First Name: Annemarie
Last Name: Mayer
Date: December 14, 2003
Homework Number: 1
Homework Title: Exercise 1.15

Problem Description:

Give specific examples to show that floating-point addition is not associative in each of the following floating-point systems:

a) The toy floating-point system of Example 1.9

b) IEEE single-precision floating-point arithmetic

Problem Solution:

The solution is to find a representable number e for which the equation

$$(1 + e) + e = 1 + (e + e)$$

does not hold.
A good way to choose e is to set it to a (representable) number slightly smaller than the machine precision ϵ_{mach}, which is defined as

$\epsilon_{mach} = \beta^{1-p}$ using rounding by chopping,
$\epsilon_{mach} = \frac{1}{2}\beta^{1-p}$ using rounding to nearest.

In this case, $1 + (e + e) = 1 + 2e$ is representable (if $\beta = 2$) and greater than 1, whereas $1 + e$ is not representable and gets always rounded to 1.
Therefore we get

$$1 \neq 1 + 2e.$$

Solutions in detail:

a) In this example system ($\beta = 2, p = 3, L = -1, U = 1$), the Underflow Level $UFL = \beta^L = \frac{1}{2}$ is greater than $\epsilon_{mach} = \frac{1}{4}$ or $\frac{1}{8}$, depending on the rounding method used.
Therefore it is not possible to find a representable number e, slightly smaller than ϵ_{mach}, for the above equation.

I then modified the equation to

$$(1 + 1) + e = 1 + (1 + e)$$

and found that this does not hold for $e = -3/4$.

operation	**operator 1**	**operator 2**	**result** (normalized, rounded)
$2 - \frac{3}{4}$	(100) e 1	$-$(110) e -1	(101) e 0
$1 - \frac{3}{4}$	(100) e 0	$-$(110) e -1	(100) e -1
$1 + \frac{1}{2}$	(100) e 0	(100) e -1	(110) e 0

The operation $1 - \frac{3}{4}$ first results in (001) e $2 = \frac{1}{4}$, but this number is not representable in a normalized format (it would be (100) e -2, but then the exponent would be out of range).
So we have to round to (100) e $-1 = \frac{1}{2}$ (using rounding to nearest).

b) Using IEEE single-precision arithmetic, as $UFL < \epsilon_{mach}$, one can find a representable number e slightly smaller than ϵ_{mach} for the equation $(1 + e) + e \neq 1 + (e + e)$.
$\epsilon_{mach} = \frac{1}{2}\beta^{1-p} = \frac{1}{2}2^{-23} = 2^{-24}$ (using rounding to nearest), therefore I chose $e = 1.5 \cdot 2^{-25}$.

operation	**operator 1**	**operator 2**	**result**
$e + e$	(110...0) e -25	(110...0) e -25	(110...0) e -24
1 + (e+e)	(10...0) e 0	(0...011) e -2	(10...01) e 0
1 + e	(10...0) e 0	(0...011) e -3	(10...0) e 0

When adding 1 and $(e + e)$, the second operator has to be transformed to a representation with exponent 0. This is not possible - the only two significant bits would be lost. But as $(e + e)$ is greater than ϵ_{mach}, the result is rounded up to the next representable number after 1, (10...01) e 0, this is $1 + 2^{-23}$.
Accordingly, as $e < \epsilon_{mach}$, $1 + e$ is always rounded down to 1.

Results:

a)
$(1 + 1) - \frac{3}{4}$ results in (101) e 0 or $\frac{5}{4}$
$1 + (1 - \frac{3}{4})$ results in (110) e 0 or $\frac{3}{2}$

b)
$1 + (1.5 \cdot 2^{-25} + 1.5 \cdot 2^{-25})$ results in (10...01) e 0 or $1 + 2^{-23}$
$(1 + 1.5 \cdot 2^{-25}) + 1.5 \cdot 2^{-25}$ results in (10...0) e 0 or 1