Ergänzungsmaterial Zahlendarstellung

G. Kemnitz

Institut für Informatik, Technische Universität Clausthal

January 25, 2013

Wertedarstellung

Datentypen

Datentypen

- Befehle, Adressen und Daten sind Bitvektoren.
- In einem Programm hat jeder Bitvektor einen Typ.
- Der Datentyp ordnen den 2^n darstellbaren Werten eines n-Bit-Vektors eine Bedeutung zu:
- Aufzählungstyp: Zuordnung symbolischer Werte.
  - Typ zur Darstellung der Wochentage:

Bitvektor	0000	0001	0010	0011	0100	0101	0110	0111
Wert	Mo	Di	Mi	Do	Fr	Sa	So	unzu

  - Typ zur Darstellung von Wahrheitswerten:

Bitvektor	0000	0001 ... 1111
Wert	falsch	wahr

- Zahlentypen: Zuordnung von Zahlenwerte.
  - Zahlentypen sind Aufzählungstypen, für die weitere Operationen definiert sind (+, -, *, >, ...)
  - Die Zuordnung ist so gewählt, dass die Operationen (+, -, *, ...) minimalen Schaltungsaufwand erfordern. (Nicht vom Programmierer festlegbar)
- zusammengesetzte Typen, z.B. Vektoren von Zahlen (in normalen Prozessoren keine direkte Hardware-Unterstützung)

Achtung:

- Den Typ und damit die Bedeutung der Bitvektoren kennt nur der Programmierer.
- Der Rechner kontrolliert im Allg. nicht, ob an der richtigen Speicherstelle ein Datenobjekt mit dem richtigen Typ steht.

Stellenwertsystem

Stellenwertsysteme

- B – Basis des Zahlensystems
- Darstellung einer Zahl durch einen Anreihung von Ziffern:

$$b_{n-1}b_{n-2}\ldots b_1b_0 \quad \text{mit } 0 \leq b_i \leq B-1$$

- Wert:

$$Z = \sum_{i=0}^{n-1} b_i \cdot B^i$$

- kleinster darstellbarer Wert: 0
- größter mit n Ziffern darstellbarer Wert: $B^n - 1$.

Beispiel 3-Ziffern-Dezimalzahl 543

- Basis: $B = 10$
- Wert: $Z = 5 \cdot 10^2 + 4 \cdot 10^1 + 3 \cdot 10^0$

Binär-, Oktal- und Hexadezimalzahlen

Binärsystem	$B = 2$	$b_i \in \{0, 1\}$
Oktalsystem	$B = 8$	$b_i \in \{0, 1, \ldots, 7\}$
Hexadezimalsystem	$B = 16$	$b_i \in \{0, 1, \ldots, 9, \mathrm{A}, \mathrm{B}, \ldots, \mathrm{F}\}$

- eine Hexadezimalziffer $\Leftrightarrow$ vier Binärziffern
- eine Oktalziffer $\Leftrightarrow$ drei Binärziffern

dezimal	binär	oktal	hexadezimal
0	0_2	0_8	0_{16}
1	1_2	1_8	1_{16}
...	...	...	...
7	111_2	7_8	7_{16}
8	1000_2	10_8	8_{16}
...	...	...	...
15	1111_2	17_8	F_{16}
16	10000_2	20_8	10_{16}
...	...	...	...

Berechnung der Ziffernfolge zu einem Wert

Umwandlung in eine Binärzahl

$$
\begin{array}{rcrl}
75 & : 2 = & 37 & \text{Rest: } b_0 = 1 \\
37 & : 2 = & 18 & \text{Rest: } b_1 = 1 \\
18 & : 2 = & 9 & \text{Rest: } b_2 = 0 \\
9 & : 2 = & 4 & \text{Rest: } b_3 = 1 \\
4 & : 2 = & 2 & \text{Rest: } b_4 = 0 \\
2 & : 2 = & 1 & \text{Rest: } b_5 = 0 \\
1 & : 2 = & 0 & \text{Rest: } b_6 = 1
\end{array}
$$

Umwandlung in eine Hexadezimalzahl

$$
\begin{array}{rcrl}
75 & : 16 = & 4 & \text{Rest: } 11 = (\mathrm{B})_{16} \\
4 & : 16 = & 0 & \text{Rest: } 4 = (4)_{16}
\end{array}
$$

Umwandlung in eine Oktalzahl

$$
\begin{array}{rcrl}
75 & : 8 = & 9 & \text{Rest: } 3 \\
9 & : 8 = & 1 & \text{Rest: } 1 \\
1 & : 8 = & 0 & \text{Rest: } 1
\end{array}
$$

Ergebnis: $(1001011)_2 = (4\mathrm{B})_{16} = (113)_8$

Zweierkomplement

Vorzeichenbehaftete Zahlen und Stellenkomplement

Statt durch Vorzeichen und Betrag werden vorzeichenbehaftete Zahlen durch »Stellenkomplement +1« dargestellt.
Mathematische Grundlage:

- Das Stellenkomplement zu einer Ziffer b_i ist die Differenz zur größten darstellbaren Ziffer mit dem Wert $B - 1$:

$$\bar{b}_i = B - 1 - b_i$$

Beispiel: $\overline{437} = 562$

- Zahl plus Stellenkomplement gleich größte darstellbare Zahl.

Beispiel: $437 + \overline{437} = 437 + 562 = 999$

- plus Eins gleich kleinste nicht darstellbare Zahl:

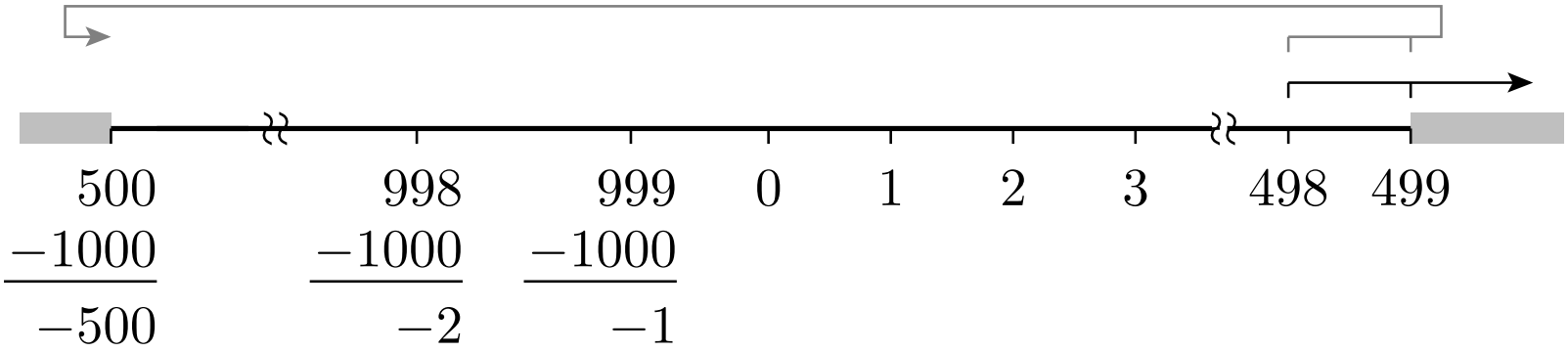
$$Z + \bar{Z} + 1 = B^n$$

$$Z + \bar{Z} + 1 = B^n$$

- Auflösung nach $-Z$:

$$-Z = \bar{Z} + 1 - \left[\underbrace{B^n}_{*}\right] \quad * \text{ nicht darstellbar}$$

- Aufteilung des Darstellungsbereichs



nicht darstellbar
→ Addition ohne Wertebereichsbegrenzung
→ Addition modulo-B^n

- unterer Bereich positiv: $Z = \sum_{i=0}^{n-1} b_i \cdot B^i$
- oberer Bereich negativ: $Z = \left(\sum_{i=0}^{n-1} b_i \cdot B^i\right) - B^n$

die Bereichsgrenze muss festgelegt sein, in der Regel Bereichsmitte

---

Subtraktion als Addition mit dem Stellenkomplement:

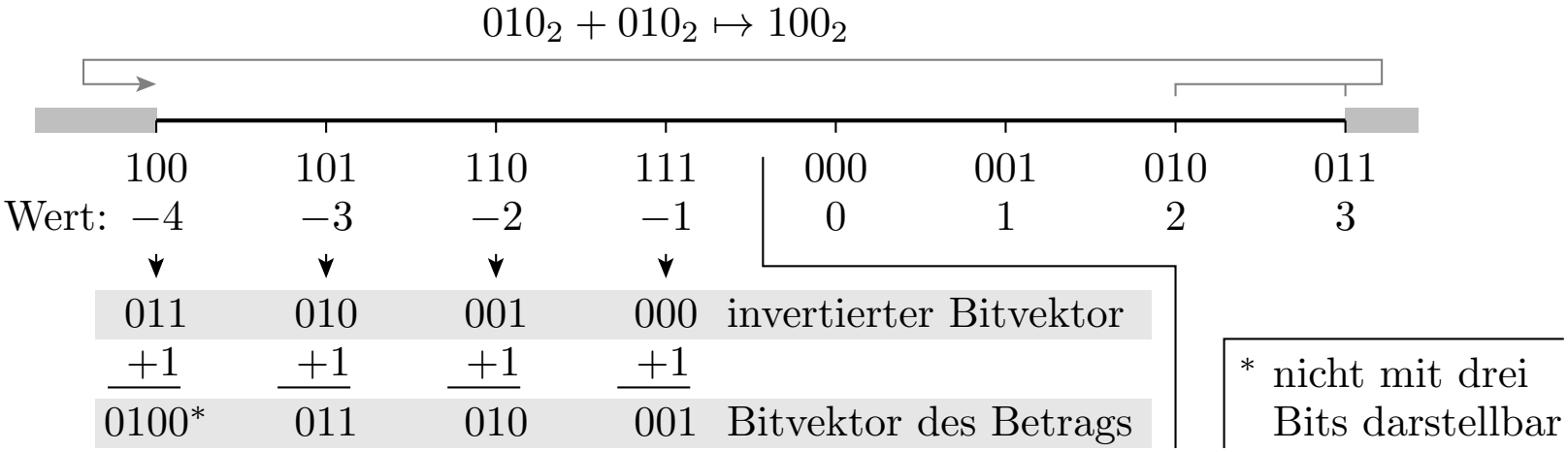
$$\begin{aligned} 231 - 008 &= 223 \\ 231 + 991 + 1 &= (1)223 \end{aligned}$$

Beispiel mit negativem Ergebnis:

$$\begin{aligned} 008 - 231 &= -223 \\ |-223| &= 223 \\ 008 + 768 + 1 &= 776 + 1 = 777 \\ |777| &= 222 + 1 = 223 \end{aligned}$$

Zweierkomplement

Stellenkomplement für Binärzahlen mit führendem Bit gleich Vorzeichenbit



- Stellenkomplement für Binärziffern: $\bar{0} \mapsto 1$; $\bar{1} \mapsto 0$ (Invertierung)
- Überlaufgrenze: 011...1 → 100...0

- Wert:

$$Z = \begin{cases} \sum_{i=0}^{n-1} b_i \cdot 2^i & \text{für } b_{n-1} = 0 \\ \sum_{i=0}^{n-1} b_i \cdot 2^i - 2^n & \text{für } b_{n-1} = 1 \end{cases}$$

  vereinfacht:

$$Z = \sum_{i=0}^{n-1} b_i \cdot 2^i - b_{n-1} \cdot 2^n = \sum_{i=0}^{n-2} b_i \cdot 2^i - b_{n-1} \cdot 2^{n-1}$$

- größter darstellbarer Wert: $Z(011\ldots1_2) = 2^{n-1} - 1$
- kleinster darstellbarer Wert: $Z(100\ldots0_2) = -2^{n-1}$

Festkommazahlen

Festkommazahlen

- Erweiterung um m Nachkommastellen
- Nachkommastellen haben einen negativen Stellenindex
- Wert vorzeichenfreier Festkommazahlen:

$$Z = \sum_{i=-m}^{n-1} b_i \cdot B^i$$

Beispiel: $23{,}89 = 2 \cdot 10^1 + 3 \cdot 10^0 + 8 \cdot 10^{-1} + 9 \cdot 10^{-2}$

- Wert negativer vorzeichenbehafteter Festkommazahlen:

$$Z = \sum_{i=-m}^{n-1} b_i \cdot B^i - B^n$$

Beispiel:
$-23{,}89 = 7 \cdot 10^1 + 6 \cdot 10^0 + 1 \cdot 10^{-1} + 1 \cdot 10^{-2} \left(-10^2\right)^*$

$(*$ – nicht darstellbarer Summand$)$

Für gebrochene Binärzahlen im Zweierkomplement ist das führende Bit das Vorzeichenbit. Wert:

$$Z = \sum_{i=-m}^{n-2} b_i \cdot 2^i - b_{n-1} \cdot 2^{n-1} \tag{1}$$

Die größte mit n Vorkomma- und m Nachkommabits darstellbare Zahl 01..1,11... hat dem Wert $2^n - 2^{-m}$ und die kleinste darstellbare Zahl 10...0,00... den Wert -2^n.

Gleitkommazahlen

Wertebereich vs. Rundungsfehler

Kommaposition	Wertebereich	Rundungsfehler
$n = 2, m = 6$	0 bis $2^2 - 2^{-6}$	$\pm 2^{-7}$
$n = 4, m = 4$	0 bis $2^4 - 2^{-4}$	$\pm 2^{-5}$
$n = 6, m = 2$	0 bis $2^6 - 2^{-2}$	$\pm 2^{-3}$

- bei einer Festkommazahl bestimmt die Anzahl
  - der Vorkommastellen den Wertebereich
  - der Nachkommastellen den Rundungsfehler
- Wahl der Kommaposition: Kompromiss zwischen der Größe des Wertebereichs und der Genauigkeit

Gleitkommazahlen (variable Kommaposition)

- Mantisse M: Wertebereich (normiert) $1 \leq Z(M) < 2$
- Charakteristik c: Kommaverschiebung, ganzzahlig; c_0 – Wert von c für Kommaverschiebung Null
- Vorzeichenbit s

Wert für $0 < c < c_{\max}$ (normierte Darstellung):

$$Z = (-1)^s \cdot (1, M_{-1} \ldots M_{-m}) \cdot 2^{c-c_0}$$

Wert für $c = 0$ (denormiert):

$$Z = (-1)^s \cdot (M_0, M_{-1} \ldots M_{-m}) \cdot 2^{-c_0}$$

Sonderwerte $c = c_{\max}$:

$$Z = \begin{cases} \infty & \text{für } s = 0 \text{ und } m = 0 \\ -\infty & \text{für } s = 1 \text{ und } m = 0 \\ \text{nan} & \text{für } m \neq 0 \end{cases}$$

(nan, not a number – ungültig; $\pm\infty$ – positiver/negativer Wertebereichsüberlauf)

- 32-Bit-Format »IEEE-754 single«

Bitvektor (31 … 24 | 23 … 16 | 15 … 8 | 7 … 0, Bitnummer ←)	Wert
s | c | M	
0 | 10000011 | 00100100000000000000000	$+1.240000_{16} \cdot 2^{83_{16}-7f_{16}}$
8 3 | 2 4 0 0 0 0	$= 18{,}25$
$+$ | $c = 83_{16}$ | $M = 1{,}240000_{16}$	
1 | 01111001 | 11001011001110100001110	$-1.CB3A1C_{16} \cdot 2^{79_{16}-7f_{16}}$
7 9 | C B 3 A 1 C	$\approx -2{,}8029 \cdot 10^{-2}$
$-$ | $c = 79_{16}$ | $M = 1{,}CB3A1C_{16}$	
0 | 00000000 | 00011001011110100011001100	$+0{,}32F464_{16} \cdot 2^{0-7f_{16}}$
, 3 2 F 4 6 4	$\approx 1{,}170 \cdot 10^{-39}$
0/denorm. | $M = 0{,}32F464_{16}$	