3 Lecture - CS302

Important Subjective

1. **What is a floating-point number, and how is it represented in computer systems?**
   Answer: A floating-point number is a numerical data representation used in computing to represent real numbers with high precision. It is represented using a significand and an exponent.

2. **What is the difference between single-precision and double-precision floating-point numbers?**
   Answer: Single-precision floating-point numbers use 32 bits to represent a number, while double-precision floating-point numbers use 64 bits. Double-precision numbers provide greater precision than single-precision numbers.

3. **How are floating-point numbers stored in memory?**
   Answer: Floating-point numbers are stored in memory using a binary representation. The bits are divided into a significand and an exponent, which are combined to represent the actual value of the number.

4. **What is the difference between normalized and denormalized floating-point numbers?**
   Answer: Normalized floating-point numbers have a leading 1 bit in the significand, while denormalized floating-point numbers have a leading 0 bit in the significand. Denormalized numbers have reduced precision and are used to represent very small numbers.

5. **What is a NaN in floating-point arithmetic?**
   Answer: NaN stands for "Not a Number" and is a special value used to indicate that a mathematical operation has resulted in an undefined or indeterminate value.

6. **How do rounding errors occur in floating-point arithmetic?**
   Answer: Rounding errors occur when a floating-point number is rounded to fit into a limited number of bits. This can result in small errors in the actual value of the number.

7. **What is the difference between relative and absolute error in floating-point arithmetic?**
   Answer: Absolute error is the difference between the actual value and the calculated value of a number, while relative error is the absolute error divided by the actual value.

8. **What is the significance of the machine epsilon in floating-point arithmetic?**
   Answer: The machine epsilon is the smallest positive floating-point number that can be added to 1 and result in a different value. It is used to determine the precision of a floating-point number.

9. **What are the advantages and disadvantages of using floating-point numbers in computing?**
   Answer: The advantages of using floating-point numbers include their ability to represent a wide range of real numbers with high precision. The disadvantages include their limited precision and potential for rounding errors.

10. **How does the IEEE 754 standard for floating-point arithmetic address the issues of precision and rounding errors?**

Answer: The IEEE 754 standard defines the format for representing floating-point numbers in binary form and specifies the rules for performing arithmetic operations on them. It includes provisions for rounding and handling of special values like NaNs.