Rechnerarithmetik

Vorlesung im Sommersemester 2008

Eberhard Zehendner

FSU Jena

Thema: Fest- und Gleitkommasysteme

Gleitkommazahlen: Allgemeiner Zahlenbereich

Allgemeiner Zahlenbereich für Gleitkommazahlen (halblogarithmische Darstellung)

$\{s \times R^e | R \in \mathbb{N}, R > 1, (s, e) \in W \subset \mathbb{Z} \times \mathbb{Z}\}$

s heißt *Signifikant*, e *Exponent*, R *Basis*.

Spezieller: reguläre Kombination von Signifikanten und Mantissen

$\{s \times R^e | R \in \mathbb{N}, R > 1, s \in S \subset \mathbb{Z}, e \in E \subset \mathbb{Z}\}$

Zusätzlich: Intervallbereiche für Signifikanten und Mantissen

$\{s \times R^e | R \in \mathbb{N}, R > 1, s \in \mathbb{Z}, s = 0 \vee s^+_{\min} \leq s \leq s^+_{\max} \vee s^-_{\min} \leq -s \leq s^-_{\max}, e \in \mathbb{Z}, e_{\min} \leq e \leq e_{\max}\}$

Symmetrische Zahlenbereiche: $s^+_{\min} = s^-_{\min}, s^+_{\max} = s^-_{\max}$

Historisch: auch (leicht) unsymmetrische Zahlenbereiche verwendet

Basis R bestimmt den dynamischen Bereich, ist fest, braucht also nicht gespeichert zu werden

R meist 2, ergänzend auch 10 (iAPX87),

seltener 8 (Manchester University Atlas, 1962; Burroughs B5500, 1964)

oder 16 (IBM System/360-370, 1964/1970; Manchester University MU5, 1972; HEP, 1982),

andere Werte nur in Ausnahmefällen, z. B. 256 (MANIAC II, Los Alamos, 1956)

Typische Gleitkommasysteme mit symmetrischem Zahlenbereich

Heutzutage dominierende Gleitkommasysteme (IEEE-754, JAVA, ...):

System normalisierter Gleitkommazahlen $\quad \mathrm{Float}(R, l, e_1, e_2) =$

$$\{0\} \cup \{v \times m \times R^e | v \in \{-1, 1\}, m \in [R^{l-1}, R^l - 1] \cap \mathbb{N}, e \in [e_1 - l, e_2 - l] \cap \mathbb{Z}\}$$

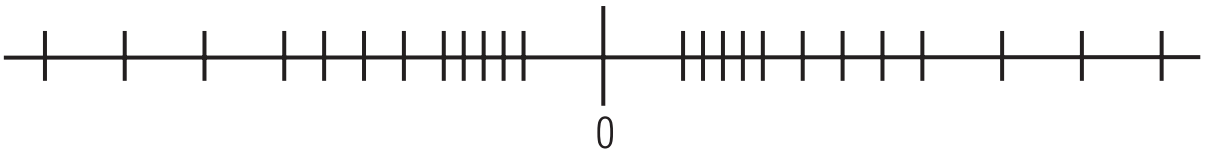
Erweitertes Gleitkommasystem $\quad \mathrm{Float}_e(R, l, e_1, e_2) =$

$$\{v \times m \times R^e | v \in \{-1, 1\}, m \in [0, R^l - 1] \cap \mathbb{N}, e \in [e_1 - l, e_2 - l] \cap \mathbb{Z}\}$$

Der Signifikant ist faktorisiert in das *Vorzeichen* v und die *Magnitude* m.

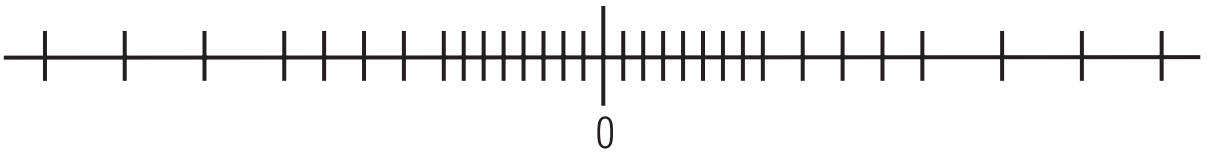
System normalisierter Gleitkommazahlen: Float(2, 3, 0, 2)

e \ m	0	1	2	3	4	5	6	7
-3	0				$\frac{4}{8}$	$\frac{5}{8}$	$\frac{6}{8}$	$\frac{7}{8}$
-2					$\frac{8}{8}$	$\frac{10}{8}$	$\frac{12}{8}$	$\frac{14}{8}$
-1					$\frac{16}{8}$	$\frac{20}{8}$	$\frac{24}{8}$	$\frac{28}{8}$



Erweitertes Gleitkommasystem: $\text{Float}_e(2, 3, 0, 2)$

e \ m	0	1	2	3	4	5	6	7
-3	0	$\frac{1}{8}$	$\frac{2}{8}$	$\frac{3}{8}$	$\frac{4}{8}$	$\frac{5}{8}$	$\frac{6}{8}$	$\frac{7}{8}$
-2	0	$\frac{2}{8}$	$\frac{4}{8}$	$\frac{6}{8}$	$\frac{8}{8}$	$\frac{10}{8}$	$\frac{12}{8}$	$\frac{14}{8}$
-1	0	$\frac{4}{8}$	$\frac{8}{8}$	$\frac{12}{8}$	$\frac{16}{8}$	$\frac{20}{8}$	$\frac{24}{8}$	$\frac{28}{8}$



Repräsentationen für Gleitkommazahlen

Es existiert eine Vielzahl unterschiedlicher Formate:

Signifikant und Exponent können jeweils in Vorzeichen-Betrag-Darstellung, Basis-Komplement oder vermindertem Basis-Komplement vorliegen; dies ergibt 9 verschiedene Grundformen.

Bei Vorzeichen-Betrag-Darstellung können Vorzeichen und Betrag jeweils separat oder in einem Feld zusammenhängend gespeichert werden.
In seltenen Fällen wurde auch das Vorzeichen von Komplement-Darstellungen abgetrennt.

Die verschiedenen Teile der Darstellung können beliebig angeordnet werden.

Die Basis R_s für den Signifikanten kann von der Basis R_e für den Exponenten abweichen, beide können wiederum von R verschieden sein (in der Praxis ist allerdings meist $R_e = 2$).

In Komplement-Darstellung vorliegende Exponenten können mit einem Bias versehen werden.

Schließlich sind noch Darstellungen für nicht normalisierte Werte (z. B. 0, ∞, $-\infty$) zu wählen.
In der Rechnerarithmetik wird häufig mit symbolischen unendlichen Elementen $\pm\infty$ operiert.
Der *Abschluss* einer Menge M ist definiert durch $M^{\pm\infty} := M \cup \{-\infty, +\infty\}$.

Vorzeichen-Betrag-Darstellung des Signifikanten

In der Vorzeichen-Betrag-Darstellung des Signifikanten wird das Vorzeichen $v \in \{-1, +1\}$ üblicherweise durch ein *Vorzeichenbit* dargestellt, mit der Codierung $0 \mathrel{\hat{=}} +1$ und $1 \mathrel{\hat{=}} -1$.

Eine Magnitude $m \in [R_s^{l-1}, R_s^l - 1] \cap \mathbb{N}$ wird codiert als Ziffernfolge $m_1\, m_2 \ldots m_l$ mit der Bedeutung $m = \sum_{i=0}^{l-1} m_{l-i} \times R_s^i$.

Für $R_s = 2$ kann wegen $m_1 = 1$ die Ziffer m_1 auch implizit sein, d. h. sie wird dann nicht in das Speicherformat aufgenommen, sondern bei der Verarbeitung der Zahlen je nach Bedarf ergänzt (*Hidden-Bit*, z. B. in IEEE-754, DEC/VAX).

Darstellung des Exponenten

Für den Exponenten $e \in \mathbb{Z}$ gibt es eine Reihe verschiedener Codierungen mit den unterschiedlichsten Eigenschaften und Intentionen:

Häufig wird zu Exponenten in einer Komplement-Darstellung ein sogenannter *Bias* addiert; dies ist eine positive Zahl mit der Eigenschaft, dass das Ergebnis der Addition nichtnegativ (in manchen Zahlensystemen auch echt positiv) ist.

Der Bias kann bei Bedarf so gewählt werden, dass unterhalb und/oder oberhalb der eigentlichen Darstellungen von Exponenten einige unbenutzte Werte auftreten, die der Kennzeichnung von Null, ∞, $-\infty$ oder dem Auftreten eines arithmetischen Fehlers, Über- oder Unterlaufs dienen.

Werden Anordnung und Darstellung der verschiedenen Teile einer Maschinenzahl sorgfältig aufeinander abgestimmt, lässt sich auf den Gleitkommazahlen ein arithmetischer Größenvergleich durch lexikalischen Vergleich (wie bei ganzen Zahlen im R-Komplement) durchführen.

Die Null kommt besonders häufig als Operand in Testoperationen vor und sollte deshalb eine leicht zu testende Darstellung besitzen (z. B. nur aus Null-Bits bestehen).

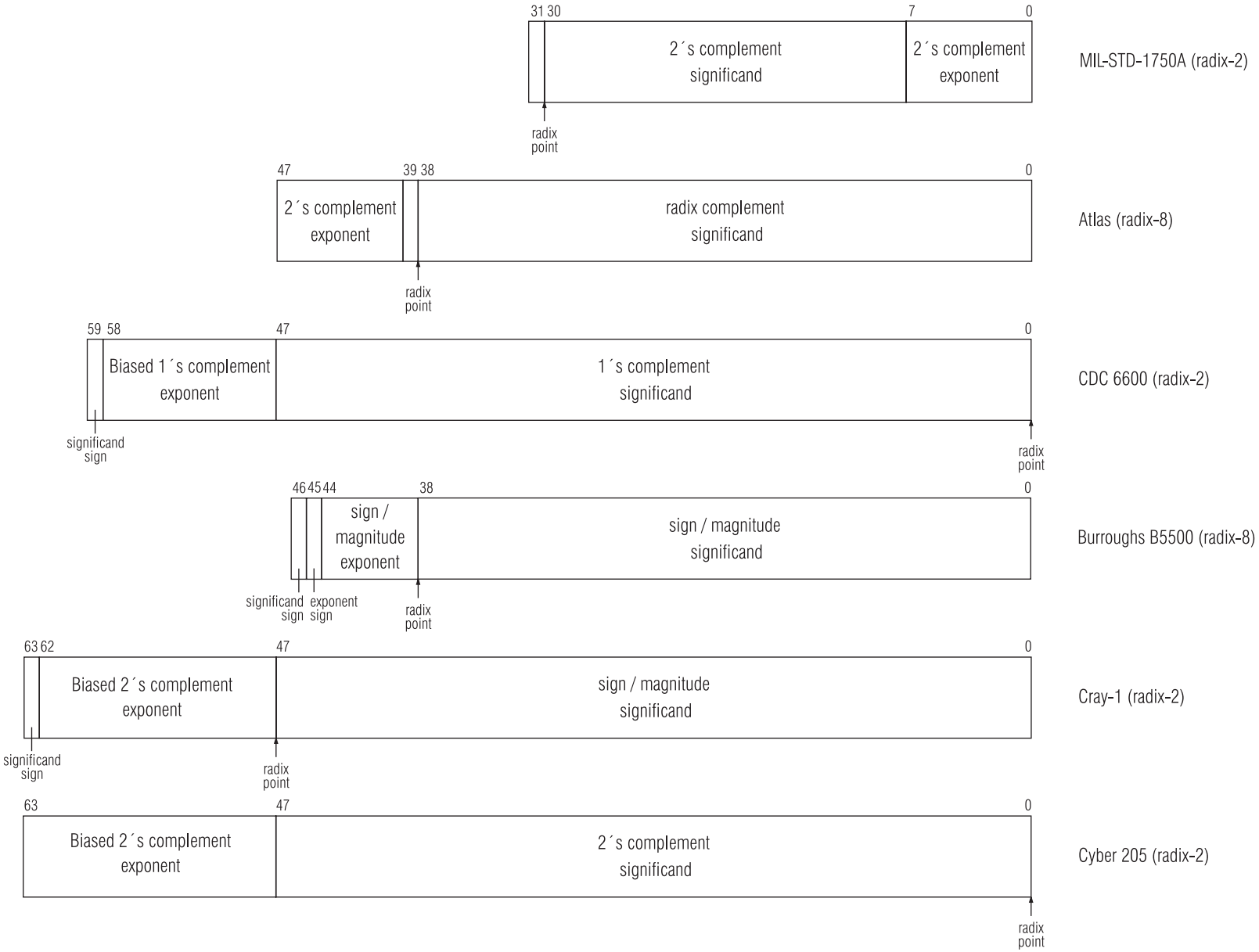
Darstellung erweiterter Gleitkommasysteme

Die Darstellung von Zahlen aus $\text{Float}_e(R, l, e1, e2)$ erfolgt im Prinzip nach dem gleichen Schema wie die der Zahlen aus $\text{Float}(R, l, e1, e2)$.

Abweichungen bestehen in folgenden Punkten:

- Für die Null ist in $\text{Float}_e(R, l, e1, e2)$ keine Sonderbehandlung nötig.
  Aus Gründen der Verträglichkeit wird die Null in $\text{Float}_e(R, l, e1, e2)$
  jedoch häufig genauso dargestellt wie in $\text{Float}(R, l, e1, e2)$.
- Die Hidden-Bit-Technik lässt sich hier nur nutzen, wenn normalisierte und denormalisierte Darstellungen unterscheidbar sind — etwa anhand ihrer abgespeicherten Exponenten.
  Für denormalisierte Zahlen gilt $m_1 = 0$, sodass für diese Zahlen ein Hidden-Bit
  mit dem Wert 0 benutzt werden kann.

Repräsentation von Gleitkommazahlen: Historische Beispiele



Festkommazahlen

Festkommasystem kann als spezielle Variante eines Gleitkommasystems gedeutet werden:

Zahlen $s \times R^e$

Basis R und Exponent e fest, Signifikant s Ganzzahl zur Basis R
R^e kann als Skalierungsfaktor gedeutet werden

$e > 0$ ist eher ungewöhnlich, $e = 0$ ergibt die Ganzzahlen, $e < 0$ typisch

Häufig $|e|$ klein (Währungen, Messungen, Anteile)
oder, wenn Mantissen l Stellen besitzen, $e = -l$ (Bruchteil) bzw. $e = 1 - l$

Festkommazahlen mit Nachkomma-Anteil

Spezifische Probleme, wenn $e < 0$:

Addition/Subtraktion von Festkommazahlen: $s \times R^e = s_1 \times R^e \pm s_2 \times R^e = (s_1 \pm s_2) \times R^e$
Möglichkeit des Überlaufs, wie bei Ganzzahlen; kein Genauigkeitsverlust

Multiplikation mit Ganzzahl: $s \times R^e = (s_1 \times R^e) \times s_2 = (s_1 \times s_2) \times R^e$
Möglichkeit des Überlaufs, wie bei Ganzzahlen; kein Genauigkeitsverlust

Multiplikation von Festkommazahlen: $s \times R^e = (s_1 \times R^e) \times (s_2 \times R^e) = (s_1 \times s_2 \times R^e) \times R^e$
$s = s_1 \times s_2 \times R^e$ im Allgemeinen keine Ganzzahl, Rundung nötig

Soll die Mantisse des Ergebnisses l Stellen besitzen und weist die Mantisse des Zwischenergebnisses der Ganzzahlmultiplikation l' Stellen auf, entstehen folgende Situationen:

$l' - e = l$: Rundung auf l Stellen; kein Überlauf möglich
$l' - e > l$: Rundung auf l Stellen; Überlauf möglich
$l' - e < l$: Rundung auf $l' - e$ Stellen, Genauigkeitsverlust; kein Überlauf möglich

Division durch Ganzzahl: $s \times R^e = (s_1 \times R^e)/s_2 = (s_1/s_2) \times R^e$
$s = s_1/s_2$ besitzt im Allgemeinen keine endliche Darstellung zur Basis R
Rundung nötig; kein Überlauf möglich

Division von Festkommazahlen: $s \times R^e = (s_1 \times R^e)/(s_2 \times R^e) = ((s_1/s_2) \times R^{-e}) \times R^e$
$s = (s_1/s_2) \times R^{-e}$ besitzt im Allgemeinen keine endliche Darstellung zur Basis R
Rundung nötig; Überlauf möglich

Anomalien

Jedes Gleitkommasystem ist endliche Approximation von $\mathbb{Q}$ bzw. $\mathbb{R}$

- Massiver Verlust der Abgeschlossenheit
- Massiver Verlust der Assoziativität der Addition
- Massiver Verlust der Assoziativität der Multiplikation
- Massiver Verlust der Distributivität
- Massives Fehlen multiplikativer Inverser

Bei echten Festkommasystemen (mit Nachkomma-Anteil) ähnlich