3.3 浮点运算

- 两个浮点数：$X=Mx\cdot 2^{Ex}$，$Y=My\cdot 2^{Ey}$
- X、Y为规格化浮点数

3.3.1 浮点加减运算

步骤：1. 对阶

2. 尾数加（减）运算

3. 规格化

4. 舍入处理

1. 对阶

- 就是小数点对齐，只有当两者的阶码相同时才能进行加减运算。
- 对阶的原则是小阶对大阶，即小阶码每增加1，其尾数右移一位，直到增大到与大阶码相同。

2. 尾数加（减）运算

- 对阶之后，尾数进行加（减）运算。
- 减法可以用加法实现。

3. 规格化

- 如果结果是非规格化数(尾数和（差）的绝对值可能小于1/2，也可能大于1)，则需要规格化。
- 有两种情况:

（1）左规:

  - 如果运算结果尾数为双符号补码11.1XX...X或者是00.0XX...X（尾数未溢出）时，规格化需将尾数左移。每左移一位，阶码减1，直到使尾数成为规格化数为止。
  - 阶码减1时，需判断是否下溢。若发生下溢出，认为结果为0。

（2）右规

- 若尾数加/减时，结果（尾数）发生溢出，即出现10.XX…X或者01.XX…X时，表明尾数有溢出，整个浮点数结果未必溢出。
- 可将尾数右移一位，阶码加1，即右规。右规最多1次。
- 阶码加1时，需判断是否上溢。若发生上溢出，认为结果为∞。

4. 舍入处理

在对阶及规格化时需要将尾数右移，右移将丢掉尾数的最低位，这就出现舍入的问题。

- 截（尾）断法

  此法最简单，就是将需丢弃的尾数低位丢弃。

- 末位恒置1法

  无论尾数右移丢弃的是0还是1，此法将保证要保留的尾数的最低位永远为1。

- 0舍1入法

  当尾数右移丢弃的是1时，要保留的最末位加1；
  当尾数右移丢弃的是0时，要保留的最末位不变。

例　两浮点数为：

$X=0.110101\times 2^{-010}$，$Y=-0.101010\times 2^{-001}$

求两数之和及差。

解：设两浮点数阶码为4位，用补码表示。尾数用8位，均用双符号位补码表示。则两数可表示为：

$[X]_{浮}=1110;\ 00.110101$

$[Y]_{浮}=1111;\ 11.010110$

① 对阶

求阶差：$[\triangle E]_{补}=[Ex]_{补}+[-Ey]_{补}=1110+0001=1111$。即X的阶码比Y的阶码小。

因此，X尾数右移一位，使两者阶码相同。这时的X为：

$[X]'_{浮}=1111;\ 00.011011$（0舍1入法）

① 对阶

求阶差：$[\Delta E]_{补}=[Ex]_{补}+[-Ey]_{补}=1110+0001=1111$。即X的阶码比Y的阶码小。

因此，X尾数右移一位，使两者阶码相同。这时的X为：

$[X]'_{浮}=1111$；00.011011（0舍1入法）

②尾数求和：

$$\begin{array}{r} 00.011011 \\ +\ 11.010110 \\ \hline 11.110001 \end{array}$$

尾数求差：

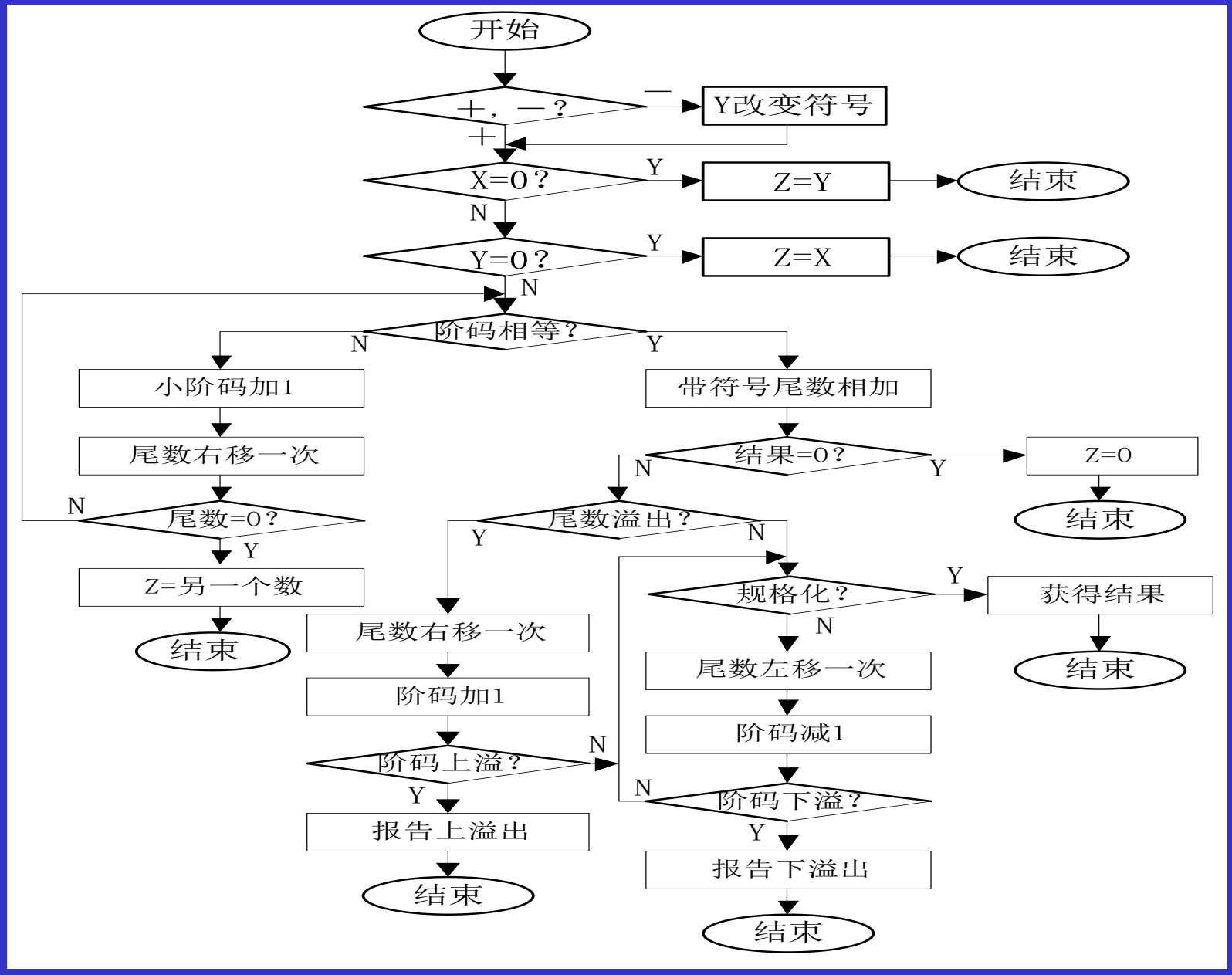
$$\begin{array}{r} 00.011011 \\ +\ 00.101010 \\ \hline 01.000101 \end{array}$$

③ 规格化

相加结果为非规格化尾数，需左规。将尾数左移2位，变为11.000100。同时，阶码减2，则阶码变为1101。两数相加结果为：$[X+Y]_{浮}=1101$；11.000100。

相减结果也为非规格化尾数，需右规。将尾数右移1位，变为00.100011，采用0舍1入法。阶码加1，则阶码为0000。两数相减结果为：$[X-Y]_{浮}=0000$；00.100011。

浮点加减运算流程


开始
+，－？
－
Y改变符号
+
X=0？
Y
Z=Y
结束
N
Y=0？
Y
Z=X
结束
N
阶码相等？
N
Y
小阶码加1
尾数右移一次
N
尾数=0？
Y
Z=另一个数
结束
带符号尾数相加
结果=0？
N
Y
Z=0
结束
尾数溢出？
Y
N
尾数右移一次
阶码加1
阶码上溢？
N
Y
报告上溢出
结束
规格化？
Y
获得结果
结束
N
尾数左移一次
阶码减1
阶码下溢？
N
Y
报告下溢出
结束

3.3.2 浮点乘除运算

1. 浮点乘法运算

- 两个规格化浮点数：$X=Mx\cdot 2^{Ex}$，$Y=My\cdot 2^{Ey}$
- 两浮点数相乘为：$Z=(Mx\cdot My)\,2^{(Ex+Ey)}$。

浮点乘法运算过程

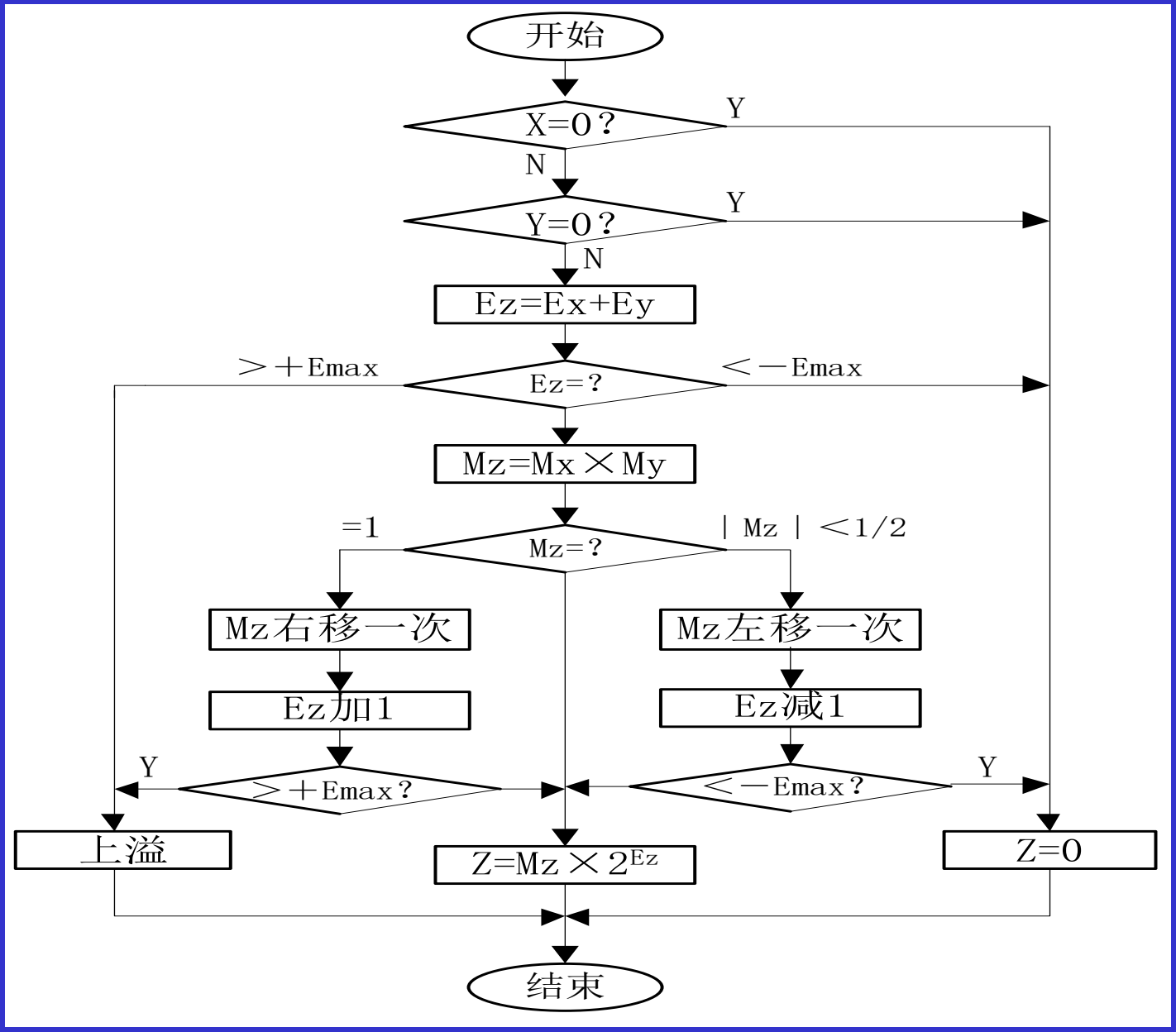
①参与运算的两浮点数一定是规格化数，且不为0。若有一个乘数为0，则乘积必为0。

②求乘积的阶码，即Ez=Ex+Ey，判断积的阶码是否溢出。

③两乘数的尾数相乘。

④规格化乘积的尾数。

浮点乘法运算流程


开始
X=0？
Y
N
Y=0？
Y
N
Ez=Ex+Ey
＞＋Emax
Ez=？
＜－Emax
Mz=Mx×My
=1
Mz=？
| Mz | ＜1/2
Mz右移一次
Mz左移一次
Ez加1
Ez减1
Y
＞＋Emax？
＜－Emax？
Y
上溢
Z=Mz×2Ez
Z=0
结束

2. 浮点除法运算

- 两个规格化浮点数：$X=Mx\cdot 2^{Ex}$，$Y=My\cdot 2^{Ey}$
- 两浮点数相除为：$Z=(Mx \div My)\ 2^{(Ex-Ey)}$

浮点除法的运算步骤为：

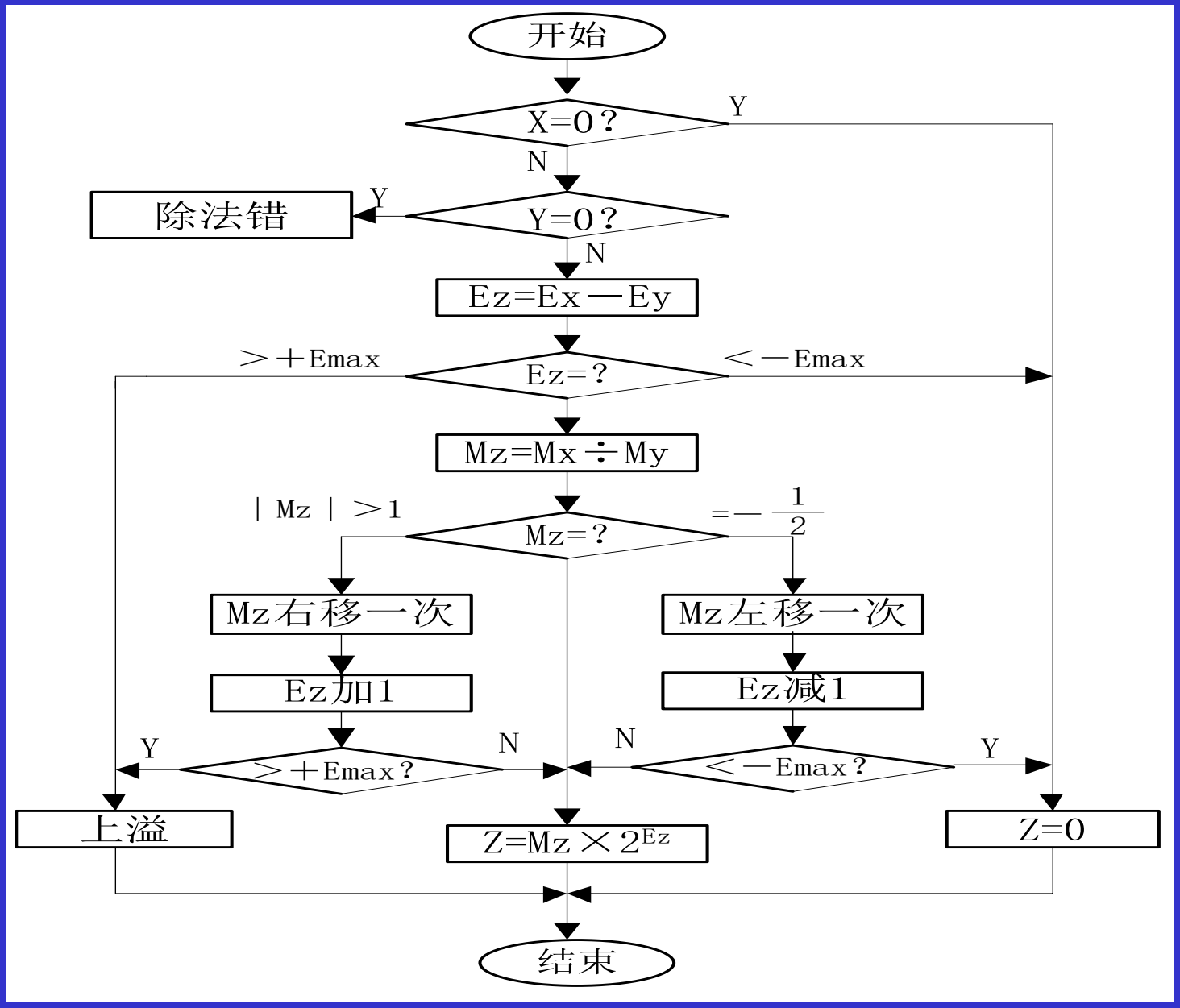
①若被除数为0，商为0。若除数为0，出错处理。

②求商的阶码，即Ez=Ex-Ey，判断商的阶码是否溢出。

③被除数尾数除以除数尾数。

④结果规格化及舍入处理。

浮点除法运算流程


开始
X=0？
Y
N
除法错
Y
Y=0？
N
Ez=Ex—Ey
＞＋Emax
Ez=？
＜—Emax
Mz=Mx÷My
| Mz | ＞1
Mz=？
=— 1/2
Mz右移一次
Mz左移一次
Ez加1
Ez减1
Y
＞＋Emax？
N
N
＜—Emax？
Y
上溢
Z=Mz×2^Ez
Z=0
结束

3.3.3 浮点数运算的实现

- 软件方法
- 配专用浮点处理器
- 在处理器中设置浮点运算部件
- 浮点运算的流水线处理

本章作业-4

第 10 题、第 26（1）题