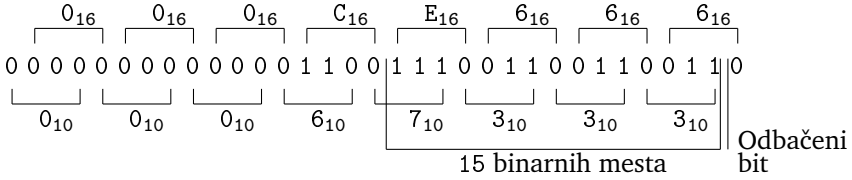
Realni brojevi i aritmetika

Realni brojevi u nepokretnom zarezu

Moguće greške:

1. Nekorektno smeštanje tačke osnove.

   Na primer, neka je pri deklaraciji navedeno da se odvaja 15 binarnih mesta za razlomljeni deo i neka je vrednost koja se zapisuje $+12.9_{10}$. Heksadekadni ekvivalent ovog broja je $+C.E666$. Kako reprezentacija zahteva 15 bitova za razlomljeni deo poslednji bit prevoda će biti odbačen, tako da se umesto očekivanog zapisa $CE666_{16}$ dobija 67333_{16}.



2. Nekorektno skaliranje pri aritmetičkim operacijama.

6	4	4	C	Originalne
0	3	5	C	vrednosti

6	4	4	C	Vrednosti poravnate
3	5	0	C	za sabiranje
9	9	4	C	Zbir

Realni brojevi u pokretnom zarezu

Predstavljaju se pomoću osnove β (koja je uvek parna) i preciznosti p.

Primer:

- β=10, p=4: broj 0.4 se predstavlja kao 4.000×10^{-1}
- β=10, p=4: broj broj 564000000000000000000000000 se predstavlja kao 5.640×10^{26}
- β=10, p=4: broj broj 564000055555555555555555555 se predstavlja kao 5.640×10^{26}
- β=2 i p=10 broj 0.4 se predstavlja kao 1.100110011×2^{-2}.

Opšti slučaj

$$\pm d_0.d_{-1}d_{-2}...d_{-(p-1)}\beta^e$$

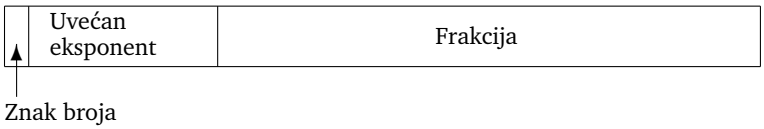
Oznake:

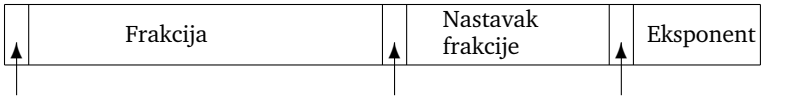
- $d_0.d_{-1}d_{-2}...d_{-(p-1)}$ – *značajni deo* (eng. *significand*). Zapisuje se u brojčanom sistemu sa osnovom β, tj. $0 \leq d_i < \beta$.
- β – *osnova*
- e – *eksponent*
- p – *preciznost*.

Zapis broja za koji važi da je $d_0 \neq 0$ se naziva **normalizovan**.

U savremenim računarima β=2, β=10 ili β=16.

Zapis brojeva u pokretnom zarezu kroz istoriju

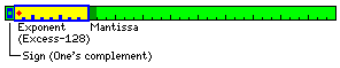




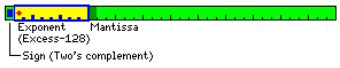
Group I

International Business Machines 704, 709, 7040, 7044, 7090, 7094, 7094 II

Univac 1107, 1108



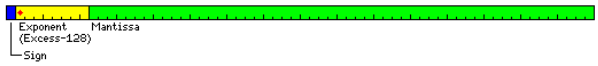
Digital Equipment Corporation PDP-6, PDP-10, DECSYSTEM-20



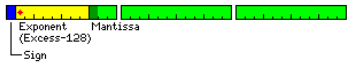
Expanded range (KL-10 only)

KA-10 Double Precision

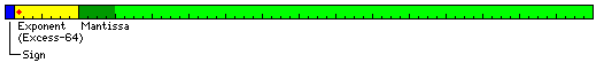
Digital Equipment Corporation PDP-11, VAX-11



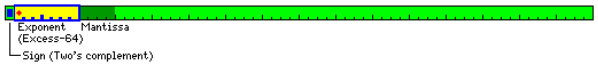
Digital Equipment Corporation PDP-8 Special (8K FORTRAN)



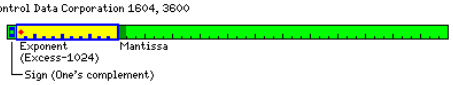
International Business Machines System/360, System/370, ESA/390, z/Architecture



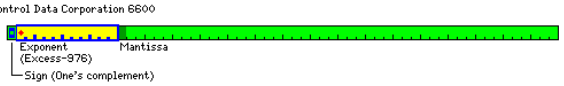
Xerox Data Systems Sigma



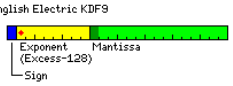
Control Data Corporation 1604, 3600

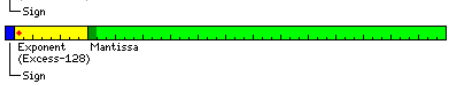


Control Data Corporation 6600

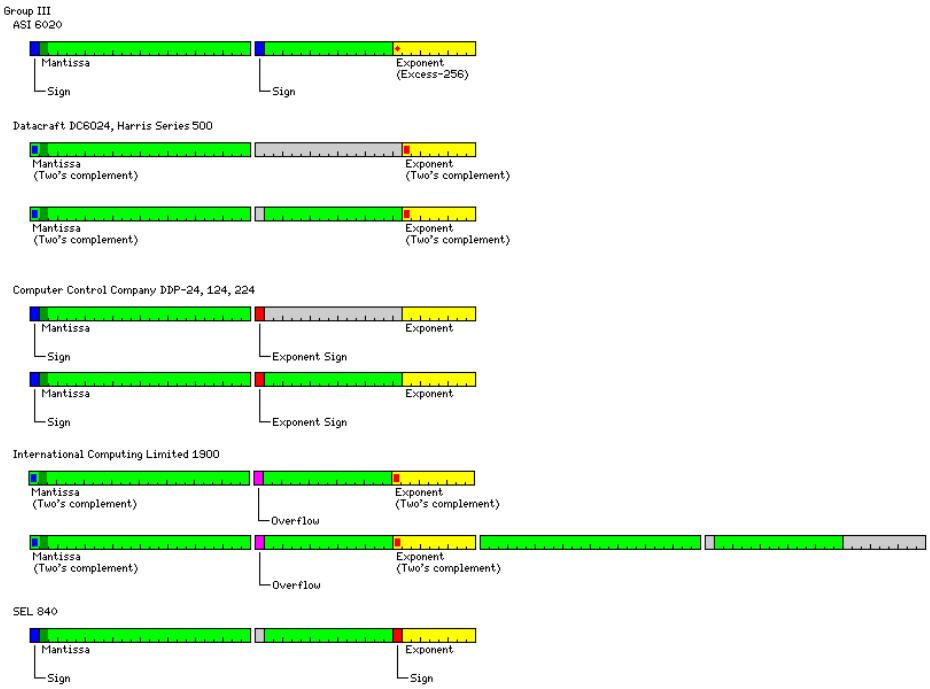


English Electric KDF9





Slika 2: Formati zapisa relanih brojeva u pokretnom zarezu kroz istoriju (nastavak)



Slika 3: Formati zapisa relanih brojeva u pokretnom zarezu kroz istoriju (nastavak)

Prethodni formati su detaljnije opisani na http://www.quadibloc.com/comp/cp0201.htm

Primer zapisa sa binarnom osnovom - PDP-11, VAX-11

Slika 4: Format zapisa realnog broja pomoću binarne osnove

Važi

- Vrednost eksponenta se povećava za 1 dok frakcija postaje

$$0.d_0 d_{-1} d_{-2} ... d_{-(p-1)}$$

- Frakcija $d_0 d_{-1} d_{-2} ... d_{-(p-1)}$ se predstavlja preko 24 bita, sa 23 binarne pozicije na mestima 0-22.
- Eksponent se zapisuje u 8 bita na pozicijama 23–30 uz uvećanje od 128.

		Znak	Eksponent	Frakcija
$+15$	=	0	10000100	11100000000000000000000
-15	=	1	10000100	11100000000000000000000
$+1/64$	=	0	01111011	00000000000000000000000
0	=	0	00000000	00000000000000000000000
$(1-2^{-24})\times 2^{+127}$	=	0	11111111	11111111111111111111111
$+1\times 2^{-128}$	=	0	00000001	00000000000000000000000

Provera vrednosti zapisa broja +15:
– Znak = +
– Eksponent = 4 (=132-128)
– Frakcija = $(0.1111)_2 = +1\times2^{-1}+1\times2^{-2}+1\times2^{-3}+1\times2^{-4}$
– Vrednost = Znak frakcija * $2^{eksponent} = (+1\times2^{-1}+1\times2^{-2}+1\times2^{-3}+1\times2^{-4})$ * 2^4
$= +1\times2^3+1\times2^2+1\times2^1+1\times2^0 = 8 + 4+ 2 +1 = +15$

Tabela 1: Zapis realnih brojeva u jednostrukoj tačnosti / binarna osnova

Za veličinu eksponenta e važi

$$-2^7 \leq e \leq 2^7 - 1$$

Za vrednost s kojom se predstavlja frakcija važi

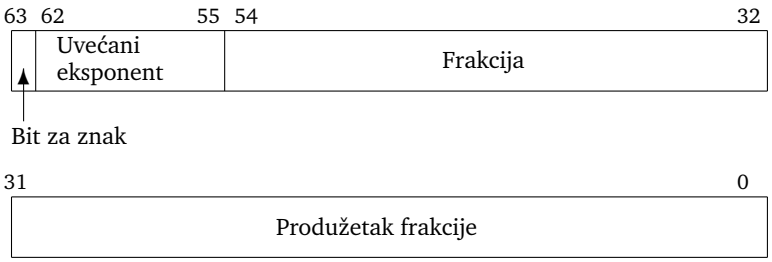
$$2^{-1} \leq |s| \leq 1 - 2^{-24}$$

Na osnovu prethodnog, mogu se zapisati brojevi u intervalu

$$2^{-1} * 2^{-128} \leq |x| \leq (1 - 2^{-24}) * 2^{+127}$$

Medjutim, kako važi da je kod za broj $2^{-1} * 2^{-128}$ predstavlja 0, opseg je

$$\begin{aligned} 2^{-128} &\leq |x| \leq (1 - 2^{-24}) * 2^{+127}, \quad \text{odnosno} \\ 2.9 * 10^{-39} &< |x| < 1.7 * 10^{+38} \end{aligned}$$



Slika 5: Format zapisa realnog broja u dvostrukoj tačnosti pomoću binarne osnove

Interval realnih brojeva koji mogu da se predstave u dvostrukoj tačnosti je

$$2^{-128} \leq |x| \leq (1-2^{-56}) * 2^{+127}$$

Zapis brojeva u pokretnom zarezu pomoću binarne osnove - IEEE754-2008)

Karakteristike izračunavanja sa realnim brojevima pre pojave standarda IEEE 754:

1. Slaba prenosivost numerički intenzivnih programa.
2. Postoje razlike u rezultatima u izvršavanju istog programa na različitim računarskim sistemima.
3. IEEE 754 poboljšava prenosivost programa propisujući
    - algoritme za operacije sabiranja, oduzimanja, množenja, deljenja i izračunavanje kvadratnog korena
    - način njihove implementacije
    - načine zaokruživanja i ponašanja u graničnim slučajevima

Neki problemi koji se javljaju pri izračunavanjima sa realnim brojevima su:

1. **Zaokruživanje.**

   β=10 i p=3. Neka je rezultat izračunavanja 5.76×10^{-2}, matematički tačna vrednost računata sa beskonačnom preciznošću 0.0574. Greška zapisa je veličine 2×jedinična vrednost na poslednjem mestu zapisa broja.

   $0.0574367 \longrightarrow 5.74\times10^{-2}$ – greška je 0.367 jedinica na poslednjem mestu.

   Veličina "jedinica na poslednjem mestu" se označava sa *ulp* (eng. *unit in the last place*.

   U opštem slučaju, ako broj $d_0.d_{-1}...d_{-(p-1)} \times \beta^e$ predstavlja broj z, tada je greška u zapisu $\mid d_0.d_{-1}...d_{-(p-1)} - (z/\beta^e) \mid \beta^{p-1}$ ulp-a.

   Relativna greška je apsolutna vrednost razlike izmedju realnog broja i njegove reprezentacije podeljena sa apsolutnom vrednošću realnog broja.

   Na primer, relativna greška pri aproksimaciji 5.74367 sa 5.74×10^0 je $0.00367/5.74367\approx0.0006$.

ulp i relativna greška zavise od tzv. *mašinskog* $\varepsilon=(\beta/2)\beta^{-p}$; Relativna greška uvek zapisuje kao faktor od ε. Na primer, u prethodnom primeru je $\varepsilon=(\beta/2)*\beta^{-p}=5*(10)^{-3}=0.005$. Tako se relativna greška može izraziti kao

$$((0.00367/5.74367)/0.005)\varepsilon \approx 0.12\varepsilon$$

Razlika izmedju *ulp* i relativne greške:

$x = 12.35$ je aproksimiran sa $x_a = 1.24 \times 10^1$.

Greška je $0.5ulp$, a relativna greška je 0.8ε.

$8 * x_a$: tačna vrednost je $8 * x = 98.8$, izračunata vrednost je $8 * x_a = 9.92 \times 10^1$. Greška merena u *ulp* je 8 puta veća, relativna greška je ista.

Ako je realan broj zapisan sa greškom od n *ulp*-a, tada je broj cifara obuhvaćenih greškom $log_\beta n$, a ako je relativna greška $n\varepsilon$ tada je broj obuhvaćenih cifara $\approx log_\beta n$.

2. **Cifre čuvari**. Neka je npr. $p = 6$, β=10 i neka treba naći razliku 10000.1-9999.93:

x =	1.00001×10^4
y =	0.99999×10^4
x-y =	0.00002×10^4

Tačan odgovor je 0.17, greška od oko 30 *ulp*-a;

Operacija sa dodatnim ciframa, npr. sa $P+1$ cifrom:

x =	1.000010×10^4
y =	0.999993×10^4
x-y =	0.000017×10^4

3. **Tačno zaokružene operacije.**

Operacije koje se izvode tako da se izračuna tačna vrednost a zatim zaokruži na najbliži broj u pokretnom zarezu se nazivaju *tačno zaokružene*.

Zaokruživanje:

- na najbližu vrednost (4,48$\longrightarrow$ 4,5; 4,34$\longrightarrow$ 4,3; 4,45$\longrightarrow$?)
- na parnu cifru (4,45$\longrightarrow$ 4,4)

Opis standarda

IEEE 754 standard propisuje:

- $\beta=2$ i $\beta=10$ kao osnove koje se koristi za zapis brojeva u pokretnom zarezu.
- Osnovne formate za zapis podataka u binarnoj i dekadnoj osnovi
- Formate za razmenu podataka zapisanih u binarnoj i dekadnoj osnovi
- Način izvodjenja operacija sabiranja, oduzimanja, množenja, deljenja, spojenog višestrukog sabiranja, dobijanja kvadratnog korena, ostatka pri deljenju, izvodjenja poredjenja, itd.
- Način konverzije izmedju celobrojnih vrednosti i vrednosti u pokretnom zarezu
- Način konverzije izmedju različitih formata brojeva u pokretnom zarezu
- Način konverzije izmedju formata brojeva u pokretnom zarezu i njihove spoljačnje reprezentacije u obliku niski karaktera
- Vrste izuzetaka koji se javljaju pri radu sa brojevima u pokretnom zarezu i načine njihove obrade

Formalno, broj u pokretnom zarezu može da bude predstavljen na jedan od sledeća tri načina:

1. Kao uredjena trojka *(znak, eksponent, značajan_deo_broja)*. U osnovi β broj u pokretnom zarezu predstavljen na ovaj način ima vrednost
   $(-1)^{znak} \times \beta^{eksponent} \times$ *značajan_deo_broja*.
2. +∞, -∞
3. qNaN (tihi NaN) i sNaN (signalni NaN)

Parametri koji odredjuju skup brojeva predstavljivih u nekom formatu:

- Osnova β=2 ili β=10
- Broj cifara u značajnom delu p. Broj cifara u značajnom delu broja odredjuje *preciznost zapisa* .
- Najveća vrednošću eksponenta *emax*.
- Najmanja vrednošću eksponenta *emin=1-emax*.

- Realni brojevi u intervalu $[\beta^{emin}, \beta^{emax} \times (\beta - \beta^{1-p})$ se nazivaju *normalni*.
- Brojevi koji su po apsolutnoj vrednosti manji od β^{emin} (ali veći od 0) se nazivaju *subnormalni*.

Specijalne vrednosti

Klase podataka propisane IEEE standardom su:

- **Normalni brojevi.**
- **NaN.**
- **Beskonačno.**
- **Označena nula.**
- **Subnormalni brojevi.**

Specijalne vrednosti se zapisuju kao različite kombinacije vrednosti bitova u zapisu broja. Način zapisa specijalnih vrednosti zavisi od osnove koja se koristi za zapis realnog broja u pokretnom zarezu.

Zašto dekadna osnova?

- Ljudi računaju u dekadnom sistemu
- Finansijske i komercijalne aplikacije, kao i aplikacije koje su orijentisane prema korisnicima (tzv. *human-centric*) imaju legitimne zahteve za aritmetikom sa brojevima u dekadnom sistemu
- Oko 55% numeričkih podataka u bazama podataka su dekadni podaci, a od preostalih još 43% su celi brojevi
- Tradicionalan način rada sa celobrojnom aritmetikom i ručnim skaliranjem nije pogodan za korišćenje i podložan je greškama

Predstavljanja brojeva u dekadnoj osnovi

1. Softverski - softver se stara o zapisu i operacijama sa takvim brojevima. Brzina rada nije zadovoljavajuća
2. Hardverski
    - Pomoću BCD zapisa.
        - fiksni zarez,
        - relativno složeno izvodjenje operacija
        - Znatno sporije od binarne osnove.
    - Pomoću zapisa realnih brojeva u pokretnom zarezu pomoću dekadne osnove (IEEE 754-2008).
        - Trenutno samo neki od procesora hardverski podržavaju ovakav zapis (IBM *z series*, Power, ...)

Zapis u registrima računara

Zahtevi

- Realne brojeve u dekadnoj osnovi u računaru treba zapisati u registrima računara, tj. u prostoru iste veličine kao i pri zapisu realnih brojeva u pokretnom zarezu pomoću binarne osnove.
- Zapis treba da bude u obliku znak, frakcija i eksponent.
- Zapis treba da bude moguć u jednostrukoj, dvostrukoj i četvorostrukoj tačnosti
- Zapis treba da bude takav omogući tačno zapisivanje dekadnih cifara bez konverzije
- Preciznost zapisa treba da bude uporediva sa preciznošću zapisa dekadnih brojeva pomoću binarne osnove

...i problemi pri tom zapisu

- Nedovoljan prostor (npr. u registru veličine 32 bita) za smeštanje svih potrebnih komponenti ako se kodiraju pomoću BCD koda
- Potrebno je primeniti kodiranje koje omogućuje zapis svih potrebnih podataka u dovoljnoj tačnosti
- Zapis treba da bude moguć u jednostrukoj, dvostrukoj i četvorostrukoj tačnosti
- Treba predvideti zapis specijalnih vrednosti
- Za takav zapis treba definisati pravila za izvodjenje operacija

Osnovni formati zapisa

IEEE 754 propisuje pet osnovnih formata zapisa:

1. Tri formata zapisa sa binarnom osnovom sa preslikavanjem vrednosti u 32, 64 i 128 bita.

2. Dva formata zapisa sa dekadnom osnovom sa preslikavanjem vrednosti u 64 i 128 bita.

Parametri	Formati				
	Sa binarnom osnovom (β=2)			Sa dekadnom osnovom (β=10)	
	binary32	binary64	binary128	decimal64	decimal128
p	24	53	113	16	34
emax	+127	+1023	+16383	+384	+6144

Tabela 2: Osnovni formati zapisa brojeva u IEEE 754-2008

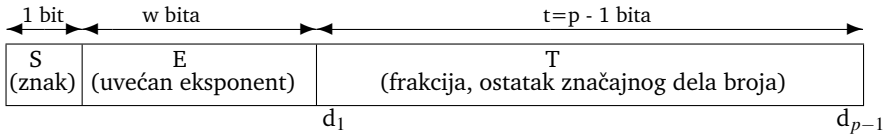
Bez obzira na format zapisa, implementacija mora da omogući predstavljanje sledećih brojeva u pokretnom zarezu:

- Označenih nula i ne-nula brojeva u pokretnom zarezu u obliku $(-1)^s \times \beta^e \times m$ gde važi:
  1. s je 0 ili 1
  2. e je proizvoljan ceo broj $emin \leq e \leq emax$
  3. m je broj predstavljen niskom cifara oblika $d_0.d_1d_2...d_{p-1}$ gde je d_i cifra $0 \leq d_i < \beta$. Odavde je $0 \leq m < \beta$.
- Dve beskonačne vrednosti -∞ i + ∞.
- Dve NaN vrednosti (qNaN i sNaN)

U nekim slučajevima je pogodno da se značajan deo broja predstavi u obliku celog umesto razlomljenog broja. U tom slučaju se konačan broj u pokretnom zarezu predstavlja kao označen nula ili ne-nula broj oblika $(-1)^s \times \beta^q \times c$ gde važi:

1. s je 0 ili 1
2. q je proizvoljan ceo broj $emin \leq q + p - 1 \leq emax$
3. c je broj predstavljen niskom cifara oblika $d_0d_1d_2...d_{p-1}$ gde je d_i cifra $0 \leq d_i < \beta$. Odavde je $0 \leq c < \beta^p$.

Formati za razmenu podataka zapisanih pomoću binarne osnove



Slika 6: Format za razmenu podataka zapisanih pomoću binarne osnove

1. Znak broja S veličine 1 bit
2. Uvećani eksponent E=e+uvećanje zapisan u w bita
3. Ostatak $T = d_1 d_2 ... d_{p-1}$ značajnog dela broja zapisan u dužini t=p-1 bita. Vodeći bit d_0 značajnog dela broja je implicitno zapisan u uvećanom eksponentu E.

		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

Provera vrednosti zapisa broja +15:

– Znak = +
– Eksponent = 3 (=130-127)
– 1,frakcija = $(1{,}111)_2 = 1\times 2^0+1\times 2^{-1}+1\times 2^{-2}+1\times 2^{-3}$
– Vrednost = Znak 1,frakcija * $2^{eksponent} = +(+1\times 2^0+1\times 2^{-1}+1\times 2^{-2}+1\times 2^{-3}) * 2^3$
$= +1\times 2^3+1\times 2^2+1\times 2^1+1\times 2^0 = 8 + 4+ 2 +1 = +15$

Tabela 3: Zapis realnih brojeva u jednostrukoj tačnosti pomoću binarne osnove prema IEEE 754 standardu

Parametri kodiranja	Formati				
	binary16	binary32	binary64	binary128	binary{k} (k≥128)
k, širina polja (bita)	16	32	64	128	umnožak od 32
p, preciznost (bita)	11	24	53	113	k-round$(4\times log_2(k))+13$
emax, najveća vrednost eksponenta e	15	127	1023	16383	$2^{(k-p-1)}-1$
emin, najmanja vrednost eksponenta e	-14	-126	-1022	-16382	$1-(2^{(k-p-1)}-1)$
Uvećanje E-e	15	127	1023	16383	emax
bit za znak (bita)	1	1	1	1	1
w, dužina polja za eksponent (bita)	5	8	11	15	round$(4\times log_2(k))$-13
t, ostatak značajnog dela broja (bita)	10	23	52	112	k-w-1

Najveće i najmanje vrednosti u formatu					
N_{max}	$(1\text{-}2^{-11})\times 2^{16}$ $=6.5504\times 10^{+4}$	$(1\text{-}2^{-24})\times 2^{128}$ $\approx 3.4\times 10^{+38}$	$(1\text{-}2^{-53})\times 2^{1024}$ $\approx 1.8\times 10^{+308}$	$(1\text{-}2^{113})\times 2^{16384}$ $\approx 1.2\times 10^{+4932}$	zavisi od k
N_{min}	1.0×2^{-14} $\approx 6.1\times 10^{-5}$	1.0×2^{-126} $\approx 1.2\times 10^{-38}$	1.0×2^{-1022} $\approx 2.2\times 10^{-308}$	1.0×2^{-16382} $\approx 3.4\times 10^{-4932}$	zavisi od k
D_{min}	1.0×2^{-24} $\approx 5.96\times 10^{-8}$	1.0×2^{-149} $\approx 1.4\times 10^{-45}$	1.0×2^{-1074} $\approx 4.9\times 10^{-324}$	1.0×2^{-16494} $\approx 6.5\times 10^{-4966}$	zavisi od k

D_{min} Najmanji (po apsolutnoj vrednosti) predstavljiv subnormalan broj
N_{max} Najveći (po apsolutnoj vrednosti) predstavljiv normalan broj
N_{min} Najmanji (po apsolutnoj vrednosti) predstavljiv normalan broj

Tabela 4: Vrednosti parametara formata za razmenu podataka

Klasa podataka	Znak	Uvećani eksponent E	Implicitni bit	Frakcija T	Reprezentacija	Vrednost
Označena nula	S	0	0	0	$(S, emin, 0)$	(-1)^S×(+0)
Normalni brojevi	S	$1 \leq E \leq 2^w - 2$	1	proizvoljno	$(S, (E-uve), (1+2^{1-p} \times T))$	$(-1)^S \times 2^{E-uve} \times (1+2^{1-p} \times T)$
Subnormalni brojevi	S	0††	0	$\neq 0$	$(S, emin, (0+2^{1-p} \times T))$	$(-1)^S \times 2^{emin} \times (0+2^{1-p} \times T)$
Beskonačno	S	$2^w - 1$	xxx	0	$(-1)^S \times (+\infty)$	$(-1)^S \times (+\infty)$
Nije broj	xxx	$2^w - 1$	xxx	d_1=1, d_r=proizvoljno $\longrightarrow$ qNaN		NaN
				d_1=0, $d_r \neq 0 \longrightarrow$ SNaN		

uve	Uvećanje eksponenta
xxx	Ne primenjuje se
††	U aritmetičkim operacijama se tretira kao da ima vrednost 1
Eksponent	za E=0 - sadržaj polja za eksponent su sve nule, a za E=$2^w - 1$ - sadržaj polja za eksponent su sve jedinice
Frakcija:	d_1 je krajnje levi bit frakcije; d_r predstavlja ostale bitovi frakcije

Tabela 5: Klase podataka u zapisu sa binarnom osnovom u IEEE 754

U slučaju da je dužina formata 64 bita ili je $\geq$128 bita, vrednosti parametara se mogu odrediti na sledeći način:

$$
\begin{array}{lcl}
k & = & 1+w+t = w+p \\
 & = & 32 \times ceiling((p+round(4 \times log_2(p+round(4 \times log_2(p))-13))-13)/32) \\
w & = & k-t-1 = k-p = round(4 \times log_2(k))-13 \\
t & = & kw-1 = p-1 = k-round(4 \times log_2(k))+12 \\
p & = & k-w = t+1 = k-round(4 \times log_2(k))+13 \\
emax & = & bias = 2^{(w-1)}-1 \\
emin & = & 1-emax = 2-2^{(w-1)}
\end{array}
$$

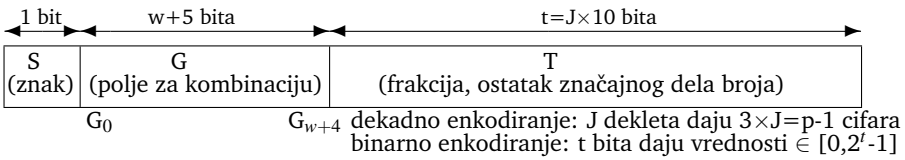
pri čemu funkcija round() vrši zaokruživanje na najbliži ceo broj.

Formati za razmenu podataka zapisanih pomoću dekadne osnove

Kohorte

- Realan broj može da ima više reprezentacija u zapisu pomoću dekadne osnove.
- Skup različitih reprezentacija u koje se preslikava bpoj u pokretnom zarezu se naziva *kohorta*.
- Primer, ako je c umnožak od 10 i q je manje od najveće dozvoljene vrednosti, tada su (s, q, c) i $(s, q+1, c/10)$ dve reprezentacije istog broja u pokretnom zarezu i članovi iste kohorte.

Enkodiranje



Slika 7: Format za razmenu podataka zapisanih pomoću dekadne osnove

1. Znak broja S veličine 1 bit
2. Polje za kombinaciju veličine $w+5$ bita koje enkodira klasifikaciju, i u slučaju da je enkodirani podataka končan broj i eksponent q i četiri bita (od kojih su 1 ili 3 implicitno zapisani) zančajnog dela broja. Uvećani eksponent E je veličine $w+2$ bita i iznosi $q+$ uvećanje, gde su vrednosti prva dva bita uvećanog eksponenta uzeta zajedno 0, 1 ili 2.
3. Ostatak T značajnog dela broja koji sadrži $t = 10 \times J$ bita. Kombinovanjem sa bitovima najveće težine značajnog dela broja koji se nalaze u polju za kombinaciju dobija se da format omogućava zapisu ukupno $p = 3 \times J + 1$ dekadnih cifara.

Reprezentacija i vrednost podatka u pokretnom zarezu zapisanog pomoću dekadne osnove se odredjuju na sledeći način:

1. Ako su 5 bitova najveće težine kombinacije ($G_0 ... G_4$) 11111, vrednost broja je NaN. Ako je bit $G_5 = 1$ tada je vrednost sNaN; u suprotnom je qNaN.

2. Ako su 5 bitova najveće težine kombinacije ($G_0 ... G_4$) jednaki 11110, reprezentacija i vrednost broja su jednaki $(-1)^S \times +\infty$. Ostalih w bitova u polju za kombinaciju i frakcija se ignorišu.

3. Ako su 5 bitova najveće težine kombinacije ($G_0 ... G_4$) u intervalu 00000-11101, tada je u pitanju konačan broj. Reprezentacija konačnog broja je jednaka $r = (S, E - \text{uvećanje}, C)$, dok je njegova vrednost jednaka $v = (-1)^S \times 10^{E-\text{uvećanje}} \times C)$ gde je C dobijeno dopisivanjem cifre najveće težine ili bitova iz polja za kombinaciju G na frakciju T(ostatka značajnog dela broja). Pri tome je uvećani eksponent E zapisan u polju za kombinaciju. Enkodiranje podataka u polju za kombinaciju:

   (a) U slučaju dekadnog enkodiranja bitovi od G_5 do G_{w+4} čine w bita najmanje težine eksponenta. Za ostale cifre veži

      i. Ako su 5 bitova kombinacije u intervalu 0xxxx-10xxx, G_0G_1 su dva bita najveće težine eksponenta, a $d_0 = 4G_2 + 2G_3 + G_4$.

      ii. Ako su 5 bitova kombinacije u intervalu 110xx-1110xx, dekadna vrednost cifre d_0 je jednaka $8 + G_4$, dok su G_2G_3 cifre najveće težine eksponenta.

      Ako je T=0 i pet cifara najveće težine u polju za kombinaciju je jednako 00000, 01000, ili 10000 tada je vrednost broja $v = (-1)^S \times (+0)$.

   Za enkodiranje cifara frakcije se koristi DPD kodiranje.

U slučaju binarnog enkodiranja (BID, eng. *binary-integer decoding*) dekadne cifre u značajnom delu broja su zapisane kao celobrojne vrednosti u binarnoj osnovi.

1. Ako su bitovi G_0 i G_1 posmatrani zajedno 00, 01 ili 10 tada se uvećani eksponent E formira od cifara G_0 do G_{w+1}, dok se značajni deo broja formira počev od bita G_{w+2} pa do kraja zapisa (uključujući i frakciju T).

2. Ako su bitovi G_0 i G_1 posmatrani zajedno 11, a G_2G_3 jedna od vrednosti 00, 01 ili 10 tada se uvećani eksponent E formira od cifara G_2 do G_{w+3}, dok se značajni deo broja formira dodajući 4 bita koji su vrednost $8+G_{w+4}$ na početak frakcije T.

$b_{(6)}, b_{(7)}, b_{(8)}, b_{(3)}, b_{(4)}$	$d_{(1)}$	$d_{(2)}$	$d_{(3)}$
$0xxxx$	$4b_{(0)} + 2b_{(1)} + b_{(2)}$	$4b_{(3)} + 2b_{(4)} + b_{(5)}$	$4b_{(7)} + 2b_{(8)} + b_{(9)}$
$100xx$	$4b_{(0)} + 2b_{(1)} + b_{(2)}$	$4b_{(3)} + 2b_{(4)} + b_{(5)}$	$8 + b_{(9)}$
$101xx$	$4b_{(0)} + 2b_{(1)} + b_{(2)}$	$8 + b_{(5)}$	$4b_{(3)} + 2b_{(4)} + b_{(9)}$
$110xx$	$8 + b_{(2)}$	$4b_{(3)} + 2b_{(4)} + b_{(5)}$	$4b_{(0)} + 2b_{(1)} + b_{(9)}$
11100	$8 + b_{(2)}$	$8 + b_{(5)}$	$4b_{(0)} + 2b_{(1)} + b_{(9)}$
11101	$8 + b_{(2)}$	$4b_{(0)} + 2b_{(1)} + b_{(5)}$	$8 + b_{(9)}$
11110	$4b_{(0)} + 2b_{(1)} + b_{(2)}$	$8 + b_{(5)}$	$8 + b_{(9)}$
11111	$8 + b_{(2)}$	$8 + b_{(5)}$	$8 + b_{(9)}$

Tabela 6: Dekodiranje 10-bitnih gusto pakovanih dekadnih brojeva zapisanih u bitovima b_0...b_9 u 3 dekadne cifre d_0,d_1 i d_2. x Označava da je vrednost na toj poziciji nevažna. Ovim preslikavanjem se svih 1024 mogućih 10-bitnih kombinacija preslikava u 1000 mogućih trocifrenih brojeva

$d_{(1,0)}, d_{(2,0)}, d_{(3,0)}$	$b_{(0)}, b_{(1)}, b_{(2)}$	$b_{(3)}, b_{(4)}, b_{(5)}$	$b_{(6)}$	$b_{(7)}, b_{(8)}, b_{(9)}$
000	$d_{(1,1:3)}$	$d_{(2,1:3)}$	0	$d_{(3,1:3)}$
001	$d_{(1,1:3)}$	$d_{(2,1:3)}$	1	$0, 0, d_{(3,3)}$
010	$d_{(1,1:3)}$	$d_{(3,1:2)}, d_{(2,3)}$	1	$0, 1, d_{(3,3)}$
011	$d_{(1,1:3)}$	$1, 0, d_{(2,3)}$	1	$1, 1, d_{(3,3)}$
100	$d_{(3,1:2)}, d_{(1,3)}$	$d_{(2,1:3)}$	1	$1, 0, d_{(3,3)}$
101	$d_{(2,1:2)}, d_{(1,3)}$	$0, 1, d_{(2,3)}$	1	$1, 1, d_{(3,3)}$
110	$d_{(3,1:2)}, d_{(1,3)}$	$0, 0, d_{(2,3)}$	1	$1, 1, d_{(3,3)}$
111	$0, 0, d_{(1,3)}$	$1, 1, d_{(2,3)}$	1	$1, 1, d_{(3,3)}$

Tabela 7: Kodiranje 3 dekadne cifre u 10-bitno gusto pakovane dekadne brojeva u 3 dekadne brojeve. Bitovi u zapisu dekadnih cifara $d_{(1)}$, $d_{(2)}$ i $d_{(3)}$ su izraženi preko drugog indeksa ($d_{(1,0:3)}$, $d_{(2,0:3)}$ i $d_{(3,0:3)}$, pri čemu je bit označen sa 0 bit najveće težine. Bitovi u dekletu u koji se vrši kodiranje su označeni sa b_0...b_9. Operacije formiraju samo 1000 kanoničkih dekleta koji su navedeni u ovoj tabeli. Nekanoničkih dekleta ima 24, sa bitskim zapisom oblika 01x11x111x, 10x11x111x, ili 11x11x111x gde 'x' označava da je vrednost na toj poziciji nevažna.

U slučaju da je dužina formata umnožak od 32, vrednosti parametara se mogu odrediti na sledeći način:

$$
\begin{array}{lcl}
k & = & 1+5+w+t = 32\times ceiling((p+2)/9) \\
w & = & kt-6 = k/16+4 \\
t & = & kw-6 = 15\times k/16-10 \\
p & = & 3t/10+1 = 9\times k/32-2 \\
emax & = & 3\times 2^{w-1} \\
emin & = & 1-emax \\
\text{uvećanje} & = & emax+p-2
\end{array}
$$

		Znak	Kombinacija	Nastavak frakcije
$+15$	=	0	01000 100101	0000 0000 0000 0001 0101
-15	=	1	01000 100101	0000 0000 0000 0001 0101
$+15.0$	=	0	01000 100100	0000 0000 0000 1101 0000
$+1/64$	=	0	01000 011111	0000 0101 0111 0010 0101
$+0$	=	0	01000 100101	0000 0000 0000 0000 0000
-0	=	1	01000 100101	0000 0000 0000 0000 0000
$+0.0$	=	0	01000 100100	0000 0000 0000 0000 0000
$+0.00$	=	0	01000 100011	0000 0000 0000 0000 0000
$+(10^7-1)\times 10^{+90}$	=	0	11101 111111	1111 1111 1111 1111 1111 nk.d.
$+(10^7-1)\times 10^{+90}$	=	0	11101 111111	0011 1111 1100 1111 1111 k.d.
$+1\times 10^{-95}$	=	0	00000 000110	0000 0000 0000 0000 0001
$+1\times 10^{-101}$	=	0	00000 000000	0000 0000 0000 0000 0001

Broj $+1\times 10^{-101}$ je zapisan na dva načina: pomoću nekanoničkog i kanoničkog dekleta.
Provera vrednosti zapisa broja +15:
– Znak = +
– Eksponent = 0 (=101 - 101)
– Frakcija = $(15)_{10}$ = 0000000000 0000010101 (kodirana u dva dekleta)
– Vrednost = Znak frakcija * $10^{eksponent}$ = +15* 10^0 = $+15_{10}$

Tabela 9: Primer zapisa realnih brojeva pomoću dekadne osnove /DPD kodiranje

Parametri kodiranja	Formati			
	decimal32	decimal64	decimal128	decimal{k} (k≥32)
k, širina polja (bita)	32	64	128	umnožak od 32
p, preciznost (cifara)	7	16	34	$9 \times k/32 - 2$
emax, najveća vrednost eksponenta e	96	384	6144	$3 \times 2^{k/16+3}$
emin, najmanja vrednost eksponenta e	-95	-383	-6143	$1 - (3 \times 2^{k/16+3})$
Uvećanje E-q	101	398	6176	emax+p-2
bit za znak (bita)	1	1	1	1
w+5, dužina polja za kombinaciju (bita)	11	13	17	k/16+9
t, ostatak značajnog dela broja (bita)	20	50	110	15×k/16 -10

Najveće i najmanje vrednosti u formatu				
N_{max}	$(10\text{-}1.0\times10^{-6})\times10^{+96}$	$(10\text{-}1.0\times10^{-15})\times10^{+384}$	$(10\text{-}1.0\times10^{-33})\times10^{+61445}$	$\approx(10\text{-}1.0\times10^{-p+1})\times10^{emax+1}$
N_{min}	1.0×10^{-95}	1.0×10^{-383}	1.0×10^{-6143}	$1.0\times10^{1-emax}$
D_{min}	1.0×10^{-101}	1.0×10^{-398}	1.0×10^{-6176}	$1.0\times10^{emax+p-2}$

D_{min} Najmanji (po apsolutnoj vrednosti) predstavljiv subnormalan broj
N_{max} Najveći (po apsolutnoj vrednosti) predstavljiv normalan broj
N_{min} Najmanji (po apsolutnoj vrednosti) predstavljiv normalan broj

Tabela 8: Vrednosti parametara formata za razmenu podataka

		Znak	Kombinacija	Nastavak frakcije
$+15$	=	0	01100101000	00000000000000001111
-15	=	1	01100101000	00000000000000001111
$+15.0$	=	0	01100100000	00000000000010010110
$+1/64$	=	0	01011111000	00000011110100001001
$+0$	=	0	01100101000	00000000000000000000
-0	=	1	01100101000	00000000000000000000
$+0.0$	=	0	01100100000	00000000000000000000
$+0.00$	=	0	01100011000	00000000000000000000
$+(10^7-1)\times 10^{+90}$	=	0	11101111111	10001001011001111111
$+1\times 10^{-95}$	=	0	00000110000	00000000000000000001
$+1\times 10^{-101}$	=	0	00000000000	00000000000000000001

Provera vrednosti zapisa broja +15:
– Znak = +
– Eksponent = 0 ($(01100101)_2$ - uvećanje=$(101)_{10}$-uvećanje = 101 - 101=0)
– Frakcija = $(15)_{10}$ = $(00000000000000000010101)_2$ (kodirana u 23 bita)
– Vrednost = Znak frakcija * $10^{eksponent}$ = +15* 10^0 = $+15_{10}$

Tabela 10: Primer zapisa realnih brojeva pomoću dekadne osnove /BID kodiranje

Prošireni i proširivi zapis

Parametri	Prošireni formati pridruženi formatu				
	binary32	binary64	binary128	decimal64	decimal128
p cifar $\geq$	32	64	128	22	40
emax	+1023	+16383	+65535	+6144	+24576

Tabela 11: Parametri proširenih formatia zapisa brojeva u IEEE 754-2008

Zaokruživanje

1. **Zaokruživanje na najbližu vrednost.**

   - **Zaokruživanje na parnu cifru**. Medjurezultat izračunavanja se zaokružuje na najbližu predstavljivu vrednost, uz zaokruživanje na parnu cifru kada je izračunati medjurezultat na sredini intervala izmedju dve predstavljive vrednosti. Ovaj način zaokruživanja je predefinisan za zapis sa binarnom osnovom, dok je za zapis sa dekadnom osnovom preporučen (ai ne i obavezan) kao predefinisan način zaokruživanja.
   - **Zaokruživanje na udaljeniju vrednost**. Medjurezultat izračunavanja se zaokružuje na najbližu predstavljivu vrednost, uz zaokruživanje na veću (po apsolutnoj vrednosti) vrednost kada je izračunati medjurezultat na sredini intervala izmedju dve predstavljive vrednosti.

2. **Usmereno zaokruživanje**

   - **Zaokruživanje ka pozitivnom.** Medjurezultat se zaokružuje na veću vrednost (moguće i $+\infty$). Ovaj način zaokruživanja se jov s naziva i zaokruživanje ka $+\infty$.
     - Ako je broj pozitivan i ako postoji bar jedna jedinica na nekoj poziciji desno od poslednje pozicije koja se čuva u zapisu, na tom mestu se dodaje jedinica.
     - Bez obzira na znak broja, odbace se sve cifre desno od poslednje pozicije koja se čuva u zapisu.
   - **Zaokruživanje ka negativnom.** Medjurezultat se zaokružuje na manju vrednost. Ovaj način zaokruživanja se jov s naziva i zaokruživanje prema -∞.
     - Ako je broj negativan i ako postoji bar jedna jedinica na nekoj poziciji desno od poslednje pozicije koja se čuva u zapisu, na tom mestu se oduzima jedinica.
     - Bez obzira na znak broja, odbace se svi bitovi desno od poslednje pozicije koja se čuva u zapisu.
   - **Zaokruživanje ka 0.** Rezultat je broj u pokretnom zarezu koji je najbliži broju, ali ne i veći po apsolutnoj vrednosti od broja koji se zaokružuje.

Aritmetika u pokretnom zarezu

1. Totalno uredjenje predstavljivih vrednosti

   - 12.5>12.50>12.500
     125×10^{-1}, 1250×10^{-2} i 12500×10^{-3}.
   - (+)qNaN > (+)sNaN >+∞> N_{max} > +konačan broj > +N_{min}> +D_{min} > 0 > -D_{min} >-N_{min} >-konačan broj > -N_{max} > -∞ > (-)sNaN > (-) qNaN

2. Izračunavanje dekadnog eksponenta.

   - izbor odgovarajućeg člana kohorte
   - ako rezultat operacije nije tačan broj da bi se dobio najveći broj značajnih cifara koristi se član kohorte sa najmanjim mogućim eksponentom
   - ako je rezultat tačna vrednost, bira se član kohorte zasnovan na ciljanom eksponentu za rezultat operacije

Specijalne vrednosti u aritmetičkim operacijama

$$-\infty < \text{bilo koji konačan broj} < +\infty$$

Operacije sa beskonačnostima su obično tačne i ne generišu izuzetke. Sa izuzetkom operacija koje proizvode `QNaN` svaka aritmetička operacija koja uključuje ∞ takodje daje ∞ kao rezultat.

x	$\pm$	$(+\infty)$	=	$\pm\infty$ (za konačno x)
x	$\pm$	(-∞)	=	$\mp\infty$ (za konačno x)
x	$\times$	(∞)	=	∞ (za konačno ili beskonačno x $\neq$0)
$(+\infty)$	+	$(+\infty)$	=	$+\infty$
$(+\infty)$	-	(-∞)	=	$+\infty$
(-∞)	+	(-∞)	=	-∞
(-∞)	-	$(+\infty)$	=	-∞
(-∞)	-	$(+\infty)$	=	-∞
		$\sqrt{+\infty}$	=	$+\infty$

ostatak pri deljenju x sa ∞ (x REM ∞) za konačno i normalano x
konverzija ∞ u ∞ u različitom formatu

- Signalni NaN označava prisustvo neinicijalizovanih promenljivih, pogrešne operacije ili proširenja u aritmetici koja nisu deo standarda.
- Tihi NaN može (u zavisnosti od implementacije) da označi postojanje informacija o pogrešnim ili nepostojećim podacima i rezultatima.
- Preporuka je da se što je moguće više takvih dijagnostičkih informacija prosledjuje kroz operacije tako što će rezultat operacija u kojima učestvuje NaN takodje biti NaN.

Operacija	`QNaN` formiran pomoću
Sabiranje, oduzimanje	$(\pm\infty) \pm (\mp\infty)$
Množenje	$0 \times \infty$
Deljenje	$0/0,\ \infty/\infty$
Ostatak	x REM $0,\ \infty$ REM y
Kvadratni koren	$\sqrt{x}$ kada je x<0

Tabela 12: Operacije koje proizvode Tihi NaN

Sabiranje i oduzimanje

Neka su brojevi x i y zapisani u pokretnom zarezu kao $x = x_s \times \beta^{x_e}, y = y_s \times \beta^{y_e}$ i neka treba izračunati njihov zbir ili razliku. U opštem slučaju dobijena vrednost će biti jednaka

$$x \pm y = \begin{cases} (x_s \times \beta^{x_e - y_e} \pm y_s) \times \beta^{y_e} & \text{ako } x_e \leq y_e \\ (x_s \pm y_s \times \beta^{y_e - x_e}) \times \beta^{x_e} & \text{ako } y_e < x_e \end{cases}$$

Pri izvodjenju operacija treba voditi računa o generisanju i propagaciji specijalnih vrednosti. Osnovni koraci u algoritmu za sabiranje i oduzimanje su:

1. Provera postojanja specijalnih vrednosti. Ukoliko je neki od argumenata operacije specijalna vrednost, rezultat se odredjuje prema odgovarajućim pravilima

2. Oduzimanje $x - y$ se se realizuje kao $x + (-y)$ uz prethodnu promenu znaka argumenta y.

3. Ukoliko je jedan od sabiraka jednak nuli, vrednost drugog sabirka je rezultat sabiranja.

4. Svodjenje sabiraka na jednake eksponente.

5. Sabiraju se frakcija sabiraka, pri čemu se uzimaju u obzir njihovi znaci. Sabiranje se vrši po pravilima za sabiranje celih brojeva u zapisu znak i apsolutna vrednost. Ukoliko je dobijeni rezultat nula, tada je ukupan zbir nula. Ako je pak pri sabiranju došlo do prekoračenja, dobijeni rezultat se pomera za jedno mesto udesno uz povećanje vrednosti eksponenta za jedan. Ako ovo povećanje vrednosti eksponenta dovede do prekoračenja (vrednosti eksponenta), ukupan rezultat sabiranja je $+\infty$ ili -∞ u zavisnosti od znaka broja.

6. Traženi zbir predstavlja broj u pokretnom zarezu čiji su znak i frakcija jednaki znaku i zbiru frakcija, a eksponent jednak eksponentu sabiraka. Ako u rezultatu sabiranja frakcija nije normalizovan (u binarnom zapisu) ili nije predstavljen pomoću ciljanog eksponenta (u dekadnom zapisu), pokušava se njegova normalizacija odnosno traži ciljani eksponet. Ukoliko je tom prilikom došlo do potkoračenja vrednosti eksponenta, ukupan zbir je nula, dok se u ostalim slučajevima po potrebi vrši zaokruživanje i formira traženi zbir.

Množenje i deljenje

Neka su brojevi x i y zapisani u pokretnom zarezu kao $x = x_s \times \beta^{x_e}, y = y_s \times \beta^{y_e}$ i neka treba izračunati njihov proizvod ili količnik. U opštem slučaju dobijena vrednost je

$$x * y = (x_s * y_s) \times \beta^{x_e + y_e}$$

$$x / y = (x_s / y_s) \times \beta^{x_e - y_e}$$

1. Proverava se postojanje specijalnih vrednosti. Ukoliko neki od argumenata operacije predstavlja specijalnu vrednost, rezultat se odredjuje prema odgovarajućim pravilima

2. Ukoliko je bar jedan od činilaca jednak nuli, rezultat je 0.

3. Saberu se vrednosti eksponenata i od dobijenog zbira oduzme uvećanje. Ako je došlo do prekoračenja pri ovom sabiranju, krajnji rezultat je $\pm\infty$ u zavisnosti od znaka brojeva x i y. Ako je pak došlo do potkoračenja vrednosti eksponenta, krajnji rezultat je pozitivna ili negativna (u zavisnosti od znaka brojeva x i y) nula.

4. Pomnože se frakcije brojeva. Množenje se vrši prema pravilima za množenje celih brojeva zapisanih pomoću znaka i apsolutne vrednosti.

5. Dobijeni rezultat se normalizuje, odnosno odredi se ciljani eksponent sličnim postupkom kao kod sabiranja.

6. Broj cifara u proizvodu je dvostruko veći od broja cifara vrednosti koje su pomnožene; cifre koje su višak se odbacuju u procesu zaokruživanja.

Osnovni koraci u algoritmu za deljenje su:

1. Proverava se postojanje specijalnih vrednosti. Ukoliko je neki od argumenata operacije specijalna vrednost, rezultat se odredjuje prema odgovarajućim pravilima
2. Ako je delilac nula, tada
   - Ako je deljenik $\neq 0$, količnik je $\pm\infty$ u zavisnosti od znaka x.
   - Ako je deljenik $=0$, tada je rezultat `NaN`.
3. Oduzmu se vrednosti eksponenata i na dobijenu razliku doda uvećanje. Ako je došlo do prekoračenja pri ovom sabiranju, krajnji rezultat je $\pm\infty$ u zavisnosti od znaka brojeva x i y. Ako je pak došlo do potkoračenja vrednosti eksponenta, krajnji rezultat je pozitivna ili negativna (u zavisnosti od znaka brojeva x i y) nula.
4. Podele se frakcije brojeva. Deljenje se vrši prema pravilima za deljenje celih brojeva zapisanih pomoću znaka i apsolutne vrednosti.
5. Dobijeni rezultat se normalizuje, odnosno odredi se ciljani eksponent sličnim postupkom kao kod sabiranja.
6. Dobijeni količnik se zaokružuje prema pravilima za zaokruživanje.

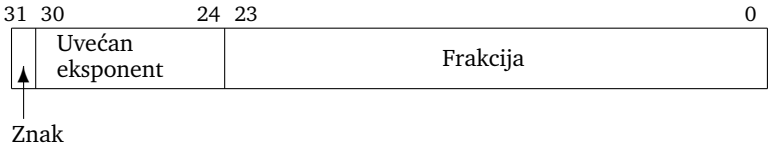
Izuzeta stanja, zastavice i zamke

- IEEE 754 standard deli izuzeta stanja u 5 klasa: prekoračenje, potkoračenje, deljenje sa nulom, pogrešna operacija i netačnost.
- Za svako od ovih izuzeća se postavlja posebna zastavica (eng. *flag*) koju korisnik može programski da ispituje.
- Standard takodje omogućuje upotrebu programskih rutina za rad sa zamkama (eng. *trap handler*).

```
do    uslov
until (x >= 100)
```

Zapis sa heksadekadnom osnovom

- znak se upisuje u bit najveće težine i ima vrednost 0 za pozitivne i 1 za negativne brojeve.
- Frakcija značajnog dela broja je normalizovana (u obliku $0.d_0...d_{-(p-1)}$) i zapisuje se sa 6 heksadekadnih cifara u 24 bita.
- Eksponent se zapisuje u 7 bita na pozicijama 24–30. Vrednosti eksponenta se zapisuju u potpunom komplementu uz uvećanje za 64.



Slika 8: Format zapisa realnog broja u jednostrukoj tačnosti pomoću heksadekadne osnove

		Znak	Eksponent	Frakcija
$+15$	=	0	1000001	11110000000000000000000
-15	=	1	1000001	11110000000000000000000
$+1/64$	=	0	0111111	01000000000000000000000
0	=	0	0000000	00000000000000000000000
$+16^{-1} \times 16^{-64}$	=	0	0000000	00010000000000000000000
$(1-16^{-6}) \times 16^{+63}$	=	0	1111111	11111111111111111111111
$+16^{-6} \times 16^{-64}$	=	0	0000000	00000000000000000000001

Provera vrednosti zapisa broja +15:
– Znak = +
– Eksponent = 1 (=65 - 64)
– Frakcija = $(F)_{16}$ = F$\times 16^{-1}$
– Vrednost = Znak frakcija * $16^{eksponent}$ = +(F$\times 16^{-1}$) * 16^{1} = $+F_{16}$ = $+15_{10}$

Tabela 13: Zapis realnih brojeva u jednostrukoj tačnosti / heksadekadna osnova

Za veličinu eksponenta e važi

$$-2^6 \leq e \leq 2^6 - 1$$

Za vrednost s kojom se predstavlja frakcija važi

$$16^{-1} \leq | s | \leq 1 - 16^{-6}$$

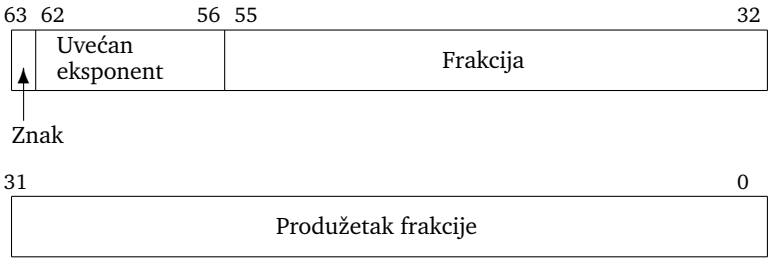
Kombinacijom ovih vrednosti dobija se da je interval brojeva koji mogu da se zapišu

$$16^{-1} * 16^{-64} \leq | x | \leq (1 - 16^{-6}) * 16^{+63}, \quad \text{odnosno}$$

$$5.4 * 10^{-79} \lessapprox | x | \lessapprox 7.2 * 10^{+75}$$

Pokušaj zapisa broja x gde važi $| x | > (1 - 16^{-6}) * 16^{+63}$ dovodi do pozitivnog ili negativnog prekoračenja

Najmanji denormalizovani broj (po apsolutnoj vrednosti) koji može da se zapiše u jednostrukoj tačnosti je $16^{-64} * 16^{-6} = 16^{-70} \approx 5.1 * 10^{-85}$

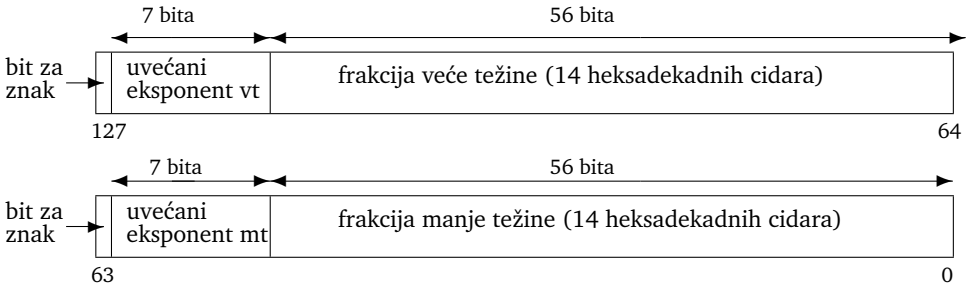


Slika 9: Format zapisa realnog broja u dvostrukoj tačnosti pomoću heksadekadne osnove

$$16^{-1} * 16^{-64} \quad \leq |x| \leq \quad (1 - 16^{-14}) * 16^{+63}, \quad \text{odnosno}$$

$$5.4 * 10^{-79} \quad \lessapprox |x| \lessapprox \quad 7.2 * 10^{+75}$$

Najmanji denormalizovani broj (po apsolutnoj vrednosti) koji može da se zapiše je $16^{-64} * 16^{-14} = 16^{-78} \approx 1.2 * 10^{-94}$



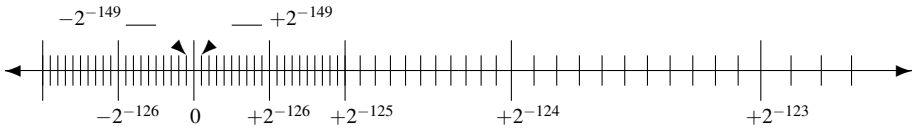
Slika 10: Format zapisa realnog broja u četvorostrukoj tačnosti pomoću heksadekadne osnove

$$16^{-1} * 16^{-64} \quad \leq |x| \leq \quad (1 - 16^{-28}) * 16^{+63}, \quad \text{odnosno}$$

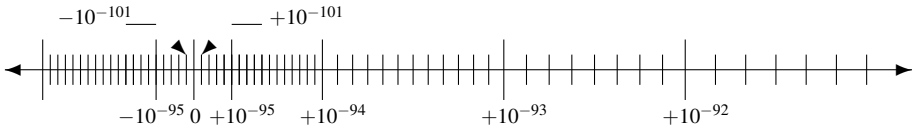
$$5.4 * 10^{-79} \quad \lessapprox |x| \lessapprox \quad 7.2 * 10^{+75}$$

Najmanji denormalizovani broj koji može da se zapiše je $16^{-64} * 16^{-28} = 16^{-92} \approx 1.7 * 10^{-111}$

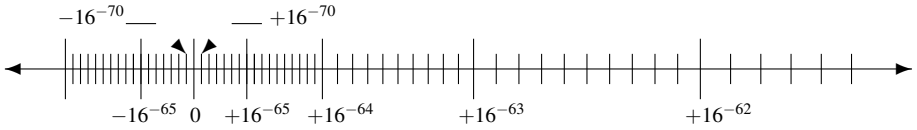
Poredjenje zapisa realnih brojeva



(b) Zapis pomoću binarne osnove - IEEE 754 format sa binarnom osnovom i uključenim subnormalnim vrednostima



(b) Zapis pomoću binarne osnove - IEEE 754 format sa dekadnom osnovom i uključenim subnormalnim vrednostima



(c) Zapis pomoću heksadekadne osnove sa uključenim subnormalnim vrednostima

Slika 11: Gustina zapisa realnih brojeva (za jednostruku tačnost)

Vrednost		S	BE or C	Ostatak frakcije	
1.0	B	0	011 1111 1	000 0000 0000 0000 0000 0000	
	H	0	100 0001	0001 0000 0000 0000 0000 0000	
	D	0	010 0010 0100	0000 0000 0000 0001 0000	[10x10^{-1}]
0.5	B	0	011 1111 0	000 0000 0000 0000 0000 0000	
	H	0	100 0000	1000 0000 0000 0000 0000 0000	
	D	0	010 0010 0100	0000 0000 0000 0000 0101	[5x10^{-1}]
1/64	B	0	011 1100 1	000 0000 0000 0000 0000 0000	
	H	0	011 1111	0100 0000 0000 0000 0000 0000	
	D	0	010 0001 1111	0000 0101 0111 0010 0101	[15625x10^{-6}]
+0	B	0	000 0000 0	000 0000 0000 0000 0000 0000	
	H	0	000 0000	0000 0000 0000 0000 0000 0000	
	D	0	010 0010 0101	0000 0000 0000 0000 0000	[+0x10^{0}]
-0	B	1	000 0000 0	000 0000 0000 0000 0000 0000	
	H	1	000 0000	0000 0000 0000 0000 0000 0000	
	D	1	010 0010 0101	0000 0000 0000 0000 0000	[-0x10^{0}]
-15.0	B	1	100 0001 0	111 0000 0000 0000 0000 0000	
	H	1	100 0001	1111 0000 0000 0000 0000 0000	
	D	1	010 0010 0100	0000 0000 0000 1101 0000	[-150x10^{-1}]
20/7	B	0	100 0000 0	011 0110 1101 1011 0110 1110	
	H	0	100 0001	0010 1101 1011 0110 1101 1011	
	D	0	010 1001 1111	1101 0111 0100 1100 0011	[2857143x10^{-6}]
2^{-126}	B	0	000 0000 1	000 0000 0000 0000 0000 0000	
	H	0	010 0001	0100 0000 0000 0000 0000 0000	
	D	0	000 0111 1001	0011 1101 0110 0101 1010	[1175494x10^{-44}]
2^{-149}	B	0	000 0000 0	000 0000 0000 0000 0000 0001	
	H	0	001 1011	1000 0000 0000 0000 0000 0000	
	D	0	000 0111 0010	1000 0000 0101 0101 1110	[1401298x10^{-51}]
$2^{128} \times F$	B	0	111 1111 0	111 1111 1111 1111 1111 1111	
$F = 1 - 2^{-24}$	H	0	110 0000	1111 1111 1111 1111 1111 1111	
	D	0	100 1100 0101	1000 0000 1001 0010 1101	[3402823x10^{+32}]
2^{-260}	B			Nula (broj je isuviše mali)	
	H	0	000 0000	0001 0000 0000 0000 0000 0000	
	D	0	001 0101 0000	0111 1110 1111 0000 0101	[5397605x10^{-85}]
$2^{248} \times F$	B			Ne može da se predstavi	
$F = 1 - 2^{-24}$	H	0	111 1110	1111 1111 1111 1111 1111 1111	
	D	0	101 0010 1001	1010 1000 1100 1010 1000	[4523128x10^{+68}]

Oznake:

- S - znak broja (frakcije)
- B - Zapis sa binarnom osnovom (IEEE 754)
- D - Zapis sa dekadnom osnovom (IEEE 754, DPD kodiranje)
- H - Zapisa sa heksadekadnom osnovom (IBM z serija)
- BE ili C Uvećani eksponent B ili H broja, kombinacije kod D broja

Efektivna širina

- Mere za rezoluciju formata zapisa: ulp, relativno rastojanje
- Relativno rastojanje može da se konvertuje u oblik koji se naziva *efektivna širina* $W_\beta = log_\beta(x/ulp)$.
- $p-1 \leq W_\beta \leq p$

Dužina /tačnost	Osnova	W_2		W_{10}	
		Max	Min	Max	Min
32 bita (4 bajta) /jednostruka	H	24.00	20.00	7.22	6.02
	B	24.00	23.00	7.22	6.92
	D	23.25	19.93	7.00	6.00
64 bita (8 bajtova) /dvostruka	H	56.00	52.00	16.86	15.65
	B	53.00	52.00	15.95	15.65
	D	53.15	49.83	16.00	15.00
128 bita (16 bajtova) /četvorostruka	H	112.00	108.00	33.72	32.51
	B	113.00	112.00	34.02	33.72
	D	112.95	109.62	34.00	33.00

Tabela 14: Najveće i najmanje efektivne širine predstavljene u osnovama 2 i 10

Primeri

Primer1: Zapisati broj +123.4 u pokretnom zarezu pomoću dekadne osnove u jednostrukoj tačnosti, koristeći

1. dekadno (DPD) kodiranje
2. binarno (BID) kodiranje

Rešenje: Broj koga treba zapisati predstavimo u obliku $+1234*10^{-1}$. Vrednosti cifara u zapisu broja su:

1. DPD kodiranje
   (a) Broj koga treba zapisati je pozitivan $\longrightarrow$ znak broja je 0.
   (b) Frakcija broja je 1234. Dopunimo frakciju do maksimalnog broja cifara za jednostruku tačnost (7). Dobijeni broj koga treba da kodiramo je 0001234. Kodiraju se 6 cifara manje težine u dva dekleta, dok se cifra najveće težine kodira u polju za kombinaciju. Primenom DPD kodiranja dobija se (npr. na osnovu tabele shematskog prikaza kodiranja):

```
  0    0    1          2    3    4         Dekadni zapis
abcd efgh ijkm       abcd efgh ijkm
0000 0000 0001       0010 0011 0100        BCD zapis

000 000 0  001       010 011 0  100        DPD dekleti
pqr stu v  wxy       pqr stu v  wxy
```

   (c) Eksponent je jednak -1, i uz uvećanje 101 dobija se 100. Prevod dobijene vrednosti u binarni brojni sistem je 01100100. Odavde je nastavak eksponenta 100100, dok se prva dva bita 01 kodiraju u kombinaciji.
   (d) Kombinacija treba da kodira cifru 0 (cifra frakcije) i dva bita najveće težine eksponenta (binarne cifre 01). Kako je 0 cifra frakcije, na osnovu pravila za predstavljanje sledi da je kombinacija = 01000.

   Odavde, zapis broja +123.4 je 0 01000 100100 0000 0000 0101 0011 0100

2. BID kodiranje
   (a) Broj koga treba zapisati je pozitivan $\longrightarrow$ znak broja je 0.
   (b) Frakcija broja je 1234. Dopunimo frakciju do maksimalnog broja cifara za jednostruku tačnost (7). Dobijeni broj koga treba da kodiramo je 0001234. Prevod ovog broja u binarni sistem je 10011010010. Posto se u u jednostrukoj tacnosti frakcija zapisuje u bar 23 bita, traženi prevod se dobija dodavanjem nula sa leve strane i iznosi 00000000000010011010010.
   (c) Eksponent je jednak -1, i uz uvećanje 101 dobija se 100. Prevod dobijene vrednosti u binarni brojni sistem je 01100100.

(d) Kombinacija treba da kodira klasifikaciju, bitove eksponenta i pocetak frakcije. Kako prevod broja u binarni sistem pocinje nulom (tj. može da se zapiše u 23 bita bez menjanja vrednosti) to su prva dva bita kombinacije ili 0x ili 10. U tom slucaju se eksponent nalazi na početku kombinacije, a za njim slede 3 bita koja se dodaju na zapis frakcije. Posto su tri bita frakcije najvece težine 000, kombinacija ima oblik 01100100000.

Odavde, zapis broja +123.4 je 0 01100100000 00000000010011010010

Primer 2: Zapisati broj -243/13 u pokretnom zarezu pomoću dekadne osnove u jednostrukoj tačnosti, koristeći

1. dekadno (DPD) kodiranje
2. binarno (BID) kodiranje

Rešenje: Dobijeni količnik zapišimo pomoću 7 dekadnih cifara u obliku znak, frakcija i eksponent. Kako je 243/13=18.69230769... pretpostavimo da je odabrano takvo pravilo zaokruživanja da je rezultat zaokružen na 7 cifara jednak 18,69230 (primedba: i za zapis pomoću dekadne osnove važe ista pravila zaokruživanja!). Predstavljajući broj pomoću ciljanog eksponenta dobija se broj koji se zapisuje: -1869230*10^{-5}.

1. DPD kodiranje
   (a) Broj koji se zapisuje je negativan $\longrightarrow$ znak broja je 1.
   (b) Frakcija broja je 1869230. Kodiraju se 6 cifara manje težine u dva dekleta, dok se cifra najveće težine kodira u polju za kombinaciju. Primenom DPD kodiranja dobija se (npr. na osnovu tabele shematskog prikaza kodiranja):

```
  8    6    9          2    3    0        Dekadni zapis
abcd efgh ijkm       abcd efgh ijkm
1000 0110 1001       0010 0011 0000        BCD zapis

110 010 1  111       010 011 0  000        DPD dekleti
pqr stu v  wxy       pqr stu v  wxy
```

   (c) Eksponent je jednak -5, i uz uvećanje 101 dobija se 96. Prevod dobijene vrednosti u binarni brojni sistem je 01100000. Odavde je nastavak eksponenta 100000, dok se prva dva bita 01 kodiraju u kombinaciji.
   (d) Kombinacija treba da kodira cifru 1 (cifra frakcije) i dva bita najveće težine eksponenta (binarne cifre 01). Kako je 1 cifra frakcije, na osnovu pravila za predstavljanje sledi da je kombinacija = 01001.

Odavde, zapis broja -243/13 je 1 01001 100000 1100 1011 1101 0011 0000

2. BID kodiranje

   (a) Broj koji se zapisuje je negativan $\longrightarrow$ znak broja je 1.

   (b) Frakcija broja je 1869230. Njen zapis u binarnom sistemu je 11100100000101101011 10. Posto se u u jednostrukoj tacnosti frakcija zapisuje u bar 23 bita, traženi prevod se dobija dodavanjem nula sa leve strane i iznosi 00111001000001011010 1110.

   (c) Eksponent je jednak -5, i uz uvećanje 101 dobija se 96. Prevod dobijene vrednosti u binarni brojni sistem je 01100000.

   (d) Kombinacija treba da kodira klasifikaciju, bitove eksponenta i pocetak frakcije. Na osnovu prevoda frakcije (počinje nulom, može da se zapiše u 23 bita) zaključuje se da su prva dva bita pocinje nulom to su prva dva bita kombinacije ili 0x ili 10. U tom slucaju se eksponent nalazi na početku kombinacije, a za njim slede 3 bita koja se dodaju na zapis frakcije. Posto su tri bita frakcije najvece težine 001, kombinacija ima oblik 01100000000.

Odavde, zapis broja -243/13 je 1 01100000001 11001000010110101110

Primer 3: Zapisati broj +982.5294*10^{42} u pokretnom zarezu pomoću dekadne osnove u jednostrukoj tačnosti, koristeći

1. dekadno (DPD) kodiranje
2. binarno (BID) kodiranje

Rešenje: Dati broj zapišimo u obliku +9825294*10^{38}. Vrednosti cifara u zapisu broja su:

1. DPD kodiranje

   (a) Broj koga treba zapisati je pozitivan $\longrightarrow$ znak broja je 0.

   (b) Frakcija broja je 9825294. Šest cifara manje težine se kodiraju u dva dekleta, dok se cifra najveće težine kodira u polju za kombinaciju. Primenom DPD kodiranja dobija se (npr. na osnovu tabele shematskog prikaza kodiranja):

```
  8    2    5          2    9    4             Dekadni zapis
abcd efgh ijkm       abcd efgh ijkm
1000 0010 0101       0010 1001 0100             BCD zapis

100 010 1  101       010 101 1  010             DPD dekleti
pqr stu v  wxy       pqr stu v  wxy
```

(c) Eksponent je jednak +38, i uz uvećanje 101 dobija se 139. Prevod dobijene vrednosti u binarni brojni sistem je 10001011. Odavde je nastavak eksponenta 001011, dok se prva dva bita 10 kodiraju u kombinaciji.

(d) Kombinacija treba da kodira cifru 9 (cifra frakcije) i dva bita najveće težine eksponenta (binarne cifre 10). Kako je 9 cifra frakcije, na osnovu pravila za predstavljanje sledi da je kombinacija = 11101.

Odavde, zapis broja $+982.5294*10^{42}$ je
1 11101 001011 1000 1011 0101 0101 1010

2. BID kodiranje

(a) Broj koga treba zapisati je pozitivan $\longrightarrow$ znak broja je 0.

(b) Frakcija broja je 9825294. Prevod frakcije u binarni sistem je 100101011110110000001110. Prevod već ima 24 binarne cifre i ne treba da se dopunjava nulama sa leve strane.

(c) Eksponent je jednak +38, i uz uvećanje 101 dobija se 139. Prevod dobijene vrednosti u binarni brojni sistem je 10001011.

(d) Kako eksponent ima 24 bita i počinje sa 1001, prva dva bita kombinacije su 11. Pošto je u pitanju konačan broj, eksponent se nalazi u produžetku kombinacije, a na kraju sledi cifra 1 (jer je $8+G_{w+4} = (1001)_2$ odakle sledi da je $G_{w+4}=1$). Dobijena kombinacija je 11100010111. Kod za frakciju čini preostalih 20 bitova prevoda.

Odavde, zapis broja $+982.5294*10^{42}$ je
0 11100010111 01011110110000001110

Primer 4: Koji dekadni broj je predstavljen brojem u pokretnom zarezu zapisanim u IEEE 754 zapisu 00111101111000000000000000110101

1. pomoću dekadne osnove (DPD kodiranje)
2. pomoću binarne osnove (BID kodiranje)

Rešenje:

1. DPD kodiranje Razdvojimo cifre u zapisu broja da bi se odredile komponente zapisa.
   0|01111|011110|0000 0000 0000 0011 0101

   Odavde se dobija:

   (a) Cifra za znak je nula $\longrightarrow$ Broj je pozitivan

(b) Prve dve cifre u kombinaciji su 01 $\longrightarrow$ broj je konačan (nije beskonačno ili NaN)

- Pošto su prve dve cifre kombinacije 01, one su i cifre najveće težine uvećanog eksponenta. Dobijena vrednost uvećanog eksponenta je $(01011110)_2$, odnosno $(94)_{10}$. Odavde je vrednost eksponenta 94-101=-7.
- Pošto su prve dve cifre kombinacije 01, preostale cifre kombinacije (111) odredjuju 7 kao dekadnu cifru najveće težine frakcije.

(c) Dekodiranjem dekleta frakcije se dobija vrednost frakcije. Dekodiranje može da se izvede pomoću tabele koja daje šematski prikaz dekodiranja na sledeći način:

```
pqr stu v  wxy       pqr stu v  wxy
000 000 0  000       000 011 0  101         DPD dekleti

0000 0000 0000       0000 0011 0101        BCD zapis
abcd efgh ijkm       abcd efgh ijkm
  0    0    0         0    3    5          Dekadni zapis
```

Primedba: u ovom slučaju dekleti su mogli jednostavnije da se dekodiraju na osnovu treće osobine navedene kao prednost DPD kodiranja u odnosu na Chen-Ho kodiranje, odakle se direktno dobija da je dekodirana vrednost dekleta 000 035.

Frakcija koja se dobija je 7000035. Broj koji je zapisan ima vrednost $+7000035*10^{-7}$, odnosno $7.000035*10^{-1}$, tj. 0.7000035

2. BID kodiranje Razdvojimo cifre u zapisu broja da bi se odredile komponente zapisa.
0|01111011110|00000000000000110101

Odavde se dobija:

(a) Cifra za znak je nula $\longrightarrow$ Broj je pozitivan

(b) Prve dve cifre u kombinaciji su 01 $\longrightarrow$ broj je konačan (nije beskonačno ili NaN)

(c) Pošto su prve dve cifre kombinacije 01, uvećani eksponent se formira od prvih 8 cifara kombinacije, dok se preostale 3 cifre dodaju na početak frakcije. Dobijena vrednost uvećanog eksponenta je $(01111011)_2$, odnosno $(123)_{10}$. Odavde je vrednost eksponenta 123-101=22.

(d) Vrednost 11000000000000000110101 je zapis frakcije 6291509 u dekadnom sistemu.

Broj koji je zapisan ima vrednost $+6291509*10^{+22}$

Zadaci za vežbu

1. Zapisati sledeće brojeve u pokretnom zarezu pomoću dekadne osnove u jednostrukoj tačnosti, koristeći dekadno (DPD) kodiranje:

   (a) -245

   (b) +245.00

   (c) +345.678*10^{65}

   (d) 177.36/31

   (e) -778.3456264556*10^{-35}. Zadatak uraditi za različite načine zaokruživanja!

2. Koji dekadni brojevi su predstavljeni brojevima u pokretnom zarezu zapisanim u IEEE 754 zapisu pomoću dekadne osnove (DPD kodiranje):

   (a) 0 00000 000000 0000 0000 0000 0000 1100

   (b) 1 00010 000000 0000 0000 0000 0000 0000

   (c) 0 10001 101101 0000 0000 0000 0000 0000

   (d) 1 01001 011110 1100 0000 0010 0000 1111

   (e) 0 10101 110010 0001 0100 0000 0110 0101

   (f) 1 11011 101011 0000 0000 0000 0000 0000

   (g) 0 11111 011100 0001 0000 0000 0000 0001

   (h) 1 11110 101011 0000 0100 1000 0000 0000