P2: Prva nedelja
- Predstavljanje realnih brojeva -

Pregled

- Veza između celih i realnih brojeva
- Konverzija realnog broja
- Predstavljanje realnih brojeva
- Opseg realnih brojeva
- Standardi za rad sa realnim brojevima
- Zaokruživanje realnih brojeva
- Primeri

Veza između celih i realnih brojeva

- **Kako predstaviti realan broj?**
  - kao odnos dva cela broja X i Y (Z := X / Y)
  - kao kombinaciju dva cela broja, gde prvi predstavlja celobrojni deo, a drugi razlomljeni (X.Y)
- **Nepraktično!**
- **Težinski sistem:**
  - $0.5_{(10)} = 5_{(10)} \cdot 10^{-1}$
  - $0.1_{(2)} = 1_{(2)} \cdot 2^{-1}$
  - $0.X_{(Y)} = X_{(Y)} \cdot Y^{-1}$
- **Prednosti**
  - jednostavna i uniformna reprezentacija
  - pristup primenjen kod celih brojeva se može i ovde primeniti
- **Nedostatak**
  - neophodno više polja: značajne cifre, stepen osnove

Konverzija realnog broja

- **Primer: $11.0011_{(2)} = 11_{(2)} + 0.0011_{(2)} = 3_{(10)} + X_{(10)}$**

$$R_{(10)} = \sum_{i=0}^{n-1} C_i \cdot B^{-(i+1)} = C_0 \cdot B^{-1} + C_1 \cdot B^{-2} + ... + C_{n-1} \cdot B^{-n}$$

$$R_{(10)} = (...(C_{n-1} / B + C_{n-2}) / B + ... + C_1) / B + C_0) / B$$

- **$0.0011_{(2)} = X_{(10)}$?**

  $\mathbf{1}/2 = 0.5$

  $(0.5+\mathbf{1})/2=0.75$

  $(0.75+\mathbf{0})/2=0.375$

  $(0.375+\mathbf{0})/2=0.1875 = X_{(10)}$

$0.1875_{(10)} = X_{(2)}$?

$0.1875 \cdot 2 = 0.375 + \boxed{0}$

$0.3750 \cdot 2 = 0.750 + \boxed{0}$

$0.750 \cdot 2 = 0.5 + \boxed{1}$

$0.5 \cdot 2 = 0.0 + \boxed{1}$

$X_{(2)} = 0.0011_{(2)}$

- **Pitanje: $0.1111111..._{(2)} = X_{(10)}$?**
  Koja je vrednost X?

Predstavljanje realnih brojeva

- **$R=(-1)^S M\ 2^E$ S-znak, M-mantisa, E-eksponent**
- **Primer**

  $0.0011 = 0.0011 \cdot 2^0 = 0.11 \cdot 2^{-2} = 1.1 \cdot 2^{-3}$
- **Normalizovana mantisa**

  $0.5 \le M < 1$ ð 0.1xxxx...xx VAX

  $1 \le M < 2$ ð 1.xxxx...xx IEEE 754

  Primetiti: dovoljno je pamtiti bite obeležene sa x. Ostali se podrazumevaju.

  ð Normalizovana mantisa sa skrivenim bitom
- **Eksponent u kodu sa viškom (nije komplement 2)**

  e = E + pomeraj ð E = e – pomeraj

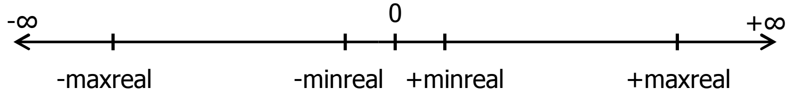
  pomeraj = 2^{k-1} VAX

  pomeraj = $2^{k-1}-1$ IEEE 754

  k – broj bitova za predstavljanje (kodiranje) eksponenta

Opseg realnih brojeva

- **Realna osa**



- **Opseg mogućih vrednosti simetričan u odnosu na nulu (pozitivne i negativne vrednosti)**
- **Ne može da predstavi brojeve veće od maxreal, ni manje od minreal a veće od 0**
- **0 se predstavlja na poseban način**

Standard ANSI/IEEE 754 (1985) za rad sa realnim brojevima

- **Broj značajnih decimalnih cifara**

  $d \approx 0.3 \cdot b$

  (b - broj bitova za predstavljanje celokupne mantise broja, b = p + 1)
- **Izgled realnog broja (ukupan broj bita: w=1+k+p)**

mašinska reč

s	eee..e	mmmmmmmm..m
1	k	p

- **$R=(-1)^S M\ 2^E$**
  - Broj je pozitivan ako bit S ima vrednost 0, a negativan ako ima vrednost 1
  - k (broj bitova za eksponent) – utiče samo na **opseg** vrednosti
  - p (broj bitova za mantisu) – utiče samo na **tačnost** brojeva

Standard ANSI/IEEE 754 (1985) za rad sa realnim brojevima

Pravila interpretacije

1. e=111...1, m≠0 ð **NaN** (**n**ot **a n**umber)
2. e=111...1, m=0 ð $(-1)^s (\infty)$ ±∞
3. 000...0<e<111...1 ð $(-1)^s\ 2^{e-v}(1.m)$ *normalizovana mantisa*
4. e=000...0, m≠0 ð $(-1)^s 2^{1-v}(0.m)$ *nenormalizovana mantisa*
5. e=0, m=0 ð $(-1)^s (0)$ ±0

- **Značajni standardni formati realnih brojeva**
  - jednostruki (32 bita): k=8, p=23 : v=+127, E_{max}=+127, E_{min}=-126
  - dvostruki (64 bita): k=11, p=52 : v=+1023, E_{max}=+1023, E_{min}=-1022

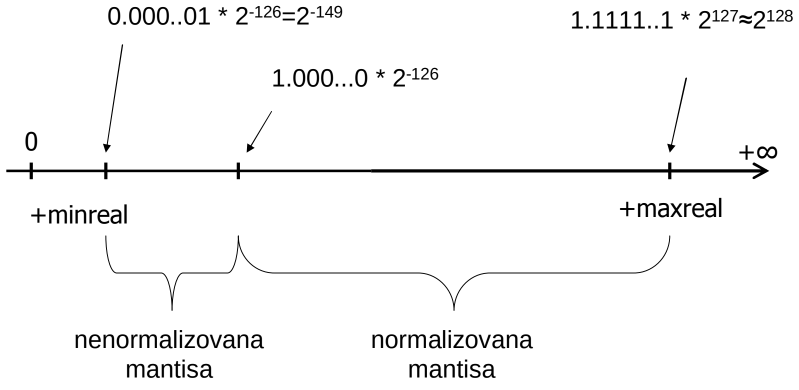
Standard ANSI/IEEE 754 (1985) Realni broj jednostruke preciznosti

- **Predstava na dužini od w=32 bita:**
  - 1 bit za znak
  - 8 bita za eksponent
  - 23 bita za mantisu
- **Opseg vrednosti**
  - Najveći normalizovani (e_{max}-127=127) $(-1)^s\, 2^{127}\, (1.111...11)$
  - Najmanji normalizovani (e_{min}-127=-126) $(-1)^s\, 2^{-126}\, (1.000...00)$
  - Najveći nenormalizovani $(-1)^s\, 2^{-126}\, (0.111...11)$
  - Najmanji nenormalizovani $(-1)^s\, 2^{-126}\, (0.000...01)$

$$\text{minreal} = (-1)^s \cdot 2^{-126} \cdot 2^{-23} = (-1)^s \cdot 2^{-149} \approx (-1)^s \cdot 1.4 \cdot 10^{-45}$$

$$\text{maxreal} = (-1)^s \cdot 2^{127} \cdot (1.111...11) \approx (-1)^s \cdot 2^{128} = (-1)^s \cdot 3.4 \cdot 10^{38}$$

Šta zapravo predstavlja nenormalizovana mantisa?



Pomoću nenormalizovane mantise se postiže proširenje opsega za izuzetno male vrednosti (bliske 0) realnih brojeva.

Zaokruživanje realnih brojeva

- **Primer (4 značajne decimlane cifre)**

R=123.4 r=0.049 ð R + r = 123.449 ≈ 123.4

R+r+r+r =?

((R+r)+r)+r ≈ (123.4+0.049)+0.049 ≈ 123.4+0.049 ≈ <u>123.4</u>

R+(r+(r+r)) = 123.4 + 0.147 = 123.547 ≈ <u>123.5</u>

ð asocijativnost? **<u>ne važi</u>**!

ð treba sumirati od brojeva manje vrednosti ka brojevima veće vrednosti!

Zaokruživanje realnih brojeva

- **Pravila zaokruživanja**
  1. x.xx|0zzzz...z ð x.xx
  2. x.xx|1yyyy...y (∃y≠0) ð x.xx + 0.01
  3. x.xx|1000....0
     - 3a. x.x0|1000....0 ð x.x0
     - 3b. x.x1|1000....0 ð x.x1 + 0.01
- **Zaokruživanje na beskonačnost**

$|R| \geq 2^{Emax} \cdot (2-2^{-p})$ ð ±∞

Zadatak 1 (IZ4)

- **Format predstave realnih brojeva pomoću reči širine 10 bita je: bit najveće težine za znak broja, četiri bita za eksponent u kôdu sa viškom 7, i pet bitova najmanje težine za normalizovanu mantisu sa skrivenim bitom (1£M<2). Rezultat sabiranja brojeva čiji je izgled $A=1606_{(8)}$ i $B=0515_{(8)}$ je:**

  a) $1633_{(8)}$

  b) $1566_{(8)}$

  c) $1766_{(8)}$

Zadatak 1

- **Rešenje:**
  Najpre je potrebno izvršiti "raspakivanje" brojeva A i B u binarni oblik:

A	B
A: 1|1100|00110	B: 0|1010|01101
$S_A=1 \Rightarrow A < 0$	$S_B=0 \Rightarrow B > 0$
$e_A = 1100 = 12$	$e_B = 1010 = 10$
$E_A = e - v = 12 - 7 = 5$	$E_B = e - v = 10 - 7 = 3$
$M_A=1.00110$	$M_B=1.01101$

$A = (-1)^{SA} \cdot 2^{EA} \cdot M_A = - 2^5 \cdot 1.00110$

$B = (-1)^{SB} \cdot 2^{EB} \cdot M_B = 2^3 \cdot 1.01101$

Zadatak 1

- **U opštem slučaju, potrebno je broj sa manjim eksponentom dovesti na red veličine broja sa većim eksponentom:**

$E_A > E_B \Rightarrow B = M_B \cdot 2^{E_B} = M_B \cdot 2^{(E_B - E_A)} \cdot 2^{E_A} = M_B \cdot 2^{-2} \cdot 2^{E_A} = M_{B'} \cdot 2^{E_A}$

$M_{B'} = M_B \cdot 2^{-2} \Rightarrow M_{B'} = 0.01011|01$ (zaokruživanje – pravilo 1)

$|M_A| > |M_{B'}| \Rightarrow A + B = -(M_A - M_{B'}) \cdot 2^{E_A} = -M_{A+B} \cdot 2^{E_A}$

$M_A - M_{B'} = 1.00110 - 0.01011 \Rightarrow M_{A+B} = 0.11011 = 1.10110 \cdot 2^{-1}$

Nakon sabiranja, mantisu M_{A+B} je potrebno normalizovati svođenjem na oblik dat postavkom zadatka (u ovom slučaju: 1.mmmmm), uz eventualno zaokruživanje nakon normalizacije.

Zadatak 1

- **Rezultat sabiranja A i B iznosi:**

$A+B = -M_{A+B} \cdot 2^{E_A} = -1.10110 \cdot 2^{-1} \cdot 2^{E_A} = -1.10110 \cdot 2^{E_A - 1} = -M_{A+B}' \cdot 2^{E_{A+B}'}$

$S_{A+B} = 1 \quad M_{A+B}' = 1.10110 \quad E_{A+B}' = E^A - 1 = 4$

$e_{A+B} = 4 + 7 = 11$

Izgled broja se dobija sklapanjem delova s, e i m u skladu sa formatom određenim postavkom zadatka.

s = 1 eeee = 1011 mmmmm = 10110

(A + B) : 1|1011|10110 = 1|101|110|110 = $1566_{(8)}$

- **Odgovor: B**

Zadatak 2 (IZ5)

- **U nekom računaru, celi brojevi brojevi su predstavljeni u drugom komplementu pomoću reči širine 10 bita, a za predstavljanje realnih brojeva je takođe predviđeno 10 bita, i to tako da bit najveće težine određuje znak broja, sledeća 4 bita su za eksponent broja u kôdu sa viškom 7, a preostalih 5 za normalizovanu mantisu sa skrivenim bitom (1£M<2). Ako je izgled realnog broja na lokaciji X: $397_{(16)}$, a izgled celog broja na lokaciji J: $310_{(16)}$, koji će biti izgled lokacije realne promenljive Y nakon izvršene operacije Y=X+J? Računanje obavljati upotrebljavajući realne brojeve.**

  a) $3FA_{(16)}$

  b) $3E5_{(16)}$

  c) $3BA_{(16)}$

Zadatak 2

- **Rešenje:**

  Realni broj X raspakujemo u binarni oblik u skladu sa zadatim formatom:

  X: 1|1100|10111

  $S_X = 1,\ X < 0,\ M_X = 1.10111,\ E_X = e - v = 12 - 7 = 5$

  Celom broju J moramo odrediti apsolutnu vrednost da bismo mogli da ga konvertujemo u realan broj:

  J: 1100010000, J<0

  $-J = 0011110000 = 1.11100|00 \cdot 2^7 = 1.11100 \cdot 2^7$

  $S_J = 1,\ J<0$

  $M_J = 1.11100$

  $E_J = 7$

Zadatak 2

- **Dalje, izvršimo svođenje eksponenta manjeg broja na eksponent većeg i izvršimo sabiranje:**

  $E_J > E_X$

  $X = -M_X \cdot 2^{-2} \cdot 2^7 = -M_X' \cdot 2^7$

  $M_X' = 0.01101|11 \approx 0.01110$ (zaokruživanje 0.01101+0.00001)

  $Y = -(M_J + M_X') \cdot 2^{E_J}$

  $M_J + M_X' = 1.11100 + 0.01110 = 10.01010 = 1.00101 \cdot 2^1 = M_Y$

  $Y = -1.00101 \cdot 2^1 \cdot 2^{E_J} = -1.00101 \cdot 2^8$

  $e = 8 + 7 = 15_{(10)} = 1111_{(2)}$

  s = 1, eeee = 1111, mmmmm = 00101

  Y: 1|1111|00101 = $3E5_{(16)}$

- **Odgovor: B**

Zadatak 3 (IZ3)

- **Format predstave realnih brojeva je seeeemmmmm – s je bit za znak broja, eeee su bitovi za eksponent broja (kôd sa viškom), a mmmmm su bitovi normalizovane mantise sa skrivenim bitom. U računaru A – predstava realnih brojeva je sa viškom 8, a u računaru B – sa viškom 7. U računaru A – mantisa je 0.5£M_A<1, a u računaru B – mantisa je 1£M_B<2. Izgled jednog broja u računaru B, u skladu sa opisanim formatom, je $3DF_{(16)}$. Kako je predstavljen realan broj iste vrednosti u računaru A?**

  a) $3DF_{(16)}$

  b) $3FF_{(16)}$

  c) $3FE_{(16)}$

Zadatak 3

- **Rešenje:**

B: $3DF_{(16)} = 1|1110|11111_{(2)}$

$E_B = eeee_B - 7 = 7$

$M_B = 1.mmmmm_B = 1.11111_{(2)}$

ð $X_B = -1.11111_{(2)} \cdot 2^7 = -11111100_{(2)}$

$X_A = X_B = -0.111111_{(2)} \cdot 2^8$

$M_A = 0.1m_A \qquad m_A = 11111_{(2)}$

$e_A = E_A + 8 = 16_{(10)} = \underline{10000}_{(2)}$

ne može se predstaviti na mašini A!

- **Odgovor: N**

Zadatak 4

- **Realni brojevi se smeštaju u 10-bitnu lokaciju prema sledećem formatu: seeeemmmmm, gde je s bit za predznak broja, e bitovi za predstavljanje eksponenta u kodu sa viškom 7, a m bitovi za predstavljanje normalizovane mantise sa skrivenim bitom (1 £ M<2). Sva zaokruživanja se obavljaju prema pravilima ANSI/IEEE standarda za realne brojeve. Razlika brojeva Z=Y-X za vrednosti X=13.5, Y=-51.0 biće:**

a) $-64.00000_{(10)}$

b) $-64.50000_{(10)}$

c) $-64.03125_{(10)}$

Zadatak 4

- **Rešenje:**

$X= 13.5 = 13 + 0.5 = 1101_{(2)} + 0.1_{(2)}$

$X = 1101.1_{(2)} = 1.1011_{(2)} \cdot 2^3 \quad (E_X = 3; M_X = 1.10110)$

$Y = -51.0$

$Y = -110011.0_{(2)} = -1.10011_{(2)} \cdot 2^5 \quad (E_Y = 5; M_Y = 1.10011)$

$M_X = 1.10110 = 0.01101|10 \cdot 2^2 \approx 0.01110 \cdot 2^2 = M_x' \cdot 2^2$

$Y < 0, X > 0 \Rightarrow (Y - X) < 0 \Rightarrow Z < 0$

$M_y + M_x' = 1.10011 + 0.01110 \Rightarrow M_{Y-X} = 10.00001$

$Z = - M_{Y-X} \cdot 2^5$

$Z = -10.00001 \cdot 2^5 = -1.00000|1 \cdot 2^6 \approx -1.0 \cdot 2^6 = -64$

- **Odgovor: A**

Zadatak 5

- **Na nekom računaru za predstavljanje normalizovane mantise sa skrivenim bitom (0.5 £ M<1) koristi se 8 bita, jedan bit je rezervisan za predstavljanje znaka, a K bitova za predstavljanje eksponenta u kodu sa viškom 2^{K-1}. Koja je najmanja vrednost K za koju važi da se broj 2^{60} može predstaviti kao regularan realni broj, dok broj 2^{70} izlazi iz opsega realnih brojeva?**

a) 6

b) 7

c) 9

Zadatak 5

- **Rešenje:**

  0.5£M<1 ð 0.1xxx...xxx ð $2^{60} = 0.1 \cdot 2^{61}$, $2^{70} = 0.1 \cdot 2^{71}$

  K=6 ð 2^{K-1}= 32 ð -32...32 (tj: -31...31) $E_{min}=1-2^{k-1}$, $E_{max}=2^{k-1}$ -1

  K=7 ð 2^{K-1}= 64 ð -64...64 (tj: -63...63)

  K=9 ð 2^{K-1}= 256 ð -256...256 (tj: -255...255)

- **Odgovor: B**

Zadatak 6

- **Na računarima A i B realni brojevi predstavljaju se sa 16 bita oblika seeeeeeemmmmmmmm gde je eeeeeee eksponent u kodu sa viškom a mmmmmmmm kod mantise. Mantisa je normalizovana i sa skrivenim bitom, pri čemu je na računaru A mantisa u opsegu [1,2), a na računaru B u opsegu [0.5,1). Eksponent se predstavlja u kodu sa viškom čija je vrednost 2^{7-1}-1 na računaru A, odnosno 2^{7-1} na računaru B. Neka je X binarna predstava realnog broja 2.72 (uz eventualno zaokruživanje) na računaru A. Kojem realnom broju odgovara ista predstava X na računaru B?**

  a) 0.6806640625

  b) 2.71875

  c) 0.6796875

Zadatak 6

- **Rešenje:**

  0.72 x 2 = 0.44 + 1
  0.44 x 2 = 0.88 + 0
  0.88 x 2 = 0.76 + 1
  0.76 x 2 = 0.52 + 1
  0.52 x 2 = 0.04 + 1
  0.04 x 2 = 0.08 + 0
  0.08 x 2 = 0.16 + 0

  2.72 = 2.0 + 0.72

  2.0 = $10_{(2)}$

  $2.72_{(10)}$= 10.1011 10000... $_{(2)}$ ≈ 1.0101 1100 · 2^1

  Na računaru A:
  s = 0; eeeeeee = 1 + 63 = 1000000; mmmmmmmm = 0101 1100
  X: 0100000001011100

  Na računaru B:
  s = 0; E = eeeeeee - 64 = 0; M = 0.1010111000
  broj = 0.1010111000
  broj = 0.5 + 0.125 + 0.03125 + 0.015625 + 0.00781250 = 0.6796875

- **Odgovor: C**

Zadatak 7

- **U nekom računaru celi brojevi se predstavljaju u 32-bitnim lokacijama u drugom (potpunom) komplementu, dok se realni brojevi predstavljaju u duhu IEEE standarda, sa normalizovanom mantisom sa skrivenim bitom (1≤M<2) koja zauzima m bita, i eksponentom u kodu sa viškom 2^{k-1}-1, koji zauzima k bita. Koje bi vrednosti za m i k bile potrebne i dovoljne da bi se svi celi brojevi na ovom računaru mogli tačno predstaviti kao realni?**

  a) k=6, m=30

  b) k=8, m=32

  c) k=5, m=31

Zadatak 7

- **Rešenje:**

  celi broj ð 32 bita ð 1 bit znaka, 31 bit za vrednost
  ð realni broj ð mantisa 31 bit (ali zbog skrivenog bita dovoljno je 30)

  k = 6 ð 2^{k-1}-1 = 31 ð opseg eskponenta -30...31

- **Odgovor: A**

Zadatak 8 (IZ34A)

- **Realni brojevi se predstavljaju sa 11 bita formatom seeeeemmmmm gde je s bit za predznak broja, eeeee (5) biti eksponenta u kodu sa viškom 8, a mmmmm (5) biti normalizovane mantise sa skrivenim bitom 0.5<=M<1. Celi brojevi predstavljaju se sa 10 bita u drugom komplementu. Ceo broj I ima izgled 177_{16} a ceo broj J ima izgled 325_{16}. Najpre se izvrši celobrojno sabiranje N=I+J. Zatim se izvrši konverzija brojeva I i J u realne brojeve, i sabiranjem tih realnih brojeva dobije broj B. Kolika će biti apsolutna vrednost razlike brojeva N i B? (Napomena: Sva zaokruživanja vrše se prema pravilima IEEE standarda, a svi koraci u sabiranju realnih brojeva vrše se na širini određenoj formatom mantise.)**
  a) 4
  b) 8
  c) 2.125

Zadatak 8

- **Rešenje:**
  Prvo ćemo izvršiti celobrojno sabiranje.
  I = 177_{16} = 01 0111 0111_2
  J = 325_{16} = 11 0010 0101_2

  N = I+J = 100 1001 1100_2

  Pošto u memoriji imamo samo 10 bita sa predstavljanje celih brojeva, jedanaesti bit koji je generisan u procesu sabiranja otpada i dobijamo:

  N = I+J = 0010011100_2 = 156_{10}

Zadatak 8

- **Realno sabiranje se vrši tako što prvo celobrojne vrednosti konvertujemo u realne, uz potrebna zaokruživanja, a potom dobijene realne vrednosti saberemo:**

  I: 0101110111 = $0.101110111 \cdot 2^9 \approx 0.101111 \cdot 2^9$

  J: 11 0010 0101, što znači da je J<0
  -J: 00 1101 1011 = $0.11011011 \cdot 2^8 \approx 0.110111 \cdot 2^8$

Zadatak 8

- **Sada je potrebno svesti ova dva broja na isti eksponent i pri tome izvršiti zaokruživanja, ako se za tim ukaže potreba:**

  $J = (-1) \cdot 0.110111 \cdot 2^8 = (-1) \cdot 0.0110111 \cdot 2^9 \approx (-1) \cdot 0.011100 \cdot 2^9$

  Pošto su brojevi različitog znaka, oduzimamo mantisu manjeg od mantise većeg broja:

  M_I: 0.101111
  $-M_J$: - 0.011100

  $M_I - M_J$: 0.010011
  $B = I+J = 0.010011 \cdot 2^9 = 0.100110 \cdot 2^8$

Zadatak 8

- **Traženi broj ima vrednost: $0.100110 \cdot 2^8$, što je jednako pozitivnom celom broju:**

  $B = 0010011000_2 = 152_{10}$

  Odavde se može zaključiti da je apsolutna razlika ova dva zbira jednaka $|N - B| = 4$.

- **Odgovor: A**
  Napomena: Kada se vrši konverzija iz celobrojnog u realni broj, vrši se zaokruživanje na po vrednosti najbliži realni broj koji se može predstaviti u memoriji datog računara. Konverzija u obrnutom smeru podrazumeva ili zaokruživanje početnog realnog broja na po vrednosti najbliži celi broj ili prosto odsecanje razlomljenog dela realnog broja. Ako nije posebno naglašeno, podrazumeva se prvi pristup.