Lecture 15

Floating Point Multiplication

Khaza Anuarul Hoque
ECE 4250/7250

Expensive Fireworks (1996)

- In 1996, code from the Ariane 4 rocket is reused in the Ariane 5, but the new rocket's faster engines trigger a bug in an arithmetic routine inside the flight computer.
- The error is in code to convert 64-bit floating-point numbers to a 16-bit signed integers. The faster engines cause the 64-bit numbers to be larger, triggering an overflow condition that crashes the flight computer.
- As a result, the rocket's primary processor overpowers the rocket's engines and causes the rocket to disintegrate only 40 seconds after launch.

Intel Pentium FDIV Bug

- Try 4195835 – 4195835 / 3145727 * 3145727.
  - In 94' Pentium, it doesn't return 0, but 256.
- Intel uses the SRT algorithm for floating point division. Five entries in the lookup table are missing.
- Cost: $500 million
- Xudong Zhao's Thesis on Word Level Model Checking

Why Bias ???

In single precision floating point, you get 8 bits in which to store the exponent. Instead of storing it as a signed two's complement number, it was decided that it'd be easier to just add 127 to the exponent (since the lowest it could be in 8 bit signed is -127) and just store it as an unsigned number. If the stored value is greater than the bias, that means the value of the exponent is positive, if it's lower than the bias, it's negative, if it's equal, it's zero.

IEEE Rounding modes

- It is desirable to round to the nearest value.
- The IEEE standard has 4 rounding modes when the number falls halfway:
  - 1. **Round up**: round toward positive infinity; round up to the next higher number.
  - 2. **Round down**: round toward negative infinity; round down to the nearest smaller number.
  - 3. **Truncate**: Ignore bits beyond the allowable number of bits.
  - 4. **Unbiased**: Known as round to even.

Rounding example

Round 1.10101 and 1.01111 to 2 binary places using the 4 different IEEE rounding modes.

Answer:

According to IEEE, roundup is toward positive infinity, and rounddown is toward negative infinity. Truncate just retains 2 bits after the binary point from the original numbers. Round-to-even (unbiased rounding) rounds up if the last bit retained is a 1, resulting in 1.01111 becoming 1.10. It should be noted that 1.10101 also results in the same value in round-to-even.

Number	Roundup	Rounddown	Truncate	Round-to-Even
1.10101	1.11	1.10	1.10	1.10
1.01111	1.10	1.01	1.01	1.10

Floating point multiplication

- Multiplication of floating point numbers F1 (with sign s1, exponent e1 and significand p1) and F2 (with sign s2, exponent e2 and significand p2) is a five step process.

Floating point multiplication (cont.)

- Step 1
  - Calculate the tentative exponent of the product by adding the biased exponents of the two numbers, subtracting the bias. The bias is 127 and 1023 for single precision and double precision IEEE data format respectively
  - e1+e2-bias

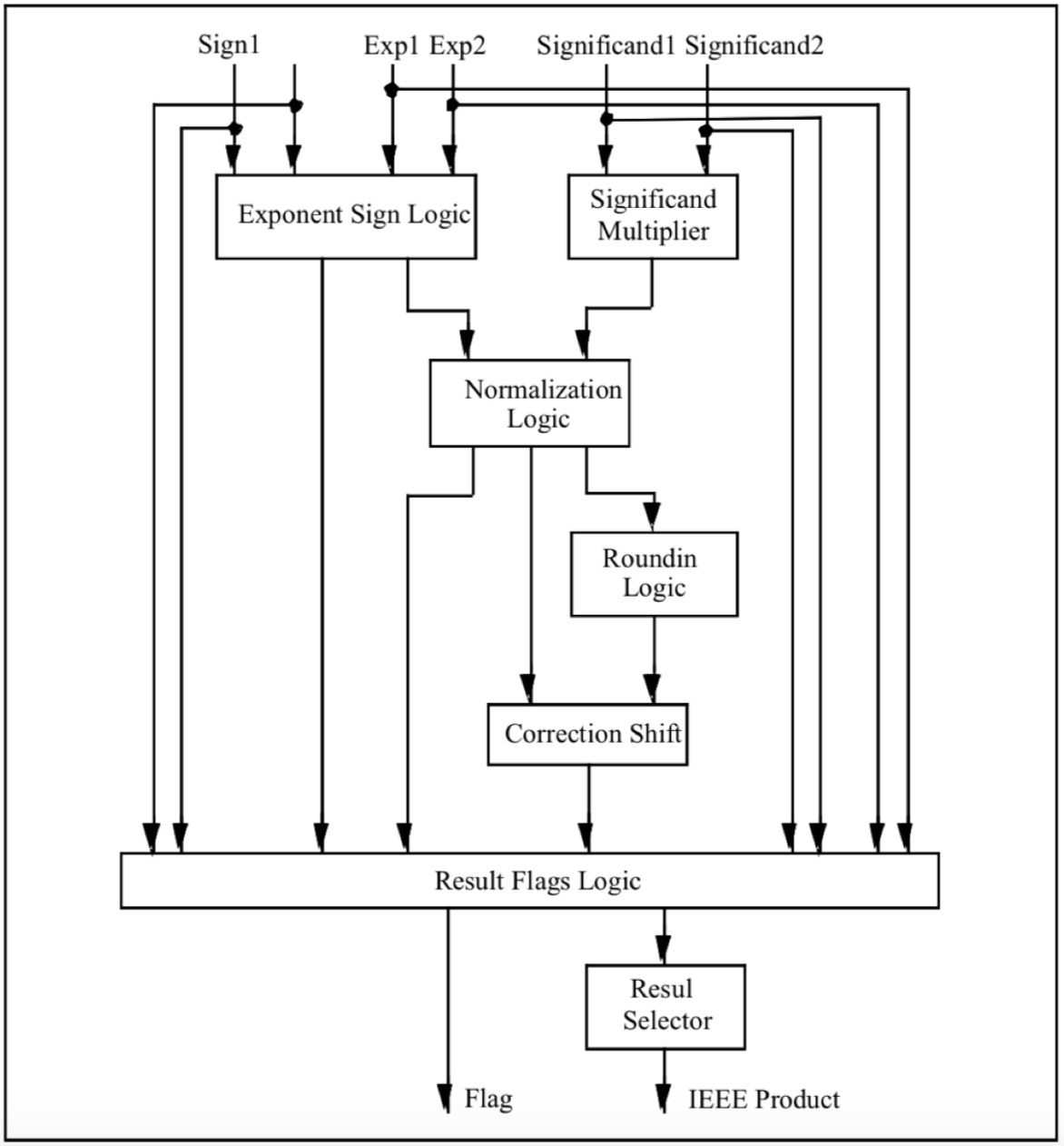
Floating point multiplication (cont.)

- Step 2
  - If the sign of two floating point numbers are the same, set the sign of product to ‘+’, else set it to ‘-’.
- Step 3
  - Multiply the two significands.
- Step 4
  - Normalize the product if MSB of the product is 1 (i.e. product of two significands) .

Floating point multiplication (cont.)

- Step 5:
  - Round the product if R(M0 + S) is true, where M0 and R represent the p-th and (p+1)th bits from the left end of normalized product and Sticky bit (S) is the logical OR of all the bits towards the right of R bit. If the rounding condition is true, a 1 is added at the p-th bit (from the left side) of the normalized product.
  - If all p MSBs of the normalized product are 1's, rounding can generate a carry- out. In that case normalization (step 4) has to be done again.

Multiplier Architecture



Example of Multiplication

- Multiply the following two numbers. Use IEEE 754 standard: A= 25.5 B= -0.375