Realni brojevi u pokretnom zarezu. IEEE 754 standard.

Jovana Kovačević

www.uoar1.matf.bg.ac.rs

Uvod u organizaciju i arhitekturu računara 1

Pregled

Pregled

Konverzija zapisa razlomljenih brojeva

- Ceo deo i razlomljeni deo broja se odvojeno prebacuju u traženu osnovu
- Ceo deo se prebacuje prema poznatim pravilima
- Razlomljeni deo se prebacuje na sledeći način:
  - pomnožimo razlomljeni deo novom osnovom
  - celobrojni deo dobijenog broja predstavlja cifru novog zapisa, a razlomljeni deo se ponovo množi novom osnovom
  - postupak ponavljamo sve dok je razlomljeni deo različit od nule ili do traženog broja decimala

Primer

Zapisati broj $(0.84375)_{10}$ u osnovi 4.

0	1	2	3
0.84375	0.375	0.5	0
0	3	1	2

$(0.84375)_{10} = (0.312)_4$

Primer

Zapisati broj $(23.71)_{10}$ u osnovi 16.
$(23)_{10} = (17)_{16}$

0	1	2	3	4	5	6
0.71	0.36	0.76	0.16	0.56	0.96	0.36
0	B	5	C	2	8	F

$(23.71)_{10} = (17.B5C28F\ldots)_{16}$

Zapis u fiksnom zarezu

- broj cifara za zapis celog dela i za zapis razlomljenog dela je fiksiran
- primeri:

Broj	Format zapisa			
	7.4	5.3	6.1	8.0
	---.----	--.---	-----.-	--------.
$(1.3543)_{10}$	␣␣1.3543	␣1.354	␣␣␣␣1.3	␣␣␣␣␣␣␣1.
$(12.7)_{10}$	␣12.7000	12.700	␣␣␣12.7	␣␣␣␣␣␣12.
$(1347)_{10}$	********	******	␣1347.0	␣␣␣␣1347.
$(123.456)_8$	123.4560	******	␣␣123.4	␣␣␣␣␣123.
$(AB.1)_{16}$	␣AB.1000	AB.100	␣␣␣AB.1	␣␣␣␣␣␣AB.
$(1011.1101)_2$	********	******	␣1011.1	␣␣␣␣1011.
$(0.1101)_2$	␣␣0.1101	␣0.110	␣␣␣␣0.1	␣␣␣␣␣␣␣0.

Mane zapisa u fiksnom zarezu

- na primer, potrebno je zapisati 1000 brojeva u fiksnom zarezu i za zapis na raspolaganju imamo 3 dekadne cifre
- ako izdvojimo jednu cifru za ceo deo i dve za razlomljeni, moći ćemo da zapišemo brojeve $[0.00, \ldots, 9.99]$
- ako izdvojimo dve cifre za ceo deo i jednu za razlomljeni, moći ćemo da zapišemo brojeve $[00.0, \ldots, 99.9]$
- šta ako je potrebno zapisati brojeve različitih redova veličine? (npr. 0.0001 i 100000)

Zapis u pokretnom zarezu

Broj se predstavlja kao $\pm m \cdot b^e$

- m - značajni deo, mantisa, frakcija
- b - osnova zapisa
- e - eksponent
- broj cifara značajnog dela zovemo preciznost (nadalje u oznaci p)

Primeri

Broj	Neki mogući zapisi		
	Zapis 1	Zapis 2	Normalizovan zapis
$(13.543)_{10}$	(13.543, 0)	(0.13543,+2)	(1.3543,+1)
$(12.7)_{10}$	(127000.0,-4)	(0.00127,+4)	(1.27,1)
$(5347)_{10}$	(53470.,-1)	(0.005347,+6)	(5.347,+3)
$(123.22)_4$	(12322.000,-2)	(0.012322,+10)	(1.2322,+2)
$(AB.1)_{16}$	(AB10.,-2)	(0.000AB1,+5)	(A.B1,+1)
$(1011.1101)_2$	(10111101,-100)	(10.111101,+10)	(1.0111101,+11)
$(0.1101)_2$	(110.10,-11)	(1101.0,-100)	(1.101,-1)

Pregled

Greške

- Zapis realnih brojeva u računaru je aproksimacija skupa realnih brojeva u određenom intervalu.
- Pri aproksimiranju na određeni broj decimala može doći do zaokruživanja, a samim tim i greške.
- Greška pri zaokruživanju se meri na dva načina:
  - apsolutna (u terminima ulp-a)
  - relativna (u terminima mašinskog ϵ)

Apsolutna greška

- Apsolutna greška se izražava u terminima ulp-a
- ULP (Unit in the last place) je najmanja vrednost za koju se mogu razlikovati dva broja u pokretnom zarezu zapisana u istoj osnovi B i sa istom preciznošću p.
- ako je broj z u računaru predstavljen kao $d_0.d_1 \ldots d_{p-1} \cdot b^e$, onda je apsolutna greška

$$\left| d_0.d_1 \ldots d_{p-1} - \frac{z}{b^e} \right| \cdot b^{p-1}$$

ulp-a

Primer

- Neka je $b = 10, p = 4$.
- Broj $z = 0.034869$ je predstavljen kao $3.487 \cdot 10^{-2}$ u osnovi B sa preciznošću p.
- Apsolutna greška je:
  $\left|3.487 - \frac{0.034869}{10^{-2}}\right| \cdot 10^3 =$
  $|3.487 - 3.4869| \cdot 10^3 =$
  $0.0001 \cdot 10^3 = 0.1$ ulp-a

Relativna greška

- Relativna greška je apsolutna vrednost razlike realnog broja i njegove reprezentacije podeljena sa apsolutnom vrednošću realnog broja
- Uvek se zapisuje u terminima mašinskog $\epsilon = \frac{b^{1-p}}{2}$.

Primer

- Neka je $b = 10, p = 4, \epsilon = 0.0005$.
- Broj $z = 0.034869$ je predstavljen kao $3.487 \cdot 10^{-2}$ u osnovi B sa preciznošću p.
- Relativna greška je:
$\frac{|0.034869 - 0.03487|}{|0.034869|} = 0.000028678 \approx 0.0574\epsilon$

Pregled

Razlozi za standardizaciju

- različiti dogovori (osnova, raspodela bitova između delova zapisa...)
- različiti algoritmi zaokruživanja, izvođenja aritmetičkih operacija, itd.
- slaba prenosivost numeričkih programa
- odgovor: IEEE 754 (1985; 2008)

IEEE 754 standard

Standard IEEE 754 propisuje zapise:

- 32 bita – jednostruka tačnost
  - 1 bit za znak
  - 8 bita za eksponent u zapisu sa uvećanjem 127
  - 23 značajne binarne cifre
- 64 bita – dvostruka tačnost
  - 1 bit za znak
  - 11 bita za eksponent u zapisu sa uvećanjem 1023
  - 52 značajne binarne cifre
- pri tome:
  - normalizovani zapisi, podrazumeva se najviša cifra 1
  - nula se zapisuje sa svim bitovima 0
  - najviši eksponent označava posebne NaN vrednosti

Zapis sa uvećanjem

- Zapis sa uvećanjem K u osnovi B sa n cifara (u oznaci $\langle x \rangle_{B,n}^{K}$) podrazumeva da se broj x predstavi odgovarajućim neoznačenim zapisom broja $x + K$
- $\langle x \rangle_{B,n}^{K}$ predstavlja zapis $\langle x + K \rangle_{B,n}^{no}$
- $I_{B,n}^{K} = [-K, B^n - 1 - K]$
- na ovaj način se negativni brojevi veći od $-K$ predstavljaju kao pozitivni, na primer broj $-K$ se zapisuje kao $0 \ldots 0$, broj $-K + 1$ kao $0 \ldots 01$ i slično

Zapis sa uvećanjem

- specijalno, za $B = 2, n = 8, K = 127$, možemo zapisati brojeve iz intervala $[-127, 128]$
- na primer, -127 se zapisuje kao $0 \ldots 0$, -126 kao $0 \ldots 01$, a 128 se zapisuje kao $1 \ldots 1$
- ovakav način zapisivanja omogućava da se vrednosti eksponenta leksikografski porede
  - zapisi brojeva -127 i 1 u potpunom komplementu su 10000001 i 00000001; $-127 < 1$, ali kada bismo leksikografski poredili njihov zapis, zaključili bismo suprotno
  - zapisi brojeva -127 i 1 sa uvećanjem 127 su 00000000 i 10000000; $-127 < 1$ i kada bismo leksikografski poredili njihov zapis, zaključili bismo isto

Normalizovani brojevi

- Normalizovani brojevi u jednostrukoj tačnosti imaju eksponent između $-126(00000001)$ i $+127(11111110)$
- Frakcija ima oblik $1.d_{-1} \ldots d_{-(p-1)}$, pri čemu se prva jedinica ne zapisuje
- Zapis frakcije može da ima bilo koju vrednost.

Primer

- Zapisati broj 13,25 prema standardu IEEE 754 u jednostrukoj tačnosti
- $(13.25)_{10} = (1101.01)_2$
- Normalizovani oblik: $(1.10101)_2 \cdot 2^3$
- Znak: 0 (+)
- Eksponent sa uvećanjem 127: $130 = (10000010)_2$
- Značajni deo sa implicitnom jedinicom: 10101
- Zapis:

$$0 \ \underbrace{10000010}_{8} \ \underbrace{10101000000000000000000}_{23}$$

Primer

- Pročitati sledeći zapis:

$$1\ 10000110\ 01001000000000000000000$$

- Znak: -
- Eksponent zapisan sa uvećanjem: $(10000110)_2 = (134)_{10}$, bez uvećanja $134 - 127 = 7$
- Značajni deo uključujući implicitnu jedinicu: $(1.01001)_2$
- Rešenje: $(-1.01001)_2 \cdot 2^7 = (-10100100)_2 = (-164)_{10}$

NaN

- NaN (Not a number) nisu brojevi i označavaju neke izuzetne situacije prilikom izračunavanja(npr. 0/0 ili $\sqrt{-1}$)
- Eksponent NaN vrednosti je maksimalan; u slučaju jednostruke tačnosti to je 128 (11111111)
- Frakcija mora biti različita od nule.
- Postoje takozvani Signalni NaN (SNaN) i Tihi NaN (QNaN).

QNaN

- Tihi NaN predstavlja pojavu nedozvoljene operacije u programu.
- Propagira se kroz izračunavanje.
- Greška ne mora biti prijavljena.
- Zapis eksponenta u jednostrukoj tačnosti je 11111111.
- Prvi bit frakcije je 1.
- Ostali bitovi frakcije su proizvoljni.
- primeri zapisa:

1 11111111 10011000000000000000000

0 11111111 11000011000000110001110

0 11111111 10000000000000000000000

SNaN

- Signalni NaN signalizira izuzetno stanje kod racunskih operacija.
- Može se koristiti za debagovanje npr. kako bi se uočio rad sa neinicijalizovanim promenljivim.
- Zapis eksponenta u jednostrukoj tačnosti je 11111111.
- Prvi bit frakcije je 0.
- Ostatak frakcije je različit od 0.
- primeri zapisa:

1 11111111 01100000000000000000000

0 11111111 00000000000110010000010

0 11111111 01111111111111111111111

Beskonačno

- IEEE 754 standard omogućava predstavljanje beskonačnih vrednosti.
- Znak određuje da li se radi o $+\infty$ ili $-\infty$
- Eksponent u jednostrukoj tačnosti ima vrednost 128(11111111).
- Frakcija je 0.

$$+\infty : 0\ 11111111\ 00000000000000000000000$$

$$-\infty : 1\ 11111111\ 00000000000000000000000$$

Označena nula

- Nula se u jednostrukoj tačnosti predstavlja eksponentom $-127(00000000)$ i frakcijom 0.
- Pošto znak može biti + ili – (0 ili 1) onda postoje i dve nule +0 i -0.
- Prema standardu važi $+0 = -0$.
- $\log(-0) = NaN$, a $\log(+0) = -\infty$

$$+0 : 0\ 00000000\ 00000000000000000000000$$

$$-0 : 1\ 00000000\ 00000000000000000000000$$

Denormalizovani brojevi

- Da bi se povećala gustina realnih brojeva oko nule i izbegla pojava potkoračenja uvode se takozvani denormalizovani brojevi.
- U jednostrukoj tačnosti važi:
  - Eksponent je 00000000 i podrazumeva se da je njegova dekadna vrednost -126
  - Frakcija je različita od nule i umesto vodeće jedinice podrazumeva se vodeća nula
  - Na primer, ako je frakcija f, onda je predstavljeni broj $0.f \cdot 2^{-126}$
- primer: sledeći zapis

$$0\ 00000000\ 00010000000000000000000$$

ima vrednost $0.0001 \cdot 2^{-126} = 2^{-4} \cdot 2^{-126} = 2^{-130}$

Interval za normalizovane brojeve

- eksponent e: $2^{-126} \leq e \leq 2^{127}$
- frakcija f: $1 \leq f \leq 0.\underbrace{1\ldots1}_{24} \cdot 2$
  - najmanja je kada $\underbrace{0\ldots0}_{23}$ što odgovara vrednosti 1.0
  - najveća je kada $\underbrace{1\ldots1}_{23}$ što odgovara vrednosti $1.\underbrace{1\ldots1}_{23} =$

    $= 0.\underbrace{1\ldots1}_{24} \cdot 2 = (1 - 2^{-24}) \cdot 2$
- interval u osnovi 2:

$$2^{-126} \leq e \cdot f \leq (1 - 2^{-24}) \cdot 2^{128}$$

- interval u osnovi 10 (približno):

$$1.2 \cdot 10^{-38} \leq e \cdot f \leq 3.4 \cdot 10^{38}$$

Interval za denormalizovane brojeve

- eksponent e: 2^{-126}
- frakcija f: $2^{-23} \leq f \leq 1 - 2^{-23}$
  - najmanja je kada $\underbrace{0 \ldots 0}_{22}1$ što odgovara vrednosti 2^{-23}
  - najveća je kada $\underbrace{1 \ldots 1}_{23}$ što odgovara vrednosti $0.\underbrace{1 \ldots 1}_{23} = 1 - 2^{-23}$
- interval u osnovi 2:

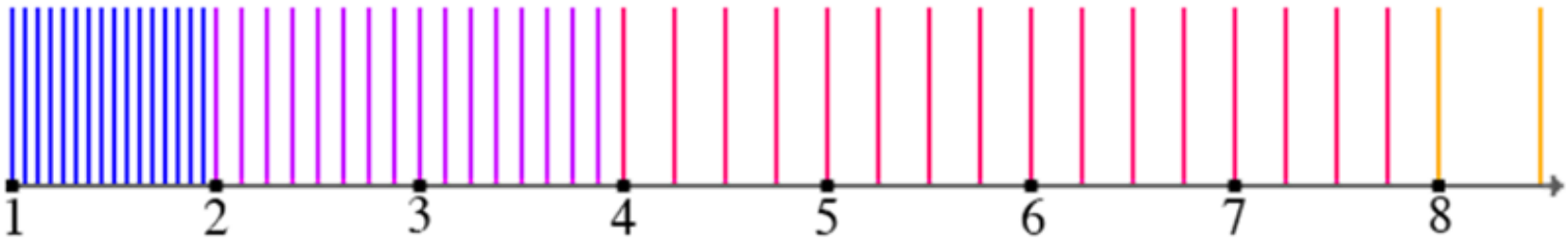
$$2^{-126-23} \leq e \cdot f \leq (1 - 2^{-23}) \cdot 2^{-126}$$

Primeri

		Znak	Eksponent	Frakcija
$+15$	$=$	0	10000010	11100000000000000000000
-15	$=$	1	10000010	11100000000000000000000
$+1/64$	$=$	0	01111001	00000000000000000000000
$+0$	$=$	0	00000000	00000000000000000000000
-0	$=$	1	00000000	00000000000000000000000
$(1-2^{-24}) \times 2^{+128}$	$=$	0	11111110	11111111111111111111111
$+1 \times 2^{-126}$	$=$	0	00000001	00000000000000000000000
$+1 \times 2^{-149}$	$=$	0	00000000	00000000000000000000001

Gustina

- jednak broj ekvidistantnih vrednosti između svaka dva stepena dvojke



- na primer, u jednostrukoj tačnosti je moguće predstaviti 2^{23} brojeva između dva stepena dvojke ($\approx$ 8 miliona)

Gustina

- problem sa predstavljanjem celobrojnih vrednosti jer gustina između dva stepena dvojke sa povećanjem stepena postaje manja od broja celih brojeva između njih
- počev od 2^{24}, u jednostrukoj tačnosti se ne mogu predstaviti svi celi brojevi i zaokružuju se na umnožak dvojke
- ako je tip int 32-bitni, tada u okviru ovog tipa možemo predstaviti cele brojeve u intervalu $[-2^{31}, 2^{31} - 1]$; ako je tip float 32-bitni a double 64-bitni, to znači da može doći do gubitka podataka prilikom konverzija int $\rightarrow$ float $\rightarrow$ int, dok prilikom konverzija int $\rightarrow$ double $\rightarrow$ int neće biti gubitka

Aritmetičke operacije

- Zaokruživanje
- Sabiranje
- Oduzimanje

Zaokruživanje

- Zaokruživanje se vrši kada rezultat operacije ne može biti tačno zapisan.
- Moguće su sledeće vrste zaokruživanja:
  - Zaokruživanje na najbližu vrednost
  - Zaokruživanje prema $+\infty$
  - Zaokruživanje prema $-\infty$
  - Zaokruživanje prema nuli.

Zaokruživanje na najbližu vrednost

- Pri ovoj vrsti zaokruživanja broj se zaokružuje na najbližu predstavljivu vrednost, uz zaokruživanje na parnu cifru kada je broj na sredini intervala između dve predstavljive vrednosti
- Ovo je predefinisani način zaokruživanja.

Zaokruživanje prema $+\infty$

Realizuje se u dva koraka:

- Ako je broj pozitivan i postoji bar jedna jedinica na nekoj poziciji desno od poslednje pozicije koja se čuva u zapisu, na poslednju poziciju se dodaje jedinica.
- Bez obzira na znak odbacuju se bitovi desno od poslednje pozicije koja se čuva u zapisu.

Zaokruživanje prema $-\infty$

Realizuje se u dva koraka:

- Ako je broj negativan i postoji bar jedna jedinica na nekoj poziciji desno od poslednje pozicije koja se čuva u zapisu, na poslednju poziciju se dodaje jedinica.
- Bez obzira na znak odbacuju se bitovi desno od poslednje pozicije koja se čuva u zapisu.

Zaokruživanje prema nuli

- Odbacuju se svi bitovi desno od poslednje pozicije koja se čuva u zapisu.

∞ u aritmetičkim operacijama

Sa izuzetkom operacija koje proizvode QNaN sve operacije koje uključuju ∞ takodje imaju ∞ kao rezultat. Međutim i dalje treba paziti na znak.

QNaN u aritmetičkim operacijama

- QNaN se propagira kroz aritmetičke operacije
- Može se pojaviti u sledećim slučajevima:
  - $(\pm\infty) - (\pm\infty)$
  - $0 \cdot \infty$
  - $0/0, \infty/\infty$
  - $x\%0, \infty\%x$
  - $\sqrt{x}, x < 0$
  - Bilo koja operacija čiji je argument SNaN.

Sabiranje i oduzimanje

- Prilikom sabiranja i oduzimanja operandi se svode na jednake eksponente.
- Manji eksponent se povećava, a cifre frakcije koja mu odgovara se pomeraju udesno za onoliko mesta za koliko je povećan eksponent.
- Ako pri pomeranju frakcija postane 0 rezultat je vrednost drugog operanda.

Sabiranje i oduzimanje

- Sabiranje i oduzimanje frakcija se vrše prema pravilima koja važe za cele brojeve u zapisu znak i apsolutna vrednost
- Eksponent rezultata je eksponent operanada posle izjednačavanja
- Ako dolazi do prekoračenja rezultat se pomera za jedno mesto udesno uz povećanje vrednosti eksponenta za jedan
- Ako povećanje vrednosti eksponenta dovede do prekoračenja rezultat je ∞, ali uzevši u obzir i znak.

Sabiranje i oduzimanje

- Ako rezultat operacije nije normalizovan, pokušava se normalizacija
- Može se dobiti i denormalizovan rezultat.
- Na kraju se vrši zaokruživanje, ako je potrebno.

Primer sabiranja I

Sabrati u pokretnom zarezu u jednostrukoj tačnosti brojeve 5.375 i 3.75

- zapis brojeva:

$$0\ 10000001\ 01011000000000000000000\ (5.375)$$

$$0\ 10000000\ 11100000000000000000000\ (3.75)$$

- eksponenti se izjednači sa većim (129) i time se frakcija broja 3.75 menja: $1.111 \cdot 2^{128} = 0.1111 \cdot 2^{129}$
- sabiranje frakcija posle izjednačavanja eksponenta:

$$1.01011 \cdot 2^{129} + 0.11110 \cdot 2^{129} = 10.01001 \cdot 2^{129}$$

Primer sabiranja II

- normalizacija rezultata:

$$10.01001 \cdot 2^{129} = 1.001001 \cdot 2^{130}$$

- zapis rezultata:

$$0\ 10000010\ 00100100000000000000000 = 9.125$$

Primer oduzimanja I

Oduzeti u pokretnom zarezu u jednostrukoj tačnosti brojeve 5.375 i 2.5

- zapis brojeva:

$$0\ 10000001\ 01011000000000000000000\ (5.375)$$

$$0\ 10000000\ 01000000000000000000000\ (2.5)$$

- eksponenti se izjednači sa većim (129) i time se frakcija broja 2.5 menja: $1.01 \cdot 2^{128} = 0.101 \cdot 2^{129}$
- oduzimanje frakcija posle izjednačavanja eksponenta:

$$1.01011 \cdot 2^{129} + 0.10100 \cdot 2^{129} = 0.10111 \cdot 2^{129}$$

Primer oduzimanja II

- normalizacija rezultata:

$$0.10111 \cdot 2^{129} = 1.0111 \cdot 2^{128}$$

- zapis rezultata:

$$0\ 10000000\ 01110000000000000000000 = 2.875$$

Slajdovi su napravljeni na osnovu materijala prof. M. Nikolića, prof. N. Mitića i A. Zečević za predmet Uoar1