Lecture 9: Floating Point

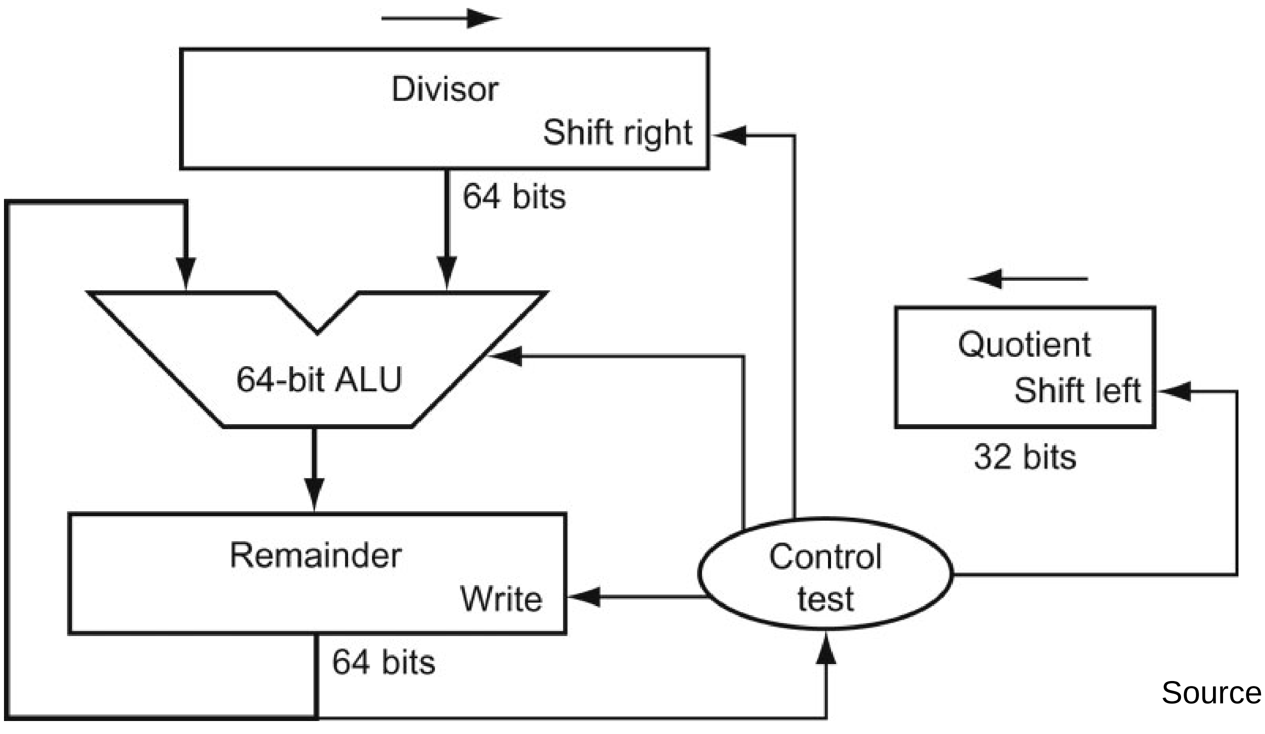
- Today's topics:
  - Division
  - IEEE 754 representations

Divide Example

- Divide 7_{ten} ($0000\ 0111_{two}$) by 2_{ten} (0010_{two})

Iter	Step	Quot	Divisor	Remainder
0	Initial values	0000	0010 0000	0000 0111
1	Rem = Rem – Div Rem < 0 ➔ +Div, shift 0 into Q Shift Div right	0000 0000 0000	0010 0000 0010 0000 0001 0000	1110 0111 0000 0111 0000 0111
2	Same steps as 1	0000 0000 0000	0001 0000 0001 0000 0000 1000	1111 0111 0000 0111 0000 0111
3	Same steps as 1	0000	0000 0100	0000 0111
4	Rem = Rem – Div Rem >= 0 ➔ shift 1 into Q Shift Div right	0000 0001 0001	0000 0100 0000 0100 0000 0010	0000 0011 0000 0011 0000 0011
5	Same steps as 4	0011	0000 0001	0000 0001

Hardware for Division

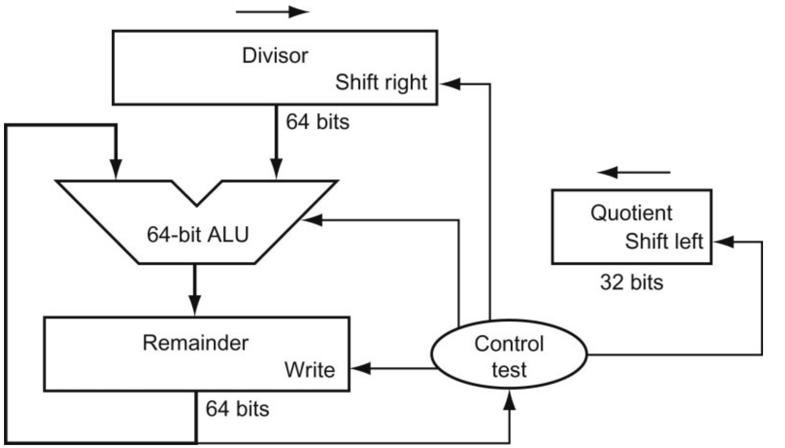


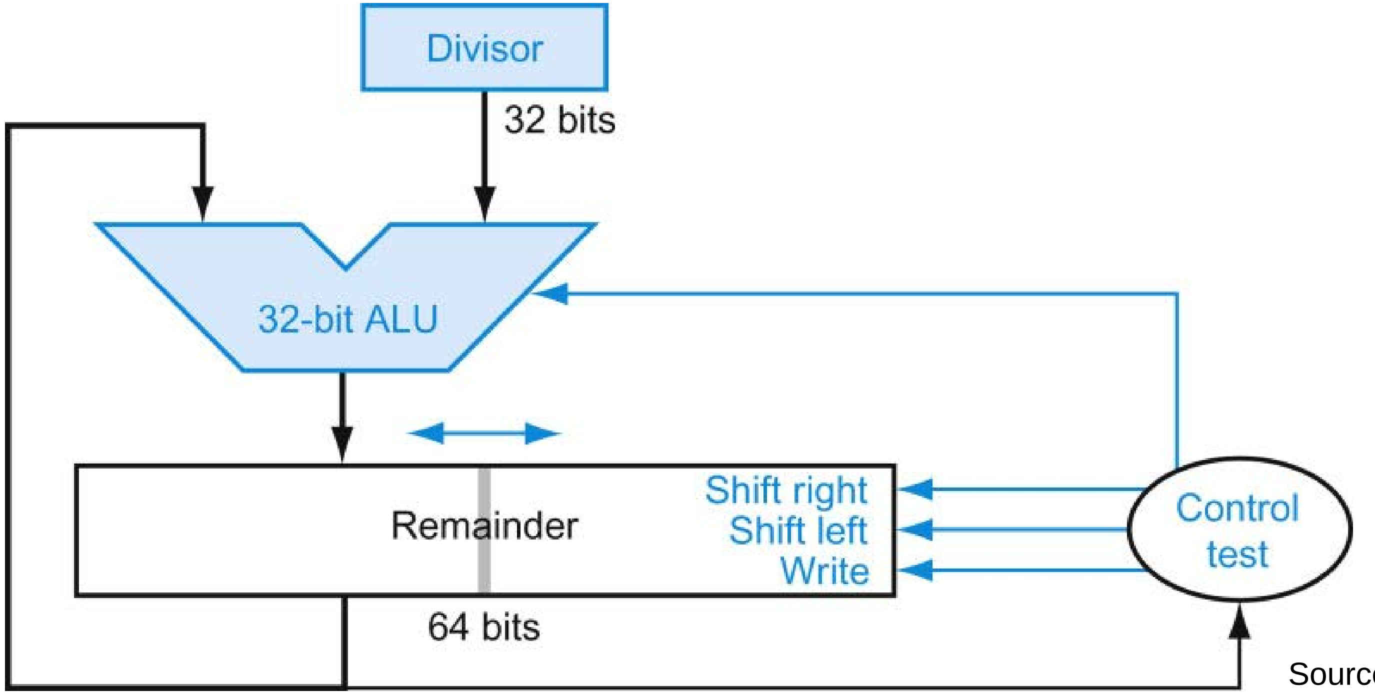
Source: H&P textbook

A comparison requires a subtract; the sign of the result is examined; if the result is negative, the divisor must be added back

Similar to multiply, results are placed in Hi (remainder) and Lo (quotient)

Efficient Division





Source: H&P textbook

Divisions involving Negatives

- Simplest solution: convert to positive and adjust sign later
- Note that multiple solutions exist for the equation:
  Dividend = Quotient x Divisor + Remainder

+7	div	+2	Quo =	Rem =
-7	div	+2	Quo =	Rem =
+7	div	-2	Quo =	Rem =
-7	div	-2	Quo =	Rem =

Divisions involving Negatives

- Simplest solution: convert to positive and adjust sign later

- Note that multiple solutions exist for the equation:
  Dividend = Quotient x Divisor + Remainder

+7 div +2		Quo = +3	Rem = +1
-7 div +2		Quo = -3	Rem = -1
+7 div -2		Quo = -3	Rem = +1
-7 div -2		Quo = +3	Rem = -1

Convention: Dividend and remainder have the same sign
Quotient is negative if signs disagree
These rules fulfil the equation above

Floating Point

- Normalized scientific notation: single non-zero digit to the left of the decimal (binary) point – example: 3.5×10^9

- $1.010001 \times 2^{-5}_{two} = (1 + 0 \times 2^{-1} + 1 \times 2^{-2} + \ldots + 1 \times 2^{-6}) \times 2^{-5}_{ten}$

- A standard notation enables easy exchange of data between machines and simplifies hardware algorithms – the IEEE 754 standard defines how floating point numbers are represented

Sign and Magnitude Representation

Sign 1 bit	Exponent 8 bits	Fraction 23 bits
S	E	F

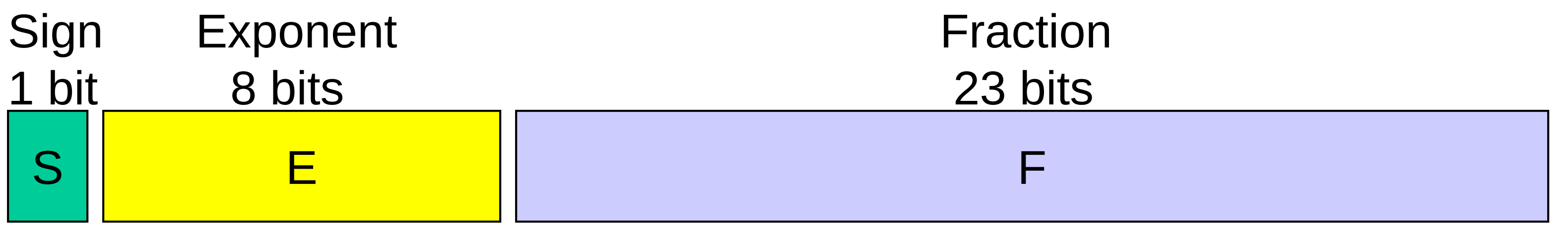
- More exponent bits ➔ wider range of numbers (not necessarily more numbers – recall there are infinite real numbers)

- More fraction bits ➔ higher precision

- Register value = $(-1)^S \times F \times 2^E$

- Since we are only representing normalized numbers, we are guaranteed that the number is of the form 1.xxxx..
  Hence, in IEEE 754 standard, the 1 is implicit
  Register value = $(-1)^S \times (1 + F) \times 2^E$

Sign and Magnitude Representation



- Largest number that can be represented:

- Smallest number that can be represented:

Sign and Magnitude Representation

Sign 1 bit	Exponent 8 bits	Fraction 23 bits
S	E	F

- Largest number that can be represented: $2.0 \times 2^{128} = 2.0 \times 10^{38}$
- Smallest number that can be represented: $1.0 \times 2^{-127} = 2.0 \times 10^{-38}$
- Overflow: when representing a number larger than the one above;
  Underflow: when representing a number smaller than the one above
- Double precision format: occupies two 32-bit registers:
  Largest: Smallest:

Sign 1 bit	Exponent 11 bits	Fraction 52 bits
S	E	F

Details

- The number "0" has a special code so that the implicit 1 does not get added: the code is all 0s
  (it may seem that this takes up the representation for 1.0, but given how the exponent is represented, we'll soon see that that's not the case)
  (see discussion of denorms (pg. 222) in the textbook)

- The largest exponent value (with zero fraction) represents +/- infinity

- The largest exponent value (with non-zero fraction) represents NaN (not a number) – for the result of 0/0 or (infinity minus infinity)

- Note that these choices impact the smallest and largest numbers that can be represented

Exponent Representation

- To simplify sort, sign was placed as the first bit

- For a similar reason, the representation of the exponent is also modified: in order to use integer compares, it would be preferable to have the smallest exponent as 00…0 and the largest exponent as 11…1

- This is the biased notation, where a bias is subtracted from the exponent field to yield the true exponent

- IEEE 754 single-precision uses a bias of 127 (since the exponent must have values between -127 and 128)…double precision uses a bias of 1023

Final representation: $(-1)^S \times (1 + \text{Fraction}) \times 2^{(\text{Exponent} - \text{Bias})}$

Examples

Final representation: $(-1)^S \times (1 + \text{Fraction}) \times 2^{(\text{Exponent} - \text{Bias})}$

- Represent -0.75_{ten} in single and double-precision formats

  Single: (1 + 8 + 23)

  Double: (1 + 11 + 52)

- What decimal number is represented by the following single-precision number?
  1 1000 0001 01000…0000

Examples

Final representation: $(-1)^S \times (1 + \text{Fraction}) \times 2^{(\text{Exponent} - \text{Bias})}$

- Represent -0.75_{ten} in single and double-precision formats

  Single: (1 + 8 + 23)
  1 0111 1110 1000…000

  Double: (1 + 11 + 52)
  1 0111 1111 110 1000…000

- What decimal number is represented by the following single-precision number?
  1 1000 0001 01000…0000
  -5.0

FP Addition

- Consider the following decimal example (can maintain only 4 decimal digits and 2 exponent digits)

$9.999 \times 10^{1} \quad + \quad 1.610 \times 10^{-1}$

Convert to the larger exponent:

$9.999 \times 10^{1} \quad + \quad 0.016 \times 10^{1}$

Add

10.015×10^{1}

Normalize

1.0015×10^{2}

Check for overflow/underflow

Round

1.002×10^{2}

Re-normalize

FP Addition

- Consider the following decimal example (can maintain only 4 decimal digits and 2 exponent digits)

$9.999 \times 10^{1} \quad + \quad 1.610 \times 10^{-1}$

Convert to the larger exponent:

$9.999 \times 10^{1} \quad + \quad 0.016 \times 10^{1}$

Add

10.015×10^{1}

Normalize

1.0015×10^{2}

If we had more fraction bits, these errors would be minimized

Check for overflow/underflow

Round

1.002×10^{2}

Re-normalize

Title

- Bullet