Programiranje 1
IEEE prikaz brojeva — sažetak

Saša Singer

singer@math.hr

web.math.pmf.unizg.hr/~singer

PMF – Matematički odsjek, Zagreb

Sadržaj predavanja

- IEEE standard za prikaz "realnih" brojeva u računalu (tzv. "floating–point" standard):
  - IEEE standard — tip single (binary32),
  - IEEE standard — tip double (binary64),
  - IEEE standard — tip extended.

Prikaz "realnih" brojeva u računalu — IEEE standard

Stvarni prikaz realnih brojeva — IEEE 754

Stvarni prikaz realnih brojeva ima tri dijela i svaki od njih ima svoju duljinu — broj bitova predviđenih za prikaz tog dijela.

- predznak s — uvijek zauzima jedan bit, i to najviši;
- karakteristika k — zauzima sljedećih w bitova (w = engl. "width", širina pomaknutog eksponenta);
- signifikand m — zauzima sljedećih t bitova (t = engl. "trailing", završni ili razlomljeni dio od m).

Po starom standardu — ako se pamti vodeći (cjelobrojni) bit mantise, on je prvi (vodeći) u m, a duljina je $t + 1$.

Još se koristi i standardna oznaka

- preciznost $p := t + 1$ — to je ukupni broj vodećih značajnih bitova cijele mantise.

Stvarni prikaz realnih brojeva — IEEE 754

Karakteristika k se interpretira kao cijeli broj bez predznaka, tako da je $k \in \{0, \ldots, 2^w - 1\}$. "Rubne" vrijednosti za k označavaju tzv. posebna stanja:

- $k = 0$ — nula i denormalizirani brojevi,
- $k = 2^w - 1$ — beskonačno i nije broj.

Sve ostale vrijednosti $k \in \{1, \ldots, 2^w - 2\}$ koriste se za prikaz normaliziranih brojeva različitih od nule.

Veza između karakteristike k i stvarnog eksponenta e je:

$$k = e + bias, \quad bias = 2^{w-1} - 1.$$

Dakle, dozvoljeni eksponenti e moraju biti između

$$e_{\min} = -(2^{w-1} - 2) \quad \text{i} \quad e_{\max} = 2^{w-1} - 1.$$

Standardni tipovi realnih brojeva — IEEE 754

Novi standard IEEE 754-2008 standard ima sljedeće tipove za prikaz realnih brojeva:

`ime tipa`	`binary32`	`binary64`	`binary128`
duljina u bitovima	32	64	128
$t =$	23	52	112
$w =$	8	11	15
$u = 2^{-p}$	2^{-24}	2^{-53}	2^{-113}
$u \approx$	$5.96 \cdot 10^{-8}$	$1.11 \cdot 10^{-16}$	$9.63 \cdot 10^{-35}$
raspon brojeva $\approx$	$10^{\pm 38}$	$10^{\pm 308}$	$10^{\pm 4932}$

Broj u je tzv. jedinična greška zaokruživanja (v. malo kasnije).

Najveći tip `binary128` još uvijek ne postoji u većini procesora.

Standardni tipovi realnih brojeva — extended

Većina PC procesora još uvijek ima posebni dio — tzv. FPU (engl. Floating–Point Unit). On stvarno koristi

- tip `extended` iz starog standarda, koji odgovara tipu `extended binary64` u novom IEEE 754-2008 standardu.

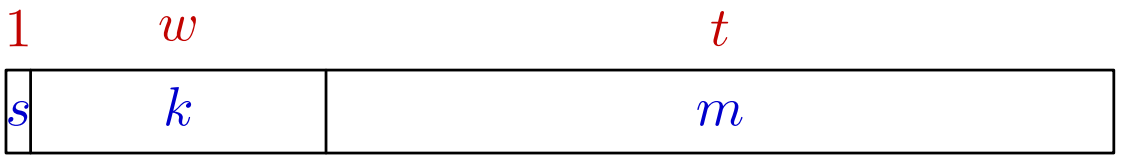
Dio primjera koje ćete vidjeti napravljen je baš u tom tipu!

`ime tipa`	`extended`
duljina u bitovima	80
$t+1=$	$63+1$
$w=$	15
$u=2^{-p}$	2^{-64}
$u\approx$	$5.42\cdot 10^{-20}$
raspon brojeva $\approx$	$10^{\pm 4932}$

Oznake

Oznake:

- Crveno — duljina odgovarajućeg polja u bitovima, bitove brojimo od 0, zdesna nalijevo (kao i obično),
- s — predznak: 0 za pozitivan broj, 1 za negativan broj,
- k — karakteristika,
- m — mantisa (signifikand).



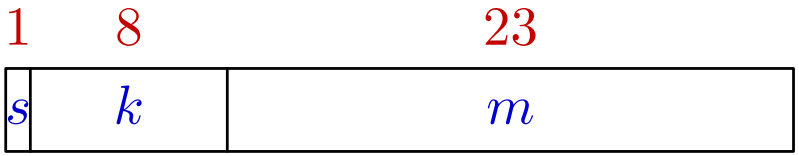
- Najznačajniji bit u odgovarajućem polju je najljeviji, a najmanje značajan bit je najdesniji.

Stvarni prikaz tipa single (binary32)

"Najkraći" realni tip je tzv. realni broj jednostruke točnosti. U C-u se taj tip zove `float`. Savjet: ne koristiti u praksi!

On ima sljedeća svojstva:

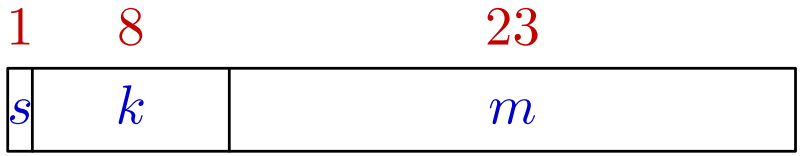
- duljina: 4 byte-a (32 bita), podijeljen u tri polja.



- u mantisi se ne pamti vodeća jedinica, ako je broj normaliziran,
- stvarni eksponent e broja, $e \in \{-126, \ldots, 127\}$,
- karakteristika $k = e + 127$, tako da je $k \in \{1, \ldots, 254\}$,
- karakteristike $k = 0$ i $k = 255$ koriste se za "posebna stanja".

Prikaz brojeva jednostruke točnosti — sažetak

IEEE tip single = `float` u C-u:



Vrijednost broja je

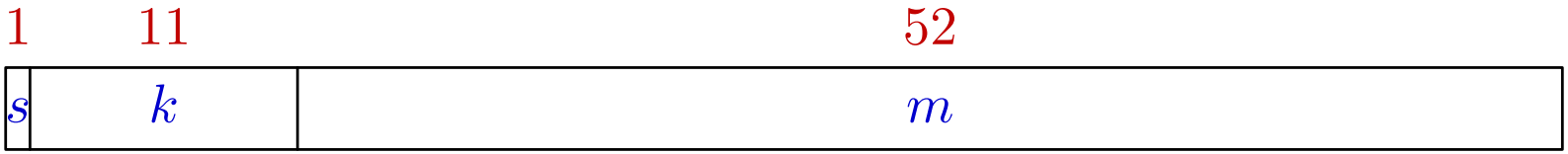
$$v = \begin{cases} (-1)^s * 2^{(k-127)} * (1.m) & \text{ako je } 0 < k < 255, \\ (-1)^s * 2^{(-126)} * (0.m) & \text{ako je } k = 0 \text{ i } m \neq 0, \\ (-1)^s * 0 & \text{ako je } k = 0 \text{ i } m = 0, \\ (-1)^s * \texttt{Inf} & \text{ako je } k = 255 \text{ i } m = 0, \\ \texttt{NaN} & \text{ako je } k = 255 \text{ i } m \neq 0. \end{cases}$$

Stvarni prikaz tipa double (binary64)

"Srednji" realni tip je tzv. realni broj dvostruke točnosti.
U C-u se taj tip zove `double`. Savjet: njega treba koristiti!

On ima sljedeća svojstva:

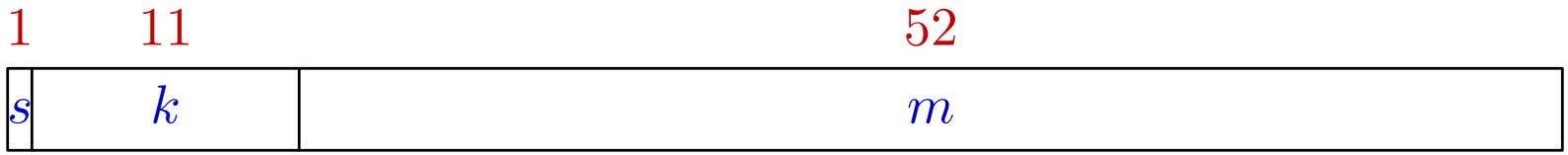
- Duljina: 8 byte-a (64 bita), podijeljen u tri polja.



- u mantisi se ne pamti vodeća jedinica, ako je broj normaliziran,
- stvarni eksponent e broja, $e \in \{-1022, \ldots, 1023\}$,
- karakteristika $k = e + 1023$, tako da je $k \in \{1, \ldots, 2046\}$,
- karakteristike $k = 0$ i $k = 2047$ — "posebna stanja".

Prikaz brojeva dvostruke točnosti — sažetak

IEEE tip double = `double` u C-u:



Vrijednost broja je

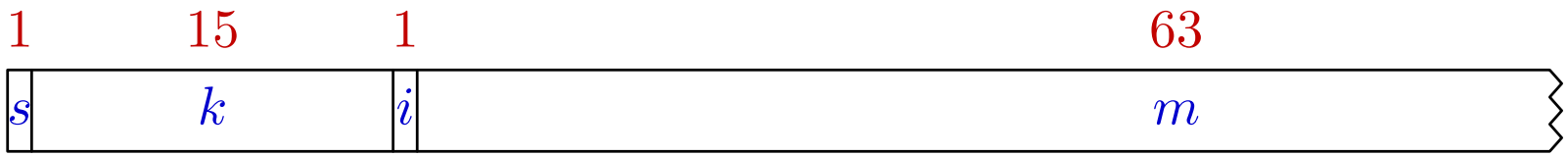
$$v = \begin{cases} (-1)^s * 2^{(k-1023)} * (1.m) & \text{ako je } 0 < k < 2047, \\ (-1)^s * 2^{(-1022)} * (0.m) & \text{ako je } k = 0 \text{ i } m \neq 0, \\ (-1)^s * 0 & \text{ako je } k = 0 \text{ i } m = 0, \\ (-1)^s * \texttt{Inf} & \text{ako je } k = 2047 \text{ i } m = 0, \\ \texttt{NaN} & \text{ako je } k = 2047 \text{ i } m \neq 0. \end{cases}$$

Tip extended

Stvarno računanje (na IA–32) se obično radi u "proširenoj" točnosti. U C-u je taj tip možda dohvatljiv kao `long double`.

On ima sljedeća svojstva:

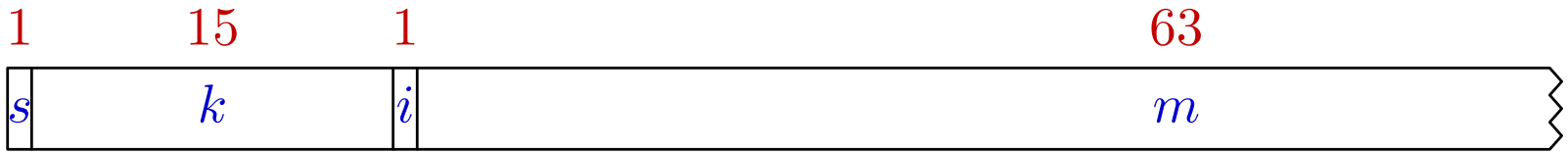
- Duljina: 10 byte-a (80 bita), podijeljen u četiri polja.



- u mantisi se pamti vodeći bit i mantise,
- stvarni eksponent e broja, $e \in \{-16382, \ldots, 16383\}$,
- karakteristika $k = e + 16383$, tako da je $k \in \{1, \ldots, 32766\}$,
- karakteristike $k = 0$ i $k = 32767$ — "posebna stanja".

Prikaz brojeva proširene točnosti — sažetak

IEEE tip extended:



Vrijednost broja je

$$v = \begin{cases} (-1)^s * 2^{(k-16383)} * (i.m) & \text{ako je } 0 \leq k < 32767, \\ (-1)^s * \texttt{Inf} & \text{ako je } k = 32767 \text{ i } m = 0, \\ \texttt{NaN} & \text{ako je } k = 32767 \text{ i } m \neq 0. \end{cases}$$

Uočite da prva mogućnost uključuje:

- $+0$, -0 i denormalizirane brojeve (za $k = 0$),

jer se pamti vodeći "cjelobrojni" bit i mantise.