**Exercises**

**Floating Point**

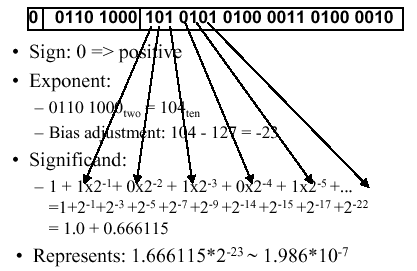
The IEEE 754 standard defines a binary representation for floating point values using three fields:

• The *sign* determines the sign of the number (0 for positive, 1 for negative)

• The *exponent* is in **biased notation** with a bias of 127

• The *significand* is akin to unsigned, but used to store a fraction instead of an integer.

**1.**The below numbers shows the bit breakdown for the single precision (32-bit) representation:



Convert the numbers from binary to decimal.

***∵ E = (01101000)B = 8+32+64 = 104***

***∵ bias = 2^7 – 1 = 127***

***∴ v = (-1)^S \* (1.significand)B \* 2^(E - bias)***

***= (-1)^0 \* (1.10101010100001101000010)B \* 2^(-23)***

***= 13976386‬ \* 2^(-23) \* 2^(-23)***

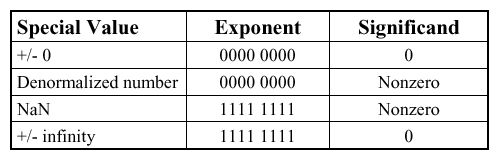
***= 1.9861639088958e-7***

**2.** There is also a double precision encoding format that uses 64 bits. Convert the above numbers from a single precision binary to a double precision binary.

***∵ bias = 2^10 – 1 = 1023***

***∴ E = -23 + bias = -23 + 1023 = 1000 = 01111101000B***

***∴ 0 011 1110 1000 1010 1010 1000 0110 1000 0100 0000 0000 0000 0000 0000 0000 0000***

**3.As Table.** Special values in the IEEE 754 standard.  


1). What is the largest finite positive value that can be stored using a single precision float?

***∵ 0 1111 1110 111 1111 1111 1111 1111 1111***

***∵ E = (1111 1110)B = 254***

***∴ v = (-1)^S \* (1.Significand)B \* 2^(E-bias)***

***= (1.11111111111111111111111)B \* 2^(127)***

***= (2 - 2^(-23)) \* 2^(127)***

***= 3.4028234663853e+38***

2). What is the smallest positive value that can be stored using a single precision float?

***∵ 0 0000 0000 000 0000 0000 0000 0000 0001***

***∴ v = (-1)^S \* (0.Significand) \* 2^(1-bias)***

***= (0.00000000000000000000001)B \* 2^(-126)***

***= 2^(-23) \* 2^(-126)***

***= 1.4012984643248e-45***

3). What is the smallest positive normalized value that can be stored using a single precision float?

***∵ 0 0000 0001 000 0000 0000 0000 0000 0000***

***∴ v = (-1)^S \* 1.Significand \* 2^(E - bias)***

***= (1)B \* 2^(-126)***

***= 1.1754943508223e-38***

4). Convert the following numbers from binary to decimal or from decimal to binary:

0x00000000 8.25 0x00000F00 39.5625 0xFF94BEEF -∞

***0x00000000 = (0)B***

***= 0***

***8.25 = 33 \* 2^(-2)***

***= (100001 \* 0.01)B***

***= (1000.01)B***

***0x00000F00 = (1111 0000 0000)B***

***= 15 \* 16^2***

***= 3 840***

***39.5625 = 39 + 0.5 + 0.0625***

***= (100111 + 0.1 + 0.0001)B***

***= (100111.1001)B***

***0xFF94BEEF = (1111 1111 1001 0100 1011 1110 1110 1111)B***

***= 15 + 14\*16 + 14\*16^2 + 11\*16^3 + 4\*16^4 +***

***9\*16^5 + 15\*16^6 + 15\*16^7***

***= 4 287 938 287***

***-∞ = (1 1111 1111 000 0000 0000 0000 0000 0000)B***

***= 2^32 - 1 - (2^23 - 1)***

***= 4 286 578 688***

**4.**Write the following numbers as powers of 2: 2 Ki, 256 Pi, 512 Ki, 64 Gi, 16 Mi, 128 Ei.

***2 Ki = 2 \* 2^10***

***= 2^11***

***256 Pi = 2^8 \* 2^50***

***= 2^58***

***512 Ki = 2^9 \* 2^10***

***= 2^19***

***64 Gi = 2^6 \* 2^30***

***= 2^36***

***16 Mi = 2^4 \* 2^20***

***= 2^24***

***128 Ei = 2^7 \* 2^60***

***= 2^67***

**5**.assume an 8 bit integer and answer each one for the case of a two's complement number, and unsigned number, indicating if it cannot be answered with a specific representation.

1). What is the largest integer? The largest integer + 1?（unsigned number, two's complement number）

***unsigned number:***

***the largest integer: 2^8 – 1 = 255***

***the largest integer + 1: (1 0000 0000)B = 0***

***two’s complement number:***

***the largest integer: 2^7 – 1 = 127***

***the largest integer + 1: (1000 0000)B = -128***

2). How do you represent the numbers 0, 1, and -1 with unsigned number and a two's complement number?

***unsigned number: 0: 0000 0000***

***1: 0000 0001***

***-1: ×***

***two’s complement number: 0: 0000 0000***

***1: 0000 0001***

***-1: 1111 1111***

3). How do you represent 17, -17 with unsigned number and a two's complement number?

***unsigned number: 17: 0001 0001***

***-17: ×***

***two’s complement number: 17: 0001 0001***

***-17: 1110 1111***