

1 Representation av tal i dator

1.1 Heltal

Exempel:

8 bitar, talet 53

$(00110101)_2$

$$2^5 + 2^4 + 2^2 + 2^0 = 32 + 16 + 4 + 1 = 53$$

Heltal vanligtvis 32 bitar

1.2 Reella tal

Liknar så kallad vetenskaplig notation.

$$43520 = 4.352 \times 10^4$$

$$0.0000642 = 6.42 \times 10^{-5}$$

$$x = m\beta^e$$

– m mantissa

– β bas

– e exponent

1.3 Flyttalssystem

$$(\beta, p, L, U)$$

– β bas

– p precision

– $[L, U]$ exponentgränser

$$x = \left(d_0 + \frac{d_1}{\beta} + \frac{d_2}{\beta^2} + \dots + \frac{d_{p-1}}{\beta^{p-1}} \right) \beta^e$$

$$0 \leq d_i < \beta, \quad i = 0, \dots, p-1$$

$$L \leq e \leq U$$

$$\text{Exempel: } 43520 = \left(4 + \frac{3}{10} + \frac{5}{10^2} + \frac{2}{10^3} \right) 10^4 = \left(1 + 0 + \frac{1}{2^2} + 0 + \frac{1}{2^4} + 0 + \frac{1}{2^6} \right) 2^{15}$$

- Exponenten lagras exakt inom övre och undre gräns.
- Mantissan måste rundas av, p är precisionen.

1.4 Normalisering

Flyttalssystem normaliseras om $d_0 \neq 0$.

$$1 \leq m < \beta$$

- Unik representation av varje tal.
- Inget slöseri med siffror på inledande nollor \rightarrow maximal noggrannhet.
- I binärt system ($\beta = 2$) är första siffran alltid 1 och behöver därför inte lagras. ("hidden bit normalization")

1.5 Litet flyttalssystem

$$(\beta, p, L, U) = (2, 3, 0, 2)$$

Möjliga värden:

$e \backslash m$	(1.00)	(1.01)	(1.10)	(1.11)
0	1	1.25	1.5	1.75
1	2	2.5	3	3.5
2	4	5	6	7

Flyttalen ej jämnt representerade – ju större tal ju glesare representation.

1.5.1 Några tester

2.3+4.4 = 6.7:

$$fl(fl(2.3) + fl(4.4)) = fl(2.5 + 4.0) = fl(6.5) = 6.0$$

Absolut fel:

$$|6.7 - 6.0| = 0.7$$

Relativt fel:

$$\frac{|6.7 - 6.0|}{|6.7|} \approx 0.104$$

(2.3+4.4)-1.2 = 5.5:

$$fl(fl(fl(2.3) + fl(4.4)) - fl(1.2)) = fl(6.0 - 1.25) = fl(4.75) = 5.0$$

2.3+(4.4-1.2) = 5.5:

$$fl(fl(2.3) + fl(fl(4.4) - fl(1.2))) = 6.0$$

Samma beräkning i annan ordning kan ge annat svar.

4.3+6.2 = 10.5:

$$fl(fl(4.3) + fl(6.2)) = fl(4 + 6) = \text{Inf} \quad \text{pga overflow}$$

1.6 Subnormala tal

Släpper på normaliseringskravet för tal mindre än minsta möjliga nollskilda normaliserade tal: $|x| < \beta^L$. Subnormala tal i det lilla flyttalssystemet:

$$(0.01) = 0.25, \quad (0.10) = 0.5, \quad (0.11) = 0.75$$

- Fortfarande unik representation av varje tal.
- Inte full precision p för de subnormala talen, men bättre än att alltid avrunda till 0 eller β^L .
- I Matlab: `realmin` minsta möjliga *normaliserade* positiva tal. Finns mindre subnormala tal.