Språkbruk.

Ett binært siffer kalles et bit.

En samling av 8 bits kalles en byte.

Datamaskiner og reelle tall.

Eks:

$$\pi \approx 3.141592654 \ldots\ldots$$

$$10^6\pi \approx 3.141592654 \ldots \cdot 10^6$$

Samme siffer, ulik størrelse.

Må lagre sifferne + størrelse.

Observasjon 4.3 (normal form).
La $a \in \mathbb{R}$ ulik 0. Da kan a entydig skrives som

$$\boxed{a = b \cdot 10^n},$$

der b er begrenset ved $\boxed{\frac{1}{10} \le |b| < 1}$

og n er et heltall.
b kalles $\boxed{\text{signifikanden}}$ til a og
n $\boxed{\text{eksponenten.}}$
Normalformen til 0 er $\boxed{0 = 0 \cdot 10^0}$

Eks: $\pi \approx 0.3142 \cdot 10^1$

$$\frac{1}{7} \approx 0.1429 \cdot 10^0$$

$$\frac{100003}{17} \approx 0.5883 \cdot 10^4$$

Binær normal form.

Anta $a \in \mathbb{R}$, $a \neq 0$. Da kan a skrives entydig som

$$a = b \cdot 2^n$$

der $\frac{1}{2} \leq |b| < 1$

$$\boxed{0 = 0 \cdot 2^0}$$

b - signifikanden

n - eksponenten

For å representere reelle tall på datamaskin lagrer vi signifikanden og eksponenten i den binære normalformen til tallet

i): 32 bit totalt:

Signifikand : 23 bit

Eksponent : 9 bit

Minste pos. tall: $1.4 \cdot 10^{-45}$

Største pos tall: $3.4 \cdot 10^{38}$

Regner omtrent 6-8 desimale siffer.

64 bit:

Signifikand: 53 bit

Eksponent: 11 bit

Minste pos. tall: $5 \cdot 10^{-324}$

Største pos. tall: $1.8 \cdot 10^{308}$

Omtrent 15-17 desimale siffer.

Noen begreper:

Et tall sies å være trunkert til m siffere hvis alle siffer bortsett fra de m lengst til venstre settes til 0.

Eks: $0.33\cancel{333}$ trunkert til 2 siffere: 0.33

Et tall er rundet av til m siffere hvis det er erstattet med det nærmeste tallet med m siffere.

Eks: 0.6666

Avrundet til 2 siffere: 0.67

Forenklet modell for flyttall (reelle tall (desimale)) på datamaskin)

Vi bruker 4 desimale siffer til signifikand og 1 desimale siffer til eksponent. Fortegn kommer i tillegg.

Addisjon av flyttall, a og b

Algoritme

1. Tallet med størst tallverdi skrives på normal form, for eks.

$$\boxed{a = \alpha \cdot 10^n} \qquad a+b=(\alpha+\beta)\cdot 10^n$$

og b skrives $\boxed{b = \beta \cdot 10^n}$ med like siffer i β som i α.

2. Adder signifikandene,

$$\gamma = \alpha + \beta$$

3. Svaret $c = \gamma \cdot 10^n$ konverteres til normal form.

Eks. $a = 5.645$, $b = 7.821$

1. b på normal form:

$b = 0.7821 \cdot 10^1$

$a = 0.5645 \cdot 10^1$

2. Adder signifikandene:

$0.7821 + 0.5645 = 1.3466$

Svar $1.3466 \cdot 10^1 = 13.466$

3. Konverter til normal form:

$1.3466 \cdot 10^1$ blir da

$0.1347 \cdot 10^2$

Eks 2: $a = 42.34$, $b = 0.0033$

1 a på normal form:

$$a = 0.4234 \cdot 10^2$$

b på samme form.

$$b = 0.000033 \cdot 10^2$$

2. Adder signifikandene

$$0.4234 + 0.000033 = 0.423433$$

Svaret er $0.423433 \cdot 10^2$

3. Konverter til normal form.

Endelig resultat $0.4234 \cdot 10^2 = 42.34$

Eks 5.11 $a = 10.34, \quad b = -10.27$

$a = 0.1034 \cdot 10^2, \quad b = -0.1027 \cdot 10^2$

$0.1034 - 0.1027 = 0.0007.$

$a + b = 0.0007 \cdot 10^2 = 0.7000 \cdot 10^{-1}$

Eks:

Anta $a = \frac{10}{7}$, $b = -1.42$

$\frac{10}{7} \approx a = 0.1429 \cdot 10^1$, $b = -0.1420 \cdot 10^1$

Adder: $0.1429 - 0.1420 = 0.0009$

Svar $0.0009 \cdot 10^1$

Konverter til normal form:

$0.9000 \cdot 10^{-2}$

Riktig svar: $0.8571 \cdot 10^{-2}$

Eks. 5.22.

Skal regne ut $\frac{1}{\sqrt{x^2+1} - x}$, $x = 10^8$

$x = 10^8$

$\sqrt{x^2+1} = 100000000.000000005 \ldots$

Det betyr at selv med 64 bits flyttall så blir $\sqrt{x^2+1} - x$ regnet ut som 0! Dermed gir $\frac{1}{\sqrt{x^2+1} - x}$ divisjon med null og Infinity

Omskriving kan redde oss her

$$\frac{1}{(\sqrt{x^2+1} - x)} \cdot \frac{\sqrt{x^2+1} + x}{(\sqrt{x^2+1} + x)} = \frac{\sqrt{x^2+1} + x}{x^2+1 - x^2}$$

$$= \sqrt{x^2+1} + x$$

$x = 10^8$ gir her riktig svar 200000000

Absolutt og relativ feil.

Hvis $\tilde{a}$ er en tilnærming til a så er den absolutte feilen gitt ved

$$|a - \tilde{a}|$$

Relativ feil er gitt ved

$$\frac{|a - \tilde{a}|}{|a|} \qquad a \neq 0.$$

Eks: $a = 1, \quad \tilde{a} = 2$

$a = 1000, \quad \tilde{a} = 1001$

Hvis den relative feilen er $\approx 10^{-m}$ vil a og $\tilde{a}$ ha omtrent m felles siffre.