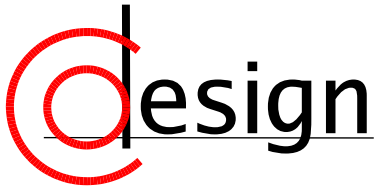
Übungen zur Grundlagen der Technischen Informatik

Übung 3 – Zahlendarstellung, -konversion und IEEE754

Florian Frank

Friedrich-Alexander-Universität Erlangen-Nürnberg

Wintersemester 2018/19



FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

TECHNISCHE FAKULTÄT

Was machen wir heute?

Aufgabe 1 – Zahlendarstellungen in der Theorie

Was machen wir heute?

Aufgabe 1 – Zahlendarstellungen in der Theorie

Aufgabe 2 – Zahlendarstellungen in der Praxis

Was machen wir heute?

Aufgabe 1 – Zahlendarstellungen in der Theorie

Aufgabe 2 – Zahlendarstellungen in der Praxis

Aufgabe 3 – Zahlendarstellungen

Was machen wir heute?

Aufgabe 1 – Zahlendarstellungen in der Theorie

Aufgabe 2 – Zahlendarstellungen in der Praxis

Aufgabe 3 – Zahlendarstellungen

Aufgabe 4 – Zahlenkonversion

Was machen wir heute?

Aufgabe 1 – Zahlendarstellungen in der Theorie

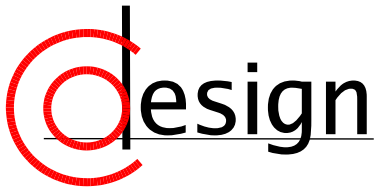
Aufgabe 2 – Zahlendarstellungen in der Praxis

Aufgabe 3 – Zahlendarstellungen

Aufgabe 4 – Zahlenkonversion

Aufgabe 5 – Konversion von Gleitkommazahlen

Aufgabe 1 – Zahlendarstellungen in der Theorie

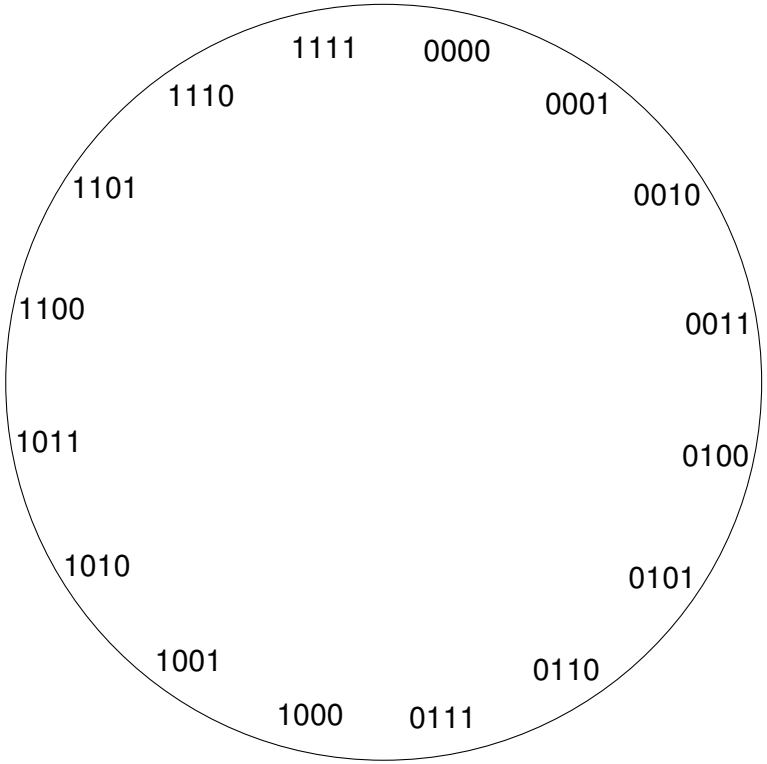


FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

TECHNISCHE FAKULTÄT

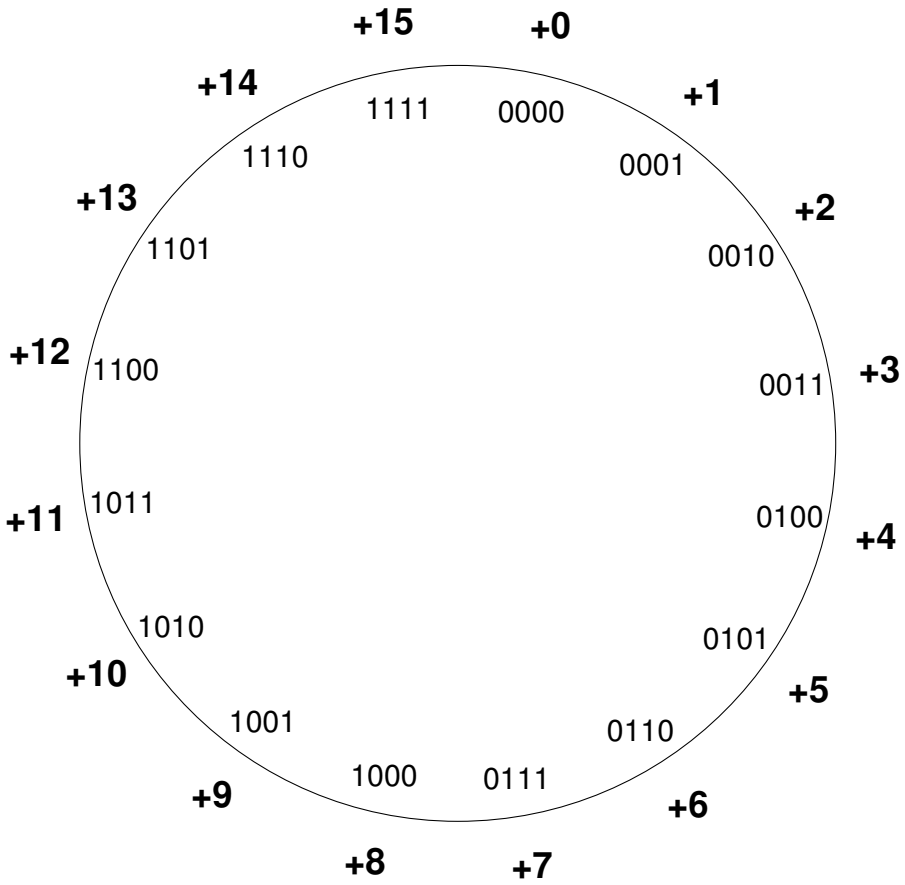
Aufgabe 1 – Zahlendarstellungen: Theorie

Vorzeichenlose Zahlendarstellung – `unsigned int`



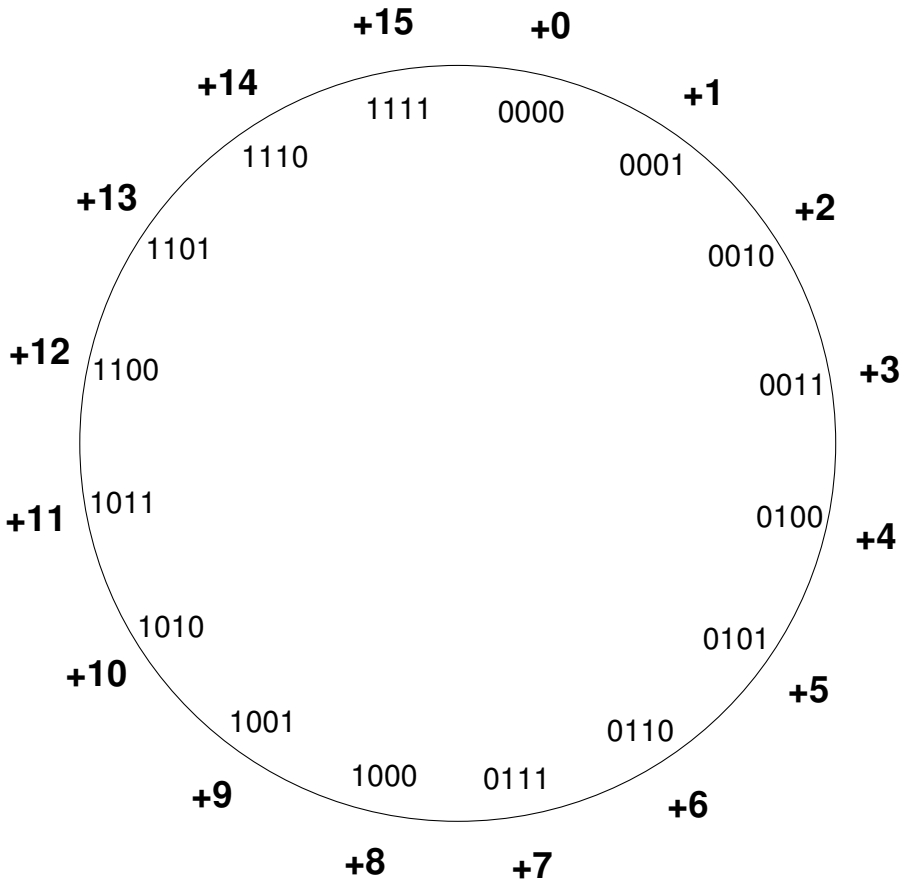
Aufgabe 1 – Zahlendarstellungen: Theorie

Vorzeichenlose Zahlendarstellung – `unsigned int`



Aufgabe 1 – Zahlendarstellungen: Theorie

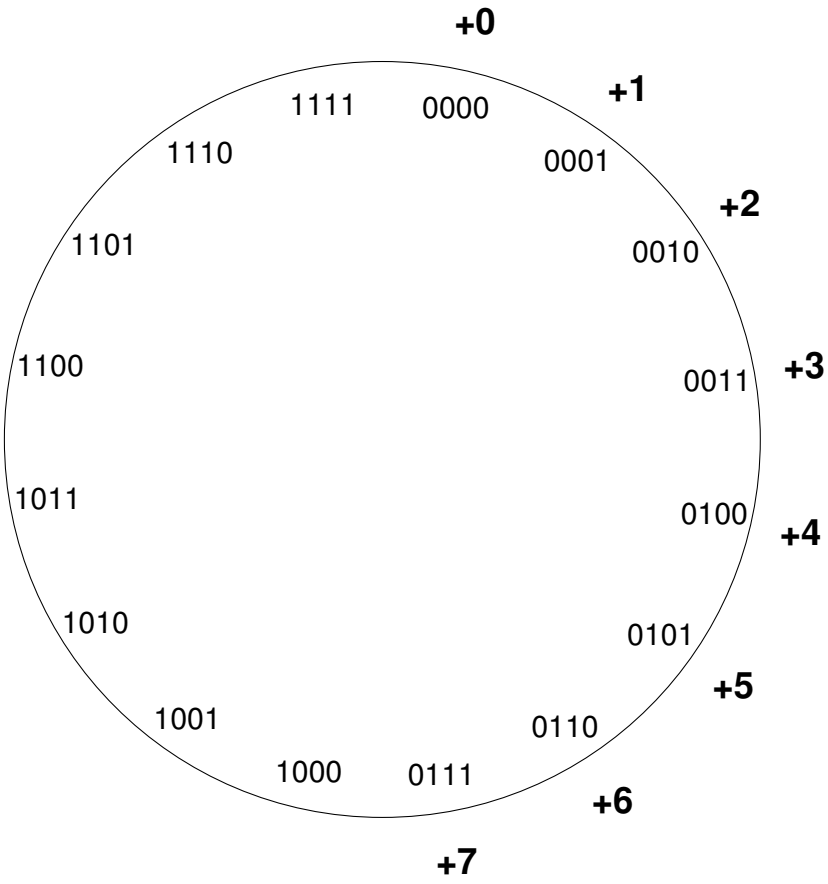
Vorzeichenlose Zahlendarstellung – `unsigned int`



Wertebereich einer n bit breiten Zahl: $[0, 2^n - 1]$

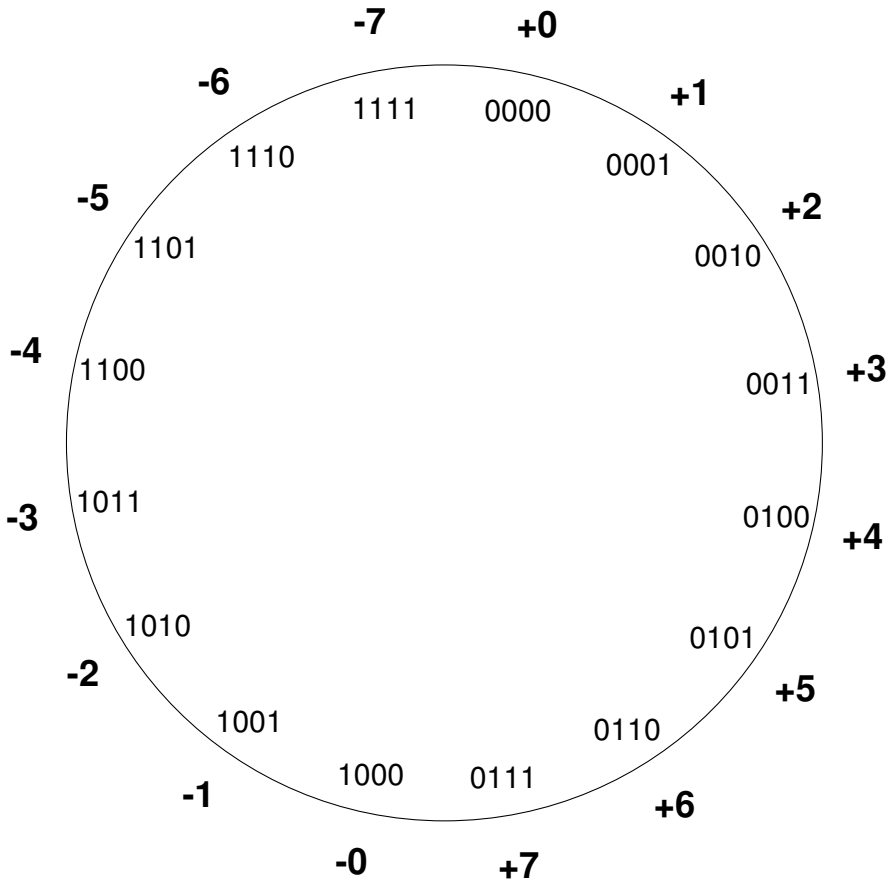
Aufgabe 1 – Zahlendarstellungen: Theorie

Vorzeichen-/Betragsdarstellung



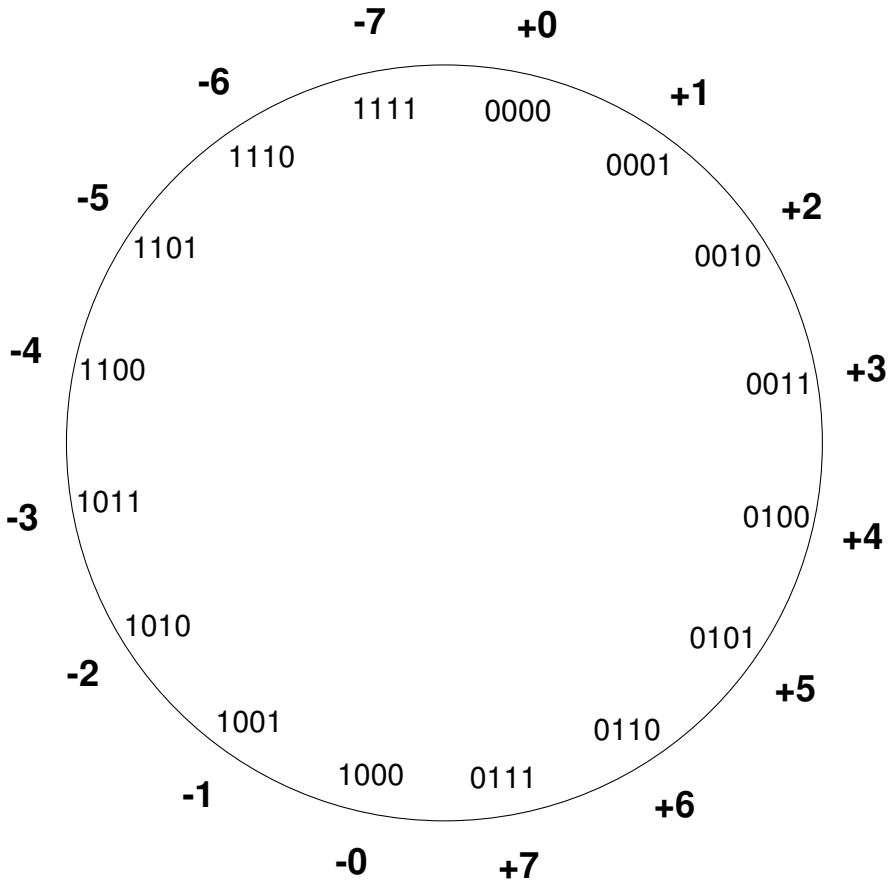
Aufgabe 1 – Zahlendarstellungen: Theorie

Vorzeichen-/Betragsdarstellung



Aufgabe 1 – Zahlendarstellungen: Theorie

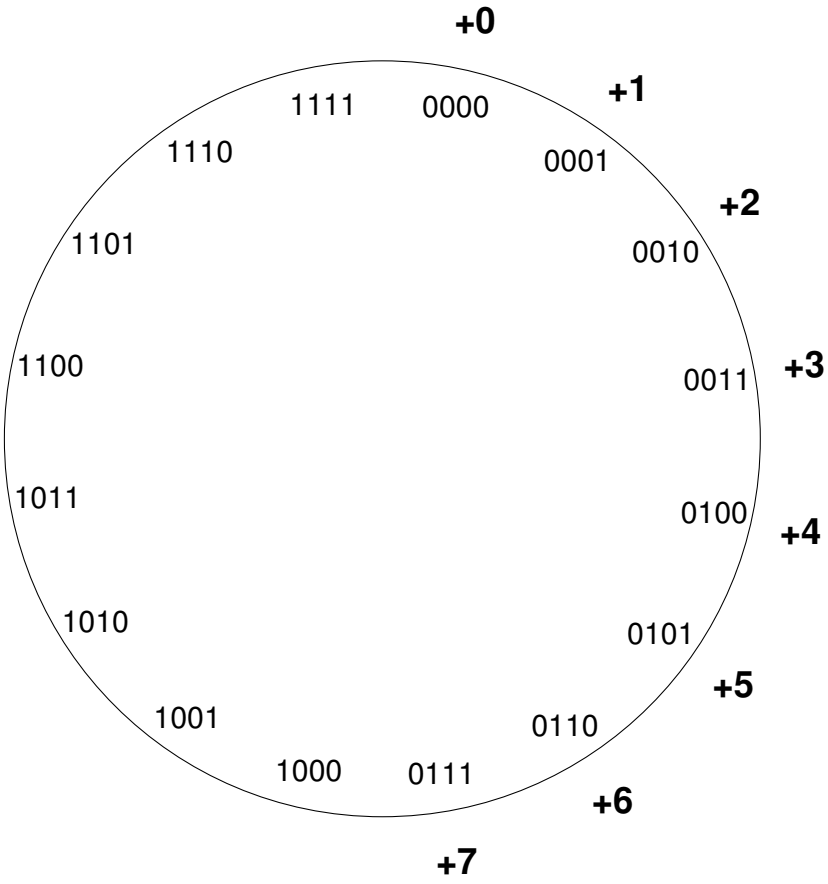
Vorzeichen-/Betragsdarstellung



Wertebereich einer n bit breiten Zahl: $\left[-2^{n-1}+1, 2^{n-1}-1\right]$

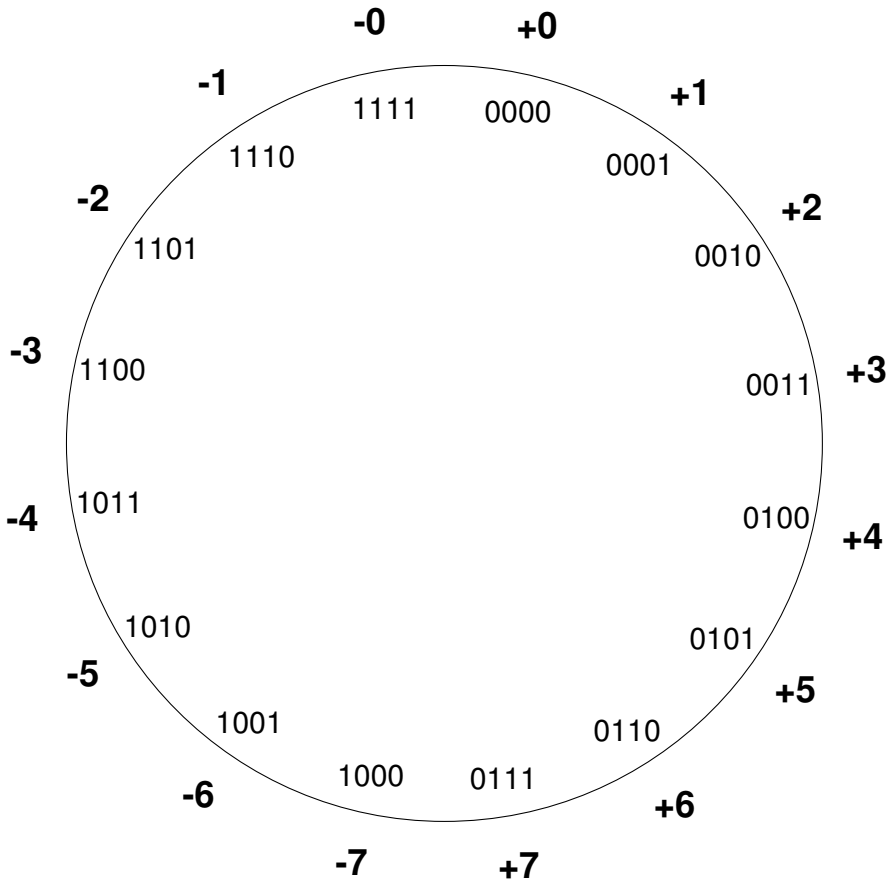
Aufgabe 1 – Zahlendarstellungen: Theorie

Einerkomplement



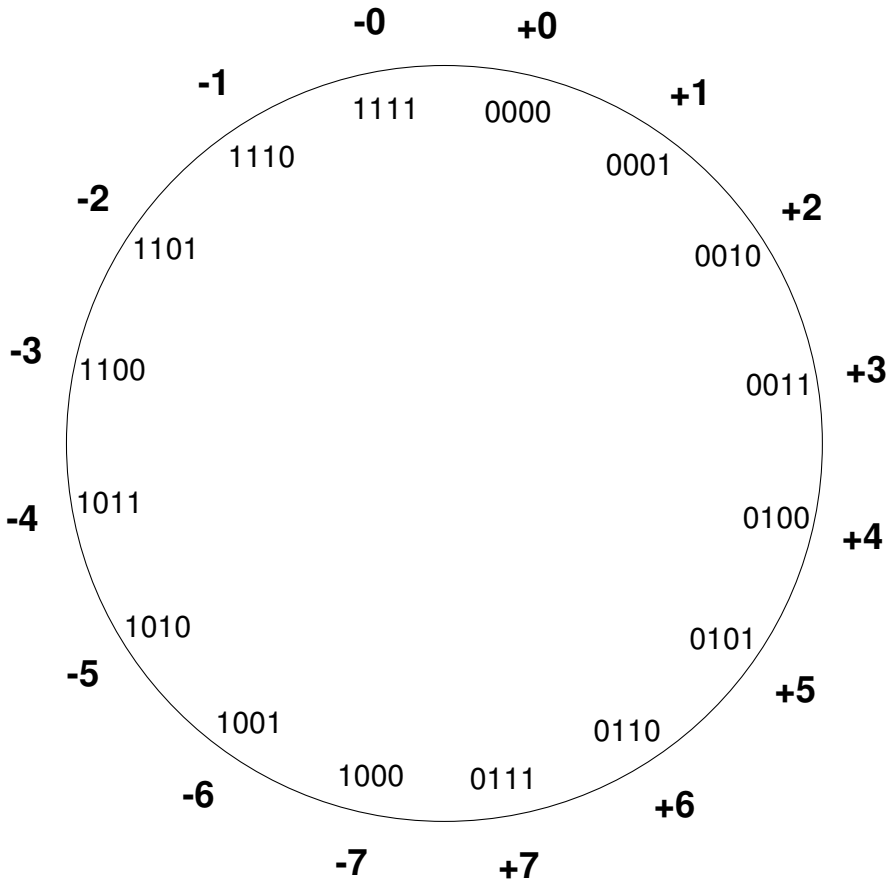
Aufgabe 1 – Zahlendarstellungen: Theorie

Einerkomplement



Aufgabe 1 – Zahlendarstellungen: Theorie

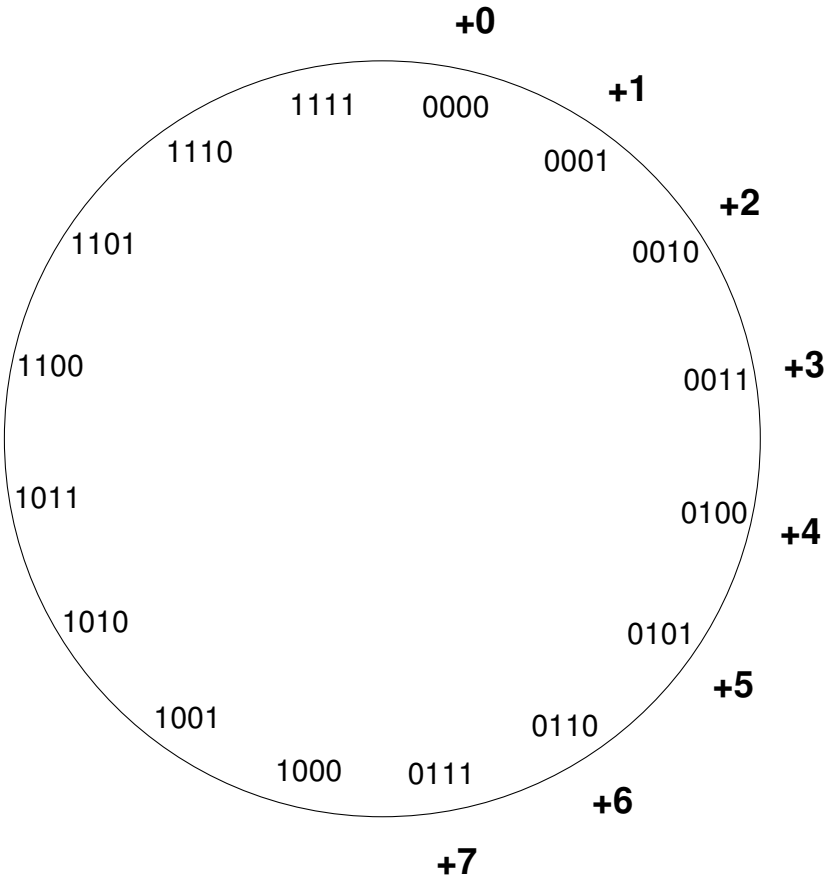
Einerkomplement



Wertebereich einer n bit breiten Zahl: $\left[-2^{n-1}+1, 2^{n-1}-1\right]$

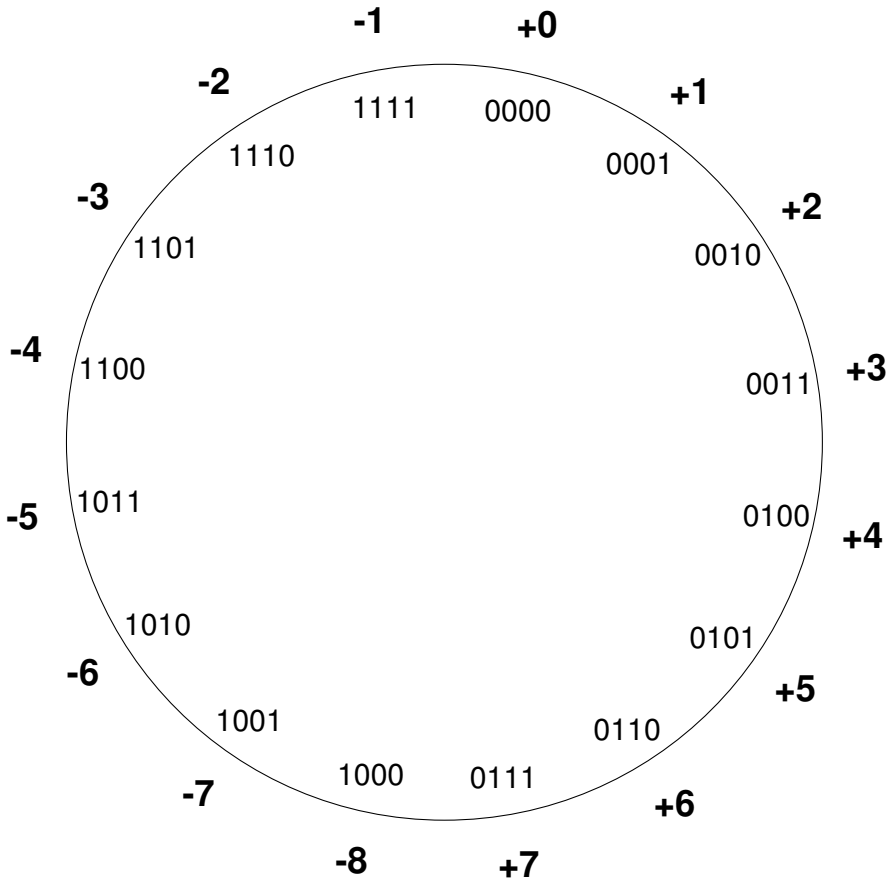
Aufgabe 1 – Zahlendarstellungen: Theorie

Zweierkomplement



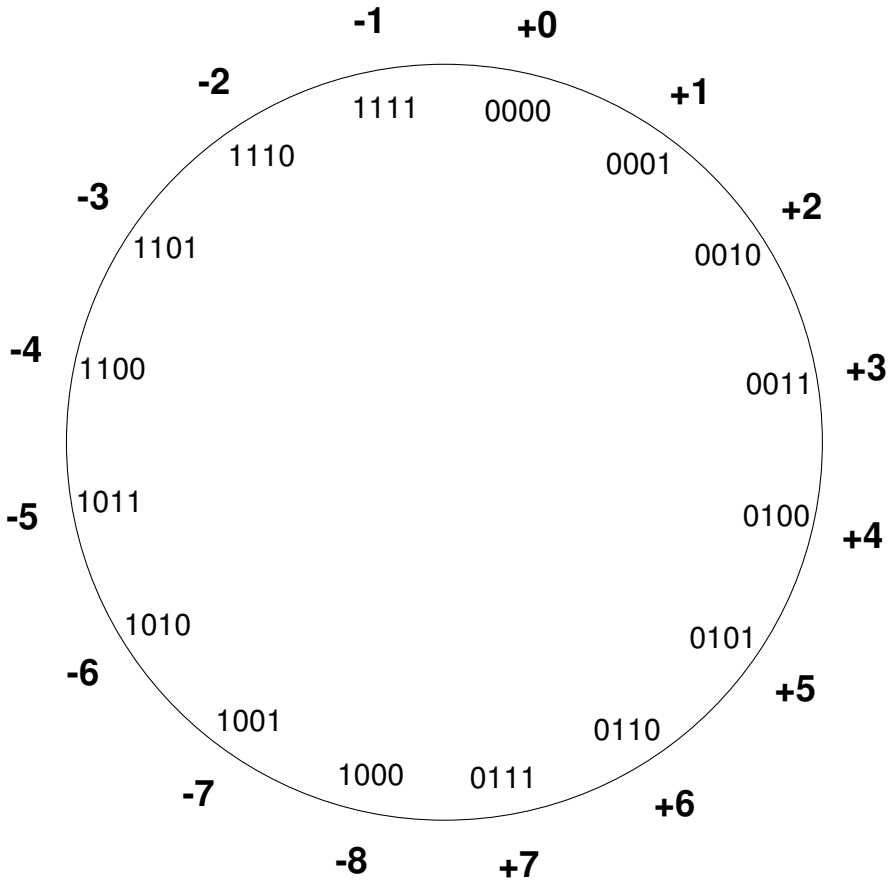
Aufgabe 1 – Zahlendarstellungen: Theorie

Zweierkomplement



Aufgabe 1 – Zahlendarstellungen: Theorie

Zweierkomplement



Wertebereich einer n bit breiten Zahl: $\left[-2^{n-1}, 2^{n-1}-1\right]$

Aufgabe 2 – Zahlendarstellungen in der Praxis

FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG
TECHNISCHE FAKULTÄT

Aufgabe 2 – Zahlendarstellungen: Praxis

Gegeben seien die folgenden Dezimalzahlen:

- 2
- 64
- 255
- -254
- -32

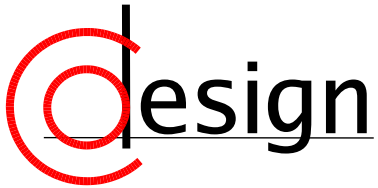
Stellen Sie diese Zahlen **jeweils** in ...

i ... Vorzeichen/Betragsdarstellung dar.

ii ... 1er-Komplementdarstellung dar.

iii ... 2er-Komplementdarstellung dar.

Aufgabe 3 – Zahlendarstellungen



FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

TECHNISCHE FAKULTÄT

Aufgabe 3 – Zahlendarstellungen

Gegeben seien die folgenden positiven Binärzahlen:

- 10_2
- 10100_2
- 1111111_2
- 1000000_2
- 11111111_2
- 11111110_2.

Stellen Sie diese Zahlen jeweils als ...

i ... Hexadezimalzahl dar.

ii ... Oktalzahl dar.

iii ... BCD-Zahl dar.

Aufgabe 3 – Begriffsklärung

BCD-Zahl

BCD-Zahl $\equiv$ **B**inary **C**oded **D**igit
Jede Dezimalziffer wird durch vier Binärstellen repräsentiert. Damit ist die Dezimalzahl leicht rekonstruierbar, die Kodierung aber vergleichsweise ineffizient.

Aufgabe 3 – Begriffsklärung

BCD-Zahl

BCD-Zahl $\equiv$ **B**inary **C**oded **D**igit
Jede Dezimalziffer wird durch vier Binärstellen repräsentiert. Damit ist die Dezimalzahl leicht rekonstruierbar, die Kodierung aber vergleichsweise ineffizient.

Oktalsystem/Hexadezimalsystem

Ein polyadisches System zur Basis 8 heißt **Oktalsystem**.
Ein polyadisches System zur Basis 16 heißt **Hexadezimalsystem**.

Aufgabe 3 – Begriffsklärung

Umwandlung verwandter Systeme

Sei eine Zahl Z aus dem System zur Basis a in das System zur Basis b (mit $a, b \in \mathbb{N}$) zu kodieren und es gilt:

Aufgabe 3 – Begriffsklärung

Umwandlung verwandter Systeme

Sei eine Zahl Z aus dem System zur Basis a in das System zur Basis b (mit $a, b \in \mathbb{N}$) zu kodieren und es gilt:

- $a = b^n$ mit $n \in \mathbb{N}$.
  Dann wird jede Stelle von Z n-äquidistant zerteilt (sprich: in n-Stellen (zur Basis b) zerlegt).

Aufgabe 3 – Begriffsklärung

Umwandlung verwandter Systeme

Sei eine Zahl Z aus dem System zur Basis a in das System zur Basis b (mit $a, b \in \mathbb{N}$) zu kodieren und es gilt:

- $a = b^n$ mit $n \in \mathbb{N}$.
  Dann wird jede Stelle von Z n-äquidistant zerteilt (sprich: in n-Stellen (zur Basis b) zerlegt).
- $a^n = b$ mit $n \in \mathbb{N}$.
  Dann werden jeweils n Stellen zu einer neuen Stelle (der Basis b) zusammengefasst.

Aufgabe 3 – Begriffsklärung

Umwandlung verwandter Systeme

Sei eine Zahl Z aus dem System zur Basis a in das System zur Basis b (mit $a, b \in \mathbb{N}$) zu kodieren und es gilt:

- $a = b^n$ mit $n \in \mathbb{N}$.
  Dann wird jede Stelle von Z n-äquidistant zerteilt (sprich: in n-Stellen (zur Basis b) zerlegt).
- $a^n = b$ mit $n \in \mathbb{N}$.
  Dann werden jeweils n Stellen zu einer neuen Stelle (der Basis b) zusammengefasst.

Umwandlung Binär $\mapsto$ BCD

Zuerst ist eine Dezimalzahl zu bilden, die dann jeweils stellenweise kodiert wird.

Aufgabe 3 – Begriffsklärung

Umwandlung verwandter Systeme

Sei eine Zahl Z aus dem System zur Basis a in das System zur Basis b (mit $a, b \in \mathbb{N}$) zu kodieren und es gilt:

- $a = b^n$ mit $n \in \mathbb{N}$.
  Dann wird jede Stelle von Z n-äquidistant zerteilt (sprich: in n-Stellen (zur Basis b) zerlegt).
- $a^n = b$ mit $n \in \mathbb{N}$.
  Dann werden jeweils n Stellen zu einer neuen Stelle (der Basis b) zusammengefasst.

Umwandlung Binär $\mapsto$ BCD

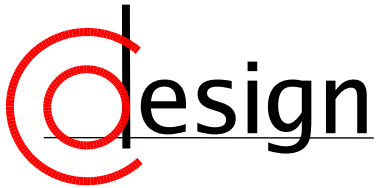
Zuerst ist eine Dezimalzahl zu bilden, die dann jeweils stellenweise kodiert wird.

Umwandlung Binär $\mapsto$ Dezimal

Für jede vorzeichenlose Binärzahl Z mit n Stellen z_i gilt:

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

Aufgabe 4 – Zahlenkonversion



FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

TECHNISCHE FAKULTÄT

Aufgabe 4 – Zahlenkonversion

a) Konvertieren Sie die Hexadezimalzahl $\mathrm{A03}_{16}$ mit sukzessiver Division unter ausschließlicher Verwendung der angegebenen Zahlensysteme ins Binär- bzw. Ternärsystem.

b) Konvertieren Sie die Binärzahl $1110\,0111_2$ unter ausschließlicher Verwendung der angegebenen Zahlensysteme ins Oktal- bzw. Ternärsystem.

c) Konvertieren Sie die Dezimalzahl $234,28125_{10}$ ins Binärformat. Verwenden Sie für die Nachkommastellen maximal 4 Bit.

Aufgabe 4 – Begriffsklärung

Umwandlung verwandter Systeme

Sei eine Zahl Z aus dem System zur Basis a in das System zur Basis b (mit $a, b \in \mathbb{N}$) zu kodieren und es gilt:

- $a = b^n$ mit $n \in \mathbb{N}$.
  Dann wird jede Stelle von Z n-äquidistant zerteilt (sprich: in n-Stellen (zur Basis b) zerlegt).
- $a^n = b$ mit $n \in \mathbb{N}$.
  Dann werden jeweils n Stellen zu einer neuen Stelle (der Basis b) zusammengefasst.

Aufgabe 4 – Begriffsklärung

Umwandlung fremder Systeme

Sei eine Zahl Z aus dem System zur Basis a in das System zur Basis b (mit $a, b \in \mathbb{N}$) zu kodieren und die Systeme nicht verwandt (wie in Aufgabe 3). Dann erfolgt die Kodierung mit folgendem Algorithmus:

Aufgabe 4 – Begriffsklärung

Umwandlung fremder Systeme

Sei eine Zahl Z aus dem System zur Basis a in das System zur Basis b (mit $a, b \in \mathbb{N}$) zu kodieren und die Systeme nicht verwandt (wie in Aufgabe 3). Dann erfolgt die Kodierung mit folgendem Algorithmus:

Schritt 1: Dividiere b im System a von Zahl s im System a

Aufgabe 4 – Begriffsklärung

Umwandlung fremder Systeme

Sei eine Zahl Z aus dem System zur Basis a in das System zur Basis b (mit $a, b \in \mathbb{N}$) zu kodieren und die Systeme nicht verwandt (wie in Aufgabe 3). Dann erfolgt die Kodierung mit folgendem Algorithmus:

Schritt 1: Dividiere b im System a von Zahl s im System a

Schritt 2: Merke das Ergebnis $e = \frac{s}{b}$ und den Rest $r = s \mod b$

Aufgabe 4 – Begriffsklärung

Umwandlung fremder Systeme

Sei eine Zahl Z aus dem System zur Basis a in das System zur Basis b (mit $a, b \in \mathbb{N}$) zu kodieren und die Systeme nicht verwandt (wie in Aufgabe 3). Dann erfolgt die Kodierung mit folgendem Algorithmus:

Schritt 1: Dividiere b im System a von Zahl s im System a

Schritt 2: Merke das Ergebnis $e = \frac{s}{b}$ und den Rest $r = s \mod b$

Schritt 3: r repräsentiert eine Stelle des Ergebnisses in System b

Aufgabe 4 – Begriffsklärung

Umwandlung fremder Systeme

Sei eine Zahl Z aus dem System zur Basis a in das System zur Basis b (mit $a, b \in \mathbb{N}$) zu kodieren und die Systeme nicht verwandt (wie in Aufgabe 3). Dann erfolgt die Kodierung mit folgendem Algorithmus:

Schritt 1: Dividiere b im System a von Zahl s im System a

Schritt 2: Merke das Ergebnis $e = \frac{s}{b}$ und den Rest $r = s \mod b$

Schritt 3: r repräsentiert eine Stelle des Ergebnisses in System b

Schritt 4: Ist $e > 0$, so setze b zu e und mache weiter mit Schritt 1.

Aufgabe 4 – Begriffsklärung

Umwandlung fremder Systeme

Sei eine Zahl Z aus dem System zur Basis a in das System zur Basis b (mit $a, b \in \mathbb{N}$) zu kodieren und die Systeme nicht verwandt (wie in Aufgabe 3). Dann erfolgt die Kodierung mit folgendem Algorithmus:

Schritt 1: Dividiere b im System a von Zahl s im System a

Schritt 2: Merke das Ergebnis $e = \frac{s}{b}$ und den Rest $r = s \mod b$

Schritt 3: r repräsentiert eine Stelle des Ergebnisses in System b

Schritt 4: Ist $e > 0$, so setze b zu e und mache weiter mit Schritt 1.

Schritt 5: Lese das Ergebnis rückwärts aus

Aufgabe 4 – Zahlenkonversion

a) Konvertieren Sie die Hexadezimalzahl $\mathrm{A03}_{16}$ mit sukzessiver Division unter ausschließlicher Verwendung der angegebenen Zahlensysteme ins Binär- bzw. Ternärsystem.

Aufgabe 4 – Zahlenkonversion

b) Konvertieren Sie die Binärzahl $1110\,0111_2$ unter ausschließlicher Verwendung der angegebenen Zahlensysteme ins Oktal- bzw. Ternärsystem.

Aufgabe 4 – Zahlenkonversion

c) Konvertieren Sie die Dezimalzahl $234,28125_{10}$ ins Binärformat. Verwenden Sie für die Nachkommastellen maximal 4 Bit.

Aufgabe 4 – Begriffsklärung

Festkommazahl

Eine Festkommazahl ist eine Zahl, die aus einer festen Anzahl von Ziffern besteht. Die Position des Kommas ist dabei **fest** vorgegeben, daher der Name.

Aufgabe 4 – Begriffsklärung

Festkommazahl

Eine Festkommazahl ist eine Zahl, die aus einer festen Anzahl von Ziffern besteht. Die Position des Kommas ist dabei **fest** vorgegeben, daher der Name.
Dabei wird das polyadische System einfach fortgeführt, es gilt also bei einer n-stelligen B-adischen Festkommazahl Z mit k Nachkommastellen:

Aufgabe 4 – Begriffsklärung

Festkommazahl

Eine Festkommazahl ist eine Zahl, die aus einer festen Anzahl von Ziffern besteht. Die Position des Kommas ist dabei **fest** vorgegeben, daher der Name.
Dabei wird das polyadische System einfach fortgeführt, es gilt also bei einer n-stelligen B-adischen Festkommazahl Z mit k Nachkommastellen:

$$Z = \sum_{i=1}^{k} z_i \cdot B^{-i} + \sum_{i=0}^{(n-1)-k} z_i \cdot B^i = \sum_{i=-k}^{(n-1)-k} z_i \cdot B^i$$

Aufgabe 4 – Begriffsklärung

Umwandlung einer rationalen Zahl in eine FESTKOMMAzahl

Sei eine Zahl $(z, n) \in \mathbb{Q}$ gegeben, wobei $z \in \mathbb{Z}, n \in \mathbb{N}$. Diese soll nun als binäre Festkommazahl mit k Nachkommastellen dargestellt werden.

Aufgabe 4 – Begriffsklärung

Umwandlung einer rationalen Zahl in eine FESTKOMMAzahl

Sei eine Zahl $(z, n) \in \mathbb{Q}$ gegeben, wobei $z \in \mathbb{Z}, n \in \mathbb{N}$. Diese soll nun als binäre Festkommazahl mit k Nachkommastellen dargestellt werden.
Dann wandle zuerst die Zahl z um

Aufgabe 4 – Begriffsklärung

Umwandlung einer rationalen Zahl in eine FESTKOMMAzahl

Sei eine Zahl $(z, n) \in \mathbb{Q}$ gegeben, wobei $z \in \mathbb{Z}, n \in \mathbb{N}$. Diese soll nun als binäre Festkommazahl mit k Nachkommastellen dargestellt werden.
Dann wandle zuerst die Zahl z um und danach die Nachkommazahl n mit dem binären Verdopplungsverfahren.

Aufgabe 4 – Begriffsklärung

Umwandlung einer rationalen Zahl in eine FESTKOMMAzahl

Sei eine Zahl $(z, n) \in \mathbb{Q}$ gegeben, wobei $z \in \mathbb{Z}, n \in \mathbb{N}$. Diese soll nun als binäre Festkommazahl mit k Nachkommastellen dargestellt werden.
Dann wandle zuerst die Zahl z um und danach die Nachkommazahl n mit dem binären Verdopplungsverfahren.

Binäres Verdopplungsverfahren

Sei eine Zahl $(z, n) \in \mathbb{Q}$ gegeben, wobei $z \in \mathbb{Z}, n \in \mathbb{N}$. Betrachte nun nur n:

Aufgabe 4 – Begriffsklärung

Umwandlung einer rationalen Zahl in eine FESTKOMMAzahl

Sei eine Zahl $(z, n) \in \mathbb{Q}$ gegeben, wobei $z \in \mathbb{Z}, n \in \mathbb{N}$. Diese soll nun als binäre Festkommazahl mit k Nachkommastellen dargestellt werden.
Dann wandle zuerst die Zahl z um und danach die Nachkommazahl n mit dem binären Verdopplungsverfahren.

Binäres Verdopplungsverfahren

Sei eine Zahl $(z, n) \in \mathbb{Q}$ gegeben, wobei $z \in \mathbb{Z}, n \in \mathbb{N}$. Betrachte nun nur n:

Schritt 1: Multipliziere n mit 2, merke das Ergebnis e.

Aufgabe 4 – Begriffsklärung

Umwandlung einer rationalen Zahl in eine FESTKOMMAzahl

Sei eine Zahl $(z, n) \in \mathbb{Q}$ gegeben, wobei $z \in \mathbb{Z}, n \in \mathbb{N}$. Diese soll nun als binäre Festkommazahl mit k Nachkommastellen dargestellt werden.
Dann wandle zuerst die Zahl z um und danach die Nachkommazahl n mit dem binären Verdopplungsverfahren.

Binäres Verdopplungsverfahren

Sei eine Zahl $(z, n) \in \mathbb{Q}$ gegeben, wobei $z \in \mathbb{Z}, n \in \mathbb{N}$. Betrachte nun nur n:

Schritt 1: Multipliziere n mit 2, merke das Ergebnis e.

Schritt 2: Ist $e > 1$, so setze e^* auf $e - 1$, notiere binär eine 1.

Aufgabe 4 – Begriffsklärung

Umwandlung einer rationalen Zahl in eine FESTKOMMAzahl

Sei eine Zahl $(z, n) \in \mathbb{Q}$ gegeben, wobei $z \in \mathbb{Z}, n \in \mathbb{N}$. Diese soll nun als binäre Festkommazahl mit k Nachkommastellen dargestellt werden.
Dann wandle zuerst die Zahl z um und danach die Nachkommazahl n mit dem binären Verdopplungsverfahren.

Binäres Verdopplungsverfahren

Sei eine Zahl $(z, n) \in \mathbb{Q}$ gegeben, wobei $z \in \mathbb{Z}, n \in \mathbb{N}$. Betrachte nun nur n:

Schritt 1: Multipliziere n mit 2, merke das Ergebnis e.

Schritt 2: Ist $e > 1$, so setze e^* auf $e - 1$, notiere binär eine 1.

Schritt 3: Ist $e < 1$, so notiere binär eine 0.

Aufgabe 4 – Begriffsklärung

Umwandlung einer rationalen Zahl in eine FESTKOMMAzahl

Sei eine Zahl $(z, n) \in \mathbb{Q}$ gegeben, wobei $z \in \mathbb{Z}, n \in \mathbb{N}$. Diese soll nun als binäre Festkommazahl mit k Nachkommastellen dargestellt werden.
Dann wandle zuerst die Zahl z um und danach die Nachkommazahl n mit dem binären Verdopplungsverfahren.

Binäres Verdopplungsverfahren

Sei eine Zahl $(z, n) \in \mathbb{Q}$ gegeben, wobei $z \in \mathbb{Z}, n \in \mathbb{N}$. Betrachte nun nur n:

Schritt 1: Multipliziere n mit 2, merke das Ergebnis e.

Schritt 2: Ist $e > 1$, so setze e^* auf $e - 1$, notiere binär eine 1.

Schritt 3: Ist $e < 1$, so notiere binär eine 0.

Schritt 4: Ist die Anzahl an notierten Stellen identisch mit k oder $e = 1$, so breche ab.

Aufgabe 4 – Begriffsklärung

Umwandlung einer rationalen Zahl in eine FESTKOMMAzahl

Sei eine Zahl $(z, n) \in \mathbb{Q}$ gegeben, wobei $z \in \mathbb{Z}, n \in \mathbb{N}$. Diese soll nun als binäre Festkommazahl mit k Nachkommastellen dargestellt werden.
Dann wandle zuerst die Zahl z um und danach die Nachkommazahl n mit dem binären Verdopplungsverfahren.

Binäres Verdopplungsverfahren

Sei eine Zahl $(z, n) \in \mathbb{Q}$ gegeben, wobei $z \in \mathbb{Z}, n \in \mathbb{N}$. Betrachte nun nur n:

Schritt 1: Multipliziere n mit 2, merke das Ergebnis e.

Schritt 2: Ist $e > 1$, so setze e^* auf $e - 1$, notiere binär eine 1.

Schritt 3: Ist $e < 1$, so notiere binär eine 0.

Schritt 4: Ist die Anzahl an notierten Stellen identisch mit k oder $e = 1$, so breche ab.

Schritt 5: Sonst wiederhole Schritt 1.

Aufgabe 4 – Zahlenkonversion

c) Konvertieren Sie die Dezimalzahl $234,28125_{10}$ ins Binärformat. Verwenden Sie für die Nachkommastellen maximal 4 Bit.

Aufgabe 4 – Zahlenkonversion: Festkommazahlen

Unstabile Polygone bei der PSX:
https://www.youtube.com/watch?v=nqw2HMUrNiA

Aufgabe 4 – Zahlenkonversion: Festkommazahlen

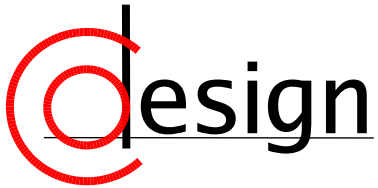
Unstabile Polygone bei der PSX: https://www.youtube.com/watch?v=nqw2HMUrNiA

Was ist das Problem?

... The real problem is that the PS1 didn't have a FPU; a co-processor for math dealing with real numbers called a Floating Point Unit, so consequently it used fixed point math that has limited precision [...]. The limited precision causes the polygon vertices themselves to jump around as the camera moves around the scene because there isn't enough bits to finely position the vertex ...

(Kommentar unter dem Video)

Aufgabe 5 – Konversion von Gleitkommazahlen

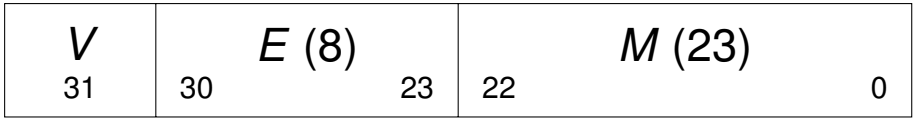


FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

TECHNISCHE FAKULTÄT

Aufgabe 5 – Konversion von Gleitkommazahlen

Das Format für Gleitkommazahlen im IEEE-Standard 754 einfacher Genauigkeit lautet:



a) Konvertieren Sie die folgenden nach obigem Standard codierten Zahlen in das Dezimalsystem:
   i) 0 1001 1001 1001 1001 1001 0000 0000 000
   i) 1 0001 1001 1001 1001 0000 0000 0000 000

b) Wandeln Sie die folgenden Zahlen in den obigen IEEE-Standard um:
   ii) $-6,25_{10} \cdot 10^{-3} = -0,000000\overline{0011}_2$
   ii) $3,14159_{10} = 11,0010\,0100\,0011\,1111\,0011\,11(1\ldots)_2$

c) Warum kann einer `float`-Variablen der Wert $1 \cdot 10^{-42}$, nicht aber der Wert $1 \cdot 10^{42}$ zugewiesen werden?

Aufgabe 5 – Begriffsklärung

IEEE

Das **Institute of Electrical and Electronics Engineers** (**IEEE**) ist ein globaler Verband von Ingenieuren – haupsächlich aus – der Elektro- und Informationstechnik. Mit ihr kommen verschiedene Gremien zur Standardisierung von Techniken, Hardware und Software zusammen.

Aufgabe 5 – Begriffsklärung

V	E (8)	M (23)
31	30 23	22 0

Gleitkommaarithmetik

Wie kann man eine Gleitkommazahl berechnen?
Es gilt:

$$Z = (-1)^V \cdot 2^{(E-B)} \cdot (1 + M)$$

- V beschreibt das Vorzeichenbit. Ist $V = 0$, so ist Z positiv, sonst negativ.
- E beschreibt den *biased exponent*[a]. Sie berechnet zusammen mit dem BIAS B den realen Exponenten.
- B beschreibt den BIAS. Er ist für Gleitkommazahlen fest und berechnet sich durch $B = 2^{|E|-1} - 1$. Mit dem BIAS ist es möglich negaive Exponenten darzustellen.
- M beschreibt die Nachkommastellen der Mantisse. Man berechnet die reale Mantisse mit $1 + M$.

[a] auch Charakteristik genannt

Aufgabe 5 – Begriffsklärung

V	E (8)	M (23)
31	30 23	22 0

Gleitkommaarithmetik – Sonderfälle

Wie kann man eine Gleitkommazahl berechnen?
Es gilt:

$$Z = (-1)^V \cdot 2^{(E-B)} \cdot (1 + M)$$

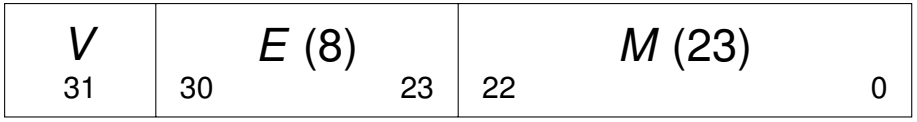
Allerdings ist bei der Berechnung von Gleitkommazahlen auf einige Sonderfälle zu achten:

E	M	Wert
$0 < E < 2^{\lvert E \rvert} - 1$	M	$(-1)^V \cdot 2^{(E-B)} \cdot (1 + M)$
$2^{\lvert E \rvert} - 1$	$\neq 0$	NaN (Not a Number)
	0	$\pm\infty$ (je nach Vorzeichenbit)
0	M	Denormalisierte Zahlen

Warum die Fehlerbehandlung wichtig sein kann ...

Aufgabe 5 – Konversion von Gleitkommazahlen

Das Format für Gleitkommazahlen im IEEE-Standard 754 einfacher Genauigkeit lautet:



a) Konvertieren Sie die folgenden nach obigem Standard codierten Zahlen in das Dezimalsystem:
   i) 0 1001 1001 1001 1001 1001 0000 0000 000
   i) 1 0001 1001 1001 1001 0000 0000 0000 000

Aufgabe 5 – Konversion von Gleitkommazahlen

Das Format für Gleitkommazahlen im IEEE-Standard 754 einfacher Genauigkeit lautet:

V	E (8)	M (23)
31	30 … 23	22 … 0

b) Wandeln Sie die folgenden Zahlen in den obigen IEEE-Standard um:

ii) $-6,25_{10} \cdot 10^{-3} = -0,000000\overline{0011}_2$

ii) $3,14159_{10} = 11,0010\,0100\,0011\,1111\,0011\,11(1\ldots)_2$

Aufgabe 5 – Algorithmus Dezimal $\mapsto$ IEEE

Sei die Dezimalzahl d_{10} gegeben:

Aufgabe 5 – Algorithmus Dezimal $\mapsto$ IEEE

Sei die Dezimalzahl d_{10} gegeben:

Schritt 1) Wandle d_{10} ins Binärsystem um

Aufgabe 5 – Algorithmus Dezimal $\mapsto$ IEEE

Sei die Dezimalzahl d_{10} gegeben:

Schritt 1) Wandle d_{10} ins Binärsystem um

Schritt 2) Normalisiere auf $(1, M)_2 \cdot 2^E$ und runde auf len(M) bit Nachkommastellen.

Aufgabe 5 – Algorithmus Dezimal $\mapsto$ IEEE

Sei die Dezimalzahl d_{10} gegeben:

Schritt 1) Wandle d_{10} ins Binärsystem um

Schritt 2) Normalisiere auf $(1, M)_2 \cdot 2^E$ und runde auf len(M) bit Nachkommastellen.

Schritt 3) Bestimme den *biased exponent*, den „voreingenommenen“ Exponenten.

Aufgabe 5 – Algorithmus Dezimal $\mapsto$ IEEE

Sei die Dezimalzahl d_{10} gegeben:

Schritt 1) Wandle d_{10} ins Binärsystem um

Schritt 2) Normalisiere auf $(1, M)_2 \cdot 2^E$ und runde auf len(M) bit Nachkommastellen.

Schritt 3) Bestimme den *biased exponent*, den „voreingenommenen“ Exponenten.

Schritt 4) Bestimme je nach Vorzeichen das Vorzeichenbit V und setze die Zahl zusammen.

Aufgabe 5 – Konversion von Gleitkommazahlen

c) Warum kann einer `float`-Variablen der Wert $1 \cdot 10^{-42}$, nicht aber der Wert $1 \cdot 10^{42}$ zugewiesen werden?

```
1 //...
2 public static void main(String[] args) {
3     DecimalIEEE dI = new DecimalIEEE((float) 1e-42); dI.printString();
4     DecimalIEEE dI2 = new DecimalIEEE((float) 1e42); dI2.printString();
5 }
6 //...
```

Aufgabe 5 – Konversion von Gleitkommazahlen

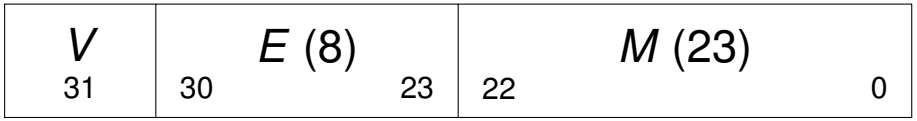
c) Warum kann einer `float`-Variablen der Wert $1 \cdot 10^{-42}$, nicht aber der Wert $1 \cdot 10^{42}$ zugewiesen werden?

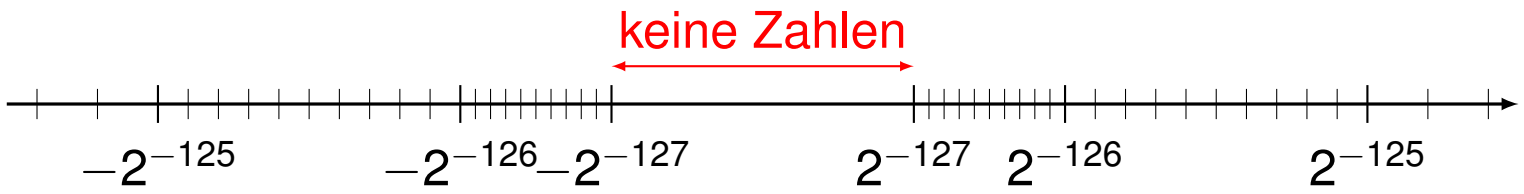
```
1 //...
2 public static void main(String[] args) {
3     DecimalIEEE dI = new DecimalIEEE((float) 1e-42); dI.printString();
4     DecimalIEEE dI2 = new DecimalIEEE((float) 1e42); dI2.printString();
5 }
6 //...
```

Ausgabe

```
usercca:~/gti$java DecimalIEEE
DecimalIEEE -- 00000000000000000000001011001010
    +1.0E-42
DecimalIEEE -- 01111111100000000000000000000000
    +Infinity
```

IEEE-Standard 754: Normalisierte Darstellung

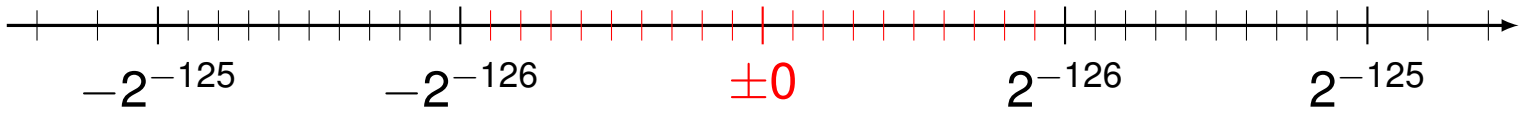




$$Z = (-1)^V \cdot 2^{(E-B)} \cdot (1 + M)$$

IEEE-Standard 754: Denormalisierte Darstellung

$\rightsquigarrow$ $E = 0$ heißt die Zahl ist denormalisiert.



$$Z = (-1)^V \cdot 2^{-126} \cdot (0 + M)$$