

信息编码与数据表示

数据

数值

编码方法

真值和编码之间的数学关系

原码

最高位为**符号位**，其他位为**数值位**

现代计算机中，原码主要用于表示浮点数的尾数

定点整数

若 $X = +X_1X_2 \dots X_n$ 则 $[X]_{\text{原}} = 0, X_1X_2 \dots X_n$

若 $X = -X_1X_2 \dots X_n$ 则 $[X]_{\text{原}} = 1, X_1X_2 \dots X_n$

定点小数

若 $X = +X_1X_2 \dots X_n$ 则 $[X]_{\text{原}} = 0.X_1X_2 \dots X_n$

若 $X = -X_1X_2 \dots X_n$ 则 $[X]_{\text{原}} = 1.X_1X_2 \dots X_n$

运算

符号位单独处理，参加运算的是绝对值。

硬件实现很困难

补码

最高位为**符号位**，其他位为**数值位**

符号位：0代表正数，1代表负数

数值位：

- 正数时与绝对值相同
- 负数时，为绝对值取反后，末位加一

定点整数

若 $X = +X_1X_2 \dots X_n$ 则 $[X]_{\text{补}} = 0, X_1X_2 \dots X_n$

若 $X = -X_1X_2 \dots X_n$ 则 $[X]_{\text{补}} = 1, \bar{X}_1\bar{X}_2 \dots \bar{X}_n + 1$

表示范围

$$-2^n \leq X \leq 2^n - 1$$

-2^n 的补码是 $1, 00 \dots 0$

定点小数

若 $X = +X_1X_2 \dots X_n$ 则 $[X]_{\text{补}} = 0.X_1X_2 \dots X_n$

若 $X = -X_1X_2 \dots X_n$ 则 $[X]_{\text{补}} = 1.\bar{X}_1\bar{X}_2 \dots \bar{X}_n + 0.00 \dots 1$

表示范围

$$-1 \leq X \leq 1 - 2^{-n}$$

-1 的补码是 $1.00 \dots 0$

移码

$$[X]_{\text{移}} = 2^n + X$$

最高位为符号位，其他位为数值位

符号位：**1代表正数，0代表负数**

数值位：

- 正数时与绝对值相同
- 负数时，为绝对值取反后，末位加一

为了方便计算，可以认为补码的符号位取反后就是移码

浮点数表示

浮点数 $N = M \times R^E$ ，阶码E，尾数M，阶码的底R

浮点数格式要素：

- 位数：阶码位数+尾数位数
- 编码：阶码和尾数采用的机器数编码
- 排列方式：数符是否在最高位

- 其他特殊的编码规则：譬如隐藏位

阶码E

阶码的位数：决定了浮点数的表示范围

阶码的编码：一般采用移码和补码

尾数M

尾数的位数：决定浮点数的精度

尾数的编码：一把采用原码和补码

阶码的底R：一般为2，8或16隐含规定

浮点数的规格化表示

充分利用尾数的二进制数位表示更多的有效数字

若 $R = 2$ ，绝对值最高有效为1。

$$\frac{1}{2} \leq |M| \leq 1$$

原码表示的尾数

- $M_1 = 1$ 规格化为x.1xx...x形式

补码表示的尾数

- $M_s \oplus M_1 = 1$ 规格化为0.1xxxx或1.0xxxx形式

IEEE754标准

$R=2$ ，格式有四种

- 单精度
- 双精度
- 扩展单精度
- 扩展双精度

数据分为**规格化数**，**非规格化数**和**特殊数值**

规格化数

优点：

- 提高了浮点数据的**精度**
- 使程序能够更**方便地交换**浮点数据
- 可以使浮点数的**运算更为简化**

阶码：是 $2^n - 1$ 的移码， $E = [E']_{\text{移}} = 2^n - +E'$

隐藏位：单，双精度浮点数，尾数规格化的"1"，放置于整数位，并将其隐藏；

临时实数无隐藏位

符号位：尾数的符号位 M_s 在最高位

非规格化数和规格化数的真值计算方法不同

无符号数和有符号数数据由指令来区分

单精度：

- 规格化数 $N = (-1)^{M_s} \times (1.M_1M_2 \dots M_n) \times 2^{E-127}$
- 非规格化数 $N = (-1)^{M_s} \times (0.M_1M_2 \dots M_n) \times 2^{E-126}$

特殊数值：

- 无穷大阶码全为1，尾数全为0
- NaN：阶码全为1，尾数不是全为0

非数值

字符编码

ASCII码

汉字：

- 键盘输入为**汉字输入码（外码）**
- 进入汉字输入法转化为**汉字内码**，和存储器进行交互
- **汉字内码**经过字形检索程序获得**汉字字形码**
- 外设通过**汉字字形码**进行显示

校验码

定义

一种具有发现某些错误或自动改正错误能力的一种数据编码方法

目的

用于检查或纠正存取，读写和传送数据的过程中可能出现的错误

构成

有效信息+校验位

原理

通过判断代码的合法性来检错

码距：一种码制的码距是指该码制中所有代码之间的最小距离

两个代码之间的距离：在任何两个代码之间逐位比较，对应位值不同的个数

只有当码距大于等于2时，校验码才有检错能力

只有当码距大于等于3时，校验码才有纠错能力

常见校验码：奇偶校验码，海明校验码，CRC校验码

二进制

用二进制表示数据

- 具有二值状态的物理器件容易实现
- 二进制数据的抗干扰性强，可靠性高
- 二进制的运算规则简单，硬件实现容易
- 具有逻辑特性，可代表“真假”

二进制码

BCD码

BCD码分为**有权码**和**无权码**

有权码

每一位有固定的权值

8421码等

无权码

没有固定的权值

格雷码，余三码等