# Day-11

## Floating-point Types and Arithmetic

### IEEE Floating-point Standard and Other Representations

### IEEE 부동 소수점 표준

- IEEE754는 부동 소수점 연산을 위한 기술 표준
- 이 표준은 binary 형식, 연산 동작, 반올림 규칙, 예외 처리 등을 정의함
- 일반적으로 C/C++은 IEEE754 부동소수점 표준을 채택함

### 32/64-bit Floating -Point

- **IEEE754 Single-precision** (32-bit) **float**
  - 부호비트(Sign) : 1-bit
  - 지수부(Exponent) : 8-bit
  - 가수부(Mantissa) : 23-bit
- **IEEE754 Double-precision** (64-bit) **double**
  - 부호비트(Sign) : 1-bit
  - 지수부(Exponent) : 11-bit
  - 가수부(Mantissa) : 52-bit

### 128/256-bit Floating-Point

- **IEEE754 Quad-Precision** (128-bit) **std::float128** C++23
  - 부호비트(Sign) : 1-bit

    - 지수부(Exponent) : 15-bit
    - 가수부(Mantissa) : 112-bit
- **IEEE754 Octuple-Precision** (256-bit) C++에서 비표준화
    - 부호비트(Sign) : 1-bit
    - 지수부(Exponent) : 19-bit
    - 가수부(Mantissa) : 236-bit

## 16-bit Floating-Point

- **IEEE754 16-bit Floating-Point**(**std::binary16**) C++23 → GPU, Arm7
- **Google 16-bit Floating-Point**(**std::bfloat16**) C++23 → TPU, GPU, Arm8

## 8-bit Floating-Point (C++/IEEE에서 비표준화)

- E4M3
- E5M2

## 기타 실수값 표현(C++/IEEE에서 비표준화)

- TensorFloat-32(TF32)
    - 딥 러닝 어플리케이션을 위한 특수 부동소수점 형식
- Posit
    - unum III 라고도 불림
    - 지수부와 가수부의 길이가 가변적인 부동소수점
- Microscaling Formats(MX)
    - AMD, Arm, Intel, Meta, Microsoft, NVIDIA 및 Qualcomm에서 정의한 저정밀 부동 소수점 형식을 위한 사양
    - FP8, FP6, (MX)INT8이 포함되어있음
- Fixed-point
    - 기수점(소수점) 뒤에 자릿수가 고정되어 있음

- 인접한 숫자간의 간격은 항상 동일
- 값의 범위가 상당히 제한적
- 임베디드 시스템에서 널리 사용됨