Wie genau ist Ihr Computer?

Gleitkommazahlen lassen nur eine begrenzte Genauigkeit bei reellen Zahlen zu. Diese Eigenschaft kann dazu benutzt werden, um die Form der Real-Repräsentation und der Real-Arithmetik zu ermitteln.

Zahlen können im Computer auf zwei prinzipiell verschiedene Arten dargestellt werden. Die einfachste Form der Zahlendarstellung sind die Festkommazahlen, bei denen die Stelle des Trennzeichens (des Kommas) hardware- oder softwaremäßig fixiert ist. Gleitkommazahlen setzen sich aus zwei Festkommazahlen zusammen, bei denen die eine die Mantisse und die andere den Exponenten einer prinzipiell frei wählbaren Basis darstellt. Typische Festkommazahlen sind die Integer-Zahlen (ganze Zahlen), bei denen das Dezimalkomma hinter der letzten Stelle angeordnet ist, aber auch Festkommazahlen mit einem vor der ersten Ziffer angeordneten Dezimalkomma sind nicht unüblich ($|z| < 1$).

Gleitkommazahlen werden in einer halblogarithmischen Form

$$x = m\, b^e \qquad (1)$$

dargestellt. Das Vorzeichen wird im allgemeinen separat gespeichert und die Null wird in einer besonderen Form abgelegt, so daß die folgenden Betrachtungen auf die Erläuterung positiver Gleitkommazahlen beschränkt bleiben können. Bei einer vorgegebenen ganzzahligen Basis $b > 1$ kann jede reelle Zahl x in der Form von Gleichung (1) dargestellt werden. Allerdings wird diese Form erst mit der Einführung von Normalisierungsbedingungen, zum Beispiel

$$1 < m \leq 1/b \qquad (2)$$

und der Forderung nach einem ganzzahligen Exponenten e eindeutig. Die Bedingung (2) bedeutet, daß die erste Stelle von m nicht Null sein darf. Aus diesem Grunde erfordert die Darstellung der Zahl Null eine Sonderbehandlung.

Im Speicher des Computers wird der Exponent »e« und eine feste Stellenzahl »l« der im Zahlensystem mit der Basis »b« geschriebenen Mantisse »m« abgelegt. Liegt nun keine ausführliche Dokumentation über den Computer und die Programmiersprache vor, so weiß man über diese Größen nichts. Die folgenden Betrachtungen gestatten es, diese interessanten Parameter zu »erfragen«. Dazu wird der Vorgang der Addition zweier reeller Zahlen x_1 und x_2

$$x_1 = m_1\, b^{e_1},\ x_2 = m_2{}^*\, b^{e_2} \text{ mit } e_1 > e_2 \qquad (3)$$

betrachtet. Zur Durchführung dieser Operation müssen beide Zahlen mit demselben Exponenten dargestellt werden. Hier wird also

$$x_2 = (m_2\, b^{e_2 - e_1})\, b^{e_1} = m_2\, b^{e_1} \qquad (4)$$

gebildet. Diese Darstellung ist nicht mehr normalisiert, vielmehr sind die ersten $e_1 - e_2$ Stellen der Mantisse m_2* Nullen. Dann erfolgt die Addition der Mantissen. Abschließend wird eventuell eine Exponentenkorrektur zur Gewährleistung der Normalisierungsbedingung nach Gleichung (2) vorgenommen. Da die zulässige Mantissenlänge nur »l« ist, gehen rechts $e_1 - e_2$ Stellen verloren. Dabei kann entweder gerundet oder abgeschnitten werden.

Ist

$$e_1 - e_2 \geq 1 \qquad (5)$$

dann bewirkt sich die Addition der zweiten Zahl nicht mehr auf das Ergebnis aus, das heißt, die Mantisse von x_2 geht vollständig verloren, und es gilt:

$$x_1 + x_2 := x_1 \qquad (6)$$

In dem Sub-Programm des Listings (Zeilen 530 bis 800), das eine Adaption von [1] ist, wird die Gleitkommazahl »Zahl« solange mit 2 multipliziert, bis die Addition von 1 nicht mehr signifikant ist. Die Schleife bricht also bei einer Zahl, die größer als b^l ist, ab. Der Exponent e_1 hat dann den Wert l + 1 (Zeilen 560 bis 600).

Die nächstgrößere darstellbare Zahl entsteht dann, wenn die letzte Stelle der Mantisse um 1 erhöht wird. Das heißt, die Mantisse wird um b^{-l} erhöht, und damit ist die nächstgrößere darstellbare Real-Zahl genau um

$$b^{-l}\, b^{l+1} = b \qquad (7)$$

also um die Basis der Zahlendarstellung größer als die Ausgangszahl (Zeilen 620 bis 660).

Da die Abarbeitung in einer Anweisung streng von links nach rechts erfolgt, liefert die Zeile 680 die erste wichtige interessierende Größe, nämlich die Basis der Zahlendarstellung »b« nach Gleichung (1). In Zeile 710 wird geprüft, ob bei der Real-Arithmetik gerundet oder abgeschnitten wird. Dazu wird die um 1 verminderte Basis zu der Zahl, die größer als b^l ist, addiert und geprüft, ob eine signifikante Addition erfolgt oder nicht erfolgt.

Im letzten Teil des Sub-Programms wird die Mantissenlänge berechnet. Dazu werden in Zeile 780 solange immer höhere Potenzen der Basis »b« gebildet, bis sich 1 nicht mehr signifikant addieren läßt. Die Zahl hat dann die Größe »b«, und die Mantissenlänge kann zur weite-

a) Ergebnisse beim HP 9816 (Serie 200)		
Basis der REAL-Variablen	=	2
Mantissenlänge der REAL-Variablen	=	53
bei der Rechnung wird abgeschnitten		
kleinste darstellbare REAL-Variable	=	2.22507385850720E-308
größte darstellbare REAL-Variable	=	1.79769313486232E+308
größte Präzision der REAL-Darstellung	=	1.11022302462516E-016
kleinste Präzision der REAL-Darstellung	=	2.22044604925031E-016
b) Ergebnisse beim HP 9845 B		
Basis der REAL-Variablen	=	10
Mantissenlänge der REAL-Variablen	=	12
bei der Rechnung wird gerundet		
kleinste darstellbare REAL-Zahl	=	1.00000000000E-99
größte darstellbare REAL-Zahl	=	9.99999999999E+99
größte Präzision der REAL-Darstellung	=	5.00000000000E-13
kleinste Präzision der REAL-Darstellung	=	5.00000000000E-12
Basis der SHORT-Variablen	=	8
Mantissenlänge der SHORT-Variablen	=	7
bei der Rechnung wird gerundet		

Gleitkommazahlen und Real-Arithmetik mit zwei Computern

ren Verarbeitung an das Hauptprogramm übergeben werden.

Aus den nun ermittelten Grunddaten der Gleitkommazahlendarstellung können weitere interessante Parameter, insbesondere die größte und die kleinste darstellbare Real-Variable und die Genauigkeit der Darstellung, abgeleitet werden. Das Listing zeigt eine Variante zur Berechnung dieser Größen für den HP 9816 (Hewlett-Packard, Serie 200). Die hier implementierten Fehlerbehandlungsbefehle gestatten eine relativ einfache Programmgestaltung.

In den Zeilen 200 bis 270 wird die kleinste darstellbare Zahl durch kontinuierliche Division, ausgehend von der kleinsten Mantisse (nur die erste Stelle von »m« ist 1), gebildet. Als Besonderheit des vorliegenden Interpreters wird hier ein Unterlauf als Fehler behandelt, so daß die LOOP-Schleife in den Zeilen 220 bis 240 keinen Ausgang benötigt. Bei den meisten Computern kommt man mit einer Konstruktion wie:

```
Zahl_min=1
REPEAT
    Zahl_min=Zahl_min/Basis
UNTIL Zahl_min/Basis=0
```

aus. Aber auch diese Schleife ist vorsichtig zu betrachten, weil unter Umständen die Prüfung Zahl_min/Basis=0 durch interne längere Zahlendarstellungen nicht zum rechtzeitigen Abbruch führt (so beim HP 9845 B). Hier hilft eine Zwischenspeicherung weiter:

```
Zahl_min=Zahl_folgende=1
REPEAT
  Zahl_min=Zahl_folgend
  Zahl_folgende=Zahl_min/Basis
UNTIL Zahl_folgende=0
```

Zur Berechnung der größten darstellbaren Zahl muß zunächst die größte Mantisse bereitgestellt werden. Diese ist vorhanden, wenn alle »1« Stellen der Mantisse mit der größten Zahl des Zahlensystems, also mit (b—1), belegt sind. In den Zeilen 290 bis 340 erfolgt diese Belegung. Dann wird diese Zahl so oft mit der Basis »b« multipliziert, bis in Zeile 320 ein Überlauf erfolgt. Dieser bewirkt dann einen Sprung zu der Marke Überlauf auf Zeile 390.

Sind in dem Computer keine Fehlerbehandlungsbefehle vorgesehen, so kann die größte darstellbare Zahl auch anders ermittelt werden. Eine Möglichkeit besteht darin, mit

$$e_{max} = -\log_b x_{min} = -\frac{\ln x_{min}}{\ln b} \quad (8)$$

den maximalen Betrag des Exponenten von Gleichung (1) zu ermitteln und zu vermuten, daß die größte darstellbare Zahl $\leq b^{e_{max}}$ ist. Eine Berechnung von $b^{e_{max}}$ führt im allgemeinen zu einem Überlauf. Auch ist die größte darstellbare Zahl oft $\leq b^{e_{max}-1}$, wenn für die Festkommadarstellung des Exponenten bei negativen Zahlen das Komplement verwendet wird und deshalb der größte positive Exponent um 1 kleiner ist als der Betrag des kleinsten negativen Exponenten. Zu einem Ergebnis kommt man nun, indem man die größte Mantisse entsprechend den Zeilen 290 bis 340 errechnet und diesen Wert dann e_{max} beziehungsweise (e_{max-1})mal mit der Basis multipliziert. Hier wird man immer etwas probieren müssen, und ein paar Überläufe treten sicher auf.

Die Tabelle auf Seite 84 zeigt die erzielten Ergebnisse bei zwei Computern von HP. Dabei ist zu beachten, daß der HP 9845 B zwei Gleitkomma-Zahlendarstellungen (Real und Short) hat. Das angegebene Programm läßt sich aber ohne große Schwierigkeiten auch auf andere Computer und Programmiersprachen übertragen und gestattet dann einen Einblick in das »Innere« des Computers und die ihm innewohnende Real-Arithmetik.

(J. Schwarz/rs)

Literatur:
[1] M.A. Malcolm: »Algorithms To Reveal Properties of Floating-Point Arithmetic«. Comm. ACM Vol. 15(1972) No. 11, pp. 949-951
[2] A.H.J. Sale: »Determine Real Number Environment«. PASCAL News No. 13, Dec. 1978, p. 33.

```
10    ! Programm zur Bestimmung der REAL-Repräsentation im HP 9816
20    !
30    ! (c)    1985 by Jürgen Schwarz        Version / Datum:     1.0 / 15.07.85
40    ! File-Name:              REAL         Speichermedium: Disketten Sz01/Sz02
50    ! Computer: HP 9816 (Serie 200)        Sprache:        HP-BASIC Version 3.0
60    !
70    REAL Zahl_min,Zahl_max,Epsilon,X
80    INTEGER Basis,Stellen,Rundung,I
90    !
100 Y:IMAGE 45A,17X,DDDD
110 Z:IMAGE 45A,Z.14DESZZZ
120   !
130   CALL Frage(Basis,Stellen,Rundung)
140   PRINT USING Y;"Basis                der REAL-Variablen      = ";Basis
150   PRINT USING Y;"Mantissenlänge   der REAL-Variablen      = ";Stellen
160   IF Rundung THEN PRINT "bei der Rechnung wird gerundet"
170   IF NOT Rundung THEN PRINT "bei der Rechnung wird abgeschnitten"
180   PRINT
190   !
200   ON ERROR GOTO Unterlauf
210   Zahl_min=1
220   LOOP
230     Zahl_min=Zahl_min/Basis
240   END LOOP
250 Unterlauf: !
260   OFF ERROR
270   PRINT USING Z;"kleinste darstellbare REAL-Variable      = ";Zahl_min
280   !
290   X=1
300   Zahl_max=0
310   FOR I=1 TO Stellen
320     X=X*Basis
330     Zahl_max=Zahl_max+(Basis-1)/X
340   NEXT I
350   ON ERROR GOTO Ueberlauf
360   LOOP
370     Zahl_max=Zahl_max*Basis
380   END LOOP
390 Ueberlauf: !
400   OFF ERROR
410   PRINT USING Z;"größte   darstellbare REAL-Variable      = ";Zahl_max
420   PRINT
430   !
440   Epsilon=1
450   FOR I=1 TO Stellen
460     Epsilon=Epsilon/Basis
470   NEXT I
480   IF Rundung THEN Epsilon=.5*Epsilon
490   PRINT USING Z;"größte   Präzision der REAL-Darstellung = ",Epsilon
500   PRINT USING Z;"kleinste Präzision der REAL-Darstellung = ",Epsilon*Basis
510   END
520   !
530   SUB Frage(INTEGER Basis,Mantissenlaenge,Rundung)
540     REAL Zahl,Zuwachs
550     !
560     !          Berechnung des kleinsten Wertes 2^n, zu dem sich 1 nicht mehr
570     Zahl=2                                          ! signifikant addieren läßt
580     WHILE Zahl+1-Zahl=1
590       Zahl=Zahl*2
600     END WHILE
610     !
620     ! Berechnung des nächstgrößeren REAL-Wertes
630     Zuwachs=2
640     WHILE Zahl+Zuwachs=Zahl
650       Zuwachs=Zuwachs*2
660     END WHILE
670     !
680     Basis=Zahl+Zuwachs-Zahl                 ! Basis der REAL-Zahlendarstellung
690     !
700     ! Prüfen, ob gerundet wird, oder nicht
710     Rundung=(Zahl+(Basis-1)>Zahl)
720     !
730     ! Berechnung der Mantissenlänge
740     Mantissenlaenge=0
750     Zahl=1
760     WHILE Zahl+1-Zahl=1
770       Mantissenlaenge=Mantissenlaenge+1
780       Zahl=Zahl*Basis
790     END WHILE
800   SUBEND
```

Programm zur Bestimmung der Real-Repräsentation