

**Question 1** True or false? Justify.

- (a)  $17 \bmod 5 = 2$                       (b)  $5 \bmod 17 = 2$                       (c)  $500 \bmod 17 = 2$   
 (d)  $17 \equiv 5 \pmod{2}$                       (e)  $5 \equiv 17 \pmod{2}$                       (f)  $500 \equiv 2 \pmod{17}$

**Question 2**

- (a) Without a calculator, complete the following table. Try to be as efficient as possible with your calculations.

$x$	$7^x \bmod 11$
1	7
2	5
3	
4	
5	
6	
7	
8	
9	
10	

- (b) Do you notice any patterns in the table? Discuss any patterns you see with your classmates.

**Question 3**

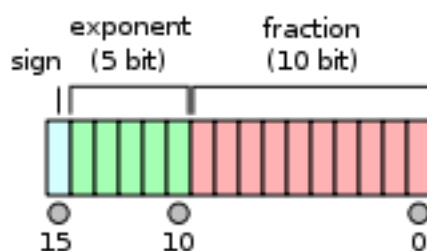
Prove the following statement: An integer is a multiple of 3 if and only if the sum of its digits is a multiple of 3.

**Question 4** Read the following algorithm:

Algorithm for converting a fraction $x$ into binary with $p$ binary places.
Inputs: fraction $x \in \mathbb{Q}$ , $0 < x < 1$ and number of places $p \in \mathbb{N}$ . Initialise: $j \leftarrow 1$ , $b_1, b_2, \dots, b_p \leftarrow 0$ Loop: if $j = p + 1$ stop. $x \leftarrow 2x$ If $x \geq 1$ [ $b_j \leftarrow 1$ , $x \leftarrow x - 1$ ] $j \leftarrow j + 1$ Outputs: bits $b_1, b_2, \dots, b_p$ such that $(0.b_1b_2\dots b_p)_2$ is the best approximation to $x$ using $p$ binary places.

- (a) Verify that the algorithm converts  $\frac{1}{6}$  to  $0.00101_2$  when  $p = 5$ . Can you spot how to express  $\frac{1}{6}$  exactly in repeating binary?  
 (b) Express  $0.45_{10}$  as accurately as possible with 8 binary places.

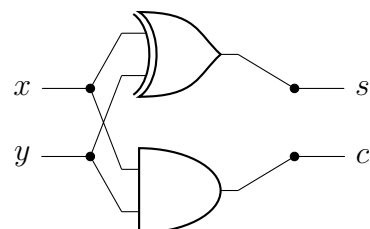
**Question 5** The shortest IEEE standard for representing rational numbers is called *half-precision floating point*. It uses a 16-bit word partitioned as in the diagram at right. (This diagram is taken from the Wikipedia article on the subject, where more details can be found.)



As described in lectures, to store a rational number  $x$  it is first represented as  $(-1)^s \times m \times 2^n$  with  $1 \leq m < 2$ . The sign bit  $s$  is stored as the left-most bit (bit 15), the mantissa  $m$  (called “significand” in IEEE parlance) is stored in the right-most 10 bits (bits 9 to 0), and the exponent  $n$  is stored in the 5 bits in between (bits 14 to 10). However:

- *Only the fractional part of  $m$  is stored.* Because  $1 \leq m < 2$ , the binary representation of  $m$  **always** has the form  $1 \cdot \star \star \star \dots$  where the stars stand for binary digits representing the fractional part of  $m$ . Hence there is no need to store the  $1 \cdot$  part.
  - *the exponent  $n$  is stored with an “offset”.* In order to allow for both positive and negative exponents, but to avoid another sign bit, the value stored is  $n + 15$ . In principle this means that the five exponent bits can store exponents in the range  $-15 \leq n \leq 16$ , but 00000 and 11111 are reserved for special purposes so in fact  $n$  is restricted to the range  $-14 \leq n \leq 15$ .
- (a) A rational number  $x$  is stored in half-precision floating point as the word  $3A6B_{16}$ . (That’s hex shorthand for the 16-bit binary word.) Write  $x$  in ordinary decimal notation.
- (b) Find the word representing  $x = -123.45_{10}$  in half-precision floating point. Use the closest approximation to  $x$  that it is possible to store in this format. Give your answer using hex shorthand, like the word you were given for (a).  
 Note: From a previous questions,  $0.45_{10} = 0.01\overline{1100}_2$ .

**Question 6** In lectures we discussed a circuit for a *half adder* for adding two 1-bit numbers  $x$  and  $y$ . Using an XOR gate  $\overline{\cup}$  to simplify that circuit, it can be drawn as shown at right. The outputs are the ‘sum bit’  $s$  and the ‘carry (out) bit’  $c$ , so that, in binary,  $x + y = cs$ .



Recall that to add multi-bit numbers we use a cascade of *full adders*. A full adder has an additional input of the the carry (out) from the prior addition of digits to the immediate right. To avoid confusion, denote this ‘carry-in’ by  $c_{\text{in}}$  and rename  $c$  as  $c_{\text{out}}$  for ‘carry-out’. (Of course, any carry bit can be 0, meaning ‘no carry’.) So, in binary,  $x + y + c_{\text{in}} = c_{\text{out}}s$ . (If you need to, see the input-output table shown in lectures.)

**Challenge**<sup>1</sup> Design a digital circuit for a full adder. Use any kind of gates that you think will help .

<sup>1</sup>This is less of a challenge if you just look it up on the internet, but try to do it yourself!