Computer Architecture & Organization (CSE2003)

(Fall 2019-20)

Prof. Anand Motwani
Faculty, SCSE, VIT Bhopal University
email-id: anand.motwani@vitbhopal.ac.in

Contents

- IEEE Standard for Floating Point Representation
- Numerical Exercise
- Quiz / Problems

Session Objectives

At the end of this session student will understand:

- to represent fixed and floating point numbers in the IEEE format
- to perform arithmetic operations with them.

By: Prof. Anand Motwani
Faculty, SCSE, VIT Bhopal University

Numerical Problem (previous lecture)

- Question: Represent +0.125 if 5 bits are used to represent exponent and 6 bits for mantissa.
- Solution steps:
- 1. Calculate bias
- 2. Calculate binary of given decimal no.
- 3. Normalize the binary no.
- 4. Calculate exponent.
- 5. Calculate Mantissa
- Represent the no. With as positive (i.e. Use 0 as sign bit)

By: Prof. Anand Motwani
Faculty, SCSE, VIT Bhopal University

Solution

- 1. Bias = 15
- 2. Binary no. = $(0.001)_2$
- 3. Normalized value = $1.0 * 2^{-3}$
- Calculate exponent: -3 + 15 = 12 ~ 1100
- Mantissa: No. to right of binary point is 0. So, mantissa is 0.
- No. is positive so sign bit is 0.
- Answer =

0	0 1 1 0 0	0 0 0 0 0 0

By: Prof. Anand Motwani
Faculty, SCSE, VIT Bhopal University

- Q. Represent 52.21875 in 32-bit binary floating point format. Exponent 8 bit and Mantissa 23
- 52.21875 = 110100.00111 =
- $.11010000111 \times 2^6$.
- Normalized 23 bit mantissa = 0.11010000111000000000000.
- As excess representation is being used for exponent, it is equal to 127 + 6 = 133.
- Thus the representation is $52.21875 = 0.11010000111 \times 2^{133} = 0.110100000111 \times 2^{10000101}$.
- The 32-bit string used to store 52.21875 in a computer will thus be

0	10000101	11010000111000000000000

By: Prof. Anand Motwani
Faculty, SCSE,
VIT Bhopal University

IEEE Standard for Floating Point Representation

- Floating point binary numbers were beginning to be used in the mid 50s.
- There was no uniformity in the formats used to represent floating point numbers and programs were not portable from one manufacturer's computer to another.
- By the mid 1980s, with the advent of personal computers, the number of bits used to store floating point numbers was standardized as 32 bits.

By: Prof. Anand Motwani
Faculty, SCSE,
VIT Bhopal University

IEEE Standard for Floating Point Representation

- This standard, called IEEE Standard 754 for floating point numbers, was adopted in 1985 by all computer manufacturers.
- It allowed porting of programs from one computer to another without the answers being different.
- The standard was updated in 2008. The current standard is IEEE 754-2008 version. It also introduced standards for representing decimal floating point numbers.

By: Prof. Anand Motwani
Faculty, SCSE,
VIT Bhopal University

IEEE Standard for Floating Point Representation

- This standard is now used by all computer manufacturers while designing floating point arithmetic units so that programs are portable among computers.

By: Prof. Anand Motwani
Faculty, SCSE,
VIT Bhopal University

Floating-Point Standards

- The IEEE has established a standard for floating-point numbers
- The IEEE-754 *single precision* floating point standard uses an 8-bit exponent (with a bias of 127) and a 23-bit mantissa (significand).
- The IEEE-754 *double precision* standard uses an 11-bit exponent (with a bias of 1023) and a 52-bit mantissa (significand).

By: Prof. Anand Motwani
Faculty, SCSE,
VIT Bhopal University

- a floating point number in the IEEE Standard is
- Bias = 127

- $(-1)\ s \times (1.\ f\ )_2 \times 2^{ex - 127}$

By: Prof. Anand Motwani
Faculty, SCSE,
VIT Bhopal University

- Thus an exponent 0 means that –127 is stored in the exponent field.
- A stored value 198 means that the exponent value is (198 – 127) = 71.
- The exponents –127 (all 0s) and + 128 (all 1s) are reserved for representing special numbers which we discuss later.

By: Prof. Anand Motwani
Faculty, SCSE,
VIT Bhopal University

- Example. Represent 52.21875 in IEEE 754 – 32-bit floating point format.
- 52.21875 = 110100.00111 = 1.1010000111 × 2^5
- Normalized significand = .1010000111.
- Exponent: (e – 127) = 5 or, e = 132.
- The bit representation in IEEE format is

0	10000100	10100001110000000000000

By: Prof. Anand Motwani
Faculty, SCSE,
VIT Bhopal University

IEEE 754-1985

- All 0s for the exponent is not allowed to be used for any other number. If the sign bit is 0 and all the other bits 0, the number is +0.
- If the sign bit is 1 and all the other bits 0, it is –0. Even though +0 and –0 have distinct representations they are assumed equal.
- All exponent bit 1 with all mantissa bits 0 represents infinity. Sign bit 0 then +~ and 1 then -~.
- All exponent bits 1 and mantissa bits non-zero is error.
- When an arithmetic operation is performed on two numbers which results in an indeterminate answer, it is called NaN (Not a Number)

By: Prof. Anand Motwani
Faculty, SCSE,
VIT Bhopal University

Computer Arithmetic

- Pseudo code for adding two nos. (say 4 bit)
- $x_3x_2x_1x_0$ and $y_3y_{2y}y_1y_0$
- int carry = 0;
- for (int i = 0; i<N; i++)

  {

  int sum = x_i + y_i + carry;

  Z_i = sum % 2;

  if (sum >=2)

  carry = 1;

  }

```
  1101
  1101
______
 11010
```

Overflow in unsigned no.

By: Prof. Anand Motwani
Faculty, SCSE,
VIT Bhopal University

Adding 2's complement No.

- Pseudo code for adding two nos. (say 4 bit)
- $x_3x_2x_1x_0$ and $y_3y_{2y}y_1y_0$
- int carry = 0;
- for (int i = 0; i<N; i++)

```
{
   int sum = x_i + y_i + carry;
   Z_i = sum % 2;
   if (sum >=2)
          carry = 1;
}
```

1101	(-3)
1101	(-3)
11010	(-6)

Overflow in 2s complement

Answer = - 6

By: Prof. Anand Motwani
Faculty, SCSE,
VIT Bhopal University

Largest and Smallest Positive Floating Point Numbers:

Largest Positive Number

0	11111110	11111111111111111111111
Sign 1 bit	Exponent 8 bits	Significand 23 bits

Significand: $1111 \ldots 1 = 1 + (1 - 2^{-23}) = 2 - 2^{-23}$.
Exponent: $(254 - 127) = 127$.
Largest Number $= (2 - 2^{-23}) \times 2^{127} \cong 3.403 \times 10^{38}$.
If the result of a computation exceeds the largest number that can be stored in the computer, then it is called an *overflow*.

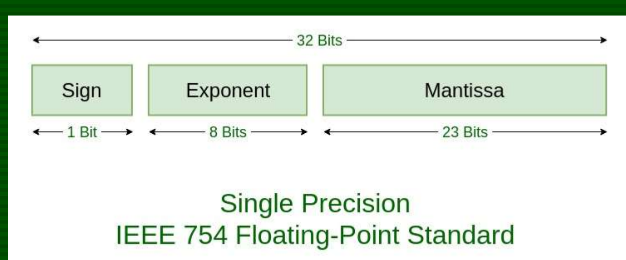
Smallest Positive Number

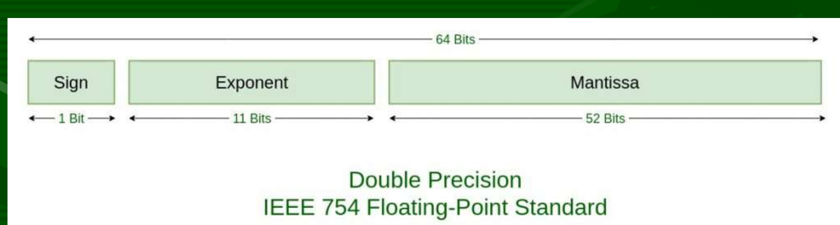
0	00000001	00000000000000000000000
Sign 1 bit	Exponent 8 bits	Significand 23 bits

Significand = 1.0.
Exponent = $1 - 127 = -126$.
The smallest normalized number is $= 2^{-126} \cong 1.1755 \times 10^{-38}$.

Subnormal Numbers (IEEE standard)

- When all the exponent bits are 0 and the leading hidden bit of the siginificand is 0, then the floating point number is called a subnormal number.
- Thus, one logical representation of a subnormal number is $(-1)^s \times 0.f \times 2^{-127}$ (all 0s for the exponent).
- where f has at least one 1 (otherwise the number will be taken as 0).





By: Prof. Anand Motwani
Faculty, SCSE,
VIT Bhopal University

Addition of Floating Point Numbers

	Sign	Exponent	Fraction
X	0	1001	110
Y	0	0111	000

- Find Normalized scientific notation for X and Y
- X is 1.110×2^2
- Y is 1.000×2^0

$(-1)^s \times (1.f)_2 \times 2^{\text{excess} - \text{bias}}$

By: Prof. Anand Motwani
Faculty, SCSE,
VIT Bhopal University

- In order to add, the exponents of two nos. must be same. To do so, just rewrite Y. Now Y is not being normalized but value is not changed.
- So Y can be re-written as:
- Y is $.0100 \times 2^2$. The readjusted value, call it Y' .
- Now add $(1.110)_2$ and $(0.01)_2$. The same is = $(10.0)_2$
- The exponent is same.
- Now shift the radix point to left by 1, and increase the exponent by 1. The result is 1.000×2^3
- Now represent in floating point.
- X + Y = 0 1010 000

By: Prof. Anand Motwani
Faculty, SCSE,
VIT Bhopal University

Numerical Exercises

- What is the normalized representation of
- $0.232 \times 10^3 = 23.2 \times 10^1 = 2.32 * 10^2$
- Ans: $011101000 = 1.1101 \times 2^7$
- Calculate Binary Representation:

0	10000110	11010000000000000000000

- What's the normalized representation of 0.0001101001110

  Ans: $1.110100111 \times 2^{-4}$
- What's the normalized representation of 00101101.101

  Ans: 1.01101101×2^5

By: Prof. Anand Motwani
Faculty, SCSE, VIT Bhopal University

Quiz

- 1. All 1s in the exponent field is assumed to represent ____________
- When all the exponent bits are 0 and the leading hidden bit of the significand is 0, then the floating point number is called _________________.

By: Prof. Anand Motwani
Faculty, SCSE, VIT Bhopal University