Gleitpunktzahlen dual nach IEEE-Standard

Mathematik mit MuPAD 4, (2011) Prof. Dr. Dörte Haftendorn 15.10.02 Version vom 15.10.02

Siehe auch Untersuchungen zur Maschinengenauigkeit und Zahldarstellung in Excel
Bin-Hex-Artithmetik.mnb MuPAD-Notebook

---

1 Bit Vorzeichen=S, 11 Bit Exponent=expo, 52 Bit Mantisse=F ohne führende1

0 100 0110 0010 001011...(52 Stellen).....0111000

Ein Wert ist dann für 0<expo<2047 gegeben durch
(-1)^S*2^(expo-1023)*1.F, der letzte Faktor in dualer Schreibweise,
d.h. dezimal, wenn man mit manti die dezimale Angabe von F bezeichnet
wert:=(-1)^S*2^(expo-1023)*(1+ manti/2^52)

Erstmal betrachten wir nur positive Werte, vorserstes Bit=0:

```
wert:=2^(expo-1023)*(1+ manti/2^52)
```

$$2^{\mathrm{expo}-1023}\cdot\left(\frac{\mathrm{manti}}{4503599627370496}+1\right)$$

will man die sofortige Auswertung der Zweierpotenzen vermeiden,
muss man **hold(...)** verwenden.

Sonderfälle:(jeweils für beide S)

expo=2047, F> 0: wert:= NaN, Not a Number

0 111 1111 1111 001011...(52 Stellen).....0111000

expo=2047, F=0: wert:=infinity, unendlich

0 111 1111 1111 000000...(52 Stellen).....0000000

1 111 1111 1111 000000...(52 Stellen).....0000000

expo=0, F=0: wert:=0,

0 000 0000 0000 000000...(52 Stellen).....0000000

1 000 0000 0000 000000...(52 Stellen).....0000000

expo=0, F>0 : wertun:=(-1)^S*2^(1-1023)*manti/2^52 (unormalisiert)

0 000 0000 0000 001100...(52 Stellen).....1011100 z.B.

Dieses ist festgelegt vom PSC (Pittsburgh Supercomputing Center)
http://www.psc.edu/general/software/packages/ieee.html
IEEE heißt Institut of Electrical and Electronic Engeneers
Dieser Standard gilt für unsere üblichen PC und viele andere,
nur wenige Großrechner haben andere Formate.
Man nennt diese Darstellung **double precision,**
64 Bit=8Byte-Darstellung
single precision arbeitet mit 1 Bit Vorz. 8 Bit Exponent, 23 Bit Mantisse
also einer 32-Bit(=4 Byte)-Darstellung
Eirweiterungen nach demselben Konzept sind gut denkbar.


```
DIGITS := 25:S=0:    k:=1023:      m:=2^52:
```

Normalisierte Werte

```
wert:=2^(expo-hold(k))*(1+manti/hold(m))
```

```
wert:=2^(expo-hold(k))*(1+manti/hold(m))
```

$$2^{\mathrm{expo}-k}\cdot\left(\frac{\mathrm{manti}}{m}+1\right)$$

Unnormalisierte Werte

```
wertun:=2^(1-hold(k))*(manti/hold(m))
```

$$\frac{2^{1-k}\cdot\mathrm{manti}}{m}$$

Einige Beispiele

Mit subs(term, x=3) wertet man einen Term für x=3 aus.

```
subs(wert,expo=1026,manti=0)
```

$$8$$

0 000 0000 0010 00000...(52 Stellen).....0000

```
subs(wert,expo=1023,manti=0)
```

$$1$$

```
subs(wert,expo=1022,manti=0)
```

$$\frac{1}{2}$$

```
int2text(1022,2)
```

"1111111110"

0 011 1111 1110 00000...(52 Stellen).....0000

```
subs(wert,expo=1022,manti=2^51)
```

$$\frac{3}{4}$$

0 011 1111 1110 10000...(52 Stellen).....0000

Die **kleinste Zahl** in diesem Gleitpunktsystem darstellbare Zahl:

0 000 0000 0000 00000...(52 Stellen).....0001

```
subs(wertun,manti=1)
```

$$\frac{1}{20240225330731061835249534671891730704955664976414211835690135802743033956799534689196038370143}$$

```
float(%)
```

$$4.9406564584124654417656879\cdot 10^{-324}$$

Kleinste Zahl der normalisierten Werte, das sind die mit Mantisse 1.F

0 000 0000 0001 00000...(52 Stellen).....0000

```
subs(wert,manti=0,expo=1)
```

$$\frac{1}{44942328371557897693232629769725618340449424473557664318357520289433168951375240783177119330601}$$

```
float(%)
```

$$2.2250738585072013830902327\cdot 10^{-308}$$

Das ist genau die in Excel beobachtete kleinste Zahl.
Daran sieht man, dass Excel die Erweiterung um die unnormalisierten Werte nicht verwendet.

Die **größte Zahl**, die noch nicht "unendlich" ist: 2

0 111 1111 1110 111111...(52 Stellen).....111111

```
subs(wert,manti=2^52-1,expo=2^11-2)
```

$$1797693134862315708145274237317043567980705675258449965989174768031572607800285387605895586327668$$

```
179769313486231570814527423731704356798070567525844996598917476803157260780028538760589558632766
```

```
float(%)
```

$$1.79769313486231570814527 4 \cdot 10^{308}$$

```
DIGITS:=20:
```

Maschinengenauigkeit eps=kleinste Zahl, deren Addition zu 1 von der Maschine gemerkt wird.

```
(1+float(subs(wert,manti=0,expo=928))-1)
```

$$2.52435489670723777 73 \cdot 10^{-29}$$

```
(1+float(subs(wert,manti=2^52-1,expo=927))-1)
```

$$0.0$$

```
Digits=30:
(1+float(subs(wert,manti=2^52-1,expo=927))-1)
```

$$0.0$$

Damit ist klar, dass in MuPAD die Maschinengenauigkeit ist:

```
2.0^(928-1023),2.0^-95
```

$$2.52435489670723777 73 \cdot 10^{-29}, 2.52435489670723777 73 \cdot 10^{-29}$$

```
eps:=(1+2.0^-95)-1
```

$$2.52435489670723777 73 \cdot 10^{-29}$$

```
(1+2.0^-96)-1
```

$$0.0$$

Also ist in MuPAD die Maschinengenauigkeit eps=2^(-95)
In Excel ist sie 2^(-49)