ZHEJIANG UNIVERSITY COLLEGE OF INFORMATION SCIENCE AND ELECTRONICS ENGINEERING

HOMEWORK #4

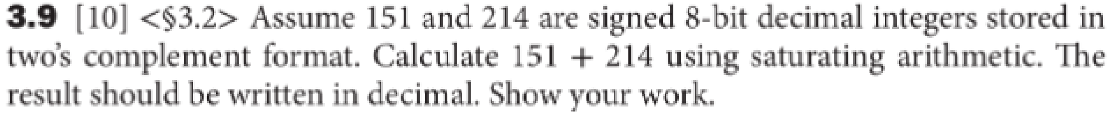
Computer Organization and Design

Name: Student ID:

Major: Electronic Science and Technology

Date: 2024 年 11 月 19 日

Problem 1.



3.9 [10] <§3.2> Assume 151 and 214 are signed 8-bit decimal integers stored in two's complement format. Calculate 151 + 214 using saturating arithmetic. The result should be written in decimal. Show your work.

Answer：

$$151_{10} = 10010111_2$$

$$214_{10} = 11010110_2$$

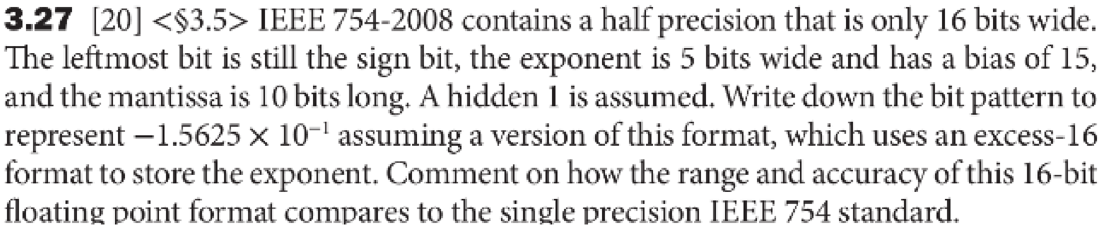
$$10010111_2 = -105_{10}$$

$$11010110_2 = -42_{10}$$

$$10010111_2 + 11010110_2 = 1\ 01101101_2 = -147_{10}$$

the result overflowed.Accoring to the saturing arithmetic, the result should be -128.

Problem 2.



3.27 [20] <§3.5> IEEE 754-2008 contains a half precision that is only 16 bits wide. The leftmost bit is still the sign bit, the exponent is 5 bits wide and has a bias of 15, and the mantissa is 10 bits long. A hidden 1 is assumed. Write down the bit pattern to represent -1.5625×10^{-1} assuming a version of this format, which uses an excess-16 format to store the exponent. Comment on how the range and accuracy of this 16-bit floating point format compares to the single precision IEEE 754 standard.

Answer :

$$-1.5625 \times 10^{-1} = -0.15625_{10} = -0.00101_2 = -1.01_2 \times 2^{-3}$$

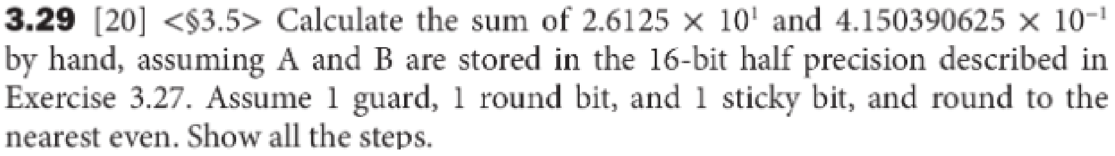
$$s = 1 \quad Exponent = -3 + 15 = 12 = 01100_2 \quad Mantissa = 0100000000_2$$

Thus,the 16-bit floating-point representation of -0.15625 is 1 01100 0100000000.

The 16-bit half precision floating point format has a smaller range and lower accuracy compared to the single precision IEEE 754 standard.

- Range: The single precision format uses 8 bits for the exponent, allowing for a larger range of representable values compared to the 5-bit exponent in the half precision format.
- Accuracy: The single precision format has a 23-bit mantissa, providing higher precision than the 10-bit mantissa in the half precision format.

Problem 3.



3.29 [20] <§3.5> Calculate the sum of 2.6125 × 10^1 and 4.150390625 × 10^{-1} by hand, assuming A and B are stored in the 16-bit half precision described in Exercise 3.27. Assume 1 guard, 1 round bit, and 1 sticky bit, and round to the nearest even. Show all the steps.

Answer :

$$2.6125 \times 10^1 = 26.125_{10} = 11010.001_2 = 1.1010001_2 \times 2^4$$

$$4.150390625 \times 10^{-1} = 0.4150390625_{10} = 0.011010100100_2 = 1.1010100100_2 \times 2^{-2}$$

$$1.1010100100_2 \times 2^{-2} \times 2^{-2} = 0.0000011010100100_2 \times 2^4$$

$$1.1010001000_2 \times 2^4 + 0.0000011010 \quad 100100_2 \times 2^4 = 1.1010100010 \quad 100100_2 \times 2^4$$

the result with guard = 1, round = 0, and sticky is set . Thus the result round up and the result is:

$$1.1010100011 \times 2^4 = 26.546875_{10} = 2.6546875 \times 10^1$$