Floating Point

1.a) # elements of $F = (1023+1+1022) \cdot 2^{51} \cdot 2$

$= 2046 \cdot 2^{52}$

b) # elements where $1 \leq x \leq 2$

$= 2^{51}$

$$\text{Fraction} = \frac{2^{51}}{2046 \cdot 2^{52}}$$

$$= \frac{1}{4092}$$

c) $\left(\frac{1}{4}\right)_{10} = (.1 \times 2^{-1})_2$

$\left(\frac{1}{2}\right)_{10} = (.1 \times 2^{0})_2$

$\therefore$ # elements where $\frac{1}{4} \leq x < \frac{1}{2}$

$= 2^{51}$

$$\text{Fraction} = \frac{2^{51}}{2046 \cdot 2^{52}}$$

$$= \frac{1}{4092}$$

d) refer to page 7.

2. $$\frac{|(fl(a) \oplus fl(b)) - (a+b)|}{|a+b|}$$

$$= \frac{|(a(1+\delta_1) \oplus b(1+\delta_2)) - (a+b)|}{|a+b|} \quad \text{where } |\delta_1|, |\delta_2| \leq \varepsilon$$

$$\approx \frac{|((a(1+\delta_1)+b(1+\delta_2))(1+\delta_3) - (a+b)|}{|a+b|} \quad \text{where } |\delta_3| \leq \varepsilon$$

$$\approx \frac{|(a+a\delta_1+b+b\delta_2)(1+\delta_3) - (a+b)|}{|a+b|}$$

$$\approx \frac{|(a+b)+a\delta_1+b\delta_2+(a+b)\delta_3+(a\delta_1+b\delta_2)\delta_3 - (a+b)|}{|a+b|}$$

$$\approx \frac{|a(\delta_1+\delta_3+\delta_1\delta_3)+b(\delta_2+\delta_3+\delta_2\delta_3)|}{|a+b|}$$

$$\leq \frac{|a||\delta_1+\delta_3+\delta_1\delta_3| + |b||\delta_2+\delta_3+\delta_2\delta_3|}{|a+b|}$$

Since $|\delta_i| \leq \varepsilon$ for $i = 1,2,3$

$$\lesssim \frac{|a|(2\varepsilon+\varepsilon^2)+|b|(2\varepsilon+\varepsilon^2)}{|a+b|} \lesssim \frac{|a|+|b|}{|a+b|}\varepsilon(2+\varepsilon)$$

3. a) $149,659,299,043,739.794$
$-\ 149,659,299,043,739.691$
0.103 mm

∴ Thickness is **0.103 mm**

b) From MATLAB: thickness = 0 mm

$$\text{Rel. Err.} = \frac{|0 - 0.103|}{|0.103|} = 1 = 100\%$$

c) From MATLAB: thickness = 0.0938 mm

$$\text{Rel. Err.} = \frac{|0.0938 - 0.103|}{|0.103|} \simeq 0.0873204$$

$$\simeq 8.9304\%$$

d) Upper Bound

$$\text{Rel. Err} \leq \frac{|149659299043739.794| + |-149659299043739.691|}{|149659299043739.794 - 149659299043739.691|} E(2+E)$$

$$\leq \frac{299318458087499.485}{0.103} E(2+E) \quad ; E = \tfrac{1}{2}B^{1-t} = 2^{-1}2^{1-52} = 2^{-52}$$

$$\leq \frac{299318458087499.485}{0.103} 2^{-52}(2+2^{-52})$$

$$\leq 1.29052522$$

$$\leq 129\%$$

4. refer to page 8

5. a) $x^2 + 100.01x + 1.2121 = 0$

Using calculator:

$r_1 = -0.012121257\underline{12}$ $\qquad$ $r_2 = -99.997878\underline{7429}$

b) $r_1 = \dfrac{-100.01 \oplus \sqrt{100.01^2 \ominus 4 \otimes 1 \otimes 1.2121}}{2 \otimes 1}$

$= \dfrac{-100.01 \oplus \sqrt{fl(10002.0001) \ominus fl(4.8484)}}{fl(2)}$

$= \dfrac{-100.01 \oplus \sqrt{\underline{10002} \ominus 4.8484}}{2}$

$= \dfrac{-100.01 \oplus \sqrt{fl(9997.1516)}}{2}$

$= \dfrac{-100.01 \oplus \sqrt{\underline{9997.2}}}{2}$

$= \dfrac{-100.01 \oplus fl(99.98599901099)}{2}$

Red underline represents loss of precision

$r_1 = \dfrac{-100.01 \oplus \underline{99.986}}{2}$

$= fl(-0.0240) \oslash 2$

$= -0.0240 \oslash 2$

$= fl(-0.0120)$

$r_1 = -0.0120$

Using the same discriminant

$r_2 = \dfrac{-100.01 \ominus \underline{99.986}}{2}$

$= fl(-199.996) \oslash 2$

$= -\underline{200.00} \oslash 2$

$= fl(-100.00)$

$r_2 = -100.00$

r_1 Rel. Err. $= \dfrac{|(-0.0120) - (-0.012121257 12)|}{0.012121257 12}$

$= 0.010003 67526 \approx 1\%$

r_2 Rel. Err. $= \dfrac{|(-100.00) - (-99.997 8787 429)|}{99.997 8787 429}$

$= 2.121302 \times 10^{-5} \approx 0.002\%$

c) $r_1 = \dfrac{-b + \sqrt{b^2-4ac}}{2a} \cdot \dfrac{-b - \sqrt{b^2-4ac}}{-b - \sqrt{b^2-4ac}}$

$= \dfrac{b^2 - (b^2 - 4ac)}{2a(-b - \sqrt{b^2-4ac})}$

$= \dfrac{4ac}{2a(-b - \sqrt{b^2-4ac})}$

$= \dfrac{2c}{-b - \sqrt{b^2-4ac}}$

$r_2 = \dfrac{-b - \sqrt{b^2-4ac}}{2a} \cdot \dfrac{-b + \sqrt{b^2-4ac}}{-b + \sqrt{b^2-4ac}}$

$= \dfrac{b^2 - (b^2 - 4ac)}{2a(-b + \sqrt{b^2-4ac})}$

$= \dfrac{4ac}{2a(-b + \sqrt{b^2-4ac})}$

$= \dfrac{2c}{-b + \sqrt{b^2-4ac}}$

d) $r_1 = \dfrac{2 \otimes 1.2121}{-100.01 \ominus \sqrt{100.01^2 - 4 \otimes 1 \otimes 1.2121}}$

using the same discriminant calculated in b)

$r_1 = \dfrac{fl(2.4242)}{-100.01 \ominus 99.986}$

$= \dfrac{2.4242}{fl(-199.996)}$

$r_2 = \dfrac{fl(2.4242)}{-100.01 \oplus 99.986}$

$= \dfrac{2.4242}{fl(-0.0240)}$

$r_1 = 2.4242 \oplus -200.00$

$= fl(-0.012121)$

$= -0.012121$

$r_2 = 2.4242 \oplus -0.0240$

$= fl(-101.008333333)$

$= -101.01$

r_1 Rel. Err. $= \dfrac{|(-0.012121) - (-0.0121212125712)|}{0.0121212125712}$

$= 2.121282 \times 10^{-5} \approx 0.002\%$

r_2 Rel. Err. $= \dfrac{|(-101.01) - (-99.9978787429)|}{99.9978787429}$

$= 0.010121427227 \approx 1\%$