Floating-point Number in IEEE 754 Format

Floating-point number in digital world

- Floating-point number in digital world
  - A. Single-precision (32 bits)
    - Representation:
      - Sign: 1 bit; exponent: 8 bits; mantissa: 23 bits
    - Decimal to IEEE 754 Conversion
      - Example: Decimal 9.75_{10} to IEEE 754

1) In binary is 1001.11_2 ; In binary scientific notation: 1.00111×2^3

2) Sign bit: positive so it is 0

3) Exponent is 3 but it must be stored with a bias of 127 (why?), so it is

3+127=130 which in binary is 10000010

5) Mantissa: the fractional part after the leading 1, which is 00111000000000000000000 (the trailing zeroes are padded to make it 23 bits).

Sign	Exponent	Mantissa
0	10000010	00111000000000000000000

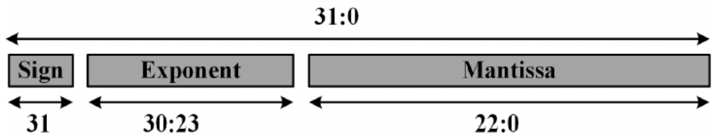

FIGURE 5.2
IEEE 754 Standard (Single Precision): $(-1)^S \times (1.0 + M) \times 2^{(E-127)}$

Floating-point number in digital world

- Floating-point number in digital world
  - A. Single-precision
    - IEEE 754 to Decimal Conversion
      - Example:

```
0 10000010 10110000000000000000000
```

1) Sign bit: 0 so it is positive
2) Exponent: 10000010_2 in decimal is 130_{10}, so the actual exponent is 130-127=3 with a bias of 127
3) Mantissa: is represented with an implicit leading 1 for normalized numbers, so the mantissa is actually 1.mantissa_bits
   In this example, therefore, it should be 1.1011_2
   Covert to decimal, it should be $1 + 2^{-1} + 2^{-3} + 2^{-4} = 1.6875$
4) In floating-point, it is
   $(-1)^{sign} \times 2^{exponent-127} \times (1.mantissa) = (-1)^0 \times 2^3 \times (1.6875) = 13.5$

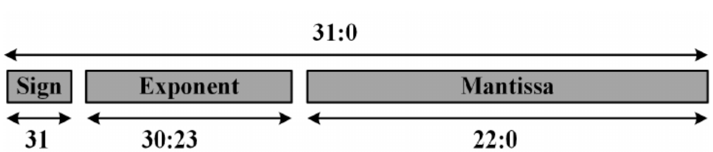

FIGURE 5.2
IEEE 754 Standard (Single Precision): $(-1)^S \times (1.0 + M) \times 2^{(E-127)}$

Floating-point number in digital world

- Floating-point number in digital world
  - A. Single-precision
    - Range: $1.175 \times 10^{-38} \sim 3.403 \times 10^{38}$
      - The largest positive value occurs when the sign bit is 0, the exponent is at its maximum value of 254 (11111110 in binary, corresponding to an actual exponent of 127), and the mantissa is all 1s (111...111).
      - The smallest positive normalized value occurs when the sign bit is 0, the exponent is at its minimum value of 1 (00000001 in binary, corresponding to an actual exponent of −126), and the mantissa is all 0s (000...000).
    - Infinity: Represented when the exponent is all 1s and the mantissa is all 0s
    - Not a Number (NaN): Represented when the exponent is all 1s and the mantissa is non-zero.

Floating-point number in digital world

- Floating-point number in digital world
  - B. Half-precision (16 bits)
    - Representation: Sign--1 bit; exponent--5 bits; mantissa--10 bits
    - Range: $6.10 \times 10^{-5} \sim 6.55 \times 10^{4}$
    - Usage: In memory-constrained applications such as machine learning and mobile graphics, where lower precision can be tolerated.
  - C. Double-precision (64 bits)
    - Representation: Sign--1 bit; exponent--11 bits; mantissa--52 bits
    - Range: $2.225 \times 10^{-308} \sim 1.8 \times 10^{308}$
    - Usage: For higher-precision computations in scientific and engineering applications.
  - D. Quadruple precision (128 bits)
    - Representation: Sign--1 bit; exponent--15 bits; mantissa--112 bits
    - Range: $3.36 \times 10^{-4932} \sim 1.18 \times 10^{4932}$
    - Usage: In applications requiring extremely high precision, such as numerical simulations, scientific computing, and high-precision financial calculations.

- SV Functions
  - IEEE 754 to SP Conversion

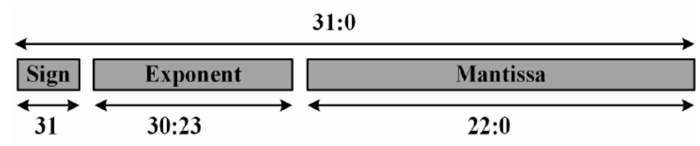

FIGURE 5.2
IEEE 754 Standard (Single Precision): $(-1)^S \times (1.0 + M) \times 2^{(E-127)}$

```
function real ieee754_to_fp (input [31:0] ieee754_data);
reg         sign     ;
reg [7:0]   exponent;
reg [22:0]  mantissa;

integer int_exp      ;
real    mantissa_val ; // Divide by 2^23
real    fp_output    ;

// Extracting sign, exponent, and mantissa bits
sign     = ieee754_data[31]    ;
exponent = ieee754_data[30:23];
mantissa = ieee754_data[22:0]  ;

// Calculating floating-point value
int_exp      = exponent-127;
mantissa_val = 1.0+(mantissa/8388608.0); // Divide by 2^23
fp_output    = (sign?-1:1)*mantissa_val*(2.0**int_exp);

return fp_output;
endfunction
```

IEEE 754 Table

- IEEE 754 V.S SP Number

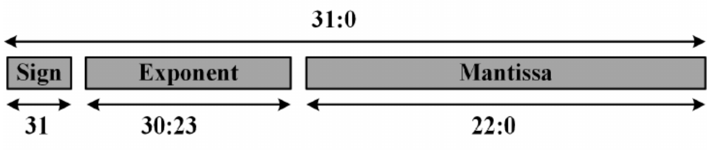

FIGURE 5.2
IEEE 754 Standard (Single Precision): $(-1)^S \times (1.0 + M) \times 2^{(E-127)}$

TABLE 8.2
IEEE 754 Format for FP Numbers

FP	**1.0**	**2.0**	**3.0**	**4.0**	**5.0**	**6.0**
HW	3f800000	40000000	40400000	40800000	40a00000	40c00000
FP	**7.0**	**8.0**	**9.0**	**10.0**	**11.0**	**12.0**
HW	40e00000	41000000	41100000	41200000	41300000	41400000
FP	**-1.0**	**-2.0**	**-3.0**	**-4.0**	**-5.0**	**-6.0**
HW	bf800000	c0000000	c0400000	c0800000	c0a00000	c0c00000
FP	**-7.0**	**-8.0**	**-9.0**	**-10.0**	**-11.0**	**-12.0**
HW	c0e00000	c1000000	c1100000	c1200000	c1300000	c1400000