Lecture 14

Floating Point Arithmetic (Chapter 7)

Khaza Anuarul Hoque
ECE 4250/7250

Introduction

- Arithmetic units for floating-point numbers are more complex than those for fixed-point numbers.
- Floating-point numbers allow very large or very small numbers to be specified.

Representation of Floating-Point Numbers: IEEE 754 Standard

- IEEE 754 is a floating-point standard established by the IEEE in 1985.
- It contains two representations for floating-point numbers:
  - **Single precision**: uses 32 bits.
  - **Double precision**: uses 64 bits.
- Designers of IEEE 754 desired a format that was easy to sort and hence adopted a **sign-magnitude system** for the **fractional part** and a **biased notation** for the **exponent**.

Representation of Floating-Point Numbers: IEEE 754 (continued)

- IEEE Single Precision Floating-Point Format:
  - 32 bits:

Sign	Exponent	Fraction
1 bit	8 bits	23 bits

- IEEE Double Precision Floating-Point Format:
  - 64 bits:

Sign	Exponent	Fraction
1 bit	11 bits	52 bits

Example 1

For an 8 bit word, determine the range of values that it represents in floating point and the accuracy of presentation for the following scenarios: (Assume a hidden 1 representation and extreme values are not reserved).

a) If 3bits are assigned to the exponents

b) If 4 bits are assigned to the exponents

S	E	M

More Examples

- Represent 21.75 in Floating point. Use the IEEE 754 standard

- Represent -0.4375 in floating point, using IEEE standard 754