



Exercise 1: (Rounding Bug)

(3+4+1 Points)

In the lecture it was mentioned that some Intel Floating-Point Units suffered from a rounding bug which occurred when exact results were rounded twice with different precisions. We assume two significant precisions p and p' such that $p' > p$. Let r and r' be rounding functions of the same rounding mode for precisions (n, p) and (n, p') respectively. Considering a real number $x \in \mathbb{R}$, lying between the two neighboring representative numbers $(x_d, x_u) \ni x$ where $x_d, x_u \in \mathcal{R}$ with respect to precision p , we now want to find values for x such that rounding first with precision p' and then with p gives a different result than just rounding x with precision p alone.

$$r(x) \neq r(r'(x))$$

- For which rounding modes and in what situation can the rounding bug occur? Give a short explanation!
- Deduce the exact range for x , such that the rounding bug occurs! Construct one example value from x_d or x_u !
- What is the minimal difference $p' - p$ for the bug to occur?

Exercise 2: (Conditional Sum Incrementer)

(4+4+4 Points)

An n -bit incrementer is a circuit with inputs $a \in \mathbb{B}^n$, $inc \in \mathbb{B}$ and outputs $s \in \mathbb{B}^n$, $c \in \mathbb{B}$ such that the property

$$\langle c \ s[n-1:0] \rangle = \langle a[n-1:0] \rangle + inc$$

holds. A fast incrementer with logarithmic delay and linear-logarithmic cost can be implemented in the same way like a Conditional Sum Adder.

- Construct an n -bit Conditional Sum Incrementer for $n = 2^k$!
- Prove the correctness of your construction!
- Give the exact delay of your Incrementer in a non-recursive formula! Prove its correctness!

Hint: More information on the Conditional Sum Adder and a description how to find and prove delay functions can be found in the SS10 System Architecture Script (see course website) and the Computer Architecture [S. M. Müller, W. J. Paul, Springer 2000] book.

Exercise 3: (Half Decoder)

(5+5 Points)

An n -half decoder (n -HDEC) is a circuit with inputs $x[n-1:0]$ and outputs $Y[2^n-1:0] = 0^{2^n-(x)}1^{(x)}$, i.e. input x activates the $\langle x \rangle$ low order bits of the output.

- Construct an n -half decoder!
- Prove the correctness of your construction!