即8位无符号整数

✓ 0080~00A0h: 控制码

- 0080~009Fh为C1控制码
- 00A0h为不间断空格

00
33
34
4D
4E
:
:
9F
A0
A5
AB
AC
:
D8
DF
E0
F8
F9
FA
FB
FC
FD
FE
FF

| | | | | | |
|---|-----------------|-------------------|------------------|-----------------------------|---------|
| A-Zone 拼音文字 See Detailed Chart | | | | | |
| CJK Unified Ideographs Extension A (新) 中日韩统一汉字扩充集 A | | | | | |
| CJK Unified Ideographs 中日韩统一汉字 | | | | | |
| Yi Syllables (A000-A48F) & Yi Radicals (A490-A4BF) 彝文音节字和字根 (新) | | | | | |
| O-Zone | | | | | |
| Hangul Extended 韩文 (新) | | | | | |
| S-Zone (For Use in UTF-16 only) (新) | | | | | |
| Private Use Area | | | | | |
| CJK Compatibility Ideographs | | | | | |
| Alphabetic Presentation Forms | | | | | |
| Arabic Presentation Forms-A | | | | | |
| | Comb. Half M'ks | CJK Compat. F' ms | Small Form Vars. | Arabic Presentation Forms-B | |
| Halfwidth And Fullwidth Forms | | | | | Special |



拼音文字区

✓ 00A1~1FFFFh: 拼音文字区

- 除基本拉丁字母以外的各种拼音文字
- 欧洲各国语言
- 希腊文
- 斯拉夫语文
- 希伯来文
- 阿拉伯文
- 亚美尼亚文
- 马来文
- 等

| | 010 | 011 | 012 | 013 | 014 | 015 | 016 | 017 |
|---|------|------|------|------|------|------|------|------|
| 0 | Ā | Đ | Ġ | İ | ı | Ŏ | Š | Ů |
| | 0100 | 0110 | 0120 | 0130 | 0140 | 0150 | 0160 | 0170 |
| 1 | ā | đ | ġ | ı | Ł | ő | š | ů |
| | 0101 | 0111 | 0121 | 0131 | 0141 | 0151 | 0161 | 0171 |
| 2 | Ǻ | Ē | Ģ | IJ | ł | Œ | Ŧ | Ū |
| | 0102 | 0112 | 0122 | 0132 | 0142 | 0152 | 0162 | 0172 |
| 3 | ǻ | ē | ġ | ij | Ń | œ | ţ | ų |
| | 0103 | 0113 | 0123 | 0133 | 0143 | 0153 | 0163 | 0173 |
| 4 | Ą | Ĕ | Ĥ | Ĵ | ń | Ŗ | Ţ | Ŵ |
| | 0104 | 0114 | 0124 | 0134 | 0144 | 0154 | 0164 | 0174 |
| 5 | ą | ĕ | ĥ | ĵ | Ņ | ŗ | ţ | ŵ |
| | 0105 | 0115 | 0125 | 0135 | 0145 | 0155 | 0165 | 0175 |
| 6 | Ć | Ė | Ħ | Ķ | ņ | Ŗ | Ŧ | Ŷ |
| | 0106 | 0116 | 0126 | 0136 | 0146 | 0156 | 0166 | 0176 |
| 7 | ć | ė | ħ | ķ | ņ | ŗ | ţ | ŷ |
| | 0107 | 0117 | 0127 | 0137 | 0147 | 0157 | 0167 | 0177 |
| 8 | Ĉ | Ė | Ĩ | κ | ñ | Ŗ | Ū | Ÿ |
| | 0108 | 0118 | 0128 | 0138 | 0148 | 0158 | 0168 | 0178 |



符号区

✓ 2000~28FFh: 符号区

- 标点符号
- 上下标
- 钱币符号
- 数字
- 箭头
- 数学符号
- 工程符号
- 光学辨识符号
-

| | 210 | 211 | 212 | 213 | 214 |
|---|------------|-----------|-------------|-----------|-----------|
| 0 | ‰ 2100 | ℑ 2110 | SM 2120 | ℰ 2130 | Σ 2140 |
| 1 | ‱ 2101 | ℐ 2111 | TEL 2121 | ℱ 2131 | Ϸ 2141 |
| 2 | ℄ 2102 | ℒ 2112 | TM 2122 | ℋ 2132 | ℓ 2142 |
| 3 | °C 2103 | ℓ 2113 | ℥ 2123 | ℳ 2133 | ℓ 2143 |
| 4 | ℥ 2104 | ℥ 2114 | ℤ 2124 | ℴ 2134 | ℵ 2144 |
| 5 | ‰ 2105 | ℕ 2115 | ℤ 2125 | ℵ 2135 | ℶ 2145 |
| 6 | ‰ 2106 | ℕ 2116 | Ω 2126 | ℷ 2136 | ℸ 2146 |
| 7 | ℰ 2107 | ℙ 2117 | ℴ 2127 | ℵ 2137 | ℶ 2147 |
| 8 | ℷ 2108 | ℸ 2118 | ℹ 2128 | ℺ 2138 | ℻ 2148 |



中日韩符号区

✓ 2E80~33FFh: 中日韩符号区

- 康熙字典部首
- 中日韩辅助部首
- 注音符号
- 日本假名和日本的假名
- 韩文音符
- 中日韩的符号
- 标点
- 带圈或带括符文数字、月份、日期、时间等

BMP Chart

| | | | | |
|----|---|-------------------|------------------|-----------------------------|
| 00 | A-Zone 拼音文字 | | | |
| 33 | See Detailed Chart | | | |
| 34 | | | | |
| 4D | CJK Unified Ideographs Extension A (新) | | | |
| 4E | 中日韩统一汉字扩充集 A | | | |
| : | CJK Unified Ideographs | | | |
| : | | | | |
| 9F | | | | |
| A0 | Yi Syllables (A000-A48F) & Yi Radicals (A490-A4BF) 彝文音节字和字根 (新) | | | |
| A5 | | | | |
| AB | O-Zone | | | |
| AC | | | | |
| : | Hangul Extended 韩文 (新) | | | |
| D8 | | | | |
| DF | S-Zone (For Use in UTF-16 only) (新) | | | |
| E0 | | | | |
| F8 | | | | |
| F9 | Private Use Area | | | |
| FA | | | | |
| FB | CJK Compatibility Ideographs | | | |
| FC | Alphabetic Presentation Forms | | | |
| FD | Arabic Presentation Forms-A | | | |
| FE | | | | |
| FF | | | | |
| | Comb. Half M'ks | CJK Compat. F' ms | Small Form Vars. | Arabic Presentation Forms-B |
| | Halfwidth And Fullwidth Forms | | | Special |



中日韩符号区

| 2E8 2E9 2EA 2EB 2EC 2ED 2FF | | | | | | 31C 31D 31E | | |
|-----------------------------|-----|-----|-----|-----|-----|-------------|---|----------------|
| 0 | 313 | 314 | 315 | 316 | 317 | 318 | 0 | 31C0 31D0 31E0 |
| 1 | | ㄹᄇᆞ | ㄱᄇᆞ | ㅍᄇᆞ | ㅊᄇᆞ | ㅇᄇᆞ | 1 | ㄱᄇᆞ 31D1 31E1 |
| 2 | ㄱᄇᆞ | ㅍᄇᆞ | ㅊᄇᆞ | ㅊᄇᆞ | ㅊᄇᆞ | ㅇᄇᆞ | 2 | ㄱᄇᆞ 31D2 31E2 |
| 3 | ㄱᄇᆞ | ㅍᄇᆞ | ㅊᄇᆞ | ㅊᄇᆞ | ㅊᄇᆞ | ㅇᄇᆞ | 3 | ㄱᄇᆞ 31D3 31E3 |
| 4 | ㄱᄇᆞ | ㅍᄇᆞ | ㅊᄇᆞ | ㅊᄇᆞ | ㅊᄇᆞ | ㅇᄇᆞ | 4 | ㄱᄇᆞ 31D4 |
| 5 | ㄱᄇᆞ | ㅍᄇᆞ | ㅊᄇᆞ | ㅊᄇᆞ | ㅊᄇᆞ | ㅇᄇᆞ | 5 | ㄱᄇᆞ 31D5 |
| 6 | ㄱᄇᆞ | ㅍᄇᆞ | ㅊᄇᆞ | ㅊᄇᆞ | ㅊᄇᆞ | ㅇᄇᆞ | 6 | ㄱᄇᆞ 31D6 |
| 7 | ㄱᄇᆞ | ㅍᄇᆞ | ㅊᄇᆞ | ㅊᄇᆞ | ㅊᄇᆞ | ㅇᄇᆞ | 7 | ㄱᄇᆞ 31D7 |



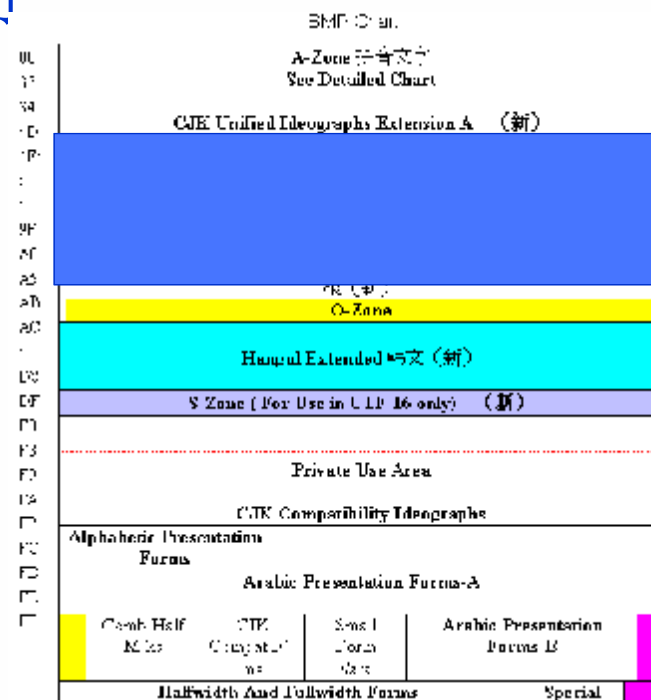
中日韩认同表意文字区

✓ 3400~4DFFh

- 中日韩认同表意文字扩充A区
- 总计收容6,582个中日韩汉字

✓ 4E00~9FFFh

- 中日韩认同表意文字区
- 收容20,902个中韩汉字





中日韩认同表意文字区

| HFX | | | | | HFX | | | | |
|---------------------|------------------|------------------|------------------|---------------------|--------------------|--------------------|--------------------|--------------------|--------------------|
| C | | | | | C | | | | |
| J | | | | | J | | | | |
| K | | | | | K | | | | |
| V | | | | | V | | | | |
| 3400 一 U+4E00 | 北 丙 U+300C | 北 丙 U+300C | 北 丙 U+300C | 3414 一 U+4E00 | 且 且 U+4E00 | 且 且 U+4E00 | 且 且 U+4E00 | 且 且 U+4E00 | 且 且 U+4E00 |
| 3401 一 U+4E01 | 北 丙 U+300C | 北 丙 U+300C | 北 丙 U+300C | 3415 一 U+4E01 | 且不 且不 U+4E01 | 且不 且不 U+4E01 | 且不 且不 U+4E01 | 且不 且不 U+4E01 | 且不 且不 U+4E01 |
| 3402 一 U+4E02 | 北 丙 U+300C | 北 丙 U+300C | 北 丙 U+300C | 3416 一 U+4E02 | 且不 且不 U+4E02 | 且不 且不 U+4E02 | 且不 且不 U+4E02 | 且不 且不 U+4E02 | 且不 且不 U+4E02 |
| 3403 一 U+4E03 | 北 丙 U+300C | 北 丙 U+300C | 北 丙 U+300C | 3417 一 U+4E03 | 且不 且不 U+4E03 | 且不 且不 U+4E03 | 且不 且不 U+4E03 | 且不 且不 U+4E03 | 且不 且不 U+4E03 |
| 3404 一 U+4E04 | 北 丙 U+300C | 北 丙 U+300C | 北 丙 U+300C | 3418 一 U+4E04 | 且不 且不 U+4E04 | 且不 且不 U+4E04 | 且不 且不 U+4E04 | 且不 且不 U+4E04 | 且不 且不 U+4E04 |
| 3405 一 U+4E05 | 北 丙 U+300C | 北 丙 U+300C | 北 丙 U+300C | 3419 一 U+4E05 | 且不 且不 U+4E05 | 且不 且不 U+4E05 | 且不 且不 U+4E05 | 且不 且不 U+4E05 | 且不 且不 U+4E05 |
| 3406 一 U+4E06 | 北 丙 U+300C | 北 丙 U+300C | 北 丙 U+300C | 341A 一 U+4E06 | 且不 且不 U+4E06 | 且不 且不 U+4E06 | 且不 且不 U+4E06 | 且不 且不 U+4E06 | 且不 且不 U+4E06 |
| 3407 一 U+4E07 | 北 丙 U+300C | 北 丙 U+300C | 北 丙 U+300C | 341B 一 U+4E07 | 且不 且不 U+4E07 | 且不 且不 U+4E07 | 且不 且不 U+4E07 | 且不 且不 U+4E07 | 且不 且不 U+4E07 |
| 3408 一 U+4E08 | 北 丙 U+300C | 北 丙 U+300C | 北 丙 U+300C | 341C 一 U+4E08 | 且不 且不 U+4E08 | 且不 且不 U+4E08 | 且不 且不 U+4E08 | 且不 且不 U+4E08 | 且不 且不 U+4E08 |
| 3409 一 U+4E09 | 北 丙 U+300C | 北 丙 U+300C | 北 丙 U+300C | 341D 一 U+4E09 | 且不 且不 U+4E09 | 且不 且不 U+4E09 | 且不 且不 U+4E09 | 且不 且不 U+4E09 | 且不 且不 U+4E09 |
| 340A 一 U+4E0A | 北 丙 U+300C | 北 丙 U+300C | 北 丙 U+300C | 341E 一 U+4E0A | 且不 且不 U+4E0A | 且不 且不 U+4E0A | 且不 且不 U+4E0A | 且不 且不 U+4E0A | 且不 且不 U+4E0A |
| 340B 一 U+4E0B | 北 丙 U+300C | 北 丙 U+300C | 北 丙 U+300C | 341F 一 U+4E0B | 且不 且不 U+4E0B | 且不 且不 U+4E0B | 且不 且不 U+4E0B | 且不 且不 U+4E0B | 且不 且不 U+4E0B |
| 340C 一 U+4E0C | 北 丙 U+300C | 北 丙 U+300C | 北 丙 U+300C | 341G 一 U+4E0C | 且不 且不 U+4E0C | 且不 且不 U+4E0C | 且不 且不 U+4E0C | 且不 且不 U+4E0C | 且不 且不 U+4E0C |
| 340D 一 U+4E0D | 北 丙 U+300C | 北 丙 U+300C | 北 丙 U+300C | 341H 一 U+4E0D | 且不 且不 U+4E0D | 且不 且不 U+4E0D | 且不 且不 U+4E0D | 且不 且不 U+4E0D | 且不 且不 U+4E0D |
| 340E 一 U+4E0E | 北 丙 U+300C | 北 丙 U+300C | 北 丙 U+300C | 341I 一 U+4E0E | 且不 且不 U+4E0E | 且不 且不 U+4E0E | 且不 且不 U+4E0E | 且不 且不 U+4E0E | 且不 且不 U+4E0E |
| 340F 一 U+4E0F | 北 丙 U+300C | 北 丙 U+300C | 北 丙 U+300C | 341J 一 U+4E0F | 且不 且不 U+4E0F | 且不 且不 U+4E0F | 且不 且不 U+4E0F | 且不 且不 U+4E0F | 且不 且不 U+4E0F |
| 3410 一 U+4E10 | 北 丙 U+300C | 北 丙 U+300C | 北 丙 U+300C | 341K 一 U+4E10 | 且不 且不 U+4E10 | 且不 且不 U+4E10 | 且不 且不 U+4E10 | 且不 且不 U+4E10 | 且不 且不 U+4E10 |
| 3411 一 U+4E11 | 北 丙 U+300C | 北 丙 U+300C | 北 丙 U+300C | 341L 一 U+4E11 | 且不 且不 U+4E11 | 且不 且不 U+4E11 | 且不 且不 U+4E11 | 且不 且不 U+4E11 | 且不 且不 U+4E11 |
| 3412 一 U+4E12 | 北 丙 U+300C | 北 丙 U+300C | 北 丙 U+300C | 341M 一 U+4E12 | 且不 且不 U+4E12 | 且不 且不 U+4E12 | 且不 且不 U+4E12 | 且不 且不 U+4E12 | 且不 且不 U+4E12 |
| 3413 一 U+4E13 | 北 丙 U+300C | 北 丙 U+300C | 北 丙 U+300C | 341N 一 U+4E13 | 且不 且不 U+4E13 | 且不 且不 U+4E13 | 且不 且不 U+4E13 | 且不 且不 U+4E13 | 且不 且不 U+4E13 |



其它区

- ✓ AC00～D7FFh: 韩文拼音组合字区
- ✓ D800～DFFFFh: S区（代理区），专门用於 UTF-16
- ✓ E000～F8FFh: 专用字区，保留供使用者自行添加
- ✓ F900～FAFFh: 中日韩相容表意文字区

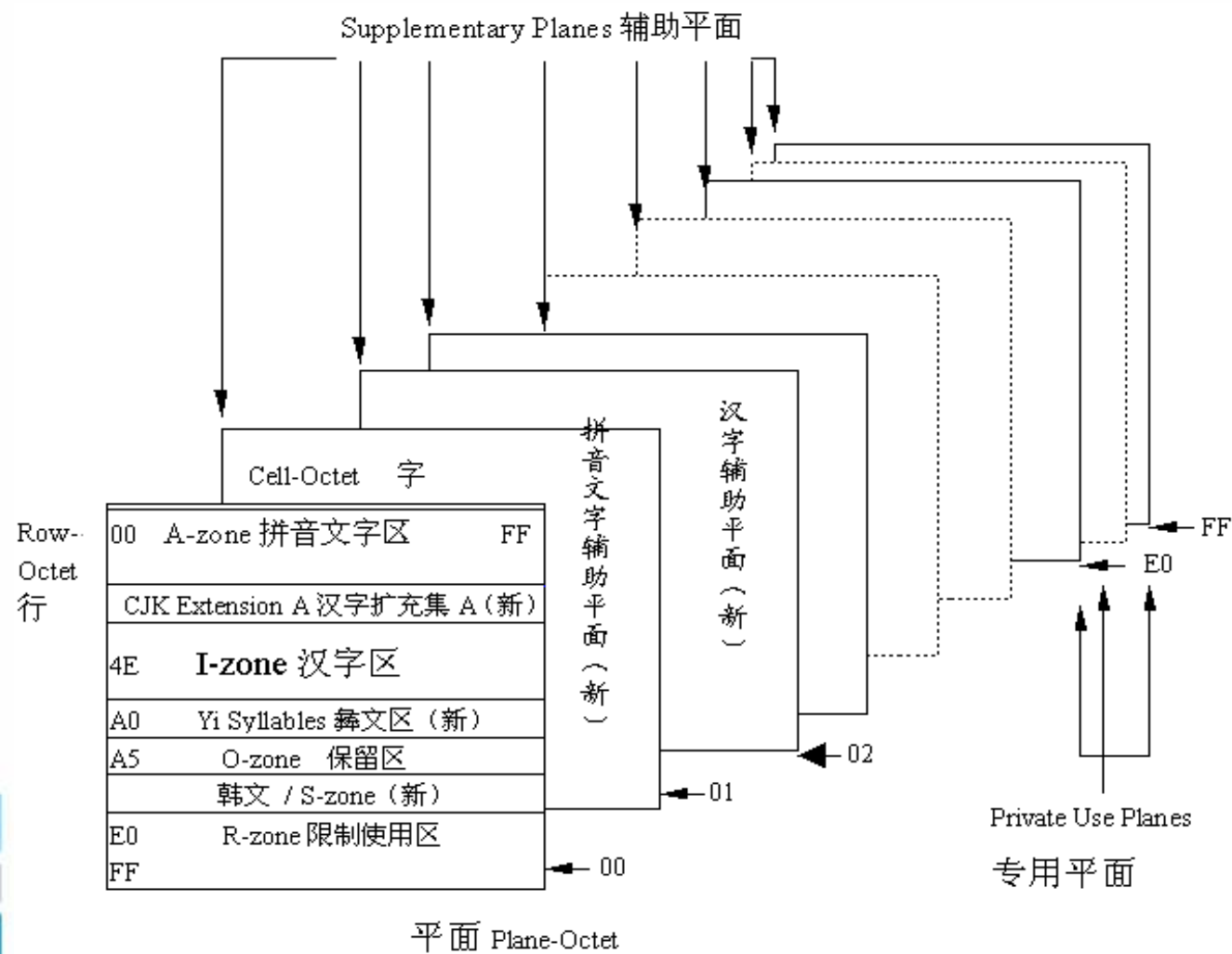


专用字面和辅助字面

- ✓ 除BMP，其余32767字面分为：
 - 专用字面(Private Use Planes)
 - ✓ 供用户自己定义ISO 10646中未收录的字符，共有8226个
 - ✓ 包括00h组的0Fh、10h、E0h—FFh号34个字面，以及60h—7Fh组内的全部字面
 - 辅助字面(Supplementary Planes)
 - ✓ 供WG2陆续定义各国文字字符
 - ✓ 24541个字面



00h组字面示意图





使用字面

- ✓ 除BMP，WG2仅使用：
 - 00h组中的01h和02h号字面
 - 第01h号字面：定义BMP内未收集的各国非表意文字和符号
 - 第02h号字面：定义BMP内未收集的各国表意文字和符号，如：
 - ✓ CJK表意文字扩充B区，共计42807个汉字
 - ✓ CNS11643兼容字符区，共计527个字符
 - Unicode明确提出，只会使用00h组前17个字面（即00h—10h号字面）



版本

✓ ISO 10646:1993

- 即：Unicode 2.0、GB13000.1
- 收录20902个汉字

✓ ISO 10646:2000

- 即：Unicode 3.0、GB13000.2
- 收录27484个汉字

✓ ISO 10646:2003

- 即：Unicode 4.0、GB13000.3
- 收录70198个汉字，加符号共96243个



版本(续)

- ✓ ISO/IEC 10646:2003
 - plus Amendment 1,2,3
 - Unicode 5.0
 - 71226汉字，加符号共98884
- ✓ ISO/IEC 10646:2011
 - Unicode 6.0
 - 75616汉字，加符号共109242
- ✓ Unicode 6.2 （最新）
 - 2012.4
 - 75619汉字，加符号共109974



内容

- ✓ ASCII码及其扩展
- ✓ 中文信息在计算机内的表示
- ✓ ISO/IEC 2022
- ✓ 汉字编码字符集
- ✓ GB2312-80
- ✓ BIG-5
- ✓ Unicode和ISO10646
- ✓ **GBK和GB18030**



8.1 GB 13000

√ 中国

- 1993年： GB13000.1-1993 (信息技术通用多
八位编码字符集 (UCS))
- 和ISO10646:1993(Unicode 2.0)在字符集上基
本一致
- 最初共收录了20902个汉字，以后将跟随
ISO10646的增补，同步进行增补。



8.2 GBK

- ✓ 2字节汉字编码
- ✓ 在内码上兼容GB2312-80
- ✓ 在字汇上兼容GB13000/ISO10646
- ✓ 是GB2312向GB13000过渡的中间代码
- ✓ 收录21886个汉字和符号
- ✓ 从8140H-FEFEH，除了xx7F一条线
- ✓ 简繁一体

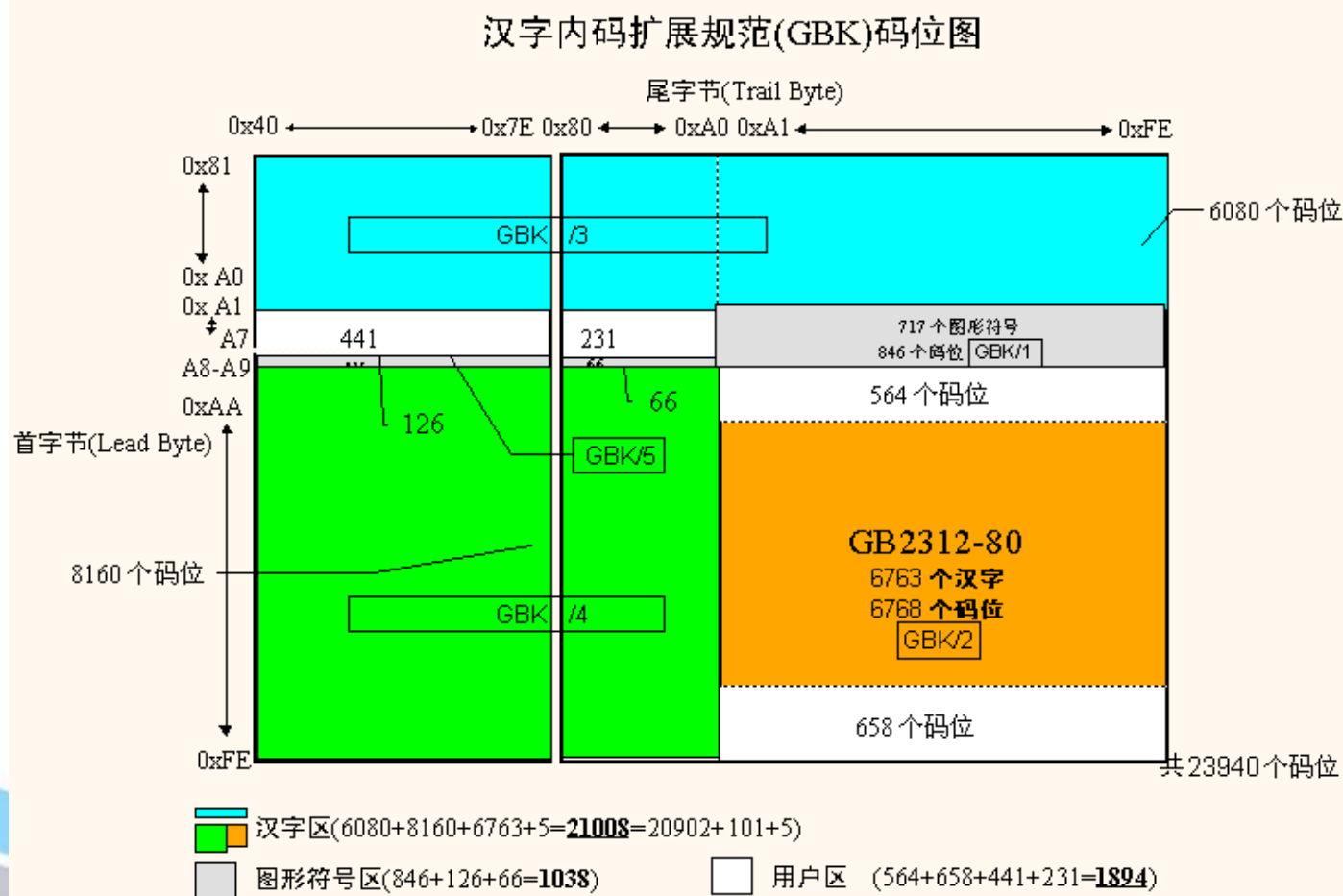


8.2.1 GBK码位分布图

| 类别 | 简称 | 范围 | 码位 | 字符 | 字符名 | 备注 |
|--------|-------|-----------|-------|-------|------|----------|
| 符号标准区 | GBK/1 | A1A1-A9FE | 846 | 717 | 图形符号 | GB2312为主 |
| | GBK/5 | A840-A9A0 | 192 | 166 | 图形符号 | BIG5结构符 |
| | 小计 | | 1038 | 883 | 图形符号 | |
| 汉字标准区 | GBK/2 | B0A1-F7FE | 6763 | 6763 | 汉字 | GB2312 |
| | GBK/3 | 8140-A0FE | 6080 | 6080 | 汉字 | GB13000 |
| | GBK/4 | AA40-FEA0 | 8160 | 8160 | 汉字 | GB13000等 |
| | 小计 | | 21008 | 21003 | 汉字 | |
| 用户自定义区 | 1区 | AAA1-AFFF | 564 | | | |
| | 2区 | F8A1-FEFE | 658 | | | |
| | 3区 | A140-A7A0 | 672 | | | 限制使用 |
| | 小计 | | 1894 | | | |
| 总计 | | | 23940 | 21886 | | |



8.2.2 GBK码位图





8.2.3 GBK字符

| | | | | | | | | | | | | | | | | | |
|---|----|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ✓ | 87 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | A | B | C | D | E | F |
| ✓ | 4 | 嘆 | 嗲 | 嗟 | 嗷 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 |
| ✓ | 5 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 |
| ✓ | 6 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 |
| ✓ | 7 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 |
| ✓ | 8 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 |
| ✓ | 9 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 |
| ✓ | A | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 |
| ✓ | B | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 |
| ✓ | C | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 |
| ✓ | D | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 |
| ✓ | E | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 |
| ✓ | F | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 | 嘍 |



8.2.4 21个兼容字

√ ISO定义字形和我国不符

√ 郎

√ 郎

| 汉字 | GBK编码 | Unicode编码 |
|----|-------|-----------|
| 郎 | FD9C | F92C |
| 凉 | FD9D | F979 |
| 季 | FD9E | F995 |
| 裏 | FD9F | F9E7 |
| 隣 | FDA0 | F9F1 |
| 兀 | FE40 | FA0C |
| 設 | FE41 | FA0D |
| 逵 | FE42 | FA0E |
| 埒 | FE43 | FA0F |
| 崎 | FE44 | FA11 |
| 栢 | FE45 | FA13 |
| 樺 | FE46 | FA14 |
| 礼 | FE47 | FA18 |
| 臈 | FE48 | FA1F |
| 藺 | FE49 | FA20 |
| 甦 | FE4A | FA21 |
| 赳 | FE4B | FA23 |
| 遐 | FE4C | FA24 |
| 鐸 | FE4D | FA27 |
| 鐸 | FE4E | FA28 |
| 隄 | FE4F | FA29 |

∴ F92C (GBK)

∴ 90DE



8.3 GB18030-2000

- ✓ 2000-3-17发布
- ✓ **2001年9月作为国家标准强制实施**
- ✓ 信息交换用汉字编码字符集基本集的扩充
 - 2000年ISO发布ISO 10646-1:2000 (Unicode 3.0)
 - 增加中日韩统一汉字Extension A的6,582个字符
- ✓ **GB18030-2000在 GBK 编码标准的基础扩充**
 - 增加四字节（32位）编码
 - 汉字后到达了**27533**个汉字
 - 总编码空间超过**150万个码位**



GB18030-2000

- ✓ GB18030标准采用
 - 单字节 (ACCII)
 - 双字节 (GBK)
 - 四字节 (Extension A的6,582个字符)
- ✓ 四字节的编码顺序为

| | | | | |
|-------|---------------|---|------|------------|
| 四字节部分 | 第一字节0x81-0x82 | 6 | 6530 | CJK统一汉字扩充A |
| | 第二字节0x30-0x39 | 5 | | |
| | 第三字节0x81-0xFE | 3 | | |
| | 第四字节0x30-0x39 | 0 | | |



8.3.1 GB18030-2000码位分布

| 字节数 | 码位空间 | | | | 码位数 |
|-----|-------------------|------|-------------------------|------|-------------|
| 单字节 | 0x00~0x7F | | | | 128 个码位 |
| 双字节 | 第一字节 0x81~0xfe | 第一字节 | 0x40~0x7e, 0x80~0xfe | | 23940 个码位 |
| 四字节 | 第一字节 | 第二字节 | 第三字节 | 第四字节 | 1587600 个码位 |



苏州大学：中文信息处理



8.3.4 GB18030-2005

√ GB18030-2005

- 信息技术中文编码字符集
- 收录了70244个汉字
- 包含多种我国少数民族文字（如藏、蒙古、傣、彝、朝鲜、维吾尔文等）的超大型中文编码字符集强制性标准

四字节部分

第一字节0x95-0x98
第二字节0x30-0x39
第三字节0x81-0xFE
第四字节0x30-0x39

42711

42711

CJK统一汉字扩充B



作业

√ P1-6