Computer Organization

605.204

Module Two

Part Four

Floating Point Workshop

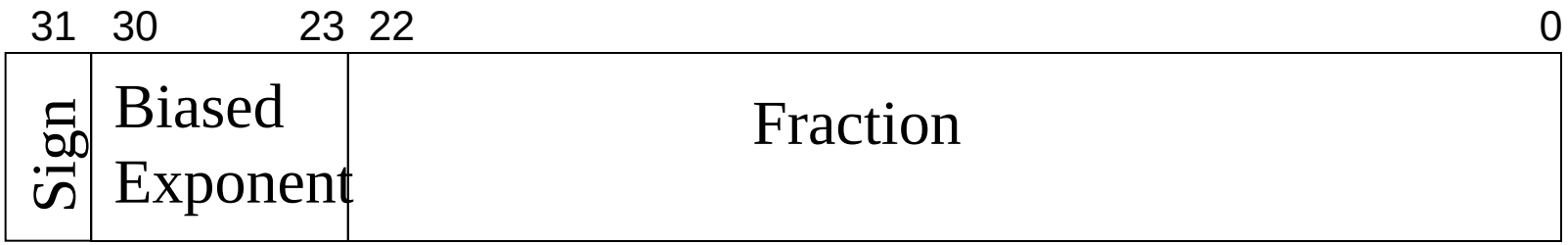
Module Two

- Part Four

- This presentation is a
- Floating Point Number Workshop

IEEE 754 Floating Point Format

- IEEE 754
  - single precision: 8 bit exponent, 23 bit fraction
  - **value**: sign(-1) x 1+fraction x 2(exponent-127)



1 01111110 100 0000 0000 0000 0000 0000

Sign ==> negative value;

Exponent = 126 - 127 = -1 ==> 2^{-1}

Fraction = $1.1_2 = 1.5_{10}$

- **Value = - 0.75**

Workshop

- Convert: 125.625 to IEEE 754 format

  – First the **integer** - 125 to binary

  125 = 64 + 32 +16 + 8 + 4 + 1

  64 + 32 +16 + 8 + 4 + 1 == > 1111101_2

Floating Point

- Convert: 125.625 to IEEE 754 format
  - Then the **fraction** - .625 to binary

0.625	x 2
1.250	x 2
0.500	x 2
1.000	x 2
0.000	

  - Therefore the fraction is .101000 …

Floating Point

- Convert: 125.625 to IEEE 754 format
  - **Normalize** the value:

1111101.101000

Adjusting the binary point left 6 places: 1.111101.101000

So now the exponent value is +6, and

$1.111101101000 * 2^6$

Floating Point

- Convert: 125.625 to IEEE 754 format
  - **Calculate** the exponent:

$1.111101101000 * 2^6$

Adding the bias (127) to the exponent ==> 127 + 6 = 133

$133_{10} = 10000101_2$

Floating Point

- Convert: 125.625 to IEEE 754 format
  - Get the fraction value: $1.111101101000 * 2^6$

    The fraction is . 111101101000

  - And, the value is greater than zero, so the sign value is 0

Floating Point

- Convert: 125.625 to IEEE 754 format

  Assemble the fields: $1.111101101000 * 2^6$

  The sign is 0

  The exponent value is 10000101_2

  The fraction is . 111101101000

  0 10000101 11110110100000000000000

  0100 0010 1111 1011 0100 0000 0000 0000

  4 2 F B 4 0 0 0

Summary

- Convert decimal value to IEEE 754 format:
  - Convert integer part to binary
  - Convert the decimal fraction to binary
  - Normalize the result
  - Calculate the biased exponent
  - Get the fraction
  - Set the sign bit
  - Assemble the fields.