



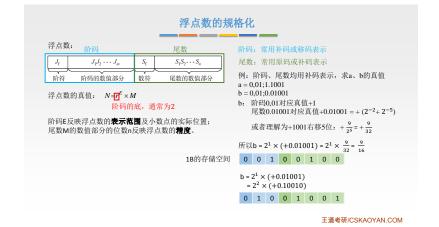
本节总览 基本格式 浮点数的表示

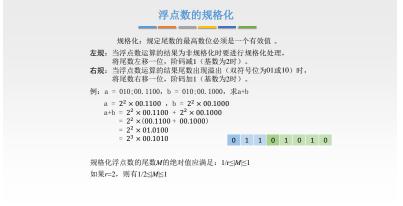
规格化
表示范围











王道考研/CSKAOYAN.COM

規格化浮点数的尾数M的绝对值应满足: 1/r≤M≤1
如果-2,则有1/2≤M≤1
1.原码规格化后:
正数为0.1××...×的形式,其最大值表示为0.11...1;最小值表示为0.10...0。
尾数的表示范围为1/2≤M≤(1-2**)。
负数为1.1××...×的形式,其最大值表示为1.10...0;最小值表示为1.11...1。
尾数的表示范围为1/2≤M≤(1-2**)。
2. 补码规格化后:
正数为0.1××...×的形式,其最大值表示为0.11...1;最小值表示为0.10...0。
尾数的表示范围为1/2≤M≤(1-2**)。
负数为1.0××...×的形式,其最大值表示为1.01...1;最小值表示为1.00...0。
尾数的表示范围为1-2≤M≤(1-2**)。

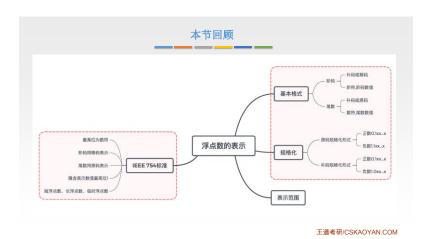
王道考研/CSKAOYAN.COM

王道考研/CSKAOYAN.COM

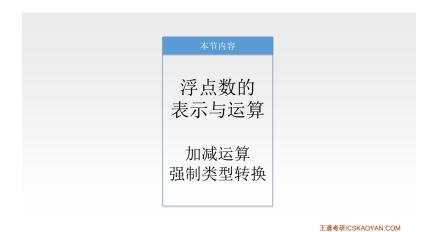








王道考研/CSKAOYAN.COM



浮点数的加减运算

浮点数加减运算步骤:

- 1. 对阶
- 2. 尾数加减
- 3. 规格化 4. 舍入
- 5. 判溢出

浮点数的加减运算

例: 己知十进制数X=-5/256、Y=+59/1024,按机器补码浮点运算规则计算X-Y,结果 用二进制表示,浮点数格式如下:阶符取2位,阶码取3位,数符取2位,尾数取9位

用补码表示阶码和尾数

扩展: 11.011000000 双符号位补码: 11.011 双符号位补码: 11011

0. 转换格式 本码: 1.011 5D = 1018, 1/256 = 2⁻⁸ → X = -101 × 2⁻⁸ = -0.101 × 2⁻⁵ = -0.101 × 2⁻¹⁰¹ 59D = 1110118, 1/1024 = 2⁻¹⁰ → Y = +111011 × 2⁻¹⁰ = +0.111011 × 2⁻⁴ = +0.111011 × 2⁻¹⁰⁰ X: 11011.11.011000000 Y: 11100.00.111011000

浮点数加减运算步骤:

- 1. 对阶
- 2. 尾数加减
- 3. 规格化
- 4. 舍入
- 5. 判溢出

王道考研/CSKAOYAN.COM

王道考研/CSKAOYAN.COM

浮点数的加减运算

例:已知十进制数X=-5/256、Y=+59/1024,按机器补码浮点运算规则计算X-Y,结果 用二进制表示,浮点数格式如下:阶符取2位,阶码取3位,数符取2位,尾数取9位

用补码表示阶码和尾数

0. 转换格式

5D = 101B, $1/256 = 2^{-8} \rightarrow X = -101 \times 2^{-8} = -0.101 \times 2^{-5} = -0.101 \times 2^{-101}$

59D = 111011B, $1/1024 = 2^{-10} \rightarrow Y = +111011 \times 2^{-10} = +0.111011 \times 2^{-4} = +0.111011 \times 2^{-100}$

X: 11011,11.011000000 Y: 11100,00.111011000

浮点数加减运算步骤:

1. 对阶 使两个数的阶码相等,小阶向大阶看齐,尾数每右移一位,阶码加1

① 求阶差: [ΔE]₃₄=11011+00100=11111, 知ΔE=-1

② 对阶: X: 11011,11.011000000 → 11100,11. 101100000 X = -0.0101 × 2⁻¹⁰⁰

- 2. 尾数加减
- 3. 规格化
- 4. 舍入
- 5. 判溢出

浮点数的加减运算

例:已知十进制数X=-5/256、Y=+59/1024,按机器补码浮点运算规则计算X-Y,结果 用二进制表示, 浮点数格式如下: 阶符取2位, 阶码取3位, 数符取2位, 尾数取9位

用补码表示阶码和尾数

0. 转换格式

5D = 101B, $1/256 = 2^{-8} \rightarrow X = -101 \times 2^{-8} = -0.101 \times 2^{-5} = -0.101 \times 2^{-101}$

59D = 111011B, $1/1024 = 2^{-10} \rightarrow Y = +111011 \times 2^{-10} = +0.111011 \times 2^{-4} = +0.111011 \times 2^{-100}$

X: 11011,11.011000000 Y: 11100,00.111011000

浮点数加减运算步骤:

1. 对阶 使两个数的阶码相等,小阶向大阶看齐,尾数每右移一位,阶码加1 ① 求阶差: [ΔE]₃₄=11011+00100=11111, 知ΔE=-1

② 对阶: X: 11011,11.011000000 → 11100,11. 101100000 X = -0.0101 × 2⁻¹⁰⁰

2. 尾数加减 -Y: 11100.11.000101000 11.101100000 X-Y

+ 11.000101000 = $(-0.0101 \times 2^{-100})$ - $(+0.111011 \times 2^{-100})$

X-Y: 11100, 10.110001000 3. 规格化

 $\begin{array}{ll} 10.110001000 & = (-0.0101 - 0.111011) \times 2^{-100} \\ & = -1.001111 \times 2^{-100} \end{array}$

4. 舍入

5. 判溢出

王道考研/CSKAOYAN.COM

王道考研/CSKAOYAN.COM

浮点数的加减运算

例:已知十进制数X=-5/256、Y=+59/1024,按机器补码浮点运算规则计算X-Y,结果 用二进制表示,浮点数格式如下:阶符取2位,阶码取3位,数符取2位,尾数取9位

用补码表示阶码和尾数

0. 转换格式

5D = 101B, $1/256 = 2^{-8} \rightarrow X = -101 \times 2^{-8} = -0.101 \times 2^{-5} = -0.101 \times 2^{-101}$

59D = 111011B, $1/1024 = 2^{-10} \rightarrow Y = +111011 \times 2^{-10} = +0.111011 \times 2^{-4} = +0.111011 \times 2^{-100}$ X: 11011.11.011000000 Y: 11100.00.111011000

浮点数加减运算步骤:

- 1. 对阶 使两个数的阶码相等,小阶向大阶看齐,尾数每右移一位,阶码加1
 - ① 求阶差: [ΔΕ] »=11011+00100=11111, 知ΔΕ=-1
 - ② 对阶: X: 11011,11.011000000 → 11100,11.101100000 X = -0.0101 × 2⁻¹⁰⁰
- 2. 尾数加减 -Y: 11100,11.000101000 X-Y: 11100, 10.110001000 + 11.000101000 $= (-0.0101 \times 2^{-100}) - (+0.111011 \times 2^{-100})$
 - 11.101100000 X-Y 10.110001000 = (-0.0101-0.111011) × 2⁻¹⁰⁰
- X-Y: 11100, 10.110001000 → 11101,11.011000100
- 4. 舍入 无舍入 5. 判溢出 常阶码, 无溢出, 结果真值为2⁻³×(-0.1001111)。

王道考研/CSKAOYAN.COM

= -1.001111 × 2⁻¹⁰⁰

 $= -0.1001111 \times 2^{-011}$

浮点数的加减运算-舍入

"0"舍"1"入法:类似于十进制数运算中的"四舍五入"法,即在尾 数右移时,被移去的最高数值位为0,则舍去;被移去的最高数值位为1, 则在尾数的末位加1。这样做可能会使尾数又溢出,此时需再做一次右规。

恒置"1"法: 尾数右移时, 不论丢掉的最高数值位是"1"还是"0", 都使右移后的尾数末位恒置"1"。这种方法同样有使尾数变大和变小的两 种可能。

浮点数加减运算步骤:

- 1. 对阶
- 2. 尾数加减 如: 加减结果为11100,10.110001011
- 3. 规格化 0舍1入: 11100,10.110001011 → 11101,11.011000101 1 → 11101.11.0110001**10** 1
 - 恒置1:11100,10.110001011 → 11101,11.011000101 1 → 11101.11.011000101 1
- 4. 舍入
- 5. 判溢出

王道考研/CSKAOYAN.COM

强制类型转换

类型	16位机器	32位机器	64位机器
char	8	8	8
short	16	16	16
int	16	32	32
long	32	32	64
long long	64	64	64
float	16	32	32
double	64	64	64

 $char \rightarrow int \rightarrow long \rightarrow double$

float → double

int → float: 可能损失精度 float → int: 可能溢出及损失精度

范围、精度从小到大,转换过程没有损失

int: 表示整数, 范围 -231~ 231-1, 有效数字32位

float:表示整数及小数,范围 $\pm[2^{-126}\sim 2^{127} imes(2-2^{-23})]$,有效数字23+1=24位

王道考研/CSKAOYAN.COM

