Arhitectura Sistemelor de Calcul November, 2021

Seminar 0x03

Cristian Rusu

1 Seminar objectives

In this seminar, we will solve problems involving:

- multiplication and division of integers;
- floating-point representation of real numbers;
- operations on numbers represented in floating-point format.

2 Exercises

1. Perform the multiplications between the following natural numbers written in binary:

   (a) 1110×0011
   (b) 1101×0101
   (c) 0010×1101
   (d) 1010×1101
   (e) 101010×001111
   (f) 011110×110011
   (g) 111000×000111
   (h) 101010×010101
   (i) 11110000×00001111
   (j) 10000001×11100011

2. Repeat the operations from the previous exercise, but consider the given numbers to be integers written in two's complement.

3. In many situations, we need to perform operations with special operands. Given a natural number a, consider the following examples and describe simplified ways to compute:

   (a) $a \times 2$
   (b) $a \times 16$
   (c) $a \times 3$
   (d) $a \times 5$
   (e) $a \times 6$
   (f) $a \times 7$
   (g) $a \div 8$
   (h) $a \mod 16$
   (i) $a \times 17$
   (j) $a \times 72$

4. What happens to the operations in the previous exercise if a is an integer?

5. Perform the divisions between the following natural numbers written in binary

   (a) $101010 \div 10$
   (b) $010101 \div 10$
   (c) $100100100 \div 100$
   (d) $111010101 \div 11$
   (e) $001110010 \div 11$
   (f) $00001111 \div 111$
   (g) $11110000 \div 111$
   (h) $01101110 \div 110$
   (i) $11101010 \div 1010$
   (j) $11111111 \div 1011$

6. Repeat the operations from the previous exercise, but consider the given numbers to be integers written in two's complement.

7. Appendix 1 reviews the structure of a number represented in IEEE 754 Floating Point format. Represent the following numbers in this format:

   (a) 0
   (b) 1.0
   (c) -1.0
   (d) 2.0
   (e) 0.67
   (f) -1313.3125
   (g) 0.1015625
   (h) 39887.5625
   (i) 0.33
   (j) 3.14

8. Given a number a represented in IEEE 754 FP format, perform the following operations using bitwise logical operations as much as possible:

   (a) compute abs(a);
   (b) change the sign of a;
   (c) divide a by 4;
   (d) multiply a by 16;
   (e) let b be another number in IEEE 754 FP format; compute $a \times b$ and compute $a + b$.

9. The following numbers are in IEEE 754 FP format. Find their decimal values:

   (a) 0xFFFFFFFF
   (b) 0x0000FFFF
   (c) 0xFFFF0000
   (d) 0xDEADBEEF
   (e) 0x44361000
   (f) 0xbe580000
   (g) 0xa3358000
   (h) 0x76650000
   (i) 0xFEEDC0DE
   (j) 0xC00010FF

10. In IEEE 754 FP format, zero is represented by 32 bits set to zero. This is a convention; what value would 32 zero bits otherwise represent in IEEE 754 FP format?

11. Given two numbers in IEEE 754 FP format, perform the addition in the following cases:

   (a) 7988747653.0 + 1.0
   (b) 3.14 + 2.72
   (c) 3.14 - 2.72
   (d) -3.14 + 2.72
   (e) 47.0 + 52.99
   (f) -13.3334 + 8.6667
   (g) 0.3185 + 0.3676
   (h) 0.1 + 0.2
   (i) 0.2 + 0.3
   (j) 0.1 - 0.2

12. When implemented using digital circuits, division is much slower than addition and multiplication. Find fast division methods for the following cases:

   (a) $a \div 8$
   (b) $a \div 19$
   (c) $a \div 30$
   (d) $a \div 10$
   (e) $a \div 24$
   (f) $a \div 60$
   (g) $a \div 100$
   (h) $a \div 256$
   (i) $a \div 3.14$
   (j) $a \div 2.72$

13. Answer the following questions about IEEE 754 FP format:

(a) how many distinct numbers can be represented in this format?

(b) how many distinct numbers between 2^{-5} and 2^{-4} can be represented in this format?

(c) how many distinct numbers between 2^{13} and 2^{14} can be represented in this format?

(d) how many distinct numbers between 2^{47} and 2^{49} can be represented in this format?

14. Given a number a in IEEE 754 Floating Point format, compute $[a]$, the integer part of a.

15. We know that $e = \sum_{i=0}^{\infty} \frac{1}{i!}$. Write a program that computes this series and displays the current value (in decimal and hexadecimal) for each partial sum.

16. We have an array containing only positive values in IEEE 754 FP format. The array is very long (it contains $N > 10^6$ elements), and we know that the elements have either very small values ($\ll 1$) or very large values ($\gg 10^9$), with most values being very small. We need to compute the sum of the array elements. What problems might arise in this situation? Propose solutions.

17. We want to compute $\sum_{i=1}^{30000} \frac{1}{i^2}$ using only the float data type (single precision). Compute the series by adding the terms in order ($i = 1, \ldots, 30000$) and then in reverse order ($i = 30000, \ldots, 1$) – since addition is commutative, the result should be identical. Are the two results identical? What is the limit of the sum when, instead of summing 30000 terms, we sum infinitely many terms of the form $\frac{1}{i^2}$? Explain the results obtained.

18. Consider IEEE 754 FP format again. Answer the following questions:

(a) what is the smallest/largest number that can be represented?

(b) what is the smallest positive number that can be represented?

(c) what is the largest number less than 1.0 that can be represented?

(d) what is the smallest number greater than 1.0 that can be represented?

(e) what is the representable number closest to π?

(f) what is the representable number closest to e?

(g) what is the representable number closest to φ?

19. In 1991, during the Gulf War (Operation "Desert Storm"), the U.S. Patriot defense system failed to intercept a SCUD missile launched by the Iraqi armed forces. Internally, the American system used an arithmetic operation to convert discrete time (an integer) into time in seconds (a real number) by multiplying the discrete time by 0.1 (to obtain the number of seconds). The system approximated 0.1 using 24 bits as follows:

$$(0.1)_{10} \approx (0.00011001100110011001100)_2.$$

Tasks:

(a) calculate the binary approximation above;

(b) what is the difference between the calculated value and 0.1;

(c) what is the (time) error after 100 hours of operation;

(d) what is the error if we represent 0.1 in IEEE 754 FP format?

(e) if SCUD missiles can reach a speed of MACH 5, what distance can the missile travel during the calculated time error?

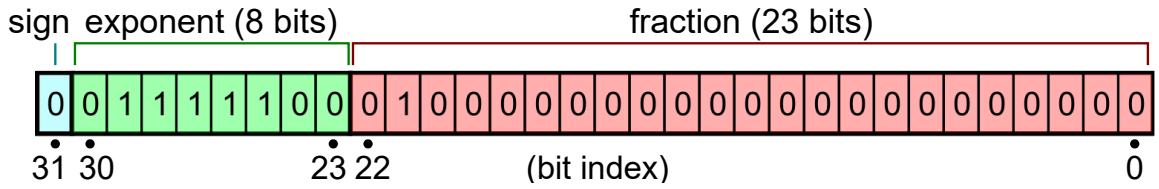
(f) what would you do to correct the problem?

20. Let a be an 8-bit integer. Find efficient procedures for computing:

(a) 1^a, $(-1)^a$, 2^a, $(-2)^a$, 4^a and $(-4)^a$;

(b) a^p, for any natural number p (hint: consider the special case a^{15}).

Appendix 1



An example of representation in IEEE 754 Floating Point format (source: Wikipedia).