SS 2009
Fakultät für Technik, Studiengänge EIT/TI
Dipl.-Ing.(FH) Felix Becker

Infoblatt

Floating point DSP

Die Signalprozessoren der TI-Familie TMS320C671x sind Gleitkomma-DSPs (Floating-Point Digital Signal Processor) mit 32 Bit Wortbreite.
Ein Gleitkomma-DSP kann natürlich auch Festkommazahlen (Fix-Point) verarbeiten.
Wenn die zu bewältigende Aufgabe keine Gleitkomma-Arithmetik erfordert, kann man durchaus auch nur mit Festkommazahlen arbeiten. Der Verzicht auf Gleitkomma-Arithmetik kann sogar zu einem Ausführungszeitvorteil führen.
Wenn Sie den DSP in Assembler programmieren würden, dann müssten Sie für die Gleitkommaverarbeitung andere Befehle verwenden als für die Festkommaverarbeitung. In C merken Sie davon wenig. Sie müssen die Variablen eben nur entsprechend deklarieren und eventuell durch Typkonvertierungen in das andere Zahlenformat bringen. Den Rest (also die Umsetzung in entsprechenden Assembler-Code) erledigt der C-Compiler für Sie.

Merke: Die Gleitkommaverarbeitung kann ein Gleitkomma-DSP also zusätzlich zur Festkommaverarbeitung.

Zahlendarstellung

Gleitkomma 32 Bit (IEEE-Standard)

Bit 32 ... 17 16 ... 1

```
sccc cccc cfff ffff ffff ffff ffff ffff
```

mit	s	Vorzeichenbit = 1 für negative Gleitkommazahlen
	c	8 Bit Charakteristik
	f	23 Bit Mantisse

$-1^{\mathbf{s}} \cdot 2^{(\mathbf{c}-127)} \cdot 1{,}\mathbf{f}$ mit $0 < c < 255$
wobei c = 0 und c = 255 eine Sonderstellung einnehmen!
(Codierung von 0, Infinities ($\pm\infty$) und NaNs (Not-a-Number))

Mit 1 < c < 254 ergibt sich also ein minimaler Exponent von -126 und ein maximaler Exponent von +127.

Dezimaler Zahlenbereich:

$$-2^{127} \cdot (1 + 2^{-1} + 2^{-2} + 2^{-3} \ldots + 2^{-23}) \ \ldots\ +2^{127} \cdot (1 + 2^{-1} + 2^{-2} + 2^{-3} \ldots + 2^{-23}) =$$
$$-2^{127} \cdot (1 + ((2^{23} - 1) / 2^{23})) \ \ldots\ +2^{127} \cdot (1 + ((2^{23} - 1) / 2^{23})) =$$
$$-2^{127} \cdot 1{,}999999880791\ldots \ \ldots\ +2^{127} \cdot 1{,}999999880791\ldots =$$
$$-3{,}40282346637\ldots E38 \ \ldots\ +3{,}40282346637\ldots E38$$

=> größte darstellbare Gleitkommazahl: $\pm 2^{127} \cdot 1{,}999999880791 = \pm 3{,}40282346637E38$
kleinste darstellbare Gleitkommazahl: $\pm 2^{-126} = \pm 1{,}17549435082E\text{-}38$

positiver Zahlenbereich: +1,17549435082E-38 ... +3,40282346637E38
negativer Zahlenbereich: -1,17549435082E-38 ... -3,40282346637E38

Auflösung für eine Gleitkommazahl FPN im Bereich $1 \leq FPN < 2$ (d.h. c = 127 => $2^{0} = 1$):

$$2^{-23} = 0{,}000000119209289550781...$$

D.h., das 32 Bit Gleitkommaformat bietet in diesem Bereich eine Genauigkeit von mindestens sechs dezimalen Nachkommastellen.

Je größer allerdings die darzustellende Zahl wird, desto schlechter wird die Auflösung:

- für eine Gleitkommazahl FPN im Bereich 4 194 304 ≤ FPN < 8 388 608 nur noch 0,5
- für eine Gleitkommazahl FPN im Bereich 8 388 608 ≤ FPN < 16 777 216 nur noch 1
- für eine Gleitkommazahl FPN im Bereich 16 777 216 ≤ FPN < 33 554 432 nur noch 2
- für eine Gleitkommazahl FPN im Bereich 33 554 432 ≤ FPN < 67 108 864 nur noch 4
  usw.
  (vgl. 32 Bit Festkommaformat, das bis 2 147 483 647 eine Auflösung von 1 bietet, die Auflösung des Gleitkommaformates ist ab 1 073 741 824 bereits auf 128 abgesunken!)

Einige Beispiele:

		Bit 32 23 17 16 1				
+ Inf. (per Def.)	=	0111 1111 1000 0000 0000 0000 0000 0000	bin	=	0x7F80 0000	hex
+ 3,4028...E38	=	0111 1111 0111 1111 1111 1111 1111 1111	bin	=	0x7F7F FFFF	hex
+ 2	=	0100 0000 0000 0000 0000 0000 0000 0000	bin	=	0x4000 0000	hex
+ 1	=	0011 1111 1000 0000 0000 0000 0000 0000	bin	=	0x3F80 0000	hex
+ 1,175...E-38	=	0000 0000 1000 0000 0000 0000 0000 0000	bin	=	0x0080 0000	hex
+ 0	=	0000 0000 0000 0000 0000 0000 0000 0000	bin	=	0x0000 0000	hex
- 0	=	1000 0000 0000 0000 0000 0000 0000 0000	bin	=	0x8000 0000	hex
- 1,175...E-38	=	1000 0000 1000 0000 0000 0000 0000 0000	bin	=	0x8080 0000	hex
- 1	=	1011 1111 1000 0000 0000 0000 0000 0000	bin	=	0xBF80 0000	hex
- 3,4028...E38	=	1111 1111 0111 1111 1111 1111 1111 1111	bin	=	0xFF7F FFFF	hex
- Inf. (per Def.)	=	1111 1111 1000 0000 0000 0000 0000 0000	bin	=	0xFF80 0000	hex

Festkomma 32 Bit K2-Format

Dezimaler Zahlenbereich: $-2^{31} \ldots 2^{31} - 1$ = - 2 147 483 648 ... + 2 147 483 647

		Bit 32 17 16 1				
+ 2147483647	=	0111 1111 1111 1111 1111 1111 1111 1111	bin	=	0x7FFF FFFF	hex
+ 2147483646	=	0111 1111 1111 1111 1111 1111 1111 1110	bin	=	0x7FFF FFFE	hex
		:				
+ 2	=	0000 0000 0000 0000 0000 0000 0000 0010	bin	=	0x0000 0002	hex
+ 1	=	0000 0000 0000 0000 0000 0000 0000 0001	bin	=	0x0000 0001	hex
0	=	0000 0000 0000 0000 0000 0000 0000 0000	bin	=	0x0000 0000	hex
- 1	=	1111 1111 1111 1111 1111 1111 1111 1111	bin	=	0xFFFF FFFF	hex
- 2	=	1111 1111 1111 1111 1111 1111 1111 1110	bin	=	0xFFFF FFFE	hex
		:				
- 2147483647	=	1000 0000 0000 0000 0000 0000 0000 0001	bin	=	0x8000 0001	hex
- 2147483648	=	1000 0000 0000 0000 0000 0000 0000 0000	bin	=	0x8000 0000	hex