

Floating Point Representation

Write the following binary numbers in normalized exponential form.

- (i) $111010011. = 0.111010011 \times 2^9$
- (ii) $0.00001110101 = 0.1110101 \times 2^{-4}$
- (iii) $1101.001101 = 0.1101001101 \times 2^4$
- (iv) $111010011.00011010111 = 0.11101001100011010111 \times 2^9$

If 7 bits in a 16-bit floating point representation are reserved for the mantissa and truncated rather than rounded at 7 bits:

1. **Identify the mantissa and the exponent and compute the characteristic if the exponent bias is $2^7 - 1$ for each of the numbers above. Write the computer representation for each;**

- (i) 0.111010011×2^9 . The bias is $10000000 - 1$ and the exponent is 1001, so the characteristic is $10001001 - 1 = 10001000$. Since the number is positive the computer representation is 0100010001110100.
- (ii) 0.1110101×2^{-4} . The bias is 01111111 and the exponent is -100 so the characteristic is 01111011 and the representation is 0011110111110101.
- (iii) 0.1101001101×2^4 . The bias is $10000000 - 1$ and the exponent 100, so the characteristic is $10000000 + 100 - 1 = 10000000 + 011 = 10000011$. The representation is 0100000111101001.
- (iv) $0.11101001100011010111 \times 2^9$. Since only 7 bits of the mantissa are stored the computer representation is the same as for (i), namely 0100010001110100.

2. **Find the computer representation for -1871 .**

$1871 = 1024 + 847 = 1024 + 512 + 335 = 1024 + 512 + 256 + 97 = 1024 + 512 + 256 + 64 + 33 = 1024 + 512 + 256 + 79 = 1024 + 512 + 256 + 64 + 15 = 1024 + 512 + 256 + 79 = 1024 + 512 + 256 + 64 + 8 + 4 + 2 + 1$, or in binary notation 11101001111 = $0.11101001111 \times 2^{11}$ The exponent in binary is 1011 and the bias is $10000000 - 1$ so the characteristic is $10001011 - 1 = 10001010$. Since the number is negative the representation is 1100010101110100.

Notice that the last four bits of the mantissa were discarded so that in fact all numbers greater than or equal to 11101000000 (1856) and less than 11101010000 (1872) will have the same representation.

3. **What will this computer's response be to the test $1869 > 1861$?**

0 the computer's representation of the two numbers will be the same.

4. **What will be the computer's result from the computation $1872 - 1862$?**

The representation of the two numbers will differ in the last bit, so the difference will be $0.0000001 \times 2^{11} = 10000_2 = 16$.

- (ii) 0.1110101×2^{-4} . The bias is 111 and the exponent is -100 so the characteristic is 011 and the representation is 0 0011 11101010000.
- (iii) 0.1101001101×2^4 . The bias is 111 and the exponent 100, so the characteristic is 11011. The representation is 0 1011 11010011010.
- (iv) $0.11101001100011010111 \times 2^9$ is also out of range.
- (b) $-1871 = 0.11101001111 \times 2^{11}$ The exponent in binary is 1011 and the bias is 111 so the characteristic requires more than 4 bits and is out of range.
- (c) **What will this computer's response be to the test $1869 > 1861$?**
Overflow
- (d) **What will be the computer's result from the computation $1872 - 1862$?**
Overflow
- (e) **What will be the computer's result from the computation $1871.9999999999 - 1871.9999999999$?**
Overflow
- (f) The integer representations have not changed.
- (g) The normalized forms of 519 and 514 are $0.1000000111 \times 2^{10}$ and 1000000010×2^{10} The characteristic in both cases is 111+1010 so again both numbers are out of range.
- (h) 45592 is obviously out of range.
- (i) **Find the largest and smallest positive numbers that can be stored in this system.**
The smallest number has the representation 0 0000 10000000000 for 0.1×2^{-7} (the characteristic being 0 means the exponent must be the negative of the bias). Thus the smallest number is $2^{-8} = 0.00390625$.
The largest number 0 1111 11111111111 is $0.1111111111 \times 2^8 = 1111111111 \times 2^{-3} = (2^{11} - 1)/8 = 2047/8 = 255.875$
The range is far too restricted with so few bits for the characteristic although we can now distinguish between 108.20 and 108.25 for example; so the accuracy has been improved.
-