Partiel d'Architecture et Système

14 novembre 2014

Duration: 2 hours. Answers can be given in either English or French. Justify all your answers. The computers in the room cannot be used during the exam.

There is a maximum of 30 points to be gained in the questions.

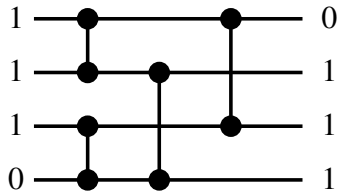
The final mark for the course will be either the average of the marks for this exam and the final exam, or the mark of the final exam, whichever is better.

1 Sorting networks

We first recall some basic facts about *sorting networks*, as treated in one of the exercises.

The purpose of a sorting network is to sort n numbers in ascending order. For a given network, n is fixed. A network is called *correct* if for all possible inputs, the output is sorted, with the lowest element at the top and the highest at the bottom. It is known that a sorting network is correct iff it is correct for all sequences consisting of 0 and 1 (zero-one principle), so we shall assume that all inputs are 0 or 1.

The basic element of a sorting network is a *comparator*. It takes two values and yields two outputs, the upper line being the smaller and the lower line the bigger of the two values. A sorting network consists of n wires, for a fixed n, connected by comparators. For convenience, we shall draw connectors simply as vertical wires. Below is a sorting network for $n = 4$ and an example of its input and output.



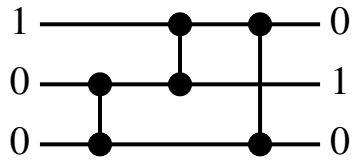
Prove or refute the following statements. It is not enough to say "yes" or "no", you need to justify your response. Every answer is worth 2 points.

(i) In every correct network there is at least one comparator between every pair of neighbouring wires.

Solution: True. Consider a network for n entries where there is no comparator between lines number i and $i+1$ and the sequence $0^{i-1}101^{n-1-i}$. For all comparators in the network the two inputs are already correctly sorted, so nothing changes. Thus, the network is not correct.

(ii) Every network that contains at least one comparator between every pair of wires is correct.

Solution: False. Consider the network shown below:

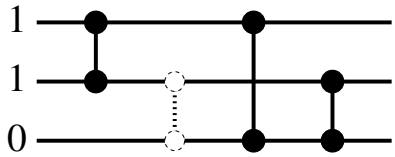


(iii) A correct network stays correct if one adds any comparator at the end.

Solution: True. All inputs are already correctly sorted before arriving at the added comparator, so the latter does not change anything.

(iv) A correct network stays correct if one adds any comparator anywhere in the network.

Solution: False. The network shown below is correct without the comparator shown in white but incorrect if it is included (for instance, for the inputs 1,1,0).



2 Floating-point numbers

We recall some basic facts about floating-point representations. A floating-point number consists of three components, the *sign*, the *exponent*, and the *mantissa*. The value of a triple (s, m, e) is thus $\pm 2^e \cdot m$.

The sign is represented by a single bit, which is 0 for positive and 1 for negative. We speak of a (k, ℓ)-format if k bits are used for the exponent and ℓ bits for the mantissa. Thus, the format used for the `float` data type in C would be a $(8, 23)$-format. As in the IEEE-754 standard we assume that if the bit pattern used for the exponent, interpreted as an unsigned integer, has value E, then $2^{k-1} - 1$ must be subtracted from E to obtain the exponent value e. Thus, in the `float` data type, a value of 128 actually means an exponent of 1. As for the mantissa, its value is in the range $[1, 2)$, where the most significant bit has a weight of $1/2$,

the second most significant bit 1/4, etc. Thus, if $\ell = 3$ and the bit pattern of the mantissa is 101, then the represented value is $1 + 1/2 + 1/8$.

Note: The representations of 0, infinity, and "not a number" will be irrelevant for the following exercises.

Due to the limited number of bits that are available for mantissa and exponent, not all real values (not even all rational values) can be represented exactly.

(a) (3 points) Assume a $(4, 7)$-format for floating-point numbers, and provide the bit patterns for the values 2.5, -42, and 12.34. If any of these numbers is not exactly representable in this format, round it to the nearest value that is.

Solution: In the following, the bit patterns are shown in the form $s.e.m$:

- $2.5 = 2^1 * 1.25$, so the solution is 0.1000.010 0000.
- $42 = (101010)_2$; this bit pattern without the leading 1 becomes the mantissa, and for the exponent $e = 5$ (since $2^5 = 32$), thus $E = 12$. Thus, we get 1.1100.010 1000.
- The least power of two smaller than 12.34 is $8 = 2^3$, thus $E = 10$. By successively comparing with smaller powers of 2, we get $12.34 = 8 + 4 + 0.25 + 0.0625 + r$, which gives 0.1010.100 0101, where $r = 0.0275$. Since r is smaller than the next power of 2, which is 0.03125, there is no need to round up.

(b) (2 points) In the $(4, 7)$-format, what is the smallest positive integer that cannot be represented exactly? What is it in the $(3, 7)$-format?

Solution: In the $(4, 7)$-format, $256 = 2^8 = (1\,0000\,0000)_2$ can still be represented by 0.1111.000 0000, i.e. $e = 15 - 7 = 8$ and $m = 1$. However, $257 = (1\,0000\,0001)_2$ cannot be exactly represented since the mantissa would have to accomodate 8 bits.

In the $(3, 7)$-format, the highest possible exponent is $e = 4$. Then, only four bits in the mantissa still give non-fractional values, and one obtains 0.111.111 1000, which yields 31. The value 32 cannot be represented exactly since it needs an exponent of 5.

(c) (2 points) Given some (k, ℓ)-format, let $N(k, \ell)$ denote the smallest positive integer that cannot be represented exactly in this format. For a fixed ℓ, give a formula for the minimal value of k such that for any $k' > k$, $N(k, \ell) = N(k', \ell)$.

Solution: We saw that $N(3, 7) = 32$ and $N(4, 7) = 257$. It is easy to see that $N(5, 7) = 257$, too, since the impossibility of representing 511 in $(4, 7)$-format is independent of the size of the exponent.

Now let ℓ be arbitrary. If $N(k, \ell)$ is independent of k, then $N(k, \ell)$ is the value whose binary representation is $10^\ell 1$, i.e. ℓ repetitions of zero between the two ones. For $N(k, \ell) - 1 = 2^{\ell+1}$ to be representable, k must be large enough to represent $\ell + 1$ in the exponent. Considering that we add $2^{k-1} - 1$ to the exponent to obtain E, we want the smallest k such that

$$(\ell + 1) + (2^{k-1} - 1) \leq 2^k - 1 \iff \ell + 1 \leq 2^{k-1} \iff k \geq 1 + \log_2(\ell + 1).$$

3 De Bruijn sequences

In this part of the exam, we will develop an efficient method to count the number of trailing zero bits in a given (unsigned) integer value x such that $x > 0$. Equivalently, we can compute the position of the least significant bit whose value is 1. Incidentally, one concrete application – when x has 64 bits – is to encode the positions of pieces on a chess board and iterate over these. Here however, we will simplify matters by assuming that x has only 8 bits.

An *index* in a bit string is identified from right to left starting at zero. E.g., for $x = (10110100)_2$, the bits of x at index 0 and 1 are 0, and the bit with index 2 is 1.

Given $x \in \mathbb{N}$ such that $0 < x < 2^8$, we will be interested in implementing a function $\ell : \{1, \ldots, 2^8 - 1\} \to \{0, \ldots, 7\}$ such that $\ell(x)$ is equal to smallest index that is set to 1 in the binary representation of x. In the example above, we have $\ell(x) = 2$.

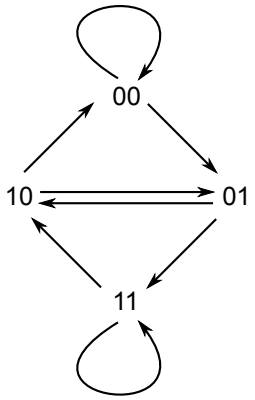
In principle, we could solve the problem using the C function below, which shifts x to the right until the least significant bit is 1.

```
unsigned int l (unsigned int x) {   // we assume 0 < x < 256
  int result = 0;
  while (x & 1 == 0) {
    result++;
    x = x >> 1;
  }
  return result;
}
```

However, the running time of this function depends on the number of bits in x. We will develop another algorithm has *constant* running time, i.e. independent of the actual number of zeros. To this end, we will study *de Bruijn sequences*. A de Bruijn sequence $s(n)$ of order n is a cyclic bit string such that every binary string of length n occurs exactly once in s. Cyclic means that once you reach the end of $s(n)$ you may continue at the beginning of $s(n)$. For example, for $n = 2$ we can set $s(n) = 0011$ since 00, 01, 10 and 11 can all be found in $s(n)$; in particular 10 starts at index 0 of $s(n)$ and then continues at index 3 of $s(n)$.

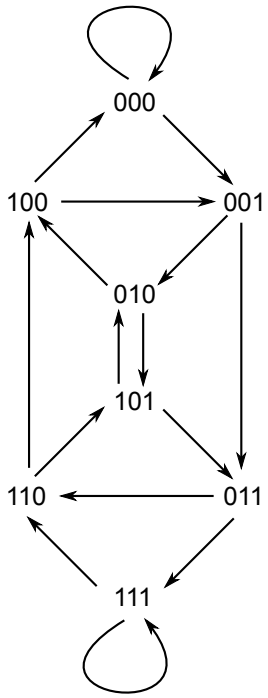
An obvious lower bound for the minimal length of a de Bruijn sequence $s(n)$ is 2^n. We will see that a sequence of this length can always be found, then use it to achieve our initial goal.

(a) (3 points) De Bruijn sequences can be obtained from paths in *de Bruijn graphs*. The vertices of a de Bruijn graph of order n are all bit strings of length n. There is a directed edge between two vertices $b_1 b_2 \cdots b_n$ and $c_1 c_2 \cdots c_n$ if and only if $b_2 = c_1, b_3 = c_2, \ldots, b_n = c_{n-1}$. The figure below depicts the de Bruijn graph of order 2.



Draw the de Bruijn graph of order 3.

Solution: The de Bruijn graph of order 3 is depicted below.



(b) (2 points) A de Bruijn sequence can be obtained from a de Bruijn graph by following a *Hamiltonian cycle* that starts and ends in the vertex $0 \cdots 0$. A Hamiltonian cycle is a cycle that visits each vertex exactly once before returning to the starting vertex. For instance, the only Hamiltonian cycle in the graph in the figure above is $00 \to 01 \to 11 \to 10 \to 00$. This cycle corresponds to the aforementioned de Bruijn sequence 0011. One can in fact prove that such a Hamiltonian cycle exists in every de Bruijn graph.

Read off two different de Bruijn sequences of order 3 by following two different Hamiltonian paths in your de Bruijn graph of order 3 starting in vertex 000.

Solution: The two possible de Bruijn sequences of order 3 are 00011101 and 00010111.

(c) (2 points) *Choose a de Bruijn sequence $s(3)$ of order 3 from (b) and complete the following table:*

Solution: In the following, we fix $s(3) = 00010111$.

bit-string	7 - index in $s(3)$
000	0
001	1
010	2
011	4
100	7
101	3
110	6
111	5

(d) (2 points) *Let $s(3)$ be the de Bruijn sequence from (c) and $0 \leq j < 8$. What is the value assigned by the table in (c) of the bit string*

$$((s(3) \ll j) \gg 5) \;\&\; 0\text{x}7$$

Here, $\ll$ and $\gg$ mean shift-left and shift-right, respectively, and & is binary AND.

Solution: The value of the expression is in fact j. This may seem useless at first, but recall that left-shifting by j positions is the same as multiplying by 2^j.

(e) (2 points) *Given an unsigned integer $k > 0$, what is the value of k & $(-k)$, where $-k$ is the two's complement of k?*

Solution: If $\ell(k) = j$, then the expression gives the binary representation of 2^j.

(f) (3 points) *Complete the following code skeleton such that it computes $\ell(x)$:*

```
const int index[8] = { 0, ... }; // the right-hand side of the table
                                 // in (c) here
const int s3 = 0b...;            // your de Bruijn sequence used in (c)

unsigned int l(unsigned int x) { // we assume 0 < x < 256
  return index[ ... ];           // complete code in the brackets
}
```

Solution: Combining the results of the previous two exercises, the expression to be placed inside the brackets is

```
((s3 * (x & -x)) >> 5) & 0x7
```

(g) (1 point) What other advantage does the algorithm above have with respect to our initial while loop, besides being constant runtime? (Think of the way assembly code is executed in a CPU, and how modern processors try to optimize that execution.)

Solution: In constrast to the while loop, there are no conditional branches in the assembly code for the function constructed above. Thus, the CPU has no need for branch prediction or speculative execution in this function and can execute more efficiently.