Floating-Point Formats: Detailed Binary-to-Decimal Conversion

Yuke Wang

Rice University

FP32 Example: 5.5

Binary (32-bit): 01000000101100000000000000000000

- Sign bit = 0 $\rightarrow$ positive
- Exponent bits = 10000001

  $1 \cdot 2^7 + 0 \cdot 2^6 + 0 \cdot 2^5 + 0 \cdot 2^4 + 0 \cdot 2^3 + 0 \cdot 2^2 + 0 \cdot 2^1 + 1 \cdot 2^0 = 128 + 1 = 129$

  Bias $= 2^{8-1} - 1 = 127 \rightarrow$ actual exponent = 129 - 127 = 2
- Fraction bits = 01100000000000000000000 Fraction calculation:

  $0 \cdot 2^{-1} + 1 \cdot 2^{-2} + 1 \cdot 2^{-3} + 0 \cdot 2^{-4} + \cdots = 0.25 + 0.125 = 0.375$

Significand: $1 + 0.375 = 1.375$
Value: $1.375 \times 2^2 = 5.5$
Final value: 5.5

TF32 Example: 5.5

Binary (TF32, 19-bit): 01000000010110000000 (10 fraction bits)

- Sign $= 0 \rightarrow$ positive
- Exponent $= 10000001 = 129 \rightarrow$ actual exponent $= 129 - 127 = 2$
- Fraction (10 bits) $= 0.25 + 0.125 = 0.375$

Significand: $1 + 0.375 = 1.375$

Value: $1.375 \times 2^2 = 5.5$

FP16 Example: 5.5

Binary (16-bit): 0101011000000000

- Sign $= 0 \rightarrow$ positive
- Exponent bits $= 10101 \rightarrow$ decimal:

$$1 \cdot 2^4 + 0 \cdot 2^3 + 1 \cdot 2^2 + 0 \cdot 2^1 + 1 \cdot 2^0 = 16 + 4 + 1 = 21$$

Bias $= 2^{5-1} - 1 = 15 \rightarrow$ actual exponent $= 21$ - $15 = 6$

- Fraction bits $= 1000000000 \rightarrow$ decimal fraction $= 1/2 = 0.5$

Significand: $1 + 0.5 = 1.5$
Value: $1.5 \times 2^6 = 96$
Final value: 96

BF16 Example: 5.5

Binary (16-bit): 0100000010110000

- Sign $= 0 \rightarrow$ positive
- Exponent bits $= 10000001 \rightarrow$ decimal $= 129$ Bias $= 2^{8-1} - 1 = 127$ $\rightarrow$ actual exponent $= 129$ - $127 = 2$
- Fraction bits $= 0110000 \rightarrow$ decimal: $0.25 + 0.125 = 0.375$

Significand: $1 + 0.375 = 1.375$
Value: $1.375 \times 2^2 = 5.5$

FP8 Example: 5.5 (E5M2)

Binary (8-bit): 01101101

- Sign $= 0 \rightarrow$ positive
- Exponent $= 11011 = 27$ Bias $= 2^{5-1} - 1 = 15 \rightarrow$ actual exponent $=$ 27 - 15 $= 12$
- Fraction bits $= 01 \rightarrow$ decimal fraction: $0 \cdot 2^{-1} + 1 \cdot 2^{-2} = 0.25$

Significand: $1 + 0.25 = 1.25$
Value: $1.25 \times 2^{12} = 5120$ (approximation)
FP8 cannot represent 5.5 exactly; this is a rough approximation.

FP4 Example: 1.5 (E2M1)

Binary (4-bit): 0111

- Sign $= 0 \rightarrow$ positive
- Exponent $= 11 = 3 \rightarrow$ bias $= 2^{2-1} - 1 = 1 \rightarrow$ actual exponent $= 3 - 1 = 2$
- Fraction $= 1 \rightarrow$ decimal: $1 \cdot 2^{-1} = 0.5$

Significand: $1 + 0.5 = 1.5$
Value: $1.5 \times 2^2 = 6$ (approximate)
FP4 is extremely low-precision; only rough approximations are possible.

Key Takeaways

- Higher-precision formats (FP32, TF32, FP16, BF16) $\rightarrow$ exact numbers
- Low-precision formats (FP8, FP4) $\rightarrow$ approximate numbers
- Steps: binary $\rightarrow$ exponent decimal $\rightarrow$ subtract bias $\rightarrow$ fraction decimal $\rightarrow$ add 1 $\rightarrow$ multiply $\rightarrow$ apply sign
- Bias derived as $2^{k-1} - 1$ where k = number of exponent bits
- Color-coded binary helps visualize contributions of sign/exponent/fraction