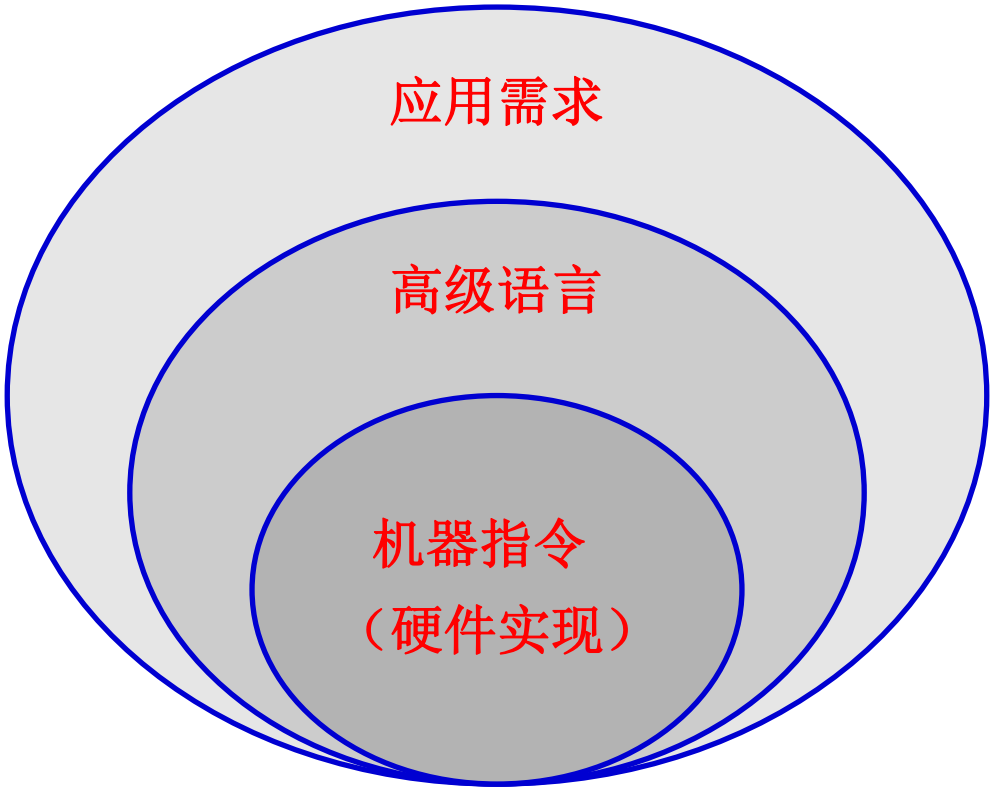
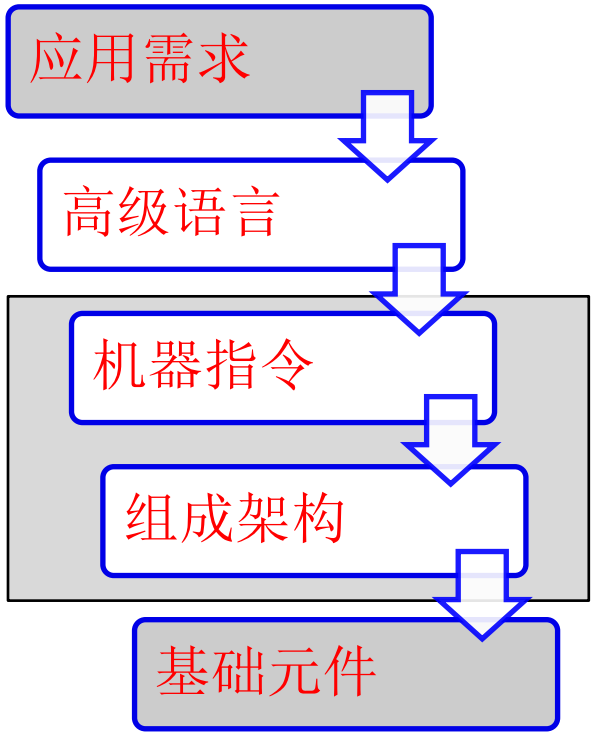
第6章 运算方法和运算部件





第6章 运算方法和运算部件

根据机器硬件的设计，来理解高级语言的实现。

第一讲 基本运算部件
第二讲 定点数运算
第二讲 浮点数运算

第一讲：基本运算部件

NJUAI 南京大学 人工智能学院

主 要 内 容

- 高级语言程序中涉及的运算（以C语言为例）
  - 整数算术运算、浮点数算术运算
  - 按位、逻辑、移位、位扩展和位截断
- 串行进位加法器
- 并行进位加法器
  - 全先行进位加法器
  - 两级/多级先行进位加法器
- 带标志加法器
- 算术逻辑部件（ALU）

如何实现高级语言源程序中的运算？

NJUAI 南京大学 人工智能学院

- C语言程序中的基本数据类型及其基本运算类型
  - 基本数据类型
    - 无符号数、带符号整数、浮点数、位串、字符（串）
  - 基本运算类型
    - 算术、按位、逻辑、移位、扩展和截断、匹配
- 计算机如何实现高级语言程序中的运算?
  - 将各类表达式编译（转换）为指令序列
  - 计算机直接执行指令来完成运算

例：C语句“f = (g+h) – (i+j);”中变量i、j、g、h由编译器分别分配在RISC-V寄存器t3~t6中，寄存器t3~t6的编号对应28~31，f分配在t0（编号5），则对应的RISC-V机器代码和汇编表示（#后为注释）如下：

```
0000000 11111 11110 000 00101 0110011   add t0, t5, t6   # g+h
0000000 11101 11100 000 00110 0110011   add t1, t3, t4   # i+j
0100000 00110 00101 000 00101 0110011   sub t0, t0, t1   # f=(g+h)–(i+j)
func7      rs2      rs1  func3 rd      opcode
```

需要提供哪些运算类指令才能支持高级语言需求呢？

数据的运算

- 高级语言程序中涉及的运算（以C语言为例）
  - 整数算术运算、浮点数算术运算
  - 按位、逻辑、移位、位扩展和位截断

逻辑运算、移位、扩展和截断等指令实现较容易，算术运算指令实现较难！

- 指令集中涉及的运算（如RISC-V指令系统提供的运算类指令）
  - 涉及的定点数运算
    - 算术运算
      - 带符号整数：取负 / 符号扩展 / 加 / 减 / 乘 / 除 / 算术移位
      - 无符号整数：0扩展 / 加 / 减 / 乘 / 除
    - 逻辑运算
      - 逻辑操作：与 / 或 / 非 / …
      - 移位操作：逻辑左移 / 逻辑右移
  - 涉及的浮点数运算：加、减、乘、除

完全能够支持高级语言对运算的所有需求

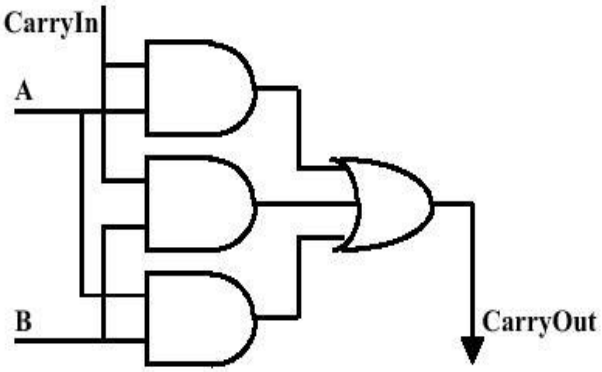
所有运算都可由ALU(或加法器+移位器+多路选择器+控制逻辑)实现！

以下介绍基本运算部件：加法器（串行→并行）→ 带标志加法器 → ALU

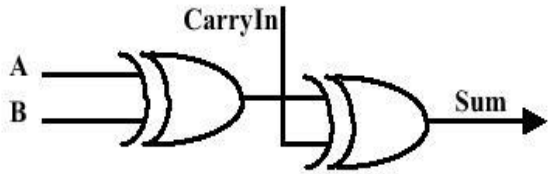
串行进位加法器

CarryOut 和 Sum 的逻辑图

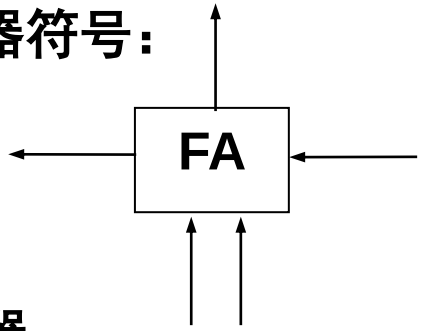
- CarryOut = B & CarryIn | A & CarryIn | A & B



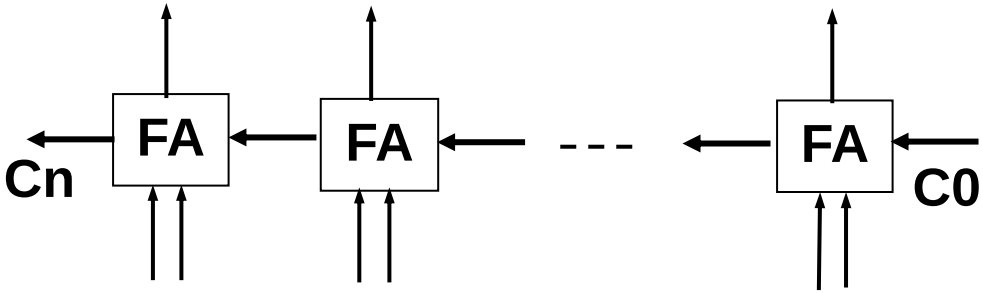
- Sum = A XOR B XOR CarryIn



全加器符号：



n位串行(行波)加法器：



串行加法器的缺点：

进位按串行方式传递，速度慢！

问题：n位串行加法器从C0到Cn的延迟时间为多少？ 2n级门延迟！

最后一位和数的延迟时间为多少？

2n+1级门延迟！ n>2

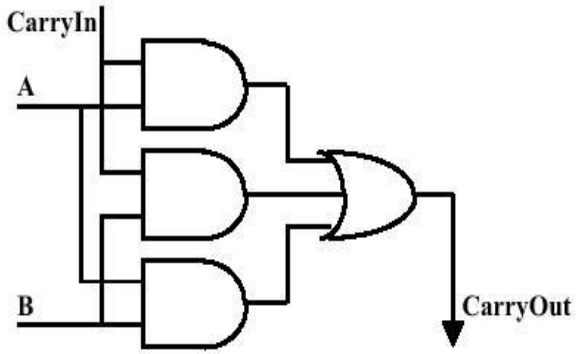
假定与/或门延迟为1ty，异或门3ty，则“和”与“进位”的延迟为多少？

Sum延迟为6ty；Carryout延迟为2ty。

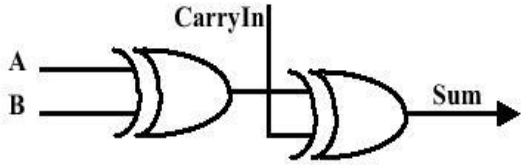
串行进位加法器

CarryOut 和 Sum 的逻辑图

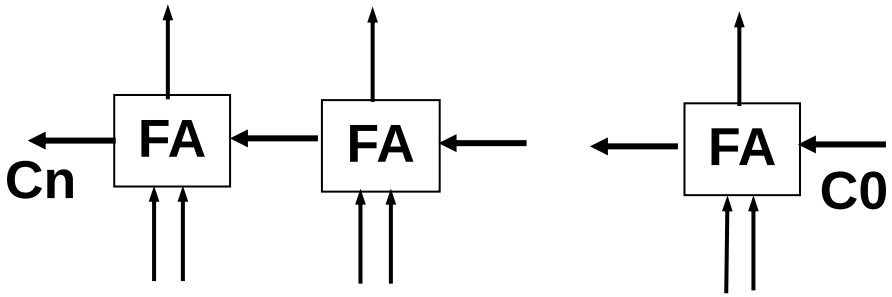
- CarryOut = B & CarryIn | A & CarryIn | A & B



- Sum = A XOR B XOR CarryIn



n位串行(行波)加法器：



4位串行进位加法器CRA的实现

```
module FA (
    input x, y, cin,
    output f, cout
);
    assign f = x ^ y ^ cin;
    assign cout = (x & y) | (x & cin) | (y & cin);
endmodule

module CRA (
    input [3:0] x, y,
    input cin,
    output [3:0] f,
    output cout
);
    wire [4:0] c;
    assign c[0] = cin;
    FA fa0(x[0], y[0], c[0], f[0], c[1]);
    FA fa1(x[1], y[1], c[1], f[1], c[2]);
    FA fa2(x[2], y[2], c[2], f[2], c[3]);
    FA fa3(x[3], y[3], c[3], f[3], c[4]);
    assign cout = c[4];
endmodule
```

并行进位加法器（CLA加法器）

NJUAI 南京大学 人工智能学院

- **为什么用先行进位方式？**

  串行进位加法器采用**串行逐级传递进位**，电路延迟与位数成正比关系，太慢了。因此，现代计算机采用一种**先行进位(Carry look ahead)方式**。

- **如何产生先行进位？**

  定义辅助函数：$G_i=A_iB_i$…进位生成函数

  $P_i=A_i+B_i$…进位传递函数（或 $P_i=A_i \oplus B_i$ ）

  通常把实现上述逻辑的电路称为**进位生成/传递部件**

  全加逻辑方程：$S_i= A_i \oplus B_i \oplus C_{i-1}$　$C_i=G_i+P_iC_{i-1}$ (i=1,…n)

  设n=4,则：$C_1=G_0+P_0C_0$

  $C_2=G_1+P_1C_1=G_1+P_1G_0+P_1P_0C_0$

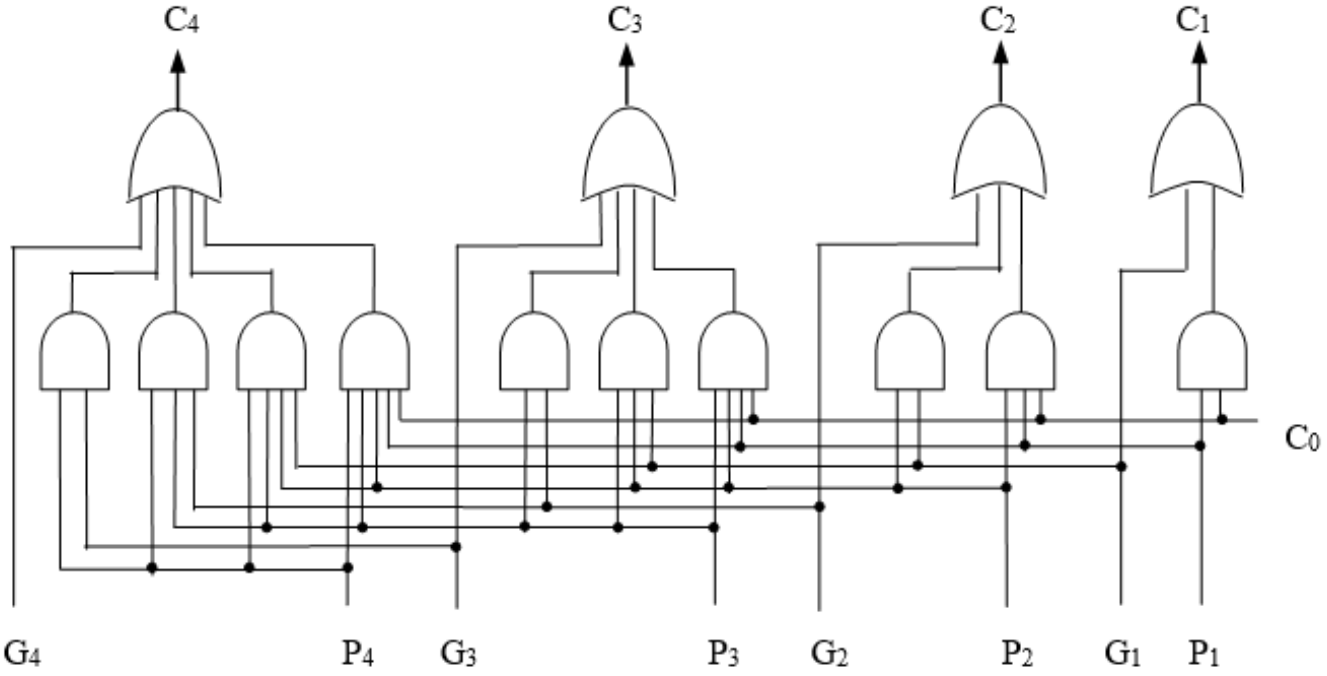
  $C_3=G_2+P_2C_2=G_2+P_2G_1+P_2P_1G_0+P_2P_1P_0C_0$

  $C_4=G_3+P_3C_3=G_3+P_3G_2+P_3P_2G_1+P_3P_2P_1G_0+P_3P_2P_1P_0C_0$

  由上式可知:各进位之间无等待，可以独立并同时产生。

  通常把实现上述逻辑的电路称为**4位CLU部件**

CLA加法器



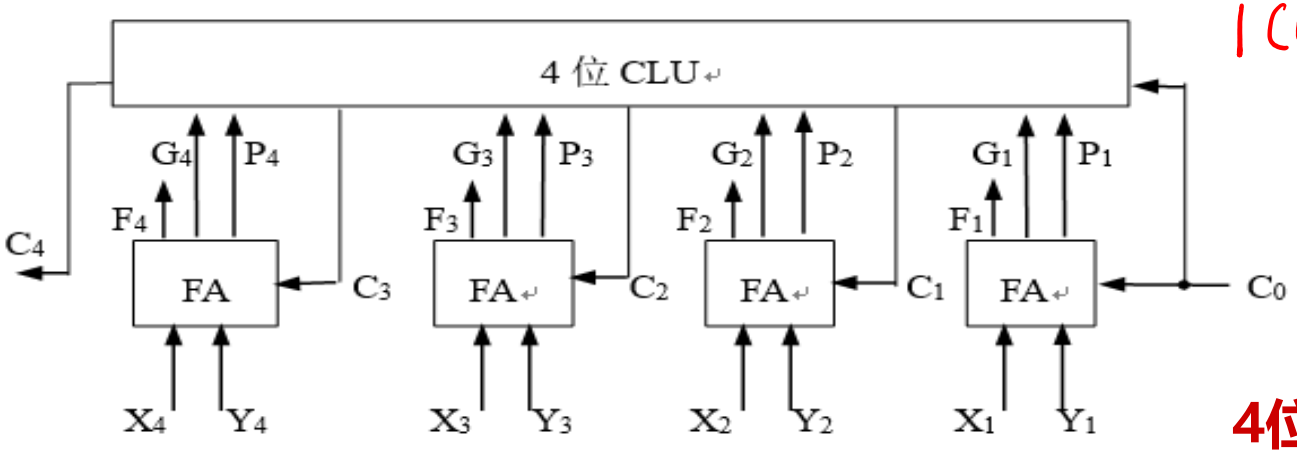
$C_1=G_0+P_0C_0$

$C_2=G_1+P_1C_1=G_1+P_1G_0+P_1P_0C_0$

$C_3=G_2+P_2C_2=G_2+P_2G_1+P_2P_1G_0+P_2P_1P_0C_0$

$C_4=G_3+P_3C_3=G_3+P_3G_2+P_3P_2G_1+P_3P_2P_1G_0+P_3P_2P_1P_0C_0$

4位CLU部件



1(GP)+2(CLU)+3(XOR)

$G_i=A_iB_i$

$P_i=A_i+B_i$ (或 $P_i=A_i\oplus B_i$)

$F_i=A_i\oplus B_i\oplus C_{i-1}$

4位CLA加法器

CLA加法器

$G_i = A_i B_i$

$P_i = A_i + B_i$ （或 $P_i = A_i \oplus B_i$ ）

$F_i = A_i \oplus B_i \oplus C_{i-1}$

```
module FA_PG (
    input x, y, cin,
    output f, p, g  );
    assign f = x ^ y ^ cin;
    assign p = x | y;
    assign g = x & y;
endmodule
```

```
module CLU (
    input [4:1] p, g,
    input c0,
    output [4:1] c );
    assign c[1] = g[1] | (p[1] & c0);
    assign c[2] = g[2] | (p[2] & g[1]) | (p[2] & p[1] & c0);
    // 以下两个表达式使用了位拼接运算和归约运算
    assign c[3] = g[3] | (p[3] & g[2]) | (&{p[3:2], g[1]}) | (&{p[3:1], c0});
    assign c[4] = g[4] | (p[4] & g[3]) | (&{p[4:3], g[2]}) | (&{p[4:2], g1}) | (&{p[4:1], c0});
endmodule
```

```
module CLA (
    input [3:0] x, y,
    input cin,
    output [3:0] f,
    output cout );
    wire [4:0] c;
    wire [4:1] p, g;
    assign c[0] = cin;
    FA_PG fa0(x[0], y[0], c[0], f[0], p[1], g[1]);
    FA_PG fa1(x[1], y[1], c[1], f[1], p[2], g[2]);
    FA_PG fa2(x[2], y[2], c[2], f[2], p[3], g[3]);
    FA_PG fa3(x[3], y[3], c[3], f[3], p[4], g[4]);
    CLU clu(p, g, c[0], c[4:1]);
    assign cout = c[4];
endmodule
```

$C_1 = G_0 + P_0 C_0$

$C_2 = G_1 + P_1 C_1 = G_1 + P_1 G_0 + P_1 P_0 C_0$

$C_3 = G_2 + P_2 C_2 = G_2 + P_2 G_1 + P_2 P_1 G_0 + P_2 P_1 P_0 C_0$

$C_4 = G_3 + P_3 C_3 = G_3 + P_3 G_2 + P_3 P_2 G_1 + P_3 P_2 P_1 G_0 + P_3 P_2 P_1 P_0 C_0$

局部（单级）先行进位加法器

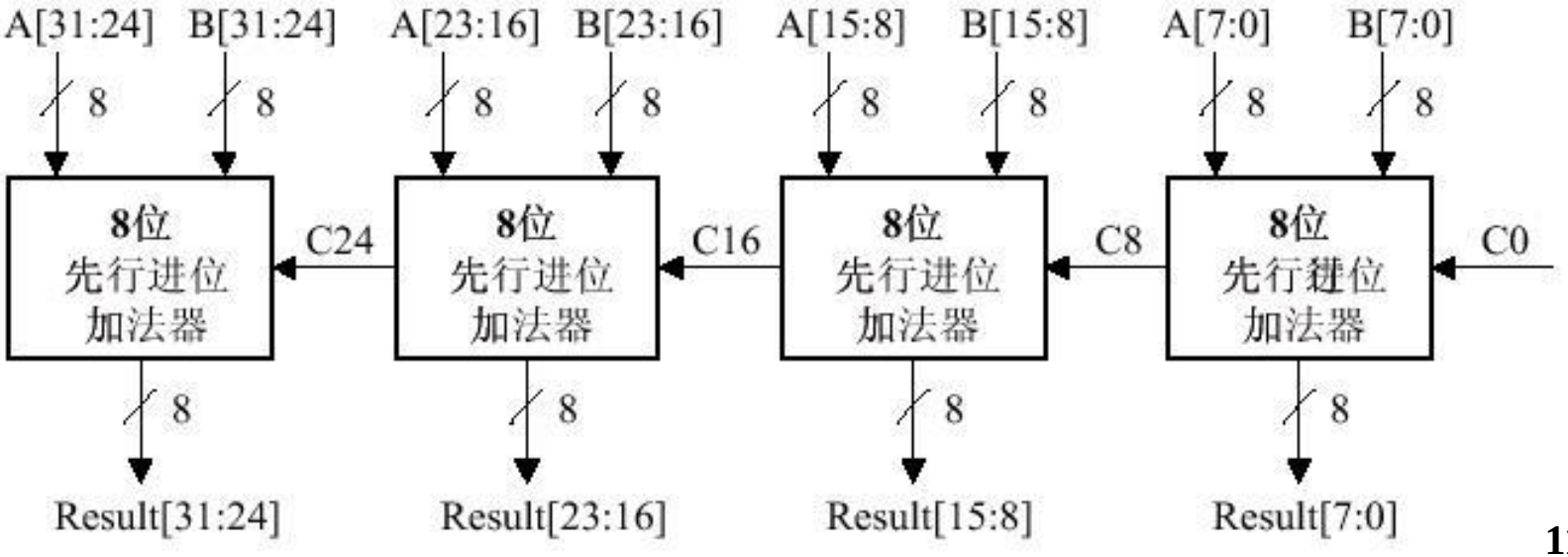
局部先行进位加法器 (Partial Carry Lookahead Adder)

或称 单级先行进位加法器

- 实现全先行进位加法器的成本太高
  - 想象 **Cin31**的逻辑方程的长度
- 一般性经验：
  - 连接一些**N**位先行进位加法器，形成一个大加法器
  - 例如：连接**4**个**8**位进位先行加法器，形成**1**个 **32**位局部先行进位加法器

1(gp)+2(cla)+2(cla)+2(cla)+2(cla)+3(xor)=12ty

问题：所有和数产生的延迟为多少？ 3+2+2+5=12ty



多级先行进位加法器

多级先行进位加法器

- 单级(局部)先行进位加法器的进位生成方式：

  “组内并行、组间串行”

- 所以，单级先行进位加法器虽然比行波加法器延迟时间短，但高位组进位依赖低位组进位，故仍有较长的时间延迟
- 通过引入组进位生成/传递函数实现“组内并行、组间并行”进位方式

设n=4,则：$C_1=G_0+P_0C_0$

$C_2=G_1+P_1C_1=G_1+P_1G_0+P_1P_0C_0$

$C_3=G_2+P_2C_2=G_2+P_2G_1+P_2P_1G_0+P_2P_1P_0C_0$

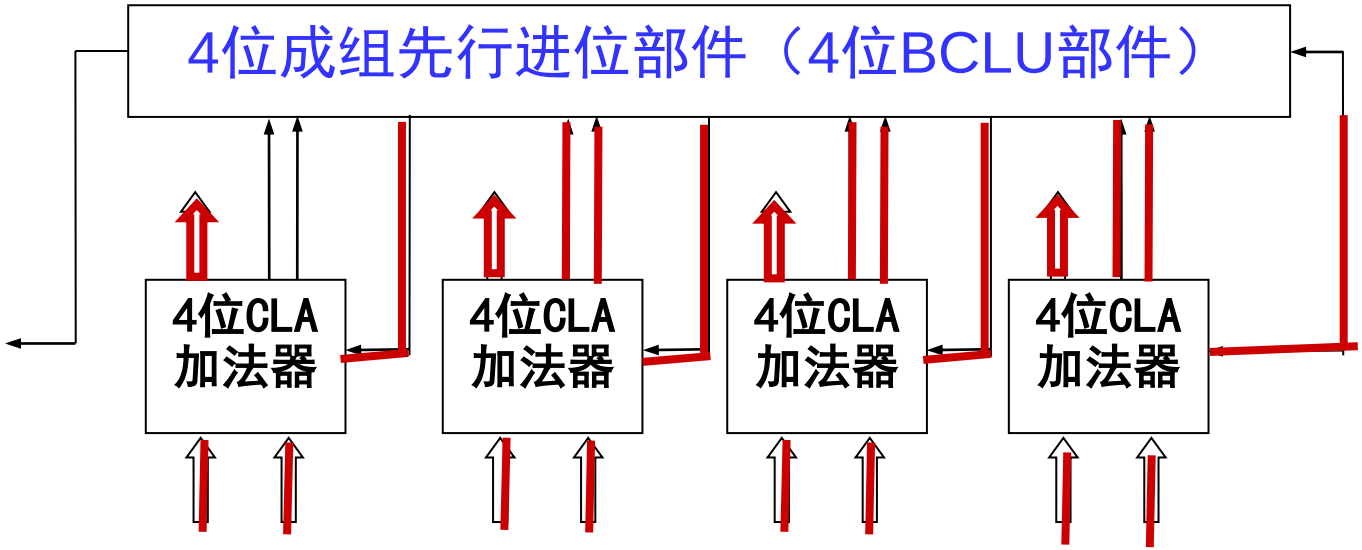
$G_3^*=G_3+P_3C_3=G_3+P_3G_2+P_3P_2G_1+P_3P_2P_1G_0$

$P_3^*=P_3P_2P_1P_0$

所以$C_4=G_3^*+P_3^*C_0$。把实现上述逻辑的电路称为4位BCLU（Block CLU）部件。

多级先行进位加法器

16位两级先行进位加法器



1(gp)+2(cla)+2(cla)+3(xor)=8ty

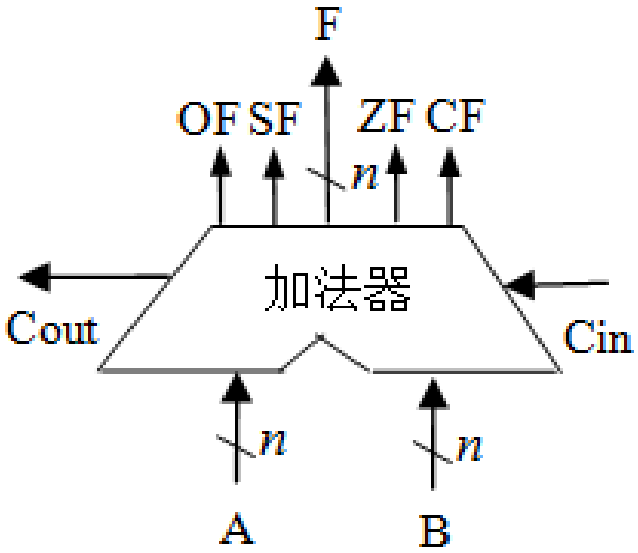
关键路径长度为多少？ 3+2+3 = 8ty

最终进位的延迟为多少？ 3+2=5ty

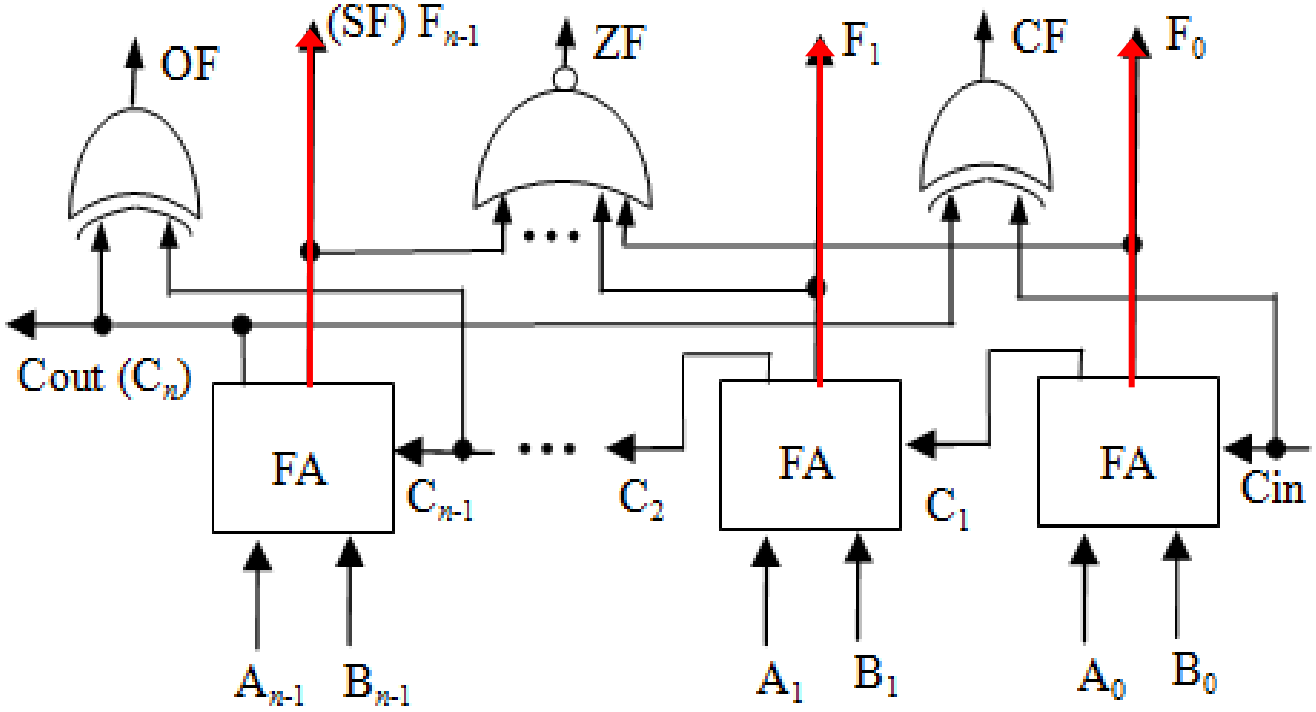
1(gp)+2(cla)+2(cla)=5ty

n位带标志加法器

- 需求：增加运算结果的**标志信息**
  - **判断是否溢出**
    - 通盘考虑：n位**带符号整数（补码）**相加
  - **比较大小**
    - 通过**（在加法器中）做减法**来判断



带标志加法器符号



带标志加法器的逻辑电路

溢出标志OF：

$OF=C_n \oplus C_{n-1}$

符号标志SF：

$SF=F_{n-1}$

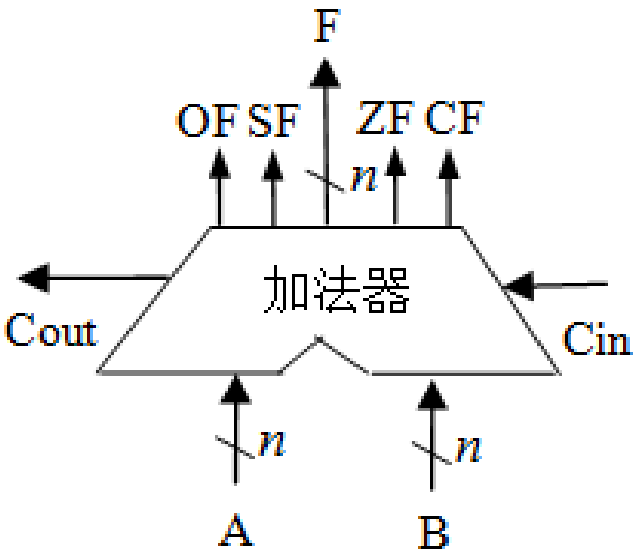
零标志ZF=1：

当且仅当F=0；

进位/借位标志CF：

$CF=Cout \oplus Cin$

n位带标志加法器



带标志加法器符号

溢出标志OF：$OF=C_n \oplus C_{n-1}$

符号标志SF：$SF=F_{n-1}$

零标志ZF=1当且仅当F=0；

进位/借位标志CF：

$CF=Cout \oplus Cin$

4位带标志先行进位加法器的实现

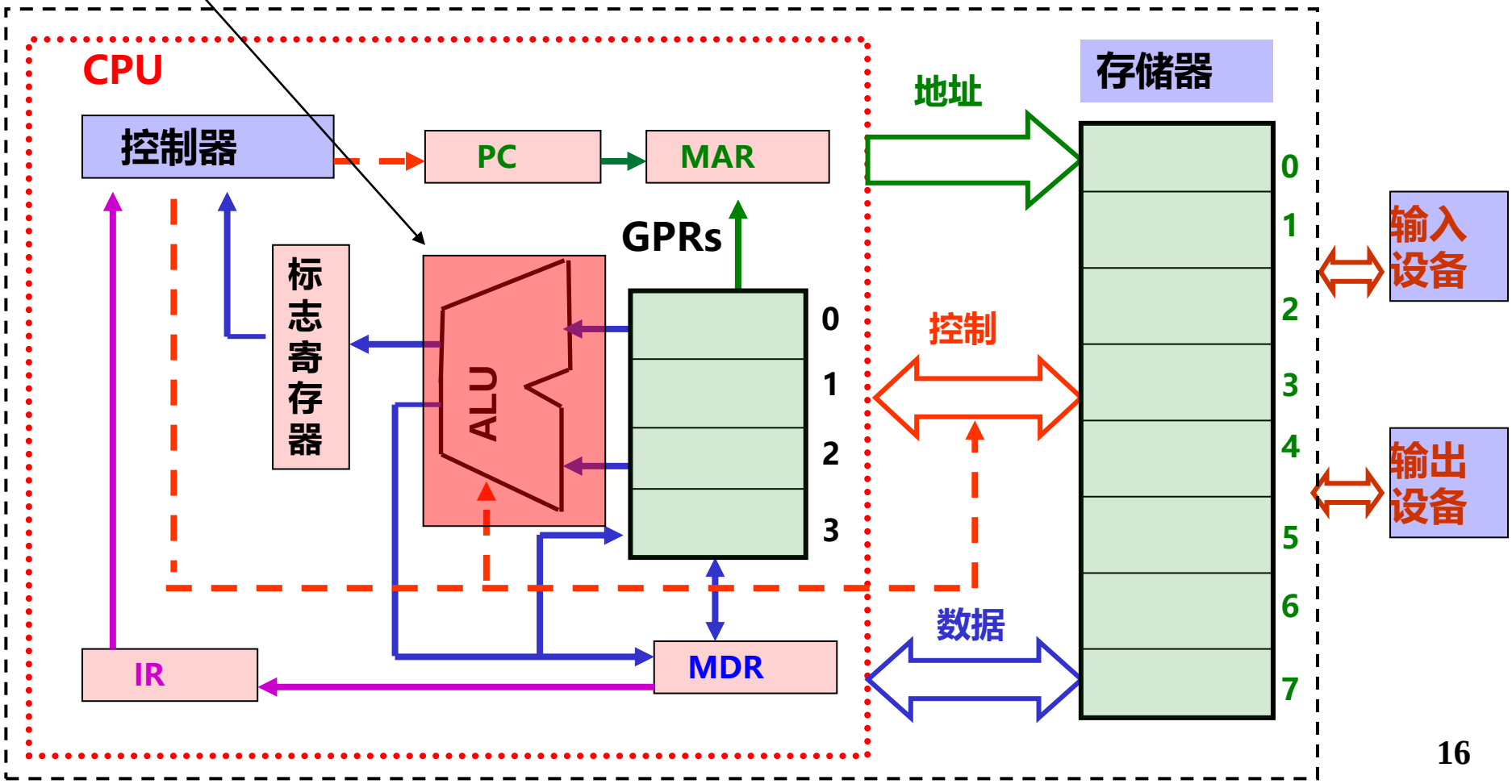
```
module CLA_FLAGS (
    input [3:0] a, b,
    input cin,
    output [3:0] f,
    output OF, SF, ZF, CF,
    output cout
);
    // 以下代码与CLA模块相同
    wire [4:0] c;
    wire [4:1] p, g;
    assign c[0] = cin;
    FA_PG fa0(a[0], b[0], c[0], f[0], p[1], g[1]);
    FA_PG fa1(a[1], b[1], c[1], f[1], p[2], g[2]);
    FA_PG fa2(a[2], b[2], c[2], f[2], p[3], g[3]);
    FA_PG fa3(a[3], b[3], c[3], f[3], p[4], g[4]);
    CLU clu(p, g, c[0], c[4:1]);
    assign cout = c[4];
   // 生成标志位
    assign OF = c[4] ^ c[3];
    assign SF = f[3];
    assign ZF = ~(|f);
    assign CF = c[4] ^ c[0];
endmodule
```

回顾：认识计算机中最基本的部件

CPU：中央处理器；PC：程序计数器；MAR：存储器地址寄存器

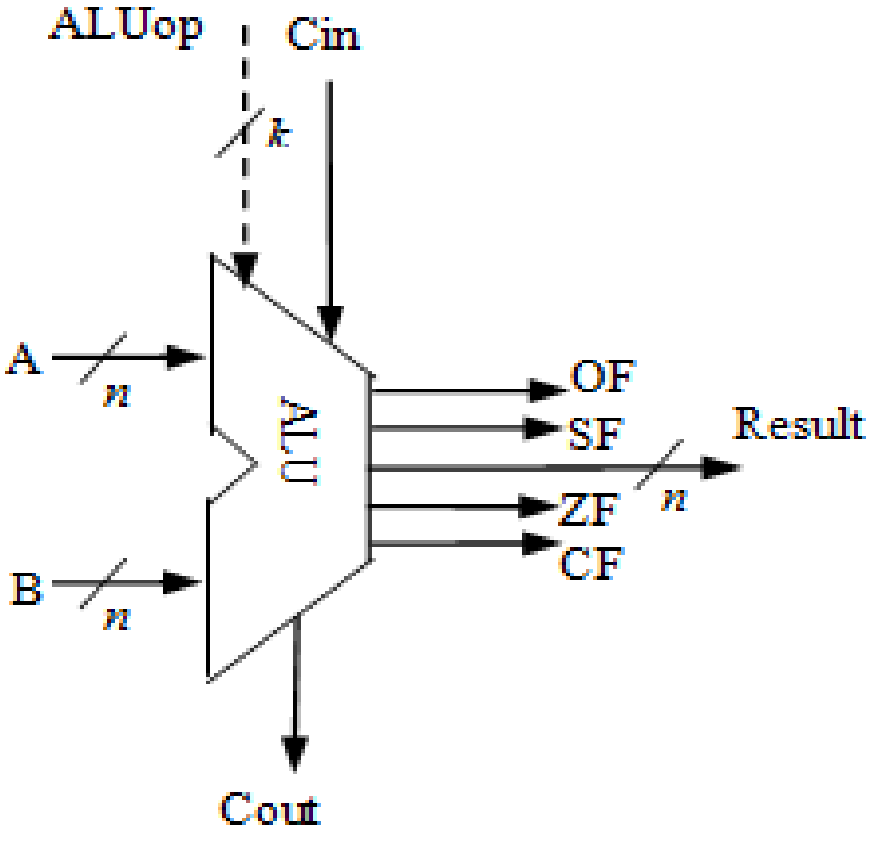
ALU：算术逻辑部件；IR：指令寄存器；MDR：存储器数据寄存器

GPRs：通用寄存器组（由若干通用寄存器组成）



算术逻辑部件（ALU）

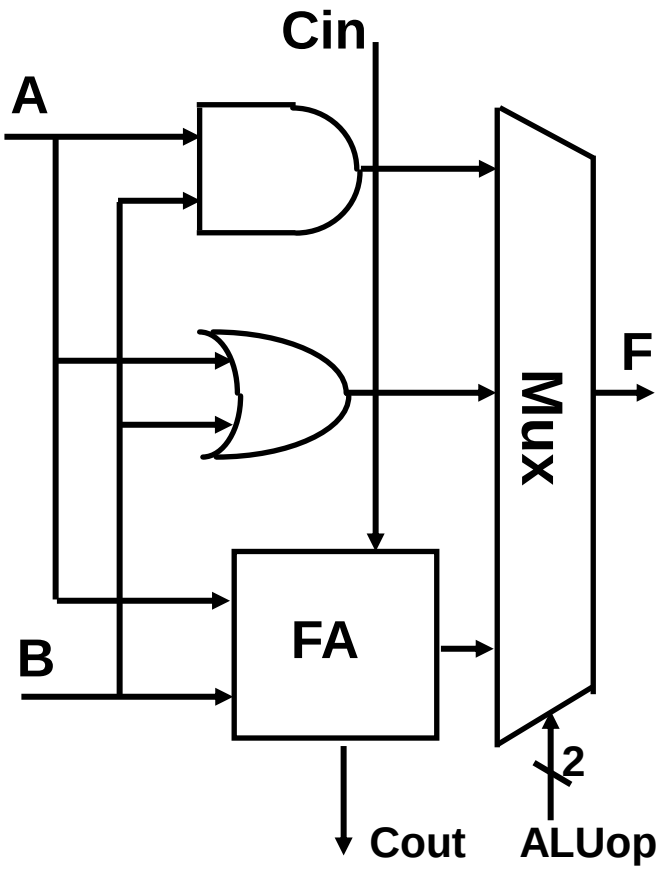
- 进行基本算术运算与逻辑运算
  - 无符号整数加、减
  - 带符号整数加、减
  - 与、或、非、异或等逻辑运算
- 核心电路是整数加/减运算部件
- 输出除和/差等，还有标志信息
- 有一个操作控制端（ALUop），用来决定ALU所执行的处理功能。ALUop的位数k决定了操作的种类例如，当位数k为3时，ALU最多只有2^3=8种操作。



ALU 符号

ALUop	Result	ALUop	Result	ALUop	Result	ALUop	Result
0 0 0	A加B	0 1 0	A与B	1 0 0	A取反	1 1 0	A
0 0 1	A减B	0 1 1	A或B	1 0 1	A⊕B	1 1 1	未用

举例：1-bit ALU和4-bit ALU



1-bit ALU

4位先行进位ALU的实现

```
module ALU (
    input [3:0] a, b,
    input [1:0] aluop,
    input cin,
    output [3:0] f,
    output OF, SF, ZF, CF,
    output cout
);
    wire [3:0] sum;
    CLA_FLAGS(a, b, cin, sum, OF, SF, ZF, CF, cout);
    always @(*) begin
        case(aluop)
            2'b00: f = a & b;
            2'b01: f = a | b;
            2'b10: f = sum;
            default: f = 0;
        endcase
    end
endmodule
```

实际的ALU中还包括减法、算术移位、逻辑移位等其他运算功能

第二讲：定点数运算

南京大学 人工智能学院

主 要 内 容

- 定点数加减运算
  - 补码加减运算
  - 原码加减运算
  - 移码加减运算
- 定点数乘法运算
  - 原码乘法运算
  - 补码乘法运算
  - 快速乘法器
- 定点数除法运算
  - 原码除法运算
  - 补码除法运算

提醒：后续的运算设计的基本思路是基于前述的ALU设计基础上进行的。

n位整数加/减运算器

先看一个C程序段：

```
int x=9, y=-6, z1, z2;
z1=x+y;
z2=x-y;
```

补码的定义　假定补码有n位，则：

$[X]_{补}=2^n+X \quad (-2^n \le X < 2^n, \mathrm{mod}\ 2^n)$

问题：上述程序段中，x和y的机器数是什么？z1和z2的机器数是什么？

回答：x的机器数为$[x]_{补}$，　y的机器数为$[y]_{补}$；

z1的机器数为$[x+y]_{补}$；

z2的机器数为$[x-y]_{补}$。

因此，计算机中需要有一个电路，能够实现以下功能：

已知 $[x]_{补}$ 和 $[y]_{补}$，计算$[x+y]_{补}$ 和 $[x-y]_{补}$。

根据补码定义，有如下公式：

$$[x+y]_{补}=2^n+x+y=2^n+x+2^n+y=[x]_{补}+[y]_{补}\ (\mathrm{mod}\ 2^n)$$

$$[x-y]_{补}=2^n+x-y=2^n+x+2^n-y=[x]_{补}+[-y]_{补}\ (\mathrm{mod}\ 2^n)$$

$$[-y]_{补}=\overline{[y]}_{补}+1$$

n位整数加/减运算器

- 补码加减运算公式

$[A+B]_{补} = [A]_{补} + [B]_{补}\ (\mathrm{mod}\ 2^n)$

$[A-B]_{补} = [A]_{补} + [-B]_{补}\ (\mathrm{mod}\ 2^n)$

– 实现减法的关键工作在于：求$[-B]_{补}$

问题：如何求$[-B]_{补}$?

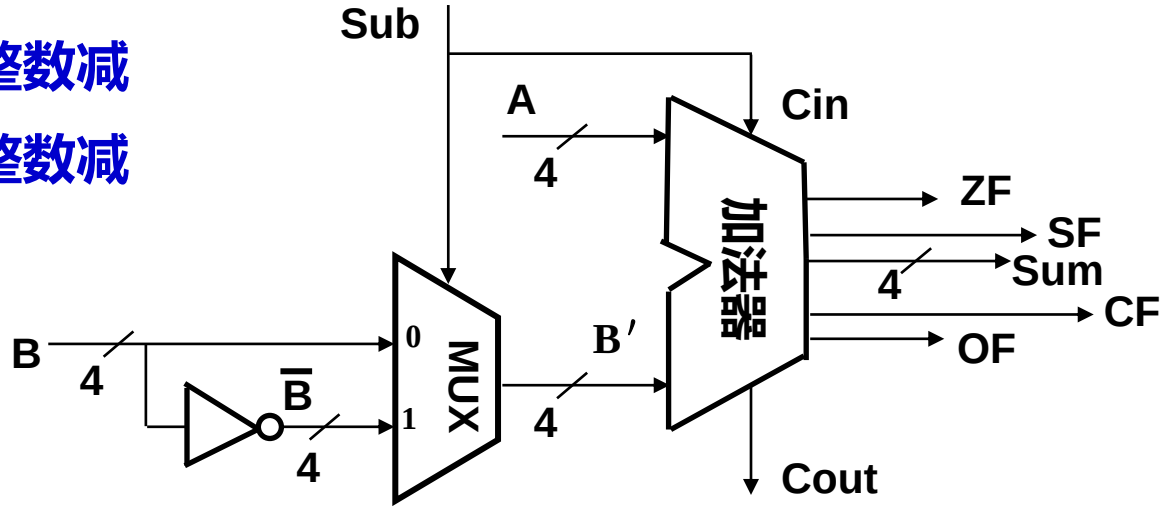
$[-B]_{补} = \overline{[B]}_{补} + 1$

- 利用带标志加法器，可构造整数加/减运算器，进行以下运算：

无符号整数加、无符号整数减

带符号整数加、带符号整数减

当Sub为1时，做减法
当Sub为0时，做加法

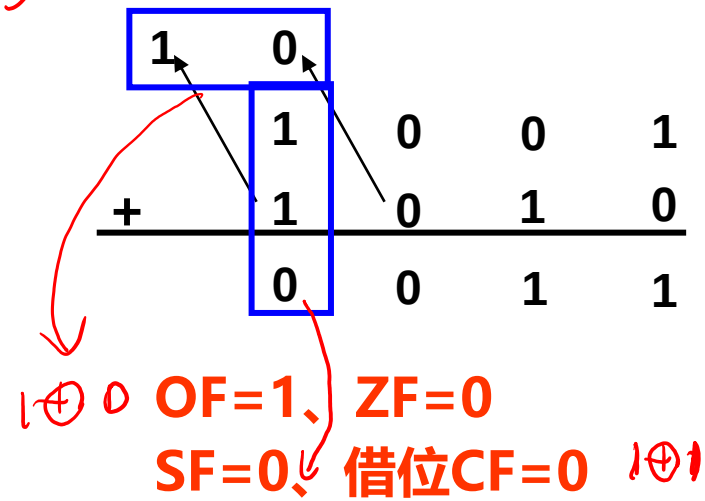


整数加/减运算部件

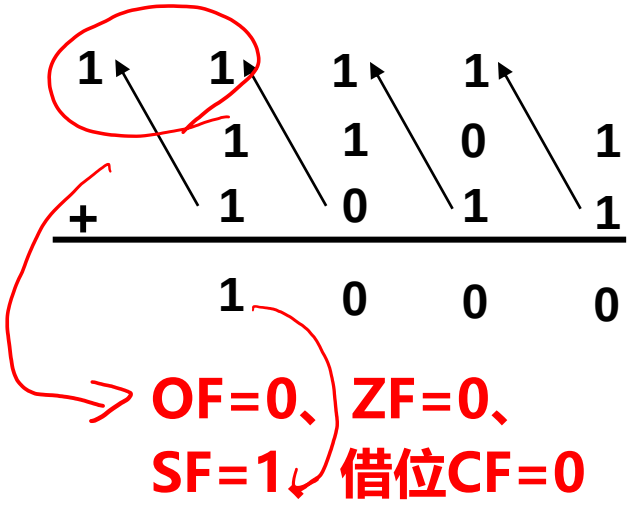
在整数加/减运算部件基础上，加上寄存器、移位器以及控制逻辑，就可实现ALU、乘/除运算以及浮点运算电路

整数减法举例（注意标志位的使用）

signed -7- 6 = -7 + (-6) = +3 ✗
unsigned 9 - 6 = 3 ✓



-3 - 5 = - 3 + (- 5) = - 8 ✓
13 - 5 = 8 ✓



带符号溢出：(1) 最高位和次高位的进位不同
(2) 和的符号位和加数的符号位不同

无符号减溢出：差为负数，即借位CF=1

做减法以比较大小，规则：
Unsigned: CF=0时，大于
Signed：OF=SF时，大于

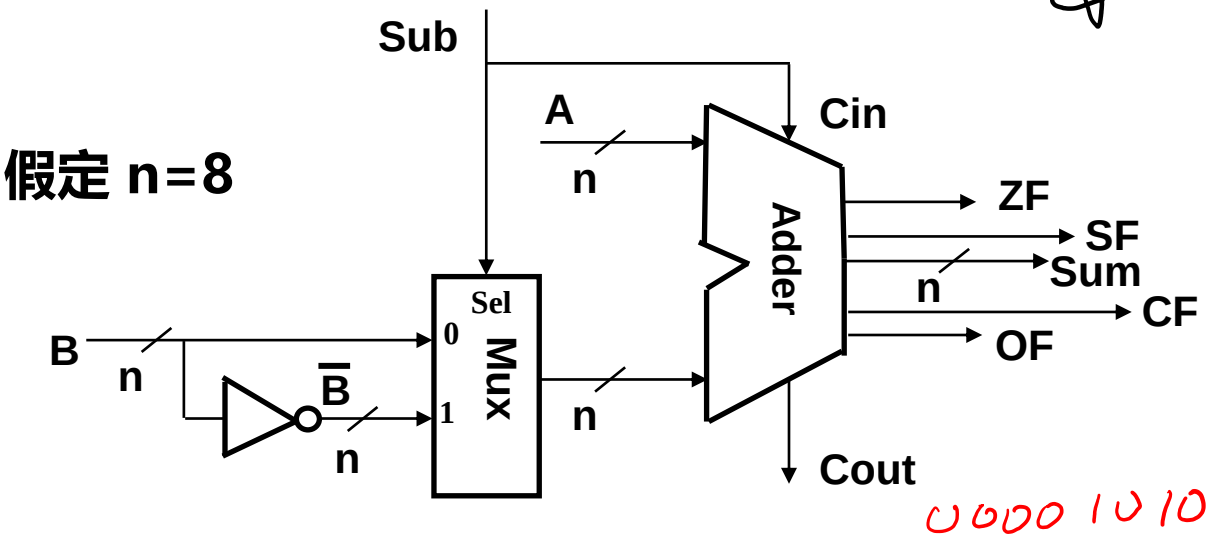
验证：9>6，故CF=0；13>5，故CF=0

验证：-7<6，故OF≠SF
-3<5，故OF≠SF

举例：整数减法（高级语言的对应）

```
unsigned int x=134;
unsigned int y=246;
int m=x;
int n=y;
unsigned int z1=x-y;
unsigned int z2=x+y;
int k1=m-n;
int k2=m+n;
```

无符号和带符号加减运算都用该部件执行



0000 1010

x和m的机器数一样：1000 0110，y和n的机器数一样：1111 0110

z1和k1的机器数一样：1001 0000，CF=1，OF=0，SF=1

z1的值为144（=134-246+256，x-y<0），k1的值为-112。

无符号减公式： CF=1 借出.

$$result=\begin{cases} x-y & (x-y>0) \\ x-y+2^n & (x-y<0) \end{cases}$$

带符号减公式： CF⊕SF 借出

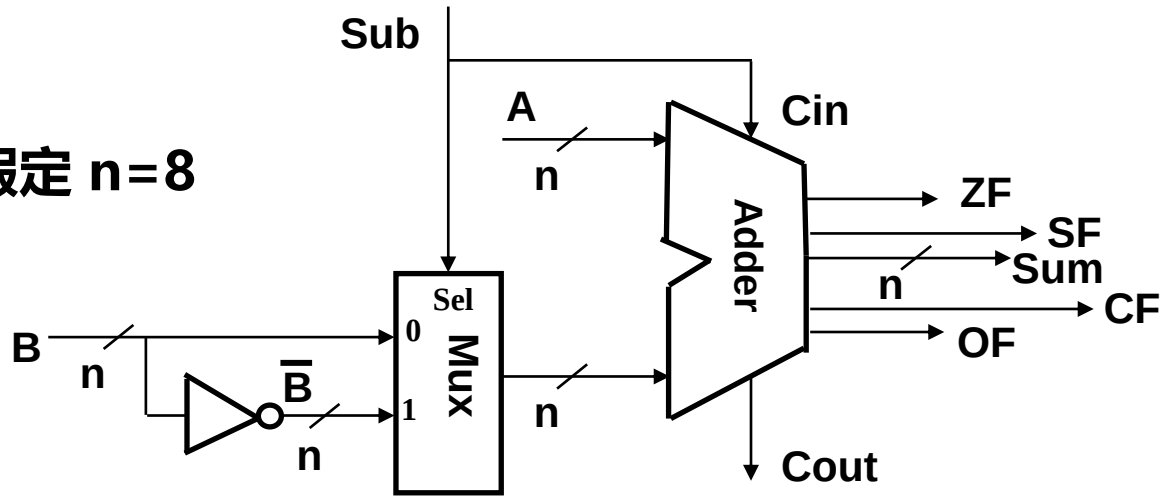
$$result=\begin{cases} x-y-2^n & (2^{n-1}\leqslant x-y) \quad \text{正溢出} \\ x-y & (-2^{n-1}\leqslant x-y<2^{n-1}) \quad \text{正常} \\ x-y+2^n & (x-y<-2^{n-1}) \quad \text{负溢出} \end{cases}$$

举例：整数加法（高级语言的对应）

```
unsigned int x=134;
unsigned int y=246;
int m=x;
int n=y;
unsigned int z1=x-y;
unsigned int z2=x+y;
int k1=m-n;
int k2=m+n;
```

无符号和带符号加减运算都用该部件执行

假定 n=8



x和m的机器数一样：1000 0110，y和n的机器数一样：1111 0110

z2和k2的机器数一样：0111 1100，CF=1，OF=1，SF=0

z2的值为124（=134+246-256，x+y>256）

k2的值为124（=134+246-256，m+n>128，即正溢出）

无符号加公式：

$$\text{result}=\begin{cases} x+y & (x+y<2^n) \\ x+y-2^n & (2^n\leqslant x+y<2^{n+1}) \end{cases}$$

带符号加公式：

$$\text{result}=\begin{cases} x+y-2^n & (2^{n-1}\leqslant x+y) \quad \text{正溢出} \\ x+y & (-2^{n-1}\leqslant x+y<2^{n-1}) \quad \text{正常} \\ x+y+2^n & (x+y<-2^{n-1}) \quad \text{负溢出} \end{cases}$$

原码加/减运算

- 用于浮点数尾数运算（注意：浮点数加减运算时，要先对齐阶码）
  - 符号位和数值部分分开处理
  - 仅对数值部分进行加减运算，符号位起判断和控制作用
- 规则如下：
  - 比较两数符号，对加法实行“同号求和，异号求差”，对减法实行“异号求和，同号求差”。
  - 求和：数值位相加，和的符号取被加数（被减数）的符号。若最高位产生进位，则结果溢出。
  - 求差：被加数（被减数）加上加数（减数）的补码。

    a）最高数值位产生进位表明加法结果为正，所得数值位正确。

    b）最高数值位没产生进位表明加法结果为负，得到的是数值位的补码形式，需对结果求补，还原为绝对值形式的数值位。
  - 差的符号位：a）情况下，符号位取被加数（被减数）的符号；

    b）情况下，符号位为被加数（被减数）的符号取反。

原码加/减运算

例1：已知 $[X]_{原} = 1.0011$，$[Y]_{原} = 1.1010$，要求计算$[X+Y]_{原}$

解：由原码加减运算规则知：同号相加，则求和，和的符号同被加数符号。

所以：和的数值位为：0011 + 1010 = 1101 （ALU中无符号数相加）

和的符号位为：1

$[X+Y]_{原} = 1.1101$

求和：直接加，有进位则溢出，符号同被加数

例2 ：已知 $[X]_{原} = 1.0011$，$[Y]_{原} = 1.1010$，要求计算$[X–Y]_{原}$

解：由原码加减运算规则知：同号相减，则求差（补码减法）

差的数值位为：$0011+(1010)_{补} = 0011 + 0110 = 1001$

最高数值位没有产生进位，表明加法结果为负，需对1001求补，还原为绝对值形式的数值位。即：$(1001)_{补} = 0111$

差的符号位为$[X]_{原}$的符号位取反，即：0

$[X–Y]_{原} = 0.0111$

求差：加补码，不会溢出，符号分情况

思考题：如何设计一个基于ALU的原码加/减法器?

移码加/减运算

- 用于浮点数阶码运算
  - 符号位和数值部分可以一起处理
- 运算公式（假定在一个n位ALU中进行加法运算）

$[E1]_{移}+[E2]_{移}=2^{n-1}+E1+2^{n-1}+E2=2^n+E1+E2=[E1+E2]_{补}$ （mod 2^n）

$[E1]_{移}-[E2]_{移}=[E1]_{移}+[-[E2]_{移}]_{补}=2^{n-1}+E1+2^n-[E2]_{移}$

$=2^{n-1}+E1+2^n-2^{n-1}-E2$

$= 2^n+E1-E2 = [E1-E2]_{补}$ （mod 2^n）

结论：移码的和、差等于和、差的补码！（需要转换成移码）

- 运算规则　　补码和移码的关系：符号位相反、数值位相同！

① 加法：直接将$[E1]_{移}$和$[E2]_{移}$进行模2^n加，然后对结果的符号取反。

② 减法：先将减数$[E2]_{移}$求补（各位取反，末位加1），然后再与被减数 $[E1]_{移}$进行模2^n相加，最后对结果的符号取反。

③ 溢出判断：进行模2^n相加时，如果两个加数的符号相同，并且与和数的符号也相同，则发生溢出。

移码加/减运算

例1： 用四位移码计算“–7+（– 6）”和“–3 + 6”的值。

解：$[-7]_{移} = 0001$　$[-6]_{移} = 0010$　$[-3]_{移} = 0101$　$[6]_{移} = 1110$

$[-7]_{移} + [-6]_{移} = 0001 + 0010 = 0011$（两个加数与结果符号都为0，溢出）

$[-3]_{移} + [6]_{移} = 0101 + 1110 = 0011$，符号取反后为 1011，其真值为+3

问题：$[-7+(-6)]_{移}$=?　$[-3+(6)]_{移}$ =?

例2： 用四位移码计算“–7 –（– 6）”和“–3 – 5”的值。

解：$[-7]_{移} = 0001$　$[-6]_{移} = 0010$　$[-3]_{移} = 0101$　$[5]_{移} = 1101$

$[-7]_{移} - [-6]_{移} = 0001 + 1110 = 1111$，符号取反后为 0111，其真值为–1。

$[-3]_{移} - [5]_{移} = 0101 + 0011 = 1000$，符号取反后为 0000，其真值为– 8。

主 要 内 容

- 定点数加减运算
  - 补码加减运算
  - 原码加减运算
  - 移码加减运算
- 定点数乘法运算
  - 原码乘法运算
  - 补码乘法运算
  - 快速乘法器
- 定点数除法运算
  - 原码除法运算
  - 补码除法运算

无符号数的乘法运算

假定：$[X]_{原}=x_0.x_1\ldots x_n$，$[Y]_{原}=y_0.y_1\ldots y_n$ ，求$[x\times Y]_{原}$

数值部分 $z_1\ldots z_{2n} = (0.x_1\ldots x_n) \times (0.y_1\ldots y_n)$

(小数点位置约定，不区分小数还是整数)

- **Paper and pencil example:**

$$X\times Y=\sum_{i=1}^{4}(X\times y_i\times 2^{-i})$$

```
Multiplicand      1000
Multiplier      x 1001
-------------------------
                  1000   X× y4× 2-4      n=4
                 0000    X× y3× 2-3
                0000     X× y2× 2-2
               1000      X× y1× 2-1
-------------------------
Product(积)  0.1001000
```

整个运算过程中用到两种操作：加法 + 左移

因而，可用ALU和移位器来实现乘法运算

无符号数的乘法运算

- 手工乘法的特点：
  - ① 每步计算：X×yi，若yi = 0，则得0；若yi = 1，则得X
  - ② 把①求得的各项结果X× yi 逐次左移，可表示为X× yi×2^{-i}
  - ③ 对②中结果求和，即 ∑ (X× yi×2^{-i})，这就是两个无符号数的乘积
- 计算机内部稍作以下改进：（i从右n向左1排列）
  - ① 每次得X×yi后，与前面所得结果累加，得到Pi，称之为部分积。因为不用等到最后一次求和，减少了保存各次相乘结果X×yi的开销。
  - ② 每次得X×yi后，不将它左移与前次部分积Pi相加，而将部分积Pi右移后与X×yi相加。因为加法运算始终对部分积中高n位进行。故用n位加法器可实现二个n位数相乘。
  - ③ 对乘数中为“1”的位执行加法和右移，对为“0”的位只执行右移，而不执行加法运算。

无符号乘法运算的算法推导

- 上述思想可写成如下数学推导过程：

$$X \times Y = X \times (0.y_1 y_2 \cdots y_n)$$

$$= X \times y_1 \times 2^{-1} + X \times y_2 \times 2^{-2} + \cdots\cdots + X \times y_{n-1} \times 2^{-(n-1)} + X \times y_n \times 2^{-n}$$

$$= 2^{-n} \times X \times y_n + 2^{-(n-1)} \times X \times y_{n-1} + \cdots\cdots + 2^{-2} \times X \times y_2 + 2^{-1} \times X \times y_1$$

$$= 2^{-1}\underbrace{(2^{-1}(2^{-1}\cdots 2^{-1}}_{n\text{个}2^{-1}}(2^{-1}(0 + X \times y_n) + X \times y_{n-1}) + \cdots\cdots + X \times y_2) + X \times y_1)$$

- 上述推导过程具有明显的递归性质，因此，无符号数乘法过程可归结为循环计算下列算式的过程：设$P_0 = 0$，每步的乘积为：

$$P_1 = 2^{-1}(P_0 + X \times y_n)$$

$$P_2 = 2^{-1}(P_1 + X \times y_{n-1})$$

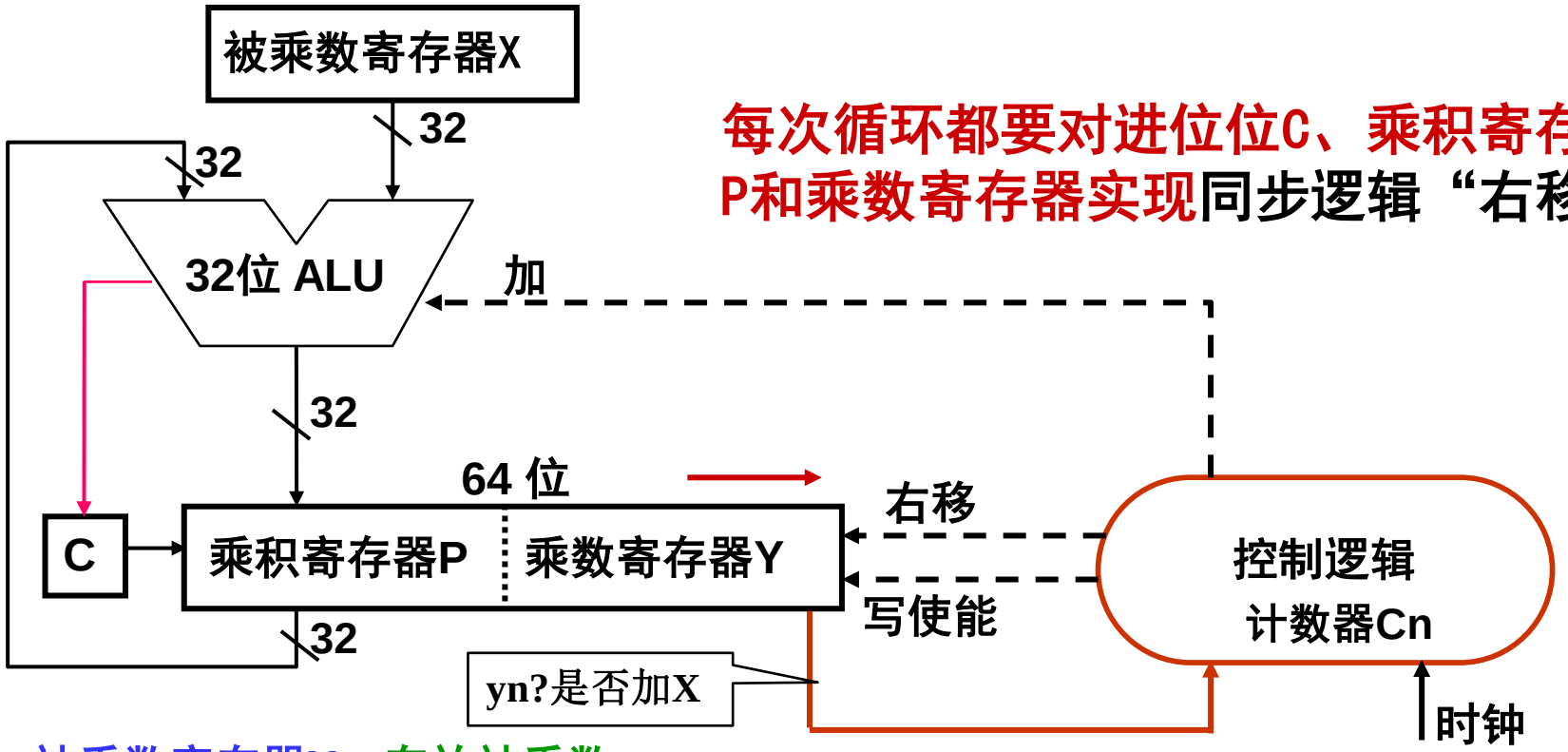
$$\cdots\cdots \quad \cdots\cdots$$

$$P_n = 2^{-1}(P_{n-1} + X \times y_1)$$

迭代过程从乘数最低位y_n和P_0=0开始，经n次“判断–加法–右移”循环，直到求出P_n为止

- 其递推公式为：$P_{i+1} = 2^{-1}(P_i + X \times y_{n-i})$　　$(i = 0, 1, 2, 3, \cdots, n-1)$
- 最终乘积$P_n = X \times Y$

32位乘法运算的硬件实现



每次循环都要对进位位C、乘积寄存器P和乘数寄存器实现同步逻辑“右移”

- 被乘数寄存器X：存放被乘数
- 乘积寄存器P：开始置初始部分积P0 = 0；结束时，存放的是64位乘积的高32位
- 乘数寄存器Y：开始时置乘数；结束时，存放的是64位乘积的低32位
- 进位触发器C：保存加法器的进位信号
- 循环次数计数器Cn：存放循环次数。初值32，每循环一次，Cn减1，Cn=0时结束
- ALU：乘法核心部件。在控制逻辑控制下，对P和X的内容“加”，在“写使能”控制下运算结果被送回P，进位位在C中

Example：无符号整数乘法运算

举例说明：若需计算z=x*y; x、y和z都是unsigned类型。

设x=1110　y=1101　应用递推公式：$P_i=2^{-1}(x*y_i+P_{i-1})$

	C	乘积P	乘数R
	0	0000	1101
	+	1110	
	0	1110	1101
→	0	0111	0110
→	0	0011	1011
	+	1110	
	1	0001	1011
→	0	1000	1101
	+	1110	
	1	0110	1101
→	0	1011	0110

可用一个双倍字长的乘积寄存器；也可用两个单倍字长的寄存器。

部分积初始为0。

保留进位位。

右移时进位、部分积和剩余乘数一起进行逻辑右移。

验证：x=14, y=13, z=x*y=182

当z取4位时，结果发生溢出，因为高4位不为全0！

原码乘法算法

- 用于浮点数尾数乘运算
  - 符号与数值分开处理：积符异或得到，数值用无符号乘法运算

例：设$[x]_{原}=0.1110$，$[y]_{原}=1.1101$，计算$[x\times y]_{原}$

解：数值部分用无符号数乘法算法计算：$1110\times1101=10110110$

符号位：$0\oplus1=1$，所以：$[x\times y]_{原}=1.10110110$

一位乘法：每次只取乘数的一位判断，需n次循环，速度慢。

两位乘法：每次取乘数两位判断，只需n/2次循环，快一倍。

- 两位乘法递推公式：

00：$P_{i+1}=2^{-2}P_i$

01：$P_{i+1}=2^{-2}(P_i+X)$

10：$P_{i+1}=2^{-2}(P_i+2X)$

11：$P_{i+1}=2^{-2}(P_i+3X)=2^{-2}(P_i+4X-X)$
$=2^{-2}(Pi-X)+X$

y_{i-1}	y_i	T	操　作	迭代公式
0	0	0	$0\rightarrow T$	$2^{-2}(P_i)$
0	0	1	$+X$　$0\rightarrow T$	$2^{-2}(P_i+X)$
0	1	0	$+X$　$0\rightarrow T$	$2^{-2}(P_i+X)$
0	1	1	$+2X$　$0\rightarrow T$	$2^{-2}(P_i+2X)$
1	0	0	$+2X$　$0\rightarrow T$	$2^{-2}(P_i+2X)$
1	0	1	$-X$　$1\rightarrow T$	$2^{-2}(P_i-X)$
1	1	0	$-X$　$1\rightarrow T$	$2^{-2}(P_i-X)$
1	1	1	$1\rightarrow T$	$2^{-2}(P_i)$

3X时，本次-X，下次+X！
部分积右移两位，相当于4X

T触发器用来记录下次是否要执行“+X”
“–X”运算用“$+[-X]_{补}$”实现！

原码两位乘法举例

已知 $[X]_{原}=0.111001$，$[Y]_{原}=0.100111$，用原码两位乘法计算$[X\times Y]_{原}$

解：先用无符号数乘法计算111001×100111，原码两位乘法过程如下：

$[|X|]_{补}=000\ 111001$，$[-|X|]_{补}=111\ 000111$

	P	Y	T	说明
	000 000000	100111	0	开始，$P_0=0$，T=0
	+111 000111			$y_5y_6T=110$，$-X$，T=1
	111 000111			P 和 Y 同时右移 2 位
	111 110001	11 1001	1	得 P_1
	+001 110010			$y_3y_4T=011$，$+2X$，T=0
	001 100011			P 和 Y 同时右移 2 位
	000 011000	1111 10	0	得 P_2
	+001 110010			$y_1y_2T=100$，$+2X$，T=0
→	010 001010			P 和 Y 同时右移 2 位
	000 100010	101111	0	得 P_3

加上符号位，得 $[X\times Y]_{原}=0.100010101111$

速度快，但代价也大

采用补码算术右移，与一位乘法不同，Why?

为模8补码形式(三位符号位)，Why?

若用模4补码，则P和Y同时右移2位时，得到的P3是负数，这显然是错误的！需要再增加一位符号。

补码乘法运算

用于对什么类型数据计算?

带符号整数！如int型

因为$[X \times Y]_{补} \neq [X]_{补} \times [Y]_{补}$，故不能直接用无符号数乘法计算。例如，若x=-5,求x*x=?

Booth's Algorithm推导如下：

假定：$[X]_{补}=x_{n-1}x_{n-2}\cdots\cdots x_1x_0$，$[Y]_{补}=y_{n-1}y_{n-2}\cdots\cdots y_1y_0$，求：$[X \times Y]_{补}=?$

基于以下补码性质：

$$Y=-y_{n-1}\cdot 2^{n-1}+y_{n-2}\cdot 2^{n-2}+\cdots\cdots y_1\cdot 2^1+y_0\cdot 2^0$$

令：$y_{-1}=0$，则：

当n=32时，$Y=-y_{31}\cdot 2^{31}+y_{30}\cdot 2^{30}+\cdots\cdots y_1\cdot 2^1+y_0\cdot 2^0+y_{-1}\cdot 2^0$

$$-y_{31}\cdot 2^{31}+(y_{30}\cdot 2^{31}-y_{30}\cdot 2^{30})+\cdots\cdots+(y_0\cdot 2^1-y_0\cdot 2^0)+y_{-1}\cdot 2^0$$

$$(y_{30}-y_{31})\cdot 2^{31}+(y_{29}-y_{30})\cdot 2^{30}+\cdots\cdots+(y_0-y_1)\cdot 2^1+(y_{-1}-y_0)\cdot 2^0$$

$$2^{-32}\cdot[X \times Y]_{补}=(y_{30}-y_{31})X\cdot 2^{-1}+(y_{29}-y_{30})X\cdot 2^{-2}+\cdots\cdots+(y_0-y_1)X\cdot 2^{-31}+(y_{-1}-y_0)X\cdot 2^{-32}$$

$$=2^{-1}(2^{-1}\ldots(2^{-1}(y_{-1}-y_0)X)+(y_0-y_1)X)+\ldots+(y_{30}-y_{31})X)$$

SKIP

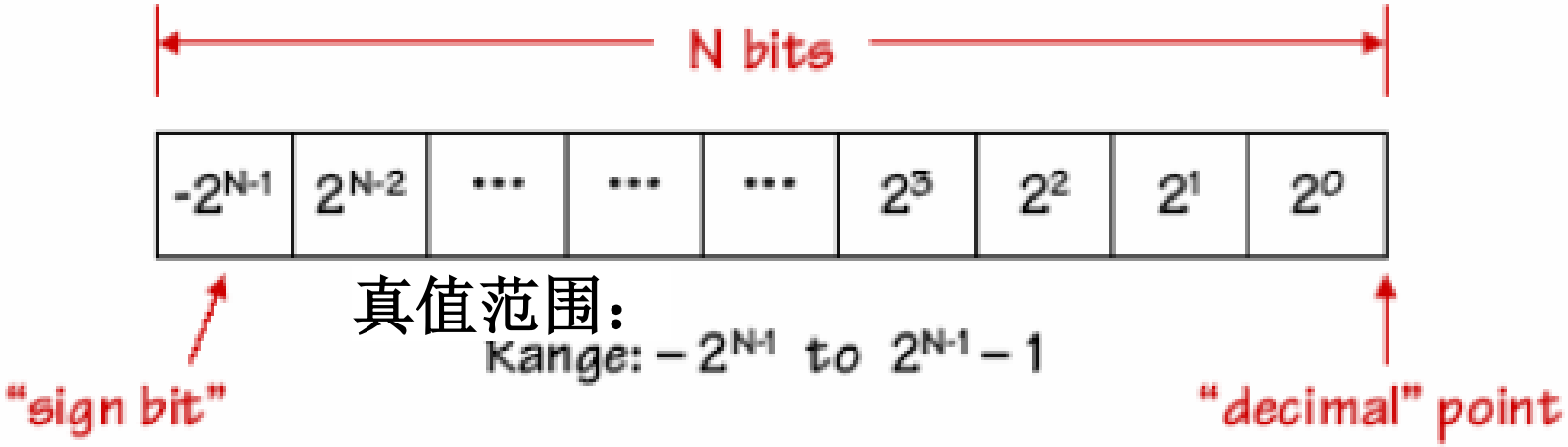
部分积公式：$P_i=2^{-1}(P_{i-1}+(y_{i-1}-y_i)X)$

符号与数值统一处理

如何求补码的真值

令：$[A]_{补}= a_{n-1}a_{n-2}\cdots\cdots a_1a_0$

则：$A= -a_{n-1}\cdot 2^{n-1}+a_{n-2}\cdot 2^{n-2}+\cdots\cdots a_1\cdot 2^1+ a_0\cdot 2^0$



8-bit 2's complement example:

$11010110 = -2^7 + 2^6 + 2^4 + 2^2 + 2^1 = -128 + 64 + 16 + 4 + 2 = -42$

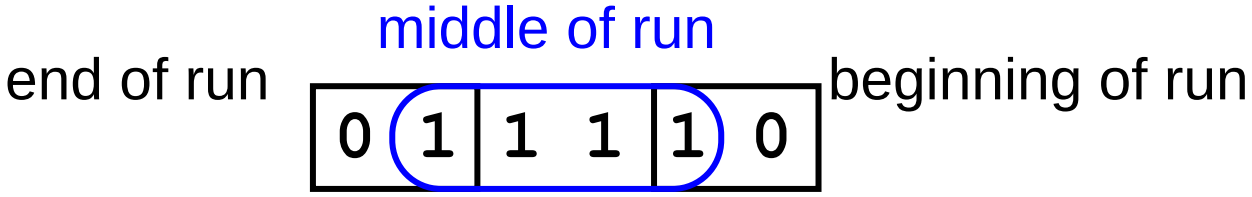
符号为0，则为正数，数值部分同

符号为1，则为负数，数值各位取反，末位加1

BACK

例如：补码“11010110”的真值为：-0101010=-(32+8+2)=-42

Booth's 算法实质



当前位	右边位	操作	Example
1	0	减被乘数	0001111000
1	1	加0 (不操作)	0001111000
0	1	加被乘数	0001111000
0	0	加0 (不操作)	0001111000

- 在“1串”中，第一个1时做减法，最后一个1做加法，其余情况只要移位。
- 最初提出这种想法是因为在Booth的机器上移位操作比加法更快！

同前面算法一样，将乘积寄存器右移一位。（这里是算术右移）

布斯算法举例

已知$[X]_{补}$ = 1 101，$[Y]_{补}$ = 0 110，计算$[X\times Y]_{补}$　$[-X]_{补}$ = 0011

X=-3，Y=6，X×Y=-18，$[X\times Y]_{补}$应等于11101110或结果溢出

P	Y	y_{-1}	说明
0 0 0 0	0 1 1 0	0	设 $y_{-1}=0$，$[P_0]_{补}=0$
	→ 1		$y_0\,y_{-1}=00$，P、Y 直接右移一位
0 0 0 0	0 0 1 1	0	得$[P_1]_{补}$
+ 0 0 1 1			$y_1\,y_0=10$，$+[-X]_{补}$
0 0 1 1	→ 1		P、Y 同时右移一位
0 0 0 1	1 0 0 1	1	得$[P_2]_{补}$
	→ 1		$y_2\,y_1=11$，P、Y 直接右移一位
0 0 0 0	1 1 0 0	1	得$[P_3]_{补}$ "11"时，直接移位
+ 1 1 0 1			$y_3\,y_2=01$，$+[X]_{补}$
1 1 0 1	→ 1		P、Y 同时右移一位
1 1 1 0	1 1 1 0	0	得$[P_4]_{补}$

如何判断结果是否溢出？

高4位是否全为符号位！

验证：当X×Y取8位时，结果 -0010010B=-18；取4位时，结果溢出

补码两位乘法（提速）

- 补码两位乘可用布斯算法推导如下：
  - $[P_{i+1}]_{补} = 2^{-1}([P_i]_{补} + (y_{i-1} - y_i)[X]_{补})$
  - $[P_{i+2}]_{补} = 2^{-1}([P_{i+1}]_{补} + (y_i - y_{i+1})[X]_{补})$

    $= 2^{-1}(2^{-1}([P_i]_{补} + (y_{i-1} - y_i)[X]_{补}) + (y_i - y_{i+1})[X]_{补})$

    $= 2^{-2}([P_i]_{补} + (y_{i-1} + y_i - 2y_{i+1})[X]_{补})$

- 开始置附加位y_{-1}为0，乘积寄存器最高位前面添加一位附加符号位0。
- 最终的乘积高位部分在乘积寄存器P中，低位部分在乘数寄存器Y中。
- 因为字长总是8的倍数，所以补码的位数n应该是偶数，因此，总循环次数为n/2。

y_{i+1}	y_i	y_{i-1}	操 作	迭代公式
0	0	0	0	$2^{-2}[P_i]_{补}$
0	0	1	$+[X]_{补}$	$2^{-2}\{[P_i]_{补}+[X]_{补}\}$
0	1	0	$+[X]_{补}$	$2^{-2}\{[P_i]_{补}+[X]_{补}\}$
0	1	1	$+2[X]_{补}$	$2^{-2}\{[P_i]_{补}+2[X]_{补}\}$
1	0	0	$+2[-X]_{补}$	$2^{-2}\{[P_i]_{补}+2[-X]_{补}\}$
1	0	1	$+[-X]_{补}$	$2^{-2}\{[P_i]_{补}+[-X]_{补}\}$
1	1	0	$+[-X]_{补}$	$2^{-2}\{[P_i]_{补}+[-X]_{补}\}$
1	1	1	0	$2^{-2}[P_i]_{补}$

补码两位乘法举例

- 已知 $[X]_{补}$ = 1 101， $[Y]_{补}$ = 0 110，用补码两位乘法计算$[X \times Y]_{补}$。
- 解：$[-X]_{补}$= 0 011，用补码二位乘法计算$[X \times Y]_{补}$的过程如下。

P_n	P	Y	y_{-1}	说明
0	0 0 0 0	0 1 1 0	0	开始，设$y_{-1} = 0$，$[P_0]_{补} = 0$
+ 0	0 1 1 0			$y_1y_0y_{-1}$ =100，+2$[-X]_{补}$
0	0 1 1 0		→ 2	P和Y同时右移二位
0	0 0 0 1	1 0 0 1	1	得$[P_2]_{补}$
+ 1	1 0 1 0			$y_3y_2y_1$ = 011，+2$[X]_{补}$
1	1 0 1 1		→ 2	P和Y同时右移二位
1	1 1 1 0	1 1 1 0		得$[P_4]_{补}$

因此 [X × Y]补=1110 1110，与一位补码乘法（布斯乘法）所得结果相同，但循环次数减少了一半。

验证：-3 ×6=-18 （-10010B）

整数的乘运算

结论：假定两个n位无符号整数x_u和y_u对应的机器数为X_u和Y_u，$p_u=x_u\times y_u$，p_u为n位无符号整数且对应的机器数为P_u；

两个n位带符号整数x_s和y_s对应的机器数为X_s和Y_s，$p_s=x_s\times y_s$，p_s为n位带符号整数且对应的机器数为P_s。

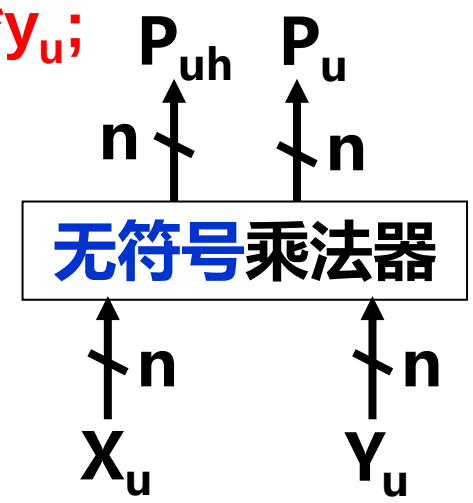
若$X_u=X_s$且$Y_u=Y_s$，则$P_u=P_s$。

无符号：若$P_{uh}=0$，则不溢出

带符号：若P_{sh}每位都等于P_s的最高位，则不溢出

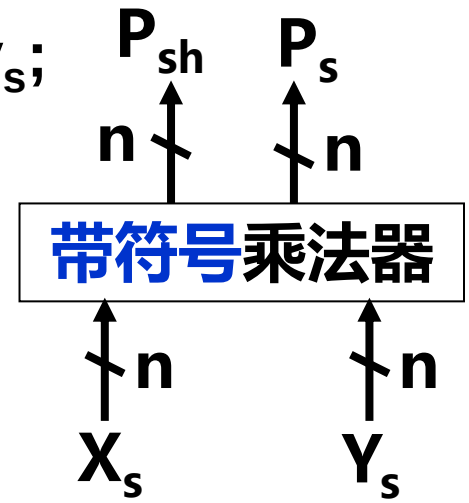
可用无符号乘来实现带符号乘，但高n位无法得到，故不能判断溢出。

$p_u=x_u*y_u$;



$P_{uh} \neq P_{sh}$

$p_s=x_s*y_s$;



整数的乘运算

- X*Y的高n位可以用来判断溢出，规则如下：
  - 无符号：若高n位全0，则不溢出，否则溢出
  - 带符号：若高n位全0或全1且等于低n位的最高位，则不溢出。

运算	x	X	y	Y	x×y	X×Y	p	P	溢出否
无符号乘	6	0110	10	1010	60	0011 1100	12	1100	溢出
带符号乘	6	0110	-6	1010	-36	1101 1100	-4	1100	溢出
无符号乘	8	1000	2	0010	16	0001 0000	0	0000	溢出
带符号乘	-8	1000	2	0010	-16	1111 0000	0	0000	溢出
无符号乘	13	1101	14	1110	182	1011 0110	6	0110	溢出
带符号乘	-3	1101	-2	1110	6	0000 0110	6	0110	不溢出
无符号乘	2	0010	12	1100	24	0001 1000	8	1000	溢出
带符号乘	2	0010	-4	1100	-8	1111 1000	-8	1000	不溢出

整数的乘运算

在计算机内部，一定有$x^2 \geq 0$吗?

若x是带符号整数，则不一定!

如x是浮点数，则一定!

例如，当 n=4 时, $5^2=-7<0$!

```
0101
×  0101
-------
0101
+ 0101
-------
00011001
```

结果溢出

只取低4位，值为-111B=-7

```
int mul(int x, int y)
{
    int z=x*y;
    return z;
}
```

高级语言程序如何判断z是正确值?

当 !x || z/x==y 为真时

编译器如何判断?

当 $-2^{n-1} \leq x*y < 2^{n-1}$ （不溢出）时

即：乘积的高n位为全0或全1，并等于低n位的最高位!

即：乘积的高n+1位为全0或全1

若x、y和z都改成unsigned类型，则判断方式为

乘积的高n位为全0，则不溢出

整数的乘运算

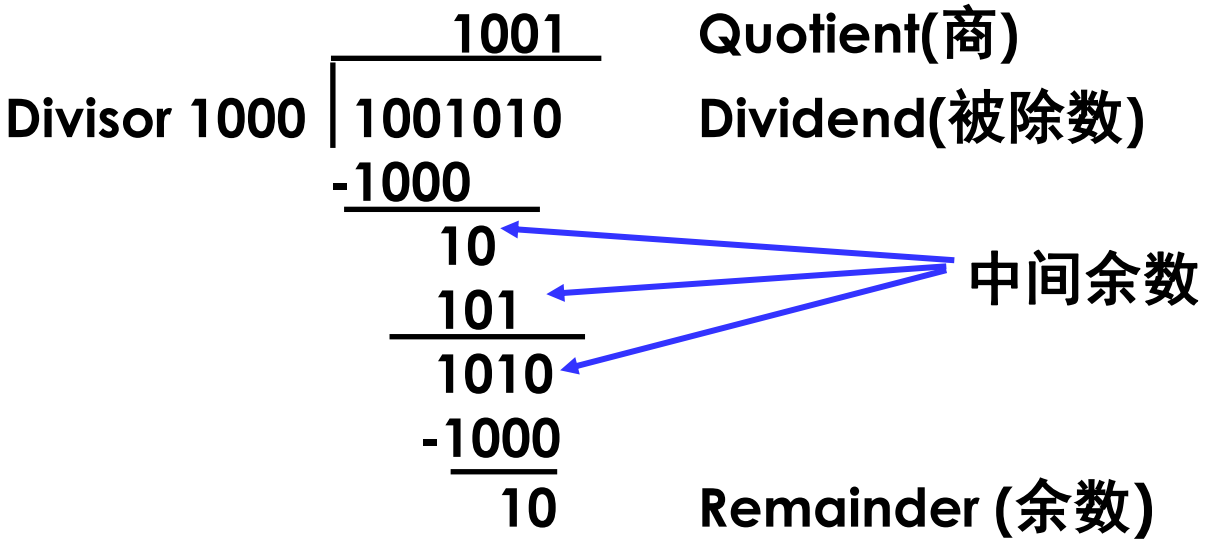
- 硬件可保留2n位乘积，故有些指令的乘积为2n位，可供软件使用
- 如果程序不采用防止溢出的措施，且编译器也不生成用于溢出处理的代码，就会发生一些由于整数溢出而带来的问题。
- 指令：分无符号数乘指令、带符号整数乘指令
- 乘法指令的操作数长度为n，而乘积长度为2n，例如：
  - IA-32中，若指令只给出一个操作数SRC，则另一个源操作数隐含在累加器AL/AX/EAX中，将SRC和累加器内容相乘，结果存放在AX（16位时）或DX-AX（32位时）或EDX-EAX（64位时）中。
  - MIPS中，mult会将两个32位带符号整数相乘，得到的64位乘积置于两个32位内部寄存器Hi和Lo中，因此，可以根据Hi寄存器中的每一位是否等于Lo寄存器中的第一位来进行溢出判断。
  - RISC-V中，用“mul rd, rs1, rs2”获得低32位乘积并存入结果寄存器rd中；mulh、mulhu指令分别将两个乘数同时按带符号整数、同时按无符号整数相乘后，得到的高32位乘积存入rd中

乘法指令可生成溢出标志，编译器也可使用2n位乘积来判断是否溢出！

主 要 内 容

- 定点数加减运算
  - 补码加减运算
  - 原码加减运算
  - 移码加减运算
- 定点数乘法运算
  - 原码乘法运算
  - 补码乘法运算
  - 快速乘法器
- 定点数除法运算
  - 原码除法运算
  - 补码除法运算

除法（Divide）：Paper & Pencil



- 手算除法的基本要点
  - ① 被除数与除数相减，够减则上商为1；不够减则上商为0。
  - ② 每次得到的差为中间余数，将除数右移后与上次的中间余数比较。用中间余数减除数，够减则上商为1；不够减则上商为0。
  - ③ 重复执行第②步，直到求得的商的位数足够为止。

定点除法运算

- 除前预处理

  ①若被除数=0且除数≠0，或定点整数除法时|被除数|<|除数|，则商为0，不再继续

  ②若被除数≠0、除数=0，则发生“除数为0”异常（浮点数时为∞）

  （若浮点除法被除数和除数都为0，则有些机器产生一个不发信号的NaN，即“quiet NaN”

  只有当被除数和除数都≠ 0，且商≠ 0时，才进一步进行除法运算。

- 计算机内部无符号数除法运算
  - 与手算一样，通过被除数（中间余数）减除数来得到每一位商

    够减上商1；不够减上商0（从msb→lsb得到各位商）
  - 基本操作为减（加）法和移位，故可与乘法合用同一套硬件

两个n位数相除的情况：

(1)定点正整数（即无符号数 ）相除：在被除数的高位添n个0

(2)定点正小数（即原码小数）相除：在被除数的低位添加n个0

这样，就将所有情况都统一为：一个2n位数除以一个n位数

第一次试商为1时的情况

NJUAI 南京大学 人工智能学院

问题：第一次试商为1，说明什么?　　**商有n+1位数，因而溢出!**

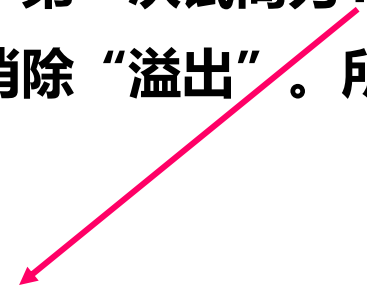
通常意义下，若是2n位除以n位的无符号整数运算，则说明将会得到n+1位的商，因而结果“溢出”（即：无法用n位表示商）。

例：1111 1111/1111 = 1 0001

若是两个n位数相除，被除数高位扩展0，则第一位商为0，肯定不会溢出

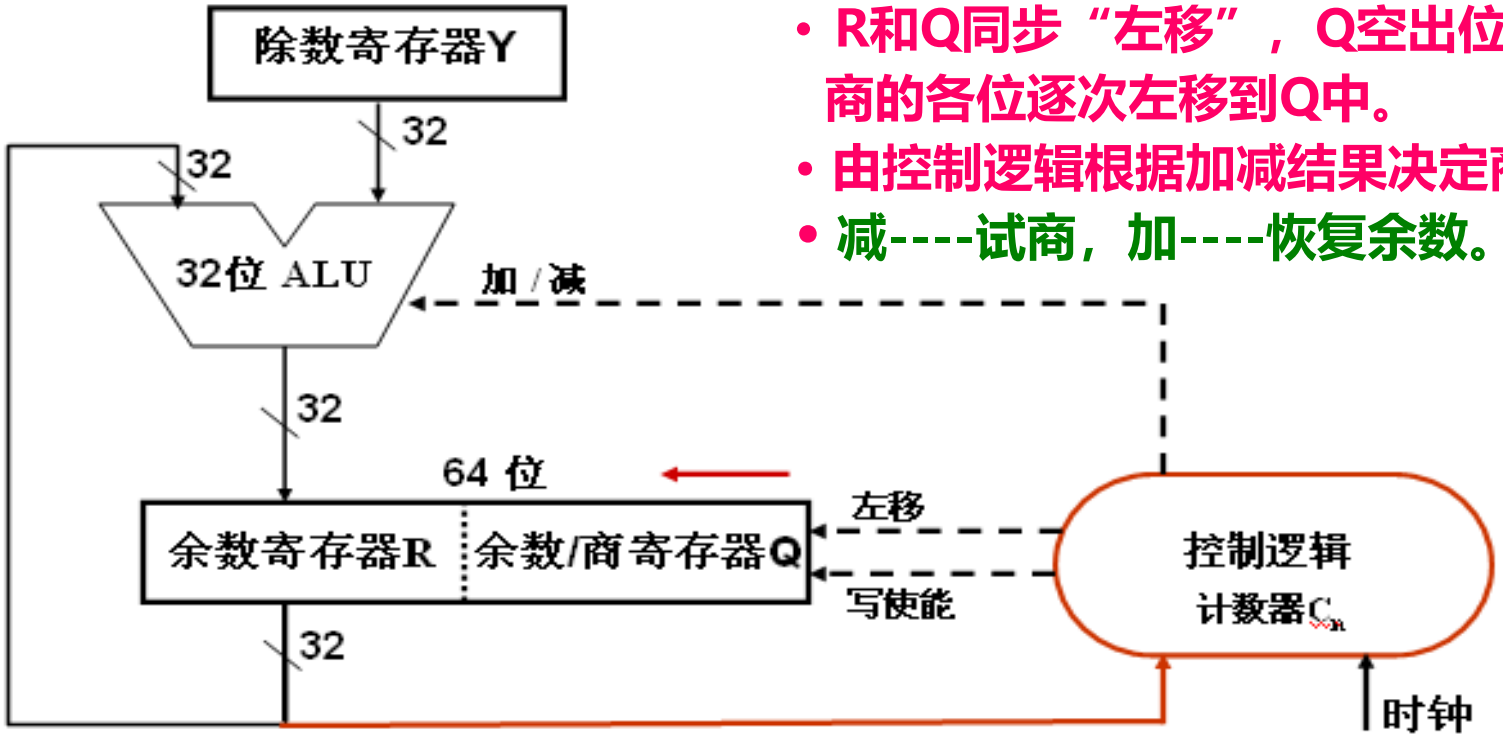
最大商为:0000 1111/0001=1111

若是浮点数中尾数原码小数运算，第一次试商为1，则说明尾数部分有“溢出”，可通过浮点数的“右规”消除“溢出”。所以，在浮点数运算器中，第一次得到的商“1” 要保留。



例：0.11110000/0.1000=+1.1110

无符号数除法算法的硬件实现



- R和Q同步“左移”，Q空出位上“商”，商的各位逐次左移到Q中。
- 由控制逻辑根据加减结果决定商为0还是1
- 减----试商，加----恢复余数。

- 除数寄存器Y：存放除数。
- 余数寄存器R：初始时高位部分为高32位被除数；结束时是余数。
- 余数/商寄存器Q：初始时为低32位被除数；结束时是32位商。
- 循环次数计数器Cn：存放循环次数。初值是32，每循环一次，Cn减1，当Cn=0时，除法运算结束。
- ALU：除法核心部件。在控制逻辑控制下，对于寄存器R和Y的内容进行“加/减”运算，在“写使能”控制下运算结果被送回寄存器R。

Divide Algorithm example

这里是两个n位无符号数相除，肯定不会溢出，故余数先左移而省略判断溢出过程。

验证：7 / 2 = 3 余 1　　–D = 1110

R：被除数（中间余数）；D：除数

	D: 0010	R: 0000 0111
Shl R	D: 0010	R: 0000 1110
R = R–D	D: 0010	R: 1110 1110
+D, sl R, 0	D: 0010	R: 0001 1100
R = R–D	D: 0010	R: 1111 1100
+D, sl R, 0	D: 0010	R: 0011 1000
R = R–D	D: 0010	R: 0001 1000
sl R, 1	D: 0010	R: 0011 0001
R = R–D	D: 0010	R: 0001 0001
sl R, 1	D: 0010	R: 0010 0011
Shr R(rh)	D: 0010	R: 0001 0011

从例子可看出：

每次上商为0时，需做加法以“恢复余数”。所以，称为“恢复余数法

也可在下一步运算时把当前多减的除数补回来。这种方法称为“不恢复余数法”，又称“加减交替法”。

开始余数先左移了一位，故最后余数需向右移一位

不恢复余数除法(加减交替法)

恢复余数法可进一步简化为“加减交替法”

根据恢复余数法(设B为除数，R_i为第i次中间余数)，有：

若$R_i<0$,则商上“0”，并做加法恢复余数，即：

- $R_{i+1}=2(R_i+2^n|B|)-2^n|B|=2R_i+2^n|B|$ （“负，0，加”）

若$R_i>=0$,则商上“1”，不需恢复余数，即：

- $R_{i+1}=2R_i-2^n|B|$ （“正，1，减”）

省去了恢复余数的过程

- 注意：最后一次上商为“0”的话，需要“纠余”处理，即把试商时被减掉的除数加回去，恢复真正的余数。

不恢复余数法也称为加减交替法

Divide Algorithm example

验证：7 / 2 = 3 余 1　　–D = 1110

R：被除数（中间余数）；D：除数

“不恢复余数法”例子

	D: 0010	R: 0000 0111
Shl R	D: 0010	R: 0000 1110
R = R–D	D: 0010	R: 1110 1110
sl R, 0	D: 0010	R: 1101 1100
R = R+D	D: 0010	R: 1111 1100
sl R, 0	D: 0010	R: 1111 1000
R = R+D	D: 0010	R: 0001 1000
sl R, 1	D: 0010	R: 0011 0001
R = R–D	D: 0010	R: 0001 0001
sl R, 1	D: 0010	R: 0010 0011
Shr R(rh)	D: 0010	R: 0001 0011

开始余数先左移了一位，故最后余数需向右移一位

带符号数除法

- 原码除法
  - 商符和商值分开处理
    - 商的数值部分由无符号数除法求得
    - 商符由被除数和除数的符号确定：同号为0，异号为 1
  - 余数的符号同被除数的符号
- 补码除法
  - 方法1：同原码除法一样，先转换为正数（类似原码表示），先用无符号数除法，然后修正商和余数。
  - 方法2：直接用补码除法，符号和数值一起进行运算，商符直接在运算中产生。

若是两个n位补码整数除法运算，则被除数进行符号扩展。

若被除数为2n位，除数为n位，则被除数无需扩展。

原码除法举例

已知 $[X]_{原}$ = 0.1011

$[Y]_{原}$ = 1.1101

用恢复余数法计算$[X/Y]_{原}$

解：商的符号位：0 ⊕ 1 = 1

减法操作用补码加法实现，是否够减通过中间余数的符号来判断，所以中间余数要加一位符号位。

$[|X|]_{补}$ = 0.1011

$[|Y|]_{补}$ = 0.1101

$[-|Y|]_{补}$ = 1.0011

小数在低位扩展0

思考：若实现无符号数相除，即1011除以1101，则有何不同？结果是什么？

被除数高位补0，1011除以1101，结果等于0

余数寄存器R	余数/商寄存器Q	说　明
01011	0000□	开始 $R_0=X$
+10011		$R_1=X-Y$
11110	00000	$R_1<0$，则 $q_4=0$
+01101		恢复余数：$R_1=R_1+Y$
01011		得 R_1
10110	0000□	$2R_1$（R和Q同时左移，空出一位商）
+10011		$R_2=2R_1-Y$
01001	00001	$R_2>0$，则 $q_3=1$
10010	0001□	$2R_2$ （R和Q同时左移，空出一位商）
+10011		$R_3=2R_2-Y$
00101	00011	$R_3>0$，则 $q_2=1$
01010	0011□	$2R_3$ （R和Q同时左移，空出一位商）
+10011		$R_3=2R_2-Y$
11101	00110	$R_4<0$，则 $q_1=0$
+01101		恢复余数：$R_4=R_4+Y$
01010	00110	得 R_4
10100	0110□	$2R_4$ （R和Q同时左移，空出一位商）
+10011		$R_5=2R_4-Y$
00111	01101	$R_5>0$，则 $q_0=1$

商的最高位为0，说明没有溢出，商的数值部分为：1101。

所以，$[X/Y]_{原}$=1.1101 (最高位为符号位)，余数为 0.0111×2^{-4}。

用于判断是否溢出

若求$[Y/X]_{原}$结果如何？

原码除法举例

已知 $[X]_{原} = 0.1011$

$[Y]_{原} = 1.1101$

用加减交替法计算$[X/Y]_{原}$

解：$[|X|]_{补} = 0.1011$

$[|Y|]_{补} = 0.1101$

$[-|Y|]_{补} = 1.0011$

余数寄存器R	余数/商寄存器Q	说　明
01011	0000□	开始 $R_0=X$
+10011		$R_1=X-Y$
11110	00000	$R_1<0$，则 $q_4=0$，没有溢出
11100	0000□	$2R_1$（R和Q同时左移，空出一位商）
+01101		$R_2=2R_1+Y$
01001	00001	$R_2>0$，则 $q_3=1$
10010	0001□	$2R_2$（R和Q同时左移，空出一位商）
+10011		$R_3=2R_2-Y$
00101	00011	$R_3>0$，则 $q_2=1$
01010	0011□	$2R_3$（R和Q同时左移，空出一位商）
+10011		$R_3=2R_2-Y$
11101	00110	$R_4<0$，则 $q_1=0$
11010	0110□	$2R_4$（R和Q同时左移，空出一位商）
+01101		$R_5=2R_4+Y$
00111	01101	$R_5>0$，则 $q_0=1$

“加减交替法”的要点：

负、0、加

正、1、减

得到的结果与恢复余数法一样！

用被除数（中间余数）减除数试商时，怎样确定是否“够减”？

中间余数的符号！（正数-正数）

补码除法能否这样来判断呢？

不能，因为符号可能不同！

补码除法

NJUAI 南京大学 人工智能学院

- 补码除法判断是否“够减”的规则

（1）当被除数（或中间余数）与除数同号时，做减法，若新余数的符号与除数符号一致表示够减，否则为不够减；

（2）当被除数（或中间余数）与除数异号时，做加法，若得到的新余数的符号与除数符号一致表示不够减，否则为够减。

上述判断规则归纳如下：

中间余数R的符号	除数Y的符号	同号：新中间余数=R–Y（同号为正商）		异号：新中间余数=R+Y（异号为负商）	
		0	1	0	1
0	0	够减	不够减		
0	1			够减	不够减
1	0			不够减	够减
1	1	不够减	够减		

总结：余数变号不够减，不变号够减

实现补码除法的基本思想

NJUAI 南京大学 人工智能学院

从上表可得到补码除法的基本算法思想：

（1）运算规则：

当被除数（或中间余数）与除数同号时，做减法；

异号时，做加法。

（2）上商规则：

若余数符号不变，则够减，商1；否则不够减，商0 。

（3）修正规则：

若被除数与除数符号一致，则商为正。 此时，“够减，商1；不够减，商0，故上商规则正确，无需修正”

若被除数与除数符号不一致，则商为负。此时，“够减，商0；不够减，商1，故上商规则相反，需修正”

即：若商为负值，则需要“各位取反，末位加1”来得到真正的商

补码除法也有：恢复余数法和不恢复余数法

SKIP

补码恢复余数法

- **两个n位带符号整数相除算法要点：**

(1) 操作数的预置：

除数装入除数寄存器Y，被除数经符号扩展后装入余数寄存器R和余数/商寄存器Q

(2) R和Q同步串行左移一位。

(3) 若R与Y同号，则R = R–Y；否则R = R+Y，并按以下规则确定商值q_0：

① 若中间余数R = 0或R操作前后符号未变，表示够减，则q_0置1，转下一步；

② 若操作前后R的符号已变，表示不够减，则q_0置0，恢复R值后转下一步；

(4) 重复第(2)和第(3)步，直到取得n位商为止。

(5) 若被除数与除数同号，则Q中就是真正的商；否则，将Q求补后是真正的商。

（即：若商为负值，则需要“各位取反，末位加1”来得到真正的商）

(6) 余数在R中。

问题：如何恢复余数？通过“做加法”来恢复吗？

无符号数（或原码）除法通过“做加法”恢复余数，但补码不是!

补码：若原来为R = R–Y，则执行R = R+Y来恢复余数；

若原来是R = R+Y，则执行R = R–Y来恢复余数。

举例：7/3=?　(-7)/3=?

被除数：0000 0111　除数 0011

	A	Q	M=0011	
	0000	0111		
←	0000	1110		
+	1101			减（同号）
	1101	1110		
+	0011			恢复(加)商0
	0000	1110		
←	0001	1100		
+	1101			减
	1110	1100		
+	0011			恢复(加)商0
	0001	1100		
←	0011	1000		
+	1101			减
	0000	1000		符同商1
←	0001	0001		
+	1101			减
	1110	0010		
+	0011			恢复(加)商0
	0001	0010		

余:0001/商:0010

验证：7/3 = 2，余数为1

被除数：1111 1001　除数 0011

	A	Q	M=0011	
	1111	1001		
←	1111	0010		
+	0011			加（异号）
	0010	0010		
+	1101			恢复(减)商0
	1111	0010		
←	1110	0100		
+	0011			加
	0001	0100		
+	1101			恢复(减)商0
	1110	0100		
←	1100	1000		
+	0011			加
	1111	1001		符同商1
←	1111	0010		
+	0011			加
	0010	0010		
+	1101			恢复(减)商0
	1111	0010		

商为负数，需求补：0010→1110

余:1111/商:1110

验证：-7/3 =- 2，余数为-1

补码不恢复余数法

- **算法要点：**

(1) 操作数的预置：

除数装入除数寄存器Y，被除数经符号扩展后装入余数寄存器R和余数/商寄存器Q。

(2) 根据以下规则求第一位商q_n：

若被除数X与Y同号，则R1=X–Y；否则R1 =X+Y，并按以下规则确定商值q_n：

① 若新的中间余数R1与Y同号，则q_n置1，转下一步；

② 若新的中间余数R1与Y异号，则q_n置0，转下一步；

q_n用来判断是否溢出，而不是真正的商。以下情况下会发生溢出：

若X与Y同号且上商q_n=1，或者，若X与Y异号且上商q_n = 0。

判断是否同号与恢复余数法不同，不是新老余数之间！而是余数和除数之间

(3) 对于i =1到n，按以下规则求出n位商：

① 若Ri与Y同号，则q_{n-i}置1，Ri+1 = 2Ri –[Y]补，i = i +1；

② 若Ri与Y异号，则q_{n-i}置0，Ri+1 =2Ri+[Y]补，i = i +1；

(4) 商的修正：最后一次Q寄存器左移一位，将最高位q_n移出，最低位置上商q_0。若被除数与除数同号， Q中就是真正的商；否则，将Q中商的末位加1。 商已经是“反码”

(5) 余数的修正：若余数符号同被除数符号，则不需修正，余数在R中；否则，按下列规则进行修正：当被除数和除数符号相同时，最后余数加除数；否则，最后余数减除数。

补码不恢复余数法也有一个六字口诀“同、1、减；异、0、加”。

其运算过程也呈加/减交替方式，因此也称为“加减交替法”。

举例：-9/2

将X=-9和Y=2分别表示成5位补码形式为：

[X]补 = 1 0111

[Y]补 = 0 0010

被除数符号扩展为：

[X]补=11111 10111

[–Y] 补 = 1 1110

同、1、减

异、0、加

X/Y= – 0100B = – 4

余数为 –0001B = –1

将各数代入公式：

“除数×商+余数= 被除数”进行验证，得

2×(–4) +(–1) = –9

余数寄存器R	余数/商寄存器Q	说　明
11111	10111	开始 $R_0=[X]$
+00010		$R_1=[X]+[Y]$
00001	10111	R_1 与[Y]同号，则 $q_5=1$
00011	01111	$2R_1$（R和Q同时左移，空出一位上商1）
+11110		$R_2=2R_1+[-Y]$
00001	01111	R_2 与[Y]同号，则 $q_4=1$，
00010	11111	$2R_2$（R和Q同时左移，空出一位上商1）
+11110		$R_3=2R_2+[-Y]$
00000	11111	R_3 与[Y]同号，则 $q_3=1$
00001	11111	$2R_3$（R和Q同时左移，空出一位上商1）
+11110		$R_4=2R_3+[-Y]$
11111	11111	R_4 与[Y]异号，则 $q_2=0$
11111	11110	$2R_4$（R和Q同时左移，空出一位上商0）
+00010		$R_5=2R_4+[Y]$
00001	11110	R_5 与[Y]同号，则 $q_1=1$，
00011	11101	$2R_5$（R和Q同时左移，空出一位上商1）
+11110		$R_6=2R_5+[-Y]$
00001	11011	R_6 与[Y]同号，则 $q_0=1$，Q左移，空出一位上商1
+11110	+ 1	商为负数，末位加1；减除数以修正余数
11111	11100	

所以，$[X/Y]_{补}$=11100。 余数为11111。

编译器遇到x/2时会如何做？ 右移一位吗！

变量与常数之间的除运算

- 不能整除时，采用朝零舍入，即截断方式
  - 无符号数、带符号正整数（地板）：移出的低位直接丢弃
  - 带符号负整数（天板）：加偏移量(2^k-1)，然后再右移k位，低位截断（这里K是右移位数）

举例：

无符号数 14/4=3：0000 1110>>2=0000 0011

带符号负整数 -14/4=-3

若直接截断，则 1111 0010 >>2=1111 1100=-4≠-3

应先纠偏，再右移: k=2, 故$(-14+2^2-1)/4=-3$

即： 1111 0010+0000 0011=1111 0101

1111 0101>>2=1111 1101=-3

-9/2: 10111+00001=11000

11000>>1=11100=-4

变量与常数之间的除运算—举例

- 假设*x*为一个int型变量，请给出一个用来计算*x*/32的值的函数div32。要求不能使用除法、乘法、模运算、比较运算、循环语句和条件语句，可以使用右移、加法以及任何按位运算。

解：若x为正数，则将x右移k位得到商；若x为负数，则x需要加一个偏移量(2^k-1)后再右移k位得到商。因为$32=2^5$，所以 k=5。

即结果为: (x>=0 ? x : (x+31))>>5

但题目要求不能用比较和条件语句，因此要找一个计算偏移量b的方式

这里，x为正时b=0，x为负时b=31. 因此，可以从x的符号得到b

x>>31 得到的是32位符号，取出最低5位，就是偏移量b。

```
int div32(int x)
{  /* 根据x的符号得到偏移量b */
    int b=(x>>31) & 0x1F;
    return (x+b)>>5;
}
```

RISC-V中整数的乘、除运算处理

- 乘法指令: mul，mulh，mulhu，mulhsu
  - mul rd, rs1, rs2：将低32位乘积存入结果寄存器rd
  - mulh、mulhu：将两个乘数同时按带符号整数（mulh）、同时按无符号整数（mulhu）相乘，高32位乘积存入rd中
  - mulhsu：将两个乘数分别作为带符号整数和无符号整数相乘后得到的高32位乘积存入rd中
  - 得到64位乘积需要两条连续的指令，其中一定有一条是mul指令，实际执行时只有一条指令
  - 两种乘法指令都不检测溢出，而是直接把结果写入结果寄存器。由软件根据结果寄存器的值自行判断和处理溢出

问题：如何判断溢出?

- 除法指令: div ，divu，rem，remu
  - div / rem：按带符号整数做除法，得到商 / 余数
  - divu / remu：按无符号整数做除法，得到商 / 余数
- RISC-V指令不检测和发出异常，而是由系统软件自行处理

定点运算部件

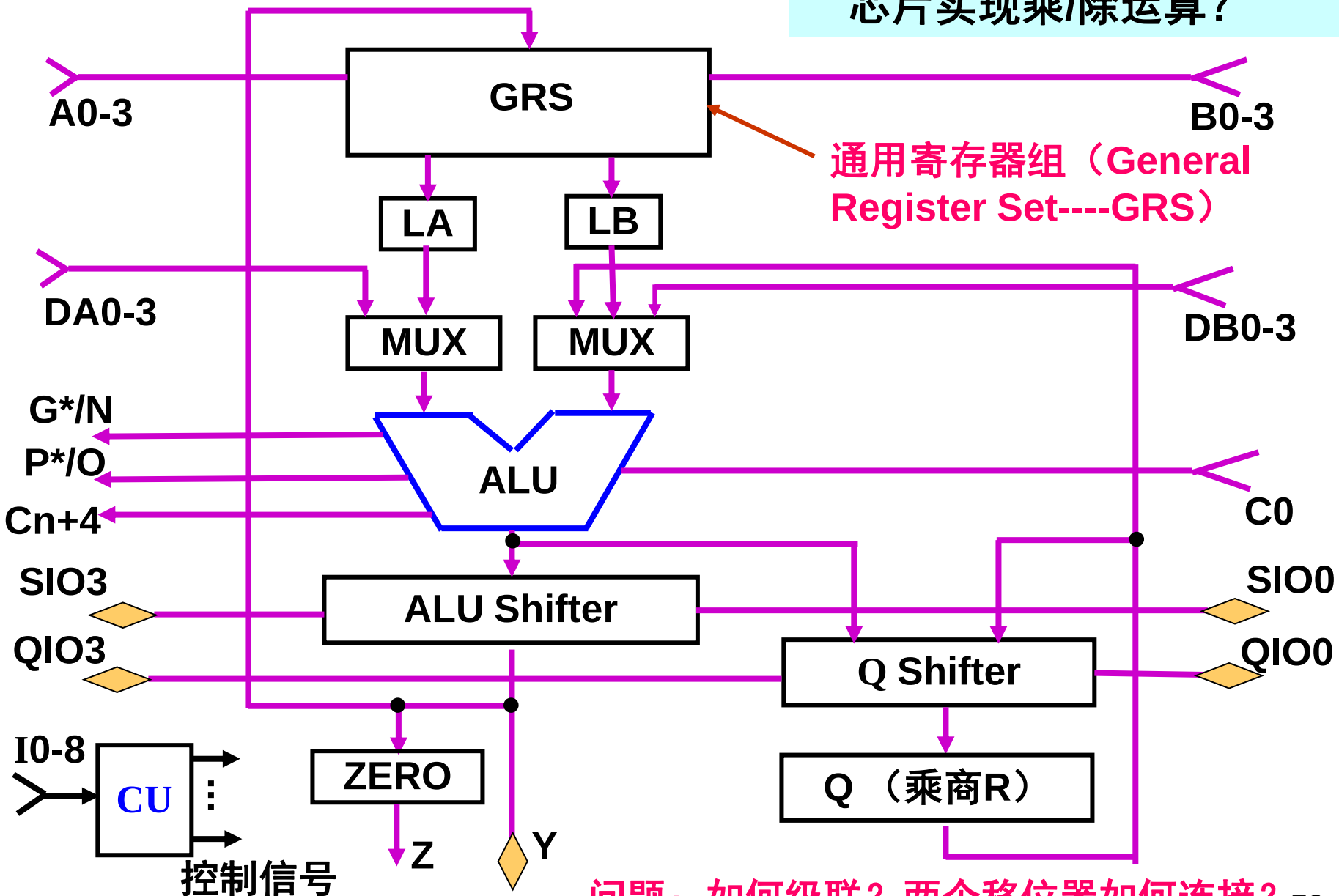
- 综合考虑各类定点运算算法后，发现：
  - 所有运算都可通过“加”和“移位”操作实现

    以一个或多个ALU（或加法器）为核心，加上移位器和存放中间临时结果的若干寄存器，在相应控制逻辑的控制下，可以实现各种运算。
- 运算部件通常指ALU、移位器、寄存器组，加上用于数据选择的多路选择器和实现数据传送的总线等构成的一个运算数据通路。
  - 可用专门运算器芯片实现（如：4位运算器芯片AM2901）
  - 可用若干芯片级联实现（如4个AM2901构成16位运算器）
  - 现代计算机把运算数据通路和控制器都做在CPU中，为实现高级流水线，CPU中有多个运算部件，通常称为“功能部件”或“执行部件”。

“运算器（Operate Unit）”、“运算部件（Operate Unit）”、“功能部件（Function Unit）”、“执行部件（Execution Unit）”和“数据通路（DataPath）”的含义基本上一样，只是强调的侧重不同

定点运算器芯片举例-AM2

问题：如何用AM2901A芯片实现乘/除运算？



问题：如何级联？两个移位器如何连接？

AM2901A的功能和结构

- 核心是ALU(含移位器)，可实现A+B、A–B、B–A和与、或、异或等
  - 进位入C0、进位出Cn+4、组进位传递/溢出P*/O、组进位生成/符号G*/N
    - 串行级联时，C0和Cn+4用来串行进位传递
    - 多级级联时，后两个信号用作组进位传递信号P*和组进位生成信号G*
  - ALU的操作数来自主存或寄存器，B输入端还可以是Q寄存器
- 4位双口GRS(16个)，一个写入口、两个读口A和B
  - A_0–A_3为读口A的编号，B_0–B_3为写口或读口B的编号
  - A和B口可同时读出，分别LA和LB送到多路选择器MUX的输入端
- 一个Q寄存器和Q移位寄存器，主要用于实现乘/除运算
  - 乘法的部分积和除法的中间余数都是双倍字长，需放到两个单倍字长寄存器中，并对其同时串行左移（除法）或右移（乘法）
  - Q寄存器就是乘数寄存器或商寄存器。因此，也被称为Q乘商寄存器
  - ALU移位器和Q移位器一起进行左移或右移，移位后，ALU移位器内容送ALU继续进行下次运算，而Q移位器内容送Q乘商寄存器。
- 将ALU的结果进行判“0”后可通过Z输出端将“零”标志信息输出，
- ALU、MUX、移位器等的控制信号来自CCU，通过对指令操作码I_0-I_8译码，得到控制信号

AM2901A的功能和结构

- 思考题：如何用AM2901A芯片实现乘/除运算?

① R0被预置为0 / 被除数高位，Q寄存器被预置为乘数 / 被除数低位，R1被预置为被乘数 / 除数

② 控制ALU执行“加”或“减”，并使ALU移位器和Q移位器同时 “右移” / “左移”

（移位后ALU移位器中是高位部分积 / 中间余数，Q移位器中是低位部分积 / 中间余数）

③ ALU移位器送ALU的A端，Q移位器送Q寄存器后，再送回Q移位器

④ 反复执行第②、③两个步骤，直到得到所有乘积位或商

BACK

第二讲小结

逻辑运算、移位运算、扩展运算等电路简单

主要考虑算术运算

- 定点运算涉及的对象

  无符号数；带符号整数(补码)；原码小数；移码整数

- 定点运算：(ALU实现基本算术和逻辑运算，ALU+移位器实现其他运算)

  补码加/减：符号位和数值位一起运算，减法用加法实现。同号相加时可能溢出

  原码加/减：符号位和数值位分开运算，用于浮点数尾数加/减运算

  移码加减：移码的和、差等于和、差的补码，用于浮点数阶码加/减运算

第二讲小结

乘法运算：

无符号数乘法：“加”+“右移”

原码（一位/两位）乘法：符号和数值分开运算，数值部分用无符号数乘法实现，用于浮点数尾数乘法运算。

补码（一位/两位）乘法：符号和数值一起运算，采用Booth算法。

快速乘法器：流水化乘法器、阵列乘法器

除法运算：

无符号数除法：用“加/减”+“左移”，有恢复余数和不恢复余数两种。

原码除法：符号和数值分开，数值部分用无符号数除法实现，用于浮点数尾数除法运算。

补码除法：符号位和数值位一起。有恢复余数和不恢复余数两种。

快速除法器：很难实现流水化除法器，可实现阵列除法器，或用乘法实现

- 定点部件：ALU、GRS、MUX、Shifter、Q寄存器等，CU控制执行

第三讲：浮点数运算

NJUAI 南京大学 人工智能学院

主　要　内　容

- 指令集中与浮点运算相关的指令（以MIPS为例）
  - 涉及到的操作数
    - 单精度浮点数
    - 双精度浮点数
  - 涉及到的运算
    - 算术运算：加 / 减 / 乘 / 除
- 浮点数加减运算
- 浮点数乘除运算
- 浮点数运算的精度问题

有关Floating-point number的问题

实现一套浮点数运算指令，要解决的问题有：

Issues:

- Representation(表示)：
  Normalized form (规格化形式) 和 Denormalized form
  单精度格式 和 双精度格式
- Range and Precision (表数范围和精度)
- Arithmetic (+, -, *, /)
- Rounding(舍入)
- Exceptions (e.g., divide by zero, overflow, underflow)
  (异常处理：如除数为0，上溢，下溢等)
- Errors (误差) 与精度控制

浮点数运算及结果

设两个规格化浮点数分别为 $A=M_a \cdot 2^{Ea}$　$B=M_b \cdot 2^{Eb}$,则:

$A \pm B = (M_a \pm M_b \cdot 2^{-(Ea-Eb)}) \cdot 2^{Ea}$　(假设Ea>=Eb)

$A*B = (M_a * M_b) \cdot 2^{Ea+Eb}$

$A/B = (M_a / M_b) \cdot 2^{Ea-Eb}$

1.5+1.5=?

1.5-1.0=?

上述运算结果可能出现以下几种情况:

SP最大指数为多少? 127

阶码上溢: 一个正指数超过了最大允许值 =〉+∞/-∞/溢出

阶码下溢: 一个负指数比最小允许值还小 =〉+0/-0

SP最小指数呢?

-126-23

(非规格数)

尾数溢出: 最高有效位有进位 =〉右规

尾数溢出，结果不一定溢出

非规格化尾数: 数值部分高位为0 =〉左规

右规或对阶时，右段有效位丢失 =〉尾数舍入　运算过程中添加保护位

IEEE建议实现时为每种异常情况提供一个自陷允许位。若某异常对应的位为1，则发生相应异常时，就调用一个特定的异常处理程序执行。

IEEE754标准规定的五种异常情况

① 无效运算（无意义）

- 运算时有一个数是非有限数，如：

  加 / 减∞、0 x ∞、 ∞/∞等

- 结果无效，如：

  源操作数是NaN、0/0、x REM 0、 ∞ REM y 等

② 除以0（即：无穷大）

③ 数太大（阶码上溢）：对于SP，阶码 E >1111 1110 （阶大于127）

④ 数太小（阶码下溢）：对于SP，阶码 E < 0000 0001（阶小于-126-23 ）

⑤ 结果不精确（舍入时引起），例如1/3，1/10等不能精确表示成浮点数

上述情况硬件可以捕捉到，因此这些异常可设定让硬件处理，也可设定让软件处理。让硬件处理时，称为硬件陷阱。

注：硬件陷阱：事先设定好是否要进行硬件处理（即挖一个陷阱），当出现相应异常时，就由硬件自动进行相应的异常处理（掉入陷阱）。

回顾：浮点数除0的问题

南京大学 人工智能学院

```
#include <conio.h>
#include <stdio.h>
int    main()
{
        int    a=1,b=0;
        printf( "Division    by    zero:%d\n ",a/b);
        getchar();
        return    0;
}
int    main()
{
        double    x=1.0,y=-1.0,z=0.0;
        printf( "division    by    zero:%f    %f\n ",x/z,y/z);
        getchar();
        return    0;
}
```

这是网上的一个帖子

为什么整数除0会发生异常?

为什么浮点数除0不会出现异常?

浮点运算中，一个有限数除以0，结果为正无穷大（负无穷大）

问题一:为什么整除int型会产生错误? 是什么错误?

二:用double型的时候结果为1.#INF00和-1.#INF00，作何解释???

回顾：浮点数加/减运算

NJUAI 南京大学 人工智能学院

- 十进制科学计数法的加法例子

$1.123 \times 10^5 + 2.560 \times 10^2$

其计算过程为：

$1.123 \times 10^5 + 2.560 \times 10^2 = 1.123 \times 10^5 + 0.002560 \times 10^5$

$=(1.123 + 0.00256) \times 10^5 = 1.12556 \times 10^5$

$=1.126 \times 10^5$

进行尾数加减运算前，必须“对阶”！最后还要考虑舍入

计算机内部的二进制运算也一样！

- “对阶”操作：目的是使两数阶码相等
  - 小阶向大阶看齐，阶小的那个数的尾数右移，右移位数等于两个阶码差的绝对值
  - IEEE 754尾数右移时，要将隐含的“1”移到小数部分，高位补0，移出的低位保留到特定的“附加位”上

浮点数加/减运算–对阶

问题：如何对阶?

通过计算$[\Delta E]_{补}$来判断两数的阶差：

$[\Delta E]_{补} = [Ex–Ey]_{补} = [Ex]_{移} + [–[Ey]_{移}]_{补} \pmod{2^n}$

问题：在ΔE为何值时无法根据$[\Delta E]_{补}$来判断阶差？ 溢出时！

例：4位移码，Ex=7，Ey=-7，则$[\Delta E]_{补}$=1111+1111=1110，ΔE<0,错

问题：对IEEE754 SP格式来说，|ΔE|大于多少时，结果就等于阶大的那个数（即小数被大数吃掉）？ 24！

1.xx…x → 0.00…01xx…x(右移24位后，尾数变为0)

问题：IEEE754 SP格式的偏置常数是127，这会不会影响阶码运算电路的复杂度？ 对计算$[Ex–Ey]_{补} \pmod{2^n}$ 没有影响

$[\Delta E]_{补} = 256+Ex–Ey=256+127+Ex–(127+Ey)$

$= 256 + [Ex]_{移} – [Ey]_{移} = [Ex]_{移}+[–[Ey]_{移}]_{补} \pmod{256}$

但$[Ex+Ey]_{移}$和 $[Ex–Ey]_{移}$的计算会变复杂！ 浮点乘除运算涉及之。

回顾：浮点数加减法基本要点

（假定：Xm、Ym分别是X和Y的尾数， Xe和Ye 分别是X和Y的阶码 ）

(1) 求阶差：Δe=Ye – Xe（若Ye > Xe，则结果的阶码为Ye)

(2) 对阶：将Xm右移Δe位，尾数变为 $Xm*2^{Xe-Ye}$（保留右移部分附加位）

(3) 尾数加减： $Xm*2^{Xe-Ye} \pm Ym$

(4) 规格化：

$0.00\ldots0001\times2^{-126}$

当尾数高位为0，则需左规：尾数左移一次，阶码减1，直到MSB为1或阶码为00000000（-126，非规格化数）

每次阶码减1后要判断阶码是否下溢（比最小可表示的阶码还要小）

当尾数最高位有进位，需右规：尾数右移一次，阶码加1，直到MSB为1

每次阶码加1后要判断阶码是否上溢（比最大可表示的阶码还要大）

阶码溢出异常处理：阶码上溢，则结果溢出；阶码下溢到无法用非规格化数表示，则结果为0

(5) 如果尾数比规定位数长（有附加位），则需考虑舍入（有多种舍入方式）

(6) 若运算结果尾数是0，则需要将阶码也置0。为什么?

尾数为0说明结果应该为0（阶码和尾数为全0）。

回顾：浮点数加法运算举例

例：用二进制浮点数形式计算 0.5 +(– 0.4375) =?

0.4375=0.25+0.125+0.0625=0.0111B

解：$0.5 = 1.000 \times 2^{-1}$，$-0.4375 = -1.110 \times 2^{-2}$

对　阶：$-1.110 \times 2^{-2} \rightarrow -0.111 \times 2^{-1}$

加　减：$1.000 \times 2^{-1} + (-0.111 \times 2^{-1}) = 0.001 \times 2^{-1}$

左　规：$0.001 \times 2^{-1} \rightarrow 1.000 \times 2^{-4}$

判溢出：无

结果为：$1.000 \times 2^{-4} = 0.0001000 = 1/16 = 0.0625$

问题：为何IEEE 754 加减运算右规时最多只需一次?

因为即使是两个最大的尾数相加，得到的和的尾数也不会达到4，故尾数的整数部分最多有两位，保留一个隐含的“1”后，最多只有一位被右移到小数部分。

IEEE 754 浮点数加法运算

在计算机内部执行上述运算时，必须解决哪些问题?

(1) 如何表示? 用IEEE754标准！

(2) 如何判断阶码的大小? 求$[\Delta E]_{补}$=?

(3) 对阶后尾数的隐含位如何处理?

右移到数值部分，高位补0，保留移出低位部分

(4) 如何进行尾数加减?

隐藏位还原后，按原码进行加减运算，附加位一起运算

(5) 何时需要规格化，如何规格化?

±1x .xx……x 形式时，则右规: 尾数右移1位, 阶码加1

± 0.0…01x…x 形式时，则左规: 尾数左移k位, 阶码减k

(6) 如何舍入?

(7) 如何判断溢出?

最终须把附加位去掉，此时需考虑舍入（IEEE754有四种舍入方式）

若最终阶码为全1，则上溢；若尾数为全0，则下溢

IEEE 754 浮点数加法运算举例（skip）

已知x=0.5, y=-0.4375, 求x+y=? (用IEEE754标准单精度格式计算)

解: $x=0.5=1/2=(0.100...0)_2=(1.00...0)_2x2^{-1}$

$y=-0.4325=(-0.01110...0)_2=(-1.110..0)_2x2^{-2}$

$[x]_{浮}$=0 01111110,00...0　$[y]_{浮}$=1 01111101,110...0

对阶: $[\Delta E]_{补}$=0111 1110 + 1000 0011=0000 0001，ΔE=1

故对y进行对阶：$[y]_{浮}$=1 0111 1110 1110...0(高位补隐藏位)

尾数相加：01.0000...0 + (10.1110...0) = 00.00100...0

(原码加法，最左边一位为符号位，符号位分开处理)

左规：　$+(0.00100...0)_2x2^{-1}=+(1.00...0)_2x2^{-4}$

(阶码减3，实际上是加了三次11111111)

$[x+y]_{浮}$=0 0111 1011 00...0

$x+y=(1.0)_2x2^{-4}=1/16=0.0625$

问题：尾数加法器最多需要多少位?

1+1+23+3=28位

原码加/减运算

- 用于浮点数尾数运算
- 符号位和数值部分分开处理
- 仅对数值部分进行加减运算，符号位起判断和控制作用
- 规则如下：
  - 比较两数符号，对加法实行“同号求和，异号求差”，对减法实行“异号求和，同号求差”。
  - 求和：数值位相加，和的符号取被加数（被减数）的符号。若最高位产生进位，则结果溢出。
  - 求差：被加数（被减数）加上加数（减数）的补码。

    a) 最高数值位产生进位表明加法结果为正，所得数值位正确。

    b) 最高数值位没产生进位表明加法结果为负，得到的是数值位的补码形式，需对结果求补，还原为绝对值形式的数值位。
  - 差的符号位：a)情况下，符号位取被加数（被减数）的符号；

    b)情况下，符号位为被加数（被减数）的符号取反。

原码加/减运算

例1：已知 $[X]_{原} = 1.0011$，$[Y]_{原} = 1.1010$，要求计算$[X+Y]_{原}$

解：由原码加减运算规则知：同号相加，则求和，和的符号同被加数符号。

和的数值位为：0011 + 1010 = 1101　（ALU中无符号数相加）

和的符号位为：1

$[X+Y]_{原} = 1.1101$

求和：直接加，有进位则溢出，符号同被

例2：已知 $[X]_{原} = 1.0011$，$[Y]_{原} = 1.1010$，要求计算$[X-Y]_{原}$

解：由原码加减运算规则知：同号相减，则求差（补码减法）

差的数值位为：$0011+(1010)_{求补} = 0011 + 0110 = 1001$

最高数值位没有产生进位，表明加法结果为负，需对1001求补，还原为绝对值形式的数值位。即：$(1001)_{补} = 0111$

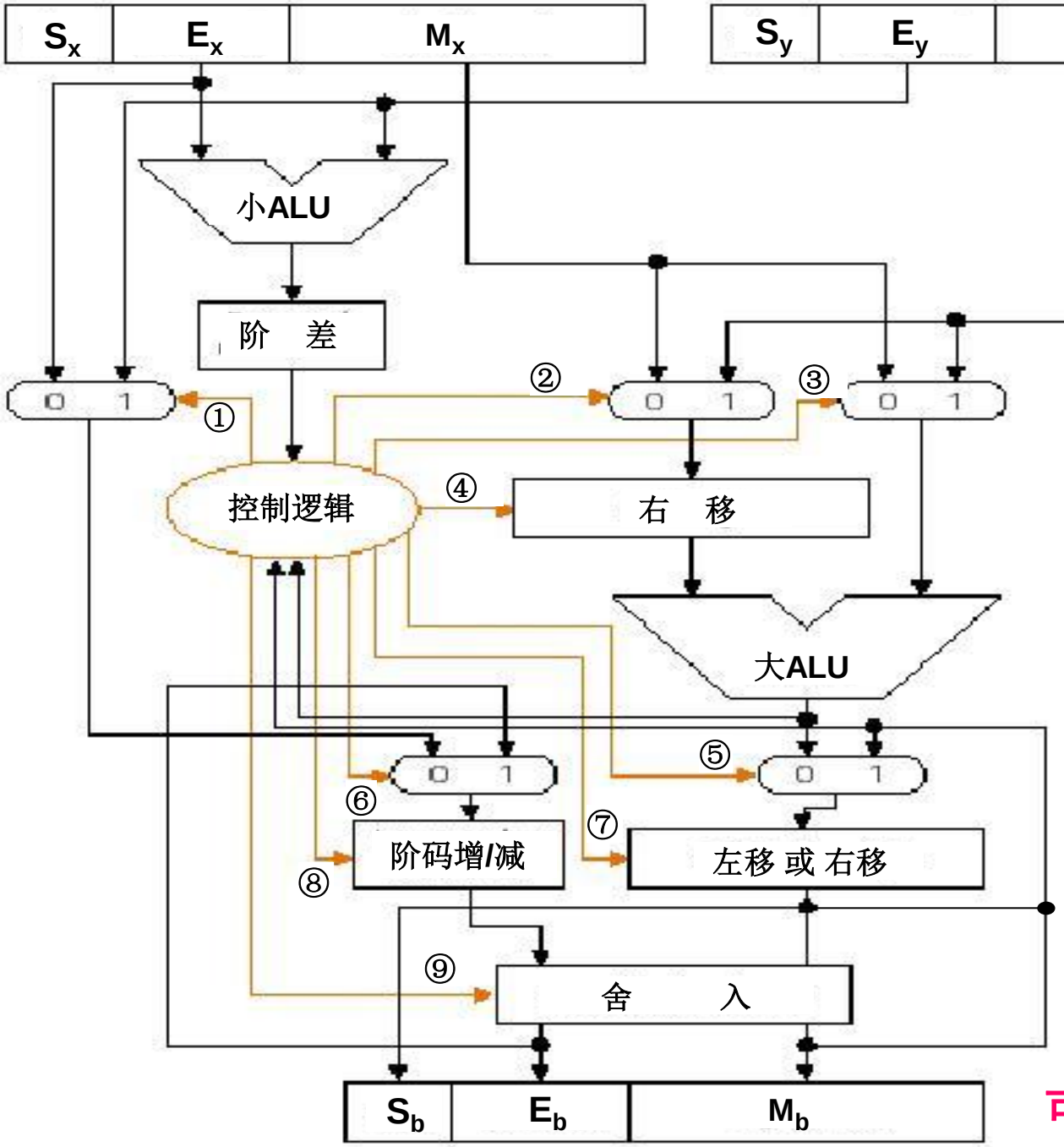
差的符号位为$[X]_{原}$的符号位取反，即：0

$[X-Y]_{原} = 0.0111$

求差：加补码，不会溢出，符号分情况

BACK

思考题：如何设计一个基于加法器的原码加/减法器？



阶码相减

阶小的数的尾数右移

尾数加/减

规格化

舍入

可用流水线方式实现！

浮点数乘/除法基本要点

- 浮点数乘法：$A*B=(M_a * M_b)\cdot 2^{Ea+Eb}$
- 浮点数除法：$A/B=(M_a / M_b)\cdot 2^{Ea-Eb}$

浮点数尾数采用原码乘/除运算

浮点数乘 / 除法步骤

（Xm、Ym分别是X和Y尾数原码， Xe和Ye 分别是X和Y阶移码 ）

(1) 求阶： Xe ± Ye ∓ 127

(2) 尾数相乘除： Xm */Ym （两个形为1.xxx的数相乘/除）

(3) 两数符号相同，结果为正；两数符号相异，结果为负；

(4) 当尾数高位为0，需左规；当尾数最高位有进位，需右规。

(5) 如果尾数比规定的长，则需考虑舍入。

(6) 若尾数是0，则需要将阶码也置0。

(7) 阶码溢出判断

不需左规！最多右规1次！

问题1：乘法运算结果最多左规几次？最多右规几次？

问题2：除法呢？ 左规次数不定！不需右规！

求阶码的和、差

设Ex和Ey分别是两个操作数的阶码，Eb是结果的阶码，则：

- 阶码加法公式为： $Eb \leftarrow Ex+Ey+129 \ (\bmod 2^8)$

$$[E1+E2]_{移} = 127 + E1 + E2 = 127 + E1 + 127 + E2 - 127$$
$$= [E1]_{移} + [E2]_{移} - 127$$
$$= [E1]_{移} + [E2]_{移} + [-127]_{补}$$
$$= [E1]_{移} + [E2]_{移} + 10000001B \ (\bmod 2^8)$$

- 阶码减法公式为： $Eb \leftarrow Ex+[-Ey]_{补}+127 \ (\bmod 2^8)$

$$[E1-E2]_{移} = 127 + E1 - E2 = 127+E1-(127+E2)+127$$
$$= [E1]_{移}-[E2]_{移} + 127$$
$$= [E1]_{移}+[-[E2]_{移}]_{补}+01111111B \ (\bmod 2^8)$$

举例

设Ex和Ey分别是两个操作数阶码的移码，Eb是结果阶码的移码表示

例：若两个阶码分别为10和-5，求10+(-5)和10-(-5)的移码。

解：Ex = 127+10 =137=1000 1001B

Ey = 127+ (–5) = 122 = 0111 1010B

[–Ey]补= 1000 0110B

将Ex和Ey代入上述公式，得：

Eb = Ex+Ey +129 = 1000 1001 + 0111 1010 +1000 0001

= 1000 0100B = 132 (mod 2^8)

其阶码的和为132 – 127 = 5，正好等于10 + (–5) = 5。

Eb = Ex+[–Ey]补+127 = 1000 1001 + 1000 0110 +0111 1111

= 1000 1110B = 142 (mod 2^8)

其阶码的差为142–127 = 15，正好等于10 – (–5) = 15。

BACK

Extra Bits(附加位)

"Floating Point numbers are like piles of sand; every time you move one you lose a little sand, but you pick up a little dirt. "

"浮点数就像一堆沙，每动一次就会失去一点‘沙’，并捡回一点‘脏’"

如何才能使失去的“沙”和捡回的“脏”都尽量少呢? 在后面加附加位!

加多少附加位才合适?

无法给出准确的答案!

Add/Sub:

	1.xxxxx	1.xxxxx	1.xxxxx	1.xxxxxxxx
+	1.xxxxx	0.001xxxxx	0.01xxxxx	-1.xxxxxxxx
	1x.xxxxy	1.xxxxxyyy	1x.xxxxyyy	0.0…0xxxx

IEEE754规定: 中间结果须在右边加2个附加位 (guard & round)

Guard (保护位): 在significand右边的位

Round (舍入位): 在保护位右边的位

BACK

附加位的作用: 用以保护对阶时右移的位或运算的中间结果。

附加位的处理: ①左规时被移到significand中; ② 作为舍入的依据。

Rounding Digits(舍入位)

举例：若十进制数最终有效位数为 3，采用两位附加位（G、R）。

问题：若没有舍入位，采用就近舍入到偶数，则结果是什么?

结果为2.36！精度没有2.37高！

$$2.3400 * 10^2$$

$$0.0253 * 10^2$$

$$2.3653 * 10^2$$

IEEE Standard: four rounding modes（用图说明）

round to nearest (default)
round towards plus infinity (always round up)
round towards minus infinity (always round down)
round towards 0

round to nearest: 称为就近舍入到偶数

round digit < 1/2 then truncate (截断、丢弃)

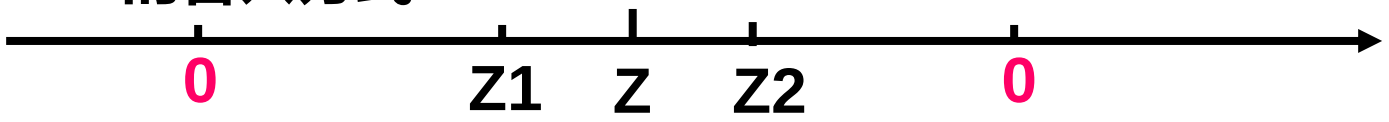
> 1/2 then round up (末位加1)

= 1/2 then round to nearest even digit（最近偶数）

可以证明默认方式得到的平均误差最小。

IEEE 754的舍入方式的说明

IEEE 754的舍入方式



（Z1和Z2分别是结果Z的最近的可表示的左、右两个数）

(1) 就近舍入：舍入为最近可表示的数

非中间值：0舍1入；

中间值：强迫结果为偶数-慢

例如：附加位为
01：舍
11：入
10：(强迫结果为偶数)

例：1.1101**11** → 1.1110;　1.1101**01** → 1.1101;
1.1101**10** → 1.1110;　1.1111**10** → 10.0000;

(2) 朝+∞方向舍入:舍入为Z2(正向舍入)

(3) 朝-∞方向舍入:舍入为Z1(负向舍入)

(4) 朝0方向舍入：截去。正数：取Z1；负数：取Z2

BACK

溢出判断

以下情况下，可能会导致阶码溢出

- **左规（阶码 - 1）时**
  - **左规（- 1）时：先判断阶码是否为全0，若是，则直接置阶码下溢；否则，阶码减1后判断阶码是否为全0，若是，则阶码下溢。**
- **右规（阶码 +1）时**
  - **右规（+ 1）时，先判断阶码是否为全1，若是，则直接置阶码上溢；否则，阶码加1后判断阶码是否为全1，若是，则阶码上溢。**

问题：机器内部如何减1？

$$+[-1]_{补} = +\ 11\ldots1$$

举例

以下情况下，可能会导致阶码溢出（续）

- 乘法运算求阶码的和时
  - 若Ex和Ey最高位皆1，而Eb最高位是0或Eb为全1，则阶码上溢
  - 若Ex和Ey最高位皆0，而Eb最高位是1或Eb为全0，则阶码下溢
- 除法运算求阶码的差时
  - 若Ex的最高位是1，Ey的最高位是0，Eb的最高位是0或Eb为全1，则阶码上溢。
  - 若Ex的最高位是0，Ey的最高位是1，Eb的最高位是1或Eb为全0，则阶码下溢。

例：若Eb = 0000 0001，则左规一次后，结果的阶码 Eb = ？

解：$Eb=Eb+[-1]_{补}$=0000 0001 + 1111 1111 = 0000 0000　阶码下溢!

例：若Ex=1111 1110，Ey=1000 0000，则乘法运算时结果的阶码 Eb=?

解：Eb=Ex+Ey+129=1111 1110+1000 0000+1000 0001=1111 1111

阶码上溢!

BACK

C语言中的浮点数类型

- C语言中有float和double类型，分别对应IEEE 754单精度浮点数格式和双精度浮点数格式
- long double类型的长度和格式随编译器和处理器类型的不同而有所不同，IA-32中是80位扩展精度格式
- 从int转换为float时，不会发生溢出，但可能有数据被舍入
- 从int或 float转换为double时，因为double的有效位数更多，故能保留精确值
- 从double转换为float和int时，可能发生溢出，此外，由于有效位数变少，故可能被舍入
- 从float 或double转换为int时，因为int没有小数部分，所以数据可能会向0方向被截断

IEEE 754 的范围和精度

南京大学 人工智能学院

- 单精度浮点数（float型）的表示范围多大?

  最大的数据: $+1.11\ldots1 \times 2^{127}$　　约 $+3.4 \times 10^{38}$

  双精度浮点数（double型）呢?　　约 $+1.8 \times 10^{308}$

- 以下关系表达式是否永真?

```
if ( i == (int) ((float) i) ) {
    printf ( "true" );
}
```

Not always true!

How about double?　True!

```
if ( f == (float) ((int) f) ) {
    printf ( "true" );
}
```

Not always true!

How about double?　False!

- 浮点数加法结合律是否正确?

FALSE!

$x = -1.5 \times 10^{38}$,　$y = 1.5 \times 10^{38}$,　$z = 1.0$

$(x+y)+z = (-1.5\times10^{38}+1.5\times10^{38}) + 1.0 = 1.0$

$x+(y+z) = -1.5\times10^{38} + (1.5\times10^{38}+1.0) = 0.0$

举例：Ariana火箭爆炸

- 1996年6月4日，Ariana 5火箭初次航行，在发射仅仅37秒钟后，偏离了飞行路线，然后解体爆炸，火箭上载有价值5亿美元的通信卫星。
- 原因是**在将一个64位浮点数转换为16位带符号整数时，产生了溢出异常。**溢出的值是火箭的水平速率，这比原来的Ariana 4火箭所能达到的速率高出了5倍。在设计Ariana 4火箭软件时，设计者确认水平速率决不会超出一个16位的整数，但在设计Ariana 5时，他们没有重新检查这部分，而是直接使用了原来的设计。
- **在不同数据类型之间转换时，往往隐藏着一些不容易被察觉的错误**，这种错误有时会带来重大损失，因此，编程时要非常小心。

举例：爱国者导弹定位错误

- 1991年2月25日，海湾战争中，美国在沙特阿拉伯达摩地区设置的爱国者导弹拦截伊拉克的飞毛腿导弹失败，致使飞毛腿导弹击中了一个美军军营，杀死了美军28名士兵。其原因是由于爱国者导弹系统时钟内的一个软件错误造成的，引起这个软件错误的原因是浮点数的精度问题。
- 爱国者导弹系统中有一内置时钟，用计数器实现，每隔0.1秒计数一次。程序用0.1的一个24位定点二进制小数x来乘以计数值作为以秒为单位的时间
- 这个x的机器数是多少呢?
- 0.1的二进制表示是一个无限循环序列：0.00011[0011]...，x=0.000 1100 1100 1100 1100 1100B。显然，x是0.1的近似表示，0.1-x

  = 0.000 1100 1100 1100 1100 1100 [1100]... -

  0.000 1100 1100 1100 1100 1100B，即为：

  =0.000 0000 0000 0000 0000 0000 1100 [1100]...B

  $=2^{-20}\times 0.1 \approx 9.54\times 10^{-8}$

  这就是机器值与真值之间的误差!

举例：爱国者导弹定位错误

已知在爱国者导弹准备拦截飞毛腿导弹之前，已经连续工作了100小时，飞毛腿的速度大约为2000米/秒，则由于时钟计算误差而导致的距离误差是多少?

100小时相当于计数了$100\times60\times60\times10=36\times10^5$次，因而导弹的时钟已经偏差了$9.54\times10^{-8}\times36\times10^5\approx0.343$秒

因此，距离误差是2000×0.343秒 $\approx$ 687米

小故事：实际上，以色列方面已经发现了这个问题并于1991年2月11日知会了美国陆军及爱国者计划办公室（软件制造商）。以色列方面建议重新启动爱国者系统的电脑作为暂时解决方案，可是美国陆军方面却不知道每次需要间隔多少时间重新启动系统一次。1991年2月16日，制造商向美国陆军提供了更新软件，但这个软件最终却在飞毛腿导弹击中军营后的一天才运抵部队。

举例：爱国者导弹定位错误

- 若x用float型表示，则x的机器数是什么？0.1与x的偏差是多少？系统运行100小时后的时钟偏差是多少？在飞毛腿速度为2000米/秒的情况下，预测的距离偏差为多少？
  - 0.1= 0.0 0011[0011]B=+1.1 0011 0011 0011 0011 0011 00B×2^{-4}，故x的机器数为0 011 1101 1 100 1100 1100 1100 1100 1100
  - Float型仅24位有效位数，后面的有效位全被截断，故x与0.1之间的误差为：|x–0.1|=0.000 0000 0000 0000 0000 0000 0000 1100 [1100]...B。这个值等于$2^{-24}\times0.1 \approx 5.96\times10^{-9}$。100小时后时钟偏差$5.96\times10^{-9}\times36\times10^{5} \approx 0.0215$秒。距离偏差$0.0215\times2000\approx43$米。比爱国者导弹系统精确约16倍。

举例：爱国者导弹定位错误

- 若用32位二进制定点小数x=0.000 1100 1100 1100 1100 1100 1100 1101 B表示0.1，则误差比用float表示误差更大还是更小?
  - 当x=0.000 1100 1100 1100 1100 1100 1100 1101 B时，与0.1之间的误差约为：|x–0.1|=0.000 0000 0000 0000 0000 0000 0000 000**0** 00 1100 [1100]...B。这个值等于$2^{-30}\times0.1 \approx 9.31\times10^{-11}$。100小时后时钟偏差$9.31\times10^{-11}\times36\times10^{5} \approx 0.000335$秒。预测的距离偏差仅为$0.000335\times2000 \approx 0.67$米。

举例：浮点数运算的精度问题

- 从上述结果可以看出：
  - 用32位定点小数表示0.1，比采用float精度高64倍
  - 用float表示在计算速度上更慢，必须先把计数值转换为IEEE 754格式浮点数，然后再对两个IEEE 754格式的数相乘，故采用float比直接将两个二进制数相乘要慢
- Ariana 5火箭和爱国者导弹的例子带来的启示

✓ 程序员应对底层机器级数据的表示和运算有深刻理解

✓ 计算机世界里，经常是“差之毫厘，失之千里”，需要细心再细心，精确再精确

✓ 不能遇到小数就用浮点数表示，有些情况下（如需要将一个整数变量乘以一个确定的小数常量），可先用一个确定的定点整数与整数变量相乘，然后再通过移位运算来确定小数点

举例：Kahan累加算法

- The Father of Floating Point

jie@debian: ~/class/ch2/3

```
 1 #include <stdio.h>
 2
 3 void main()
 4 {
 5     float tem = 0.1f;
 6     float sum = tem;
 7     float c = 0;
 8     float y, t;
 9     int i;
10     for( i=1;i<4000000;i++)
11     {
12         y = tem -c;
13         t = sum +y;
14         c = (t-sum)-y;
15         sum = t;
16     }
17     printf("%f\n",s
18 }
~
~
<ecision.c[1] [c] uni
```

```
h2/3$ ./tt

h2/3$
```

jie@debian: ~/class/ch2/3

```
jie@debian:~/class/ch2/3$ ./precision
400000.000000
jie@debian:~/class/ch2/3$
```

举例： Kahan累加算法（续）

```
function KahanSum(input)
    var sum = 0.0
    var c = 0.0
    for i = 1 to input.length do
        var y = input[i] - c
        var t = sum + y
        c = (t - sum) - y
        sum = t
return sum
```

将每次累加因舍入造成的截断误差保存起来，再加入到下一次的累加中。

```
y = 3.14159 - 0
t  = 10000.0 + 3.14159
   = 10003.1
c = (10003.1 - 10000.0) - 3.14159
   = 3.10000 - 3.14159
   = -.0415900
sum = 10003.1
```

```
 y = 2.71828 - -.0415900
    = 2.75987
 t = 10003.1 + 2.75987
    = 10005.9
c  = (10005.9 - 10003.1) - 2.75987
     = 2.80000 - 2.75987
     = .040130
sum = 10005.9
```

实例: PowerPC和80x86中的浮点部件

- PowerPC中的浮点运算
  - 比MIPS多一条浮点指令：乘累加指令
    - 将两个操作数相乘，再与另一个操作数相加，作为结果操作数
    - 可用一条乘累加指令代替两条MIPS浮点指令
    - 可为中间结果多保留几位，得到最后结果后再考虑舍入，精度高
    - 利用它来实现除法运算和平方根运算
  - 浮点寄存器的数量多一倍（32xSPR, 32xDPR)
- 80x86中的浮点运算
  - 采用寄存器堆栈结构：栈顶两个数作为操作数
  - 寄存器堆栈的精度为80位 (MIPS和PowerPC是32位或64位)
  - 所有浮点运算都转换为80位扩展浮点数进行运算，写回存储器时，再转换位32位（float）或64位（double），编译器需小心处理
  - 由浮点数访存指令自动完成转换
  - 指令类型：访存、算术、比较、函数（正弦、余弦、对数等）

第三讲小结

- 浮点数的表示（IEEE754标准）
  - 单精度SP（float）和双精度DP（double）
    - 规格化数(SP)：阶码1~254，尾数最高位隐含为1
    - 0(阶为全0，尾为全0)
    - ∞(阶为全1，尾为全0)
    - NaN(阶为全0，尾为非0)
    - 非规数(阶为全1，尾为非0)
- 浮点数加减运算
  - 对阶、尾数加减、规格化（上溢/下溢处理）、舍入
- 浮点数乘除运算
  - 求阶、尾数乘除、规格化（上溢/下溢处理）、舍入
- 浮点数的精度问题
  - 中间结果加保护位、舍入位（和粘位）
  - 最终进行舍入（有四种舍入方式）
    - 就近（中间值强迫为偶数）、+∞方向、-∞方向、0方向
    - 默认为“就近”舍入方式

本章总结（1）

定点数运算：由ALU + 移位器实现各种定点运算

- 移位运算
  - 逻辑移位：对无符号数进行，左（右）边补0，低（高）位移出
  - 算术移位：对带符号整数进行，移位前后符号位不变，编码不同，方式不同。
  - 循环移位：最左（右）边位移到最低（高）位，其他位左（右）移一位。
- 扩展运算
  - 零扩展：对无符号整数进行高位补0
  - 符号扩展：对补码整数在高位直接补符
- 加减运算
  - 补码加/减运算：用于整数加/减运算。符号位和数值位一起运算，减法用加法实现。同号相加时，若结果的符号不同于加数的符号，则会发生溢出。
  - 原码加/减运算：用于浮点数尾数加/减运算。符号位和数值位分开运算，同号相加，异号相减；加法直接加；减法用加负数补码实现。
- 乘法运算：用加法和右移实现。
  - 补码乘法：用于整数乘法运算。符号位和数值位一起运算。采用Booth算法。
  - 原码乘法：用于浮点数尾数乘法运算。符号位和数值位分开运算。数值部分用无符号数乘法实现。
- 除法运算：用加/减法和左移实现。
  - 补码除法：用于整数除法运算。符号位和数值位一起运算。
  - 原码除法：用于浮点数尾数除法运算。符号位和数值位分开运算。数值部分用无符号数除法实现。

本章总结（2）

- 浮点数运算：由多个ALU + 移位器实现
  - 加减运算
    - 对阶 、尾数相加减、规格化处理、舍入、判断溢出
  - 乘除运算
    - 尾数用定点原码乘/除运算实现，阶码用定点数加/减运算实现。
  - 溢出判断
    - 当结果发生阶码上溢时，结果发生溢出，发生阶码下溢时，结果为0。
  - 精确表示运算结果
    - 中间结果增设保护位、舍入位、粘位
    - 最终结果舍入方式：就近舍入 / 正向舍入 / 负向舍入 / 截去四种方式。
- ALU的实现
  - 算术逻辑单元ALU：实现基本的加减运算和逻辑运算。
  - 加法运算是所有定点和浮点运算（加/减/乘/除）的基础，加法速度至关重要
  - 进位方式是影响加法速度的重要因素
  - 并行进位方式能加快加法速度
  - 通过“进位生成”和“进位传递”函数来使各进位独立、并行产生