EE 3755

Computer Arithmetic

Handout # 6

EE [redacted] 3755

Floating Point Arithmetic: Multiplication, Division

Here we will show how floating point multiplication and division can take place. We will assume binary FLP arithmetic ($r=2$). All our FLP numbers will be consisting of normalized fractions and biased exponents.

Floating Point Multiplication

The problem here is defined as follows: Consider two binary ($r=2$) FLP numbers $A_1 = (-1)^{s_1} \times f_1 \times 2^{e_1}$ and $A_2 = (-1)^{s_2} \times f_2 \times 2^{e_2}$ where s_1 and s_2 are the sign bits of A_1 and A_2, f_1 and f_2 their n-bit unsigned normalized fractions and e_1 and e_2 their m-bit exponents which are considered to be in biased form. We are interested in computing $A_3 = A_1 \times A_2 =$ $= (-1)^{s_3} \times f_3 \times 2^{e_3}$ where s_3 is the sign bit, f_3 the n-bit unsigned normalized fraction and e_3 the m-bit biased exponent of the product A_3.

The following show the steps involved in a floating point multiplication:

1. Determine the product sign:

Clearly if numbers of the same sign are multiplied then the product is positive while if numbers of different signs are multiplied then the product is negative. In other words, if $s_1 = s_2$ then $s_3 = 0$ while if $s_1 \neq s_2$ then $s_3 = 1$. This means that $s_3 = s_1 \oplus s_2$ where $\oplus$ is the exlusive-OR operator.

2 f

2. Multiply fractions; add exponents; subtract bias

- The two n-bit fractions f_1 and f_2 have to be multiplied. The resulting product will have length 2n bits. Keep the left most n-bit part as the fraction product and truncate the right most n-bit part
- With respect to the exponents, the exponents e_1 and e_2 should be added and the bias should be subtracted (or $e_1 + e_2 - \text{bias}$ should be performed).

The reason of why we subtract the bias is the following: Since e_1 and e_2 are in biased form $e_1 = e_{1\,\text{unbiased}} + \text{bias}$, $e_2 = e_{2\,\text{unbiased}} + \text{bias}$. This means that $e_1 + e_2 = e_{1\,\text{unbiased}} + e_{2\,\text{unbiased}} + 2 \times \text{bias}$. Since $e_1 + e_2$ includes a double bias, one bias amount should be subtracted.

3. Postnormalize the result if necessary.

Here the following observations are in place:
The resulting product of the fractions $f_3 = f_1 \times f_2$ will be a fraction (or $f_3 < 1$) since $f_1 < 1$ and $f_2 < 1$. This f_3 could be either of the form $f_3 = .1x\cdots x$ (already normalized) or of the form $f_3 = .01x\cdots x$ (not normalized but having only one leading zero). Observe that $(f_3)_{min} = (f_1 \times f_2)_{min} = (f_1)_{min} \times (f_2)_{min} = (.100\cdots0) \times (.100\cdots0) = .0100\cdots0$. In other words the smallest possible product of fractions will have only one leading zero. So if the product of the fractions is not normalized, postnormalization means a 1-bit left shift of the fraction and a decrementation of the exp. by one.

4. Check for exponent ovf. and exp. underflow

In FLP multiplication either one of exp. ovf or exp. underflow can occur.

Some examples on FLP multiplication follow.

Example 1: Consider the following two floating point numbers:

	s_1	e_1	f_1
A_1:	0	1010	11001

	s_2	e_2	f_2
A_2:	1	1011	11110

Compute $A_3 = A_1 \times A_2$. All fractions should be considered normalized and all exponents biased.

Solution: The sign of the product is $s_3 = s_1 \oplus s_2 = 1$.

The product of the fractions is $(.11001) \times (.11110) = .1011101110$. Truncating the right most 5-bit part we get product of fractions $= .10111$.

Also $e_1 + e_2 - \text{bias} = 10 + 11 - 8 = 21 - 8 = (13)_{10} = (1101)_2$.

So the product is

	s_3	e_3	f_3
A_3:	1	1101	10111

. Observe that the fraction is normalized so no postnormalization is needed. No exp. ovf or exp undfl. occured.

Example 2: Consider two FLP numbers $A_1 = (-1)^{s_1} \times f_1 \times 2^{e_1}$ and $A_2 = (-1)^{s_2} \times f_2 \times 2^{e_2}$ with normalized fractions f_1 and f_2 and 4-bit biased exponents $e_1 = (1101)_2$ and $e_2 = (1110)_2$. Do you expect to have an exponent overflow or exponent underflow as a result of the multiplication operation $A_1 \times A_2$? The product should be consisting of a normalized fraction and biased exponent.

Solution : The exponent of the product will be $e_1 + e_2 - \text{bias}$ if no postnormalization is necessary, while it will be $e_1 + e_2 - \text{bias} - 1$ if postnormalization is needed. But $e_1 = (1101)_2 = (13)_{10}$, $e_2 = (1110)_2 = (14)_{10}$ and bias $= 2^3 = 8$. Also the dynamic range of 4-bit biased exps is $[0 \;\; 15]$. So, if postnormalization is not needed, the exponent of the product will be $e_1 + e_2 - \text{bias}$ $= 13 + 14 - 8 = 19 > 15$ which means that an exp. ovf will occur. In the case that postnormalization is needed the exp. of the product will be $e_1 + e_2 - \text{bias} - 1$ $= 13 + 14 - 8 - 1 = 18 > 15$ (again exp. ovf occurs).

Example 3 : Repeat example 2 but now let the biased exponents be $e_1 = (1011)_2$ and $e_2 = (1101)_2$.

Solution : If no postnormalization is necessary, the product's exponent will be $e_1 + e_2 - \text{bias} = 11 + 13 - 8$ $= 24 - 8 = 16 > 15$ and an exp. ovf occurs. However, if postnormalization is needed, the exponent of the product will be $e_1 + e_2 - \text{bias} - 1 = 11 + 13 - 8 - 1 = 15 \in [0 \;\; 15]$ and in this case neither an exp. ovf nor an exp. unfl. occurs.

Example 4 : Repeat example 2 with biased exponents being $e_1 = (1010)_2$ and $e_2 = (1001)_2$.

Solution : If no postnormalization is necessary, the desired exp. is $e_1 + e_2 - \text{bias} = 10 + 9 - 8 = 11 \in [0 \;\; 15]$. If postnormalization is needed the desired exp. is $e_1 + e_2 - \text{bias} - 1 = 10 + 9 - 8 - 1 = 10 \in [0 \;\; 15]$. So in this case we never have exp. ovf or exp. undf.

Example 5: Repeat example 2 with biased exponents being $e_1 = (0010)_2$ and $e_2 = (0011)_2$.

Solution: Here $e_1+e_2-\text{bias} = 2+3-8 = -3 < 0$ (exp. undfl.) and $e_1+e_2-\text{bias}-1 = -4 < 0$ (exp. undfl.). So for the data of example 5, exp. undfl. will always occur.

Floating Point Division

The problem here is defined as follows. Consider two binary FLP numbers: the dividend $A_1 = (-1)^{s_1} \times f_1 \times 2^{e_1}$ and the divisor $A_2 = (-1)^{s_2} \times f_2 \times 2^{e_2}$ where s_1 and s_2 are the sign bits of A_1 and A_2, f_1 and f_2 their n-bit unsigned normalized fractions and e_1 and e_2 their m-bit exponents which are considered to be in biased form. We are interested in computing the quotient $A_3 = \frac{A_1}{A_2} = (-1)^{s_3} \times f_3 \times 2^{e_3}$ where s_3 is the sign bit of the quotient A_3, f_3 its n-bit unsigned normalized fraction and e_3 its m-bit biased exponent.

Obviously, one task that has to take place is the division of the dividend's fraction f_1 by the divisor's fraction f_2 in order to get the quotient fraction $f_3 = \frac{f_1}{f_2}$ (while we want f_3 to be an n-bit unsigned normalized fraction). We can have two cases with respect to the fractions f_1 and f_2.

(i) $f_1 \geq f_2$:

In this case $\frac{f_1}{f_2} \geq 1$ or $\frac{f_1}{f_2} = \underbrace{1.xx\cdots x}_{n+1 \text{ bits}}$

So in this case we would expect to have a division overflow since the quotient's fraction $\frac{f_1}{f_2}$ exceeds its n-bit space. Such a division overflow can be prevented, however, by shifting the dividend's fraction f_1 by one bit to the right and incrementing its exponent e_1 by one. This is called "alignment of dividend". After alignment the dividend A_1 becomes $A_1 = (-1)^{s_1} \times \frac{f_1}{2} \times 2^{e_1+1}$ (recall that a 1-bit right shift implies division by two). We will then perform the division $f_3 = \frac{f_1/2}{f_2}$ in order to get the quotient frac. f_3. Let's now show that $\frac{f_1/2}{f_2}$ is a normalized fraction. Since f_1 is normalized $f_1 = .1xx\cdots x$. So $\frac{f_1}{2} = .01x\cdots x$. Since f_2 is normalized then $f_2 = .1x\cdots xx$. Obviously $\frac{f_1}{2} < f_2$ which implies $\frac{f_1/2}{f_2} < 1$ (or $\frac{f_1/2}{f_2}$ is fraction). Also, since $f_1 \geqslant f_2$ then $\frac{f_1}{f_2} \geqslant 1$ or $\frac{f_1/2}{f_2} \geqslant \frac{1}{2}$ (which means that $\frac{f_1/2}{f_2}$ is normalized.)

So the overall conclusion here is that if $f_1 \geqslant f_2$ we first align the dividend. Then the division of the fraction of the aligned dividend by the fraction of the divisor will return a normalized fraction and postnormalization will not be needed.

(ii) $f_1 < f_2$

In this case $\frac{f_1}{f_2} < 1$ (or $\frac{f_1}{f_2}$ is fraction). Let's now prove that $\frac{f_1}{f_2}$ is also normalized. Observe that

$$\left(\frac{f_1}{f_2}\right)_{min} = \frac{(f_1)_{min}}{(f_2)_{max}} = \frac{.100\cdots0}{.11\cdots11} = \frac{1/2}{1-2^{-n}} > \frac{1}{2} .$$

This means that any $\frac{f_1}{f_2} > \frac{1}{2}$ or any $\frac{f_1}{f_2}$ is normalized.

So the conclusion here is that if $f_1 < f_2$ then $\frac{1}{2} < \frac{f_1}{f_2} < 1$ (or normalized fraction) and postnormalization is not needed.

The above presentation makes it clear that the steps involved in a floating point division are the following:

1. Determine the sign of the quotient

Clearly if dividend and divisor are of the same sign then the quotient is positive, else it is negative. So again $S_3 = S_1 \oplus S_2$.

2. Align dividend if necessary (that is if $f_1 \geq f_2$)

3. Divide fractions; subtract exponents; add bias

With respect to the exponents, $e_1 - e_2 + \text{bias}$ should be performed. The reason of why we need to add the bias is the following: Since e_1 and e_2

are in biased form $e_1 = e_{1unbiased} + bias$, $e_2 = e_{2unbiased} + bias$. This means that $e_1 - e_2 = e_{1unbiased} + bias - e_{2unbiased} - bias = e_{1unb.} - e_{2unb.}$

Since $e_1 - e_2$ includes no bias at all, the bias amount should be added.

4. **Check for exponent ovf. and exp. underflow.**

In FLP division both exp. ovfs and exp. unfls are possible occurences.

Example 6: Consider two FLP numbers $A_1 = (-1)^{S_1} \times f_1 \times 2^{e_1}$ and $A_2 = (-1)^{S_2} \times f_2 \times 2^{e_2}$ with normalized fractions f_1 and f_2 and 4-bit biased exponents $e_1 = (1010)_2$ and $e_2 = (1011)_2$. Do you expect to have an exponent overflow or exponent underflow as a result of the division operation $\frac{A_1}{A_2}$? The quotient should be consisting of a normalized fraction and biased exponent.

Solution: The exponent of the quotient will be $e_1 - e_2 + bias$ if dividend alignment is not needed while it will be $e_1 + 1 - e_2 + bias$ if alignment of dividend is needed.

Observe that $e_1 - e_2 + bias = 10 - 11 + 8 = 7 \in [0 \quad 15]$ and $e_1 + 1 - e_2 + bias = 10 + 1 - 11 + 8 = 8 \in [0 \quad 15]$. So in this case (of example 6) exp. ovf or exp. undflow can never occur.

Example 7: Repeat example 6 with biased exponents being $e_1 = (0011)_2$ and $e_2 = (1100)_2$.

Solution: Here, if no dividend alignment is needed, the exponent of the quotient will be $e_1 - e_2 + bias =$ $= 3 - 12 + 8 = -1 < 0$ so an exp. underflow will occur. However, if dividend alignment is needed (say $f_1 = .1110$, $f_2 = .1100$) the exp. of the quotient will be $e_1 + 1 - e_2 + bias = 3 + 1 - 12 + 8 = 0 \in [0 \;\; 15]$ and in this case no exp. ovf or underflow occur

Example 8: Repeat example 6 with biased exponents being $e_1 = (1110)_2$ and $e_2 = (0001)_2$.

Solution: If no dividend alignment is needed the exp. of quot. will be $e_1 - e_2 + bias = 14 - 1 + 8 = 21 > 15$ (so exp. ovf occurs). If dividend alignment is needed the situation is even worse since $e_1 + 1 - e_2 + bias = 22$ (exp ovf)

Example 9: Repeat example 6 with biased exponents being $e_1 = (1011)_2$ and $e_2 = (0100)_2$.

Solution: Here $e_1 - e_2 + bias = 11 - 4 + 8 = 15 \in [0 \;\; 15]$ (so exp. is safe if dividend alignment not needed). However, if divid. alignment was needed $e_1 + 1 - e_2 + bias = 16$ (which indicates exp. ovf).

Example 10: Repeat example 6 with biased exp. $e_1 = (0001)_2$, $e_2 = (1110)_2$.

Solution: Here $e_1 - e_2 + bias = 1 - 14 + 8 = -5 < 0$ (exp. undf) and $e_1 + 1 - e_2 + bias = -4$ (still exp. undfl). So for the data of this example, exp. undf. will always occur.