Übungen zur Grundlagen der Technischen Informatik

Übung 4 – Binär-, Hexadezimal- und Gleitkommaarithmetik

Florian Frank

Friedrich-Alexander-Universität Erlangen-Nürnberg

Wintersemester 2018/19

FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG
TECHNISCHE FAKULTÄT

Was machen wir heute?

Prüfungsanmeldung

Was machen wir heute?

Prüfungsanmeldung

Aufgabe 1 – Binär- und Hexadezimalarithmetik

Was machen wir heute?

Prüfungsanmeldung

Aufgabe 1 – Binär- und Hexadezimalarithmetik

Blatt 3, Aufgabe 5 – Konversion von Gleitkommazahlen

Was machen wir heute?

Prüfungsanmeldung

Aufgabe 1 – Binär- und Hexadezimalarithmetik

Blatt 3, Aufgabe 5 – Konversion von Gleitkommazahlen

Aufgabe 2 – Addition und Subtraktion von Gleitkommazahlen

Was machen wir heute?

Prüfungsanmeldung

Aufgabe 1 – Binär- und Hexadezimalarithmetik

Blatt 3, Aufgabe 5 – Konversion von Gleitkommazahlen

Aufgabe 2 – Addition und Subtraktion von Gleitkommazahlen

Aufgabe 3 – Multiplikation von Gleitkommazahlen

Was machen wir heute?

Prüfungsanmeldung

Aufgabe 1 – Binär- und Hexadezimalarithmetik

Blatt 3, Aufgabe 5 – Konversion von Gleitkommazahlen

Aufgabe 2 – Addition und Subtraktion von Gleitkommazahlen

Aufgabe 3 – Multiplikation von Gleitkommazahlen

Aufgabe 4 – Assoziativität bei Gleitkommadarstellung

Prüfungsanmeldung

FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG
TECHNISCHE FAKULTÄT

Achtung – Prüfungsanmeldung

Prüfungsanmeldung

Im Wintersemester 2018/19 findet die Anmeldung zu den Prüfungen des Semesters ab dem 19. November 2018, 0:01 Uhr über „mein campus“ statt. Die Prüfungsanmeldephase endet am 07. Dezember um 12:00 Uhr.
Ihr müsst euch für die Prüfungen anmelden, an denen ihr teilnehmen wollt.
Nach dem 07. Dezember um 12:00 Uhr ist **keine** Anmeldung mehr möglich!

Achtung – Prüfungsanmeldung

Prüfungsanmeldung

Im Wintersemester 2018/19 findet die Anmeldung zu den Prüfungen des Semesters ab dem 19. November 2018, 0:01 Uhr über „mein campus“ statt. Die Prüfungsanmeldephase endet am 07. Dezember um 12:00 Uhr.
Ihr müsst euch für die Prüfungen anmelden, an denen ihr teilnehmen wollt.
Nach dem 07. Dezember um 12:00 Uhr ist **keine** Anmeldung mehr möglich!

Achtung – Prüfungsanmeldung

Prüfungsanmeldung

Im Wintersemester 2018/19 findet die Anmeldung zu den Prüfungen des Semesters ab dem 19. November 2018, 0:01 Uhr über „mein campus“ statt. Die Prüfungsanmeldephase endet am 07. Dezember um 12:00 Uhr.
Ihr müsst euch für die Prüfungen anmelden, an denen ihr teilnehmen wollt.
Nach dem 07. Dezember um 12:00 Uhr ist **keine** Anmeldung mehr möglich!

Achtung – Prüfungsanmeldung

Prüfungsanmeldung

Im Wintersemester 2018/19 findet die Anmeldung zu den Prüfungen des Semesters ab dem 19. November 2018, 0:01 Uhr über „mein campus" statt. Die Prüfungsanmeldephase endet am 07. Dezember um 12:00 Uhr.

Ihr müsst euch für die Prüfungen anmelden, an denen ihr teilnehmen wollt.

Nach dem 07. Dezember um 12:00 Uhr ist **keine** Anmeldung mehr möglich!


Prüfungsan- und abmeldung

Wichtige Informationen zur Prüfungsanmeldung. Bitte sorgfältig lesen!

Bitte beachten Sie:

Sie sind verpflichtet, die ordnungsgemäße Erfassung Ihrer An- oder Abmeldung rechtzeitig vor den Prüfungen durch Einsichtnahme in *mein campus* zu kontrollieren. Setzen Sie sich bei Unstimmigkeiten bitte sofort mit dem Prüfungsamt in Verbindung und legen Sie den Ausdruck über die angemeldeten Prüfungen vor.

Die Anmeldung zu den einzelnen Modulprüfungen oder Modulteilprüfungen erfolgt gegebenenfalls unter dem Vorbehalt einer endgültig verabschiedeten Prüfungsordnung.

Die Anmeldungen zu Modulen, die als Schlüsselqualifikationen gewertet werden sollen, erfolgt seit diesem Semester ebenfalls über *mein campus*.
Nicht bestandene Prüfungen sind innerhalb der in der geltenden Prüfungsordnung genannten Fristen zu wiederholen.
Die Anmeldung zur Wiederholungsprüfung ist zu kontrollieren. Soweit eine Anmeldung für die Wiederholungsprüfung nicht eingetragen ist und Sie sich auch nicht selbst für diese Prüfung über *mein campus* anmelden können, verständigen Sie bitte das Prüfungsamt.
Für Wiederholungen gelten die Vorschriften der jeweiligen Prüfungsordnung.

Hiermit erkläre ich, dass ich in dem oder einem verwandten Studiengang, für den Prüfungen angemeldet werden (Diplom-, Bachelor- oder Master- Studium), noch keine Prüfung endgültig nicht bestanden habe bzw. unter Verlust des Prüfungsanspruches exmatrikuliert wurde. Soweit ich mich in einem schwebenden Prüfungsverfahren an einer anderen Hochschule befinde, lege ich umgehend eine Bestätigung der bisherigen Hochschule im Prüfungsamt vor. Bis zu diesem Zeitpunkt erfolgt die Anmeldung bzw. Zulassung zu Prüfungen unter Vorbehalt.

Von den Bestimmungen der Prüfungsordnungen habe ich Kenntnis genommen.

Erst wenn Sie diesen Hinweis durch Anklicken des unten stehenden Feldes akzeptiert haben, können Sie mit Ihren gewünschten Aktionen fortfahren. Klicken Sie dazu mit der linken Maustaste auf den "Weiter"-Button und wählen anschließend den entsprechenden Link aus.

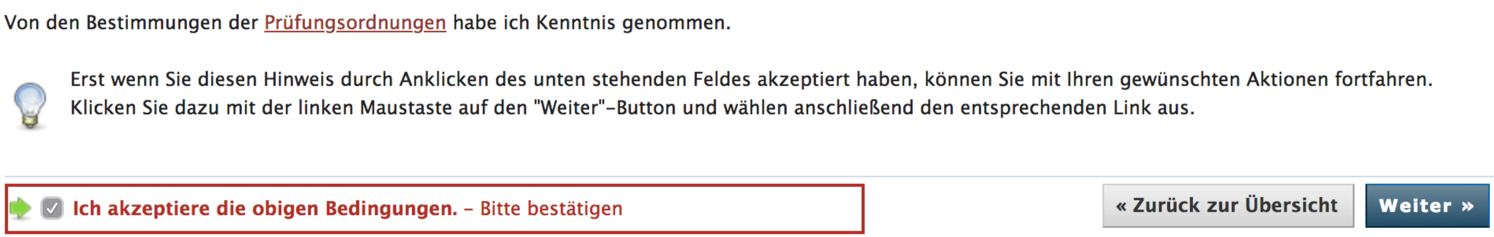
Ich akzeptiere die obigen Bedingungen. – Bitte bestätigen

« Zurück zur Übersicht | Weiter »

Achtung – Prüfungsanmeldung

Prüfungsanmeldung

Im Wintersemester 2018/19 findet die Anmeldung zu den Prüfungen des Semesters ab dem 19. November 2018, 0:01 Uhr über „mein campus“ statt. Die Prüfungsanmeldephase endet am 07. Dezember um 12:00 Uhr.
Ihr müsst euch für die Prüfungen anmelden, an denen ihr teilnehmen wollt.
Nach dem 07. Dezember um 12:00 Uhr ist **keine** Anmeldung mehr möglich!


Von den Bestimmungen der Prüfungsordnungen habe ich Kenntnis genommen.

Erst wenn Sie diesen Hinweis durch Anklicken des unten stehenden Feldes akzeptiert haben, können Sie mit Ihren gewünschten Aktionen fortfahren.
Klicken Sie dazu mit der linken Maustaste auf den "Weiter"-Button und wählen anschließend den entsprechenden Link aus.

Ich akzeptiere die obigen Bedingungen. – Bitte bestätigen

« Zurück zur Übersicht

Weiter »

Achtung – Prüfungsanmeldung

Prüfungsanmeldung

Im Wintersemester 2018/19 findet die Anmeldung zu den Prüfungen des Semesters ab dem 19. November 2018, 0:01 Uhr über „mein campus“ statt. Die Prüfungsanmeldephase endet am 07. Dezember um 12:00 Uhr.

Ihr müsst euch für die Prüfungen anmelden, an denen ihr teilnehmen wollt.

Nach dem 07. Dezember um 12:00 Uhr ist **keine** Anmeldung mehr möglich!

3020 Implementierung von Datenbanksystemen

TEC 30201 Implementierung von Datenbanksystemen (ECTS: 5.0)

Datum: 16.02.2018, Prüfer: Richard Lenz, **Prüfung anmelden**

Aufgabe 1 – Binär- und Hexadezimalarithmetik

FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG
TECHNISCHE FAKULTÄT

Aufgabe 1 – Binär- und Hexadezimalarithmetik

Führen Sie die folgenden Berechnungen im angegebenen Zahlensystem aus, ohne die Zahlen ins Dezimalsystem umzuwandeln.

a) Addition im Binär- und Hexadezimalsystem:

$$111010100110_2 + 010101111110_2$$
$$\mathrm{B674FC12}_{16} + \mathrm{2DA9D4B2}_{16}$$

Aufgabe 1 – Binär- und Hexadezimalarithmetik

Wie können wir zwei Zahlen in einem fremdem System zur Basis *a* addieren?
Erinnerung, wie haben wir das Addieren in der Grundschule gelernt?

$$\begin{array}{rr} & 5\,9 \\ + & 6\,2 \\ \hline = & \end{array}$$

Aufgabe 1 – Binär- und Hexadezimalarithmetik

Wie können wir zwei Zahlen in einem fremdem System zur Basis *a* addieren?
Erinnerung, wie haben wir das Addieren in der Grundschule gelernt?

$$
\begin{array}{rrr}
 & 5 & 9 \\
+ & 6 & 2 \\
{\scriptstyle +} & {\scriptstyle 1} & \\
\hline
= & & 1
\end{array}
$$

Aufgabe 1 – Binär- und Hexadezimalarithmetik

Wie können wir zwei Zahlen in einem fremdem System zur Basis *a* addieren?
Erinnerung, wie haben wir das Addieren in der Grundschule gelernt?

$$
\begin{array}{cccc}
 & & 5 & 9 \\
+ & & 6 & 2 \\
{\scriptstyle +} & {\scriptstyle 1} & {\scriptstyle 1} & \\
\hline
= & & 2 & 1
\end{array}
$$

Aufgabe 1 – Binär- und Hexadezimalarithmetik

Wie können wir zwei Zahlen in einem fremdem System zur Basis *a* addieren?
Erinnerung, wie haben wir das Addieren in der Grundschule gelernt?

$$
\begin{array}{cccc}
 & & 5 & 9 \\
+ & & 6 & 2 \\
\scriptstyle + & \scriptstyle 1 & \scriptstyle 1 & \\
\hline
= & 1 & 2 & 1
\end{array}
$$

Aufgabe 1 – Binär- und Hexadezimalarithmetik

Wie können wir zwei Zahlen in einem fremdem System zur Basis *a* addieren?
Erinnerung, wie haben wir das Addieren in der Grundschule gelernt?

$$
\begin{array}{rccc}
 & & 5 & 9 \\
+ & & 6 & 2 \\
{\scriptstyle +} & {\scriptstyle 1} & {\scriptstyle 1} & \\
\hline
= & 1 & 2 & 1
\end{array}
$$

(1) Addiere jeweils **stellenweise**!

Aufgabe 1 – Binär- und Hexadezimalarithmetik

Wie können wir zwei Zahlen in einem fremdem System zur Basis *a* addieren?
Erinnerung, wie haben wir das Addieren in der Grundschule gelernt?

$$
\begin{array}{rccc}
 & & 5 & 9 \\
+ & & 6 & 2 \\
{\scriptstyle +} & {\scriptstyle 1} & {\scriptstyle 1} & \\
\hline
= & 1 & 2 & 1
\end{array}
$$

(1) Addiere jeweils **stellenweise**!
Nimm auch notfalls die Hände dazu ...

Aufgabe 1 – Binär- und Hexadezimalarithmetik

Wie können wir zwei Zahlen in einem fremdem System zur Basis *a* addieren?
Erinnerung, wie haben wir das Addieren in der Grundschule gelernt?

$$\begin{array}{rrrr} & & 5 & 9 \\ + & & 6 & 2 \\ {\scriptstyle +} & {\scriptstyle 1} & {\scriptstyle 1} & \\ \hline = & 1 & 2 & 1 \end{array}$$

(1) Addiere jeweils **stellenweise**!
Nimm auch notfalls die Hände dazu ...

(2) Hat das Ergebnis der Addition **mehrere** Stellen, so schreibe nur die letzte Stelle nieder!

Aufgabe 1 – Binär- und Hexadezimalarithmetik

Wie können wir zwei Zahlen in einem fremdem System zur Basis *a* addieren?
Erinnerung, wie haben wir das Addieren in der Grundschule gelernt?

$$
\begin{array}{rccc}
 & & 5 & 9 \\
+ & & 6 & 2 \\
{\scriptstyle +} & {\scriptstyle 1} & {\scriptstyle 1} & \\
\hline
= & 1 & 2 & 1
\end{array}
$$

(1) Addiere jeweils **stellenweise**!
Nimm auch notfalls die Hände dazu ...

(2) Hat das Ergebnis der Addition **mehrere** Stellen, so schreibe nur die letzte Stelle nieder!

(3) Der Rest *e* wandert eine Stelle weiter als Übertrag (*carry*)

Aufgabe 1 – Binär- und Hexadezimalarithmetik

Wie können wir zwei Zahlen in einem fremdem System zur Basis *a* addieren?
Erinnerung, wie haben wir das Addieren in der Grundschule gelernt?

$$
\begin{array}{rrrr}
 & & 5 & 9 \\
+ & & 6 & 2 \\
{\scriptstyle +} & {\scriptstyle 1} & {\scriptstyle 1} & \\
\hline
= & 1 & 2 & 1
\end{array}
$$

(1) Addiere jeweils **stellenweise**!
Nimm auch notfalls die Hände dazu ...

(2) Hat das Ergebnis der Addition **mehrere** Stellen, so schreibe nur die letzte Stelle nieder!

(3) Der Rest *e* wandert eine Stelle weiter als Übertrag (*carry*)

(4) Berechne die nächste Stelle dann durch Addition der einzelnen Stellen mit dem Übertrag von der letzten Stelle.

Aufgabe 1 – Binär- und Hexadezimalarithmetik

Wie können wir zwei Zahlen in einem fremdem System zur Basis *a* addieren?
Erinnerung, wie haben wir das Addieren in der Grundschule gelernt?

$$
\begin{array}{rr}
 & 5\,9 \\
+ & 6\,2 \\
{\scriptstyle +} & {\scriptstyle 1\;\;1\;\;} \\
\hline
= & 1\,2\,1
\end{array}
$$

(1) Addiere jeweils **stellenweise**!
Nimm auch notfalls die Hände dazu ...

(2) Hat das Ergebnis der Addition **mehrere** Stellen, so schreibe nur die letzte Stelle nieder!

(3) Der Rest *e* wandert eine Stelle weiter als Übertrag (*carry*)

(4) Berechne die nächste Stelle dann durch Addition der einzelnen Stellen mit dem Übertrag von der letzten Stelle.

(5) Wiederhole diesen Prozess, bis man beim Ende der Ziffern angekommen ist.

Aufgabe 1 – Binär- und Hexadezimalarithmetik

Führen Sie die folgenden Berechnungen im angegebenen Zahlensystem aus, ohne die Zahlen ins Dezimalsystem umzuwandeln.

a) Addition im Binär- und Hexadezimalsystem:

$$111010100110_2 + 010101111110_2$$

$$\mathrm{B674FC12}_{16} + \mathrm{2DA9D4B2}_{16}$$

Einschub — BCD-Addition

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.
Neben den 10 gültigen Repräsentanten gibt es noch 6 sogenannte Pseudotetraden.

Pseudotetraden

Binärdarstellung von nicht dezimalen Ziffern, also 1010 bis 1111 (10 bis 15).

Die Addition im BCD-System funktioniert im Prinzip genau so wie die Addition bei Binärzahlen, jedoch mit folgenden **wesentlichen** Ausnahmen:

- Stellen (je 4 bit) werden getrennt addiert und der Übertrag in die nächste Stelle gezogen werden.
- Pseudotetraden müssen korrigiert werden.
- Ein Überlauf muss korrigiert werden, wenn er nicht durch einen Korrekturschritt verursacht wurde.
- Ein Korrekturschritt ist die Addtion von 6 (0110_{BCD})

Einschub — BCD-Addition: Beispiel

Berechne $775_{10} + 498_{10}$:

		0	1	1	1	0	1	1	1	0	1	0	1	(775_{10})
+		0	1	0	0	1	0	0	1	1	0	0	0	(498_{10})
+		1	1	1	1	1	1	1						
=		1	1	0	0	0	0	0	0	1	1	0	1	Pseudotetraden
+		0	1	1	0					0	1	1	0	Korrektur
+	1	1							1	1				
	1	$\boxed{0}$	$\boxed{0}$	$\boxed{1}$	$\boxed{0}$	$\boxed{0}$	$\boxed{0}$	$\boxed{0}$	$\boxed{1}$	0	0	1	1	Überlauf
+						0	1	1	0					Korrektur
	1	0	0	1	0	0	1	1	1	0	0	1	1	(1273_{10})

Keine Korrektur
Verursacht durch Korrekturschritt

Korrektur
Überlauf

Aufgabe 1 – Binär- und Hexadezimalarithmetik

Führen Sie die folgenden Berechnungen im angegebenen Zahlensystem aus, ohne die Zahlen ins Dezimalsystem umzuwandeln.

b) Subtraktion im Binärsystem:

$$11010010_2 - 10110101_2$$
$$01110110_2 - 10011001_2$$

Aufgabe 1 – Binär- und Hexadezimalarithmetik

Führen Sie die folgenden Berechnungen im angegebenen Zahlensystem aus, ohne die Zahlen ins Dezimalsystem umzuwandeln.

c) Multiplikation im Binärsystem:

$$11101010_2 \cdot 1011_2$$

Aufgabe 1 – Binär- und Hexadezimalarithmetik

Führen Sie die folgenden Berechnungen im angegebenen Zahlensystem aus, ohne die Zahlen ins Dezimalsystem umzuwandeln.

d) Prozessoren arbeiten mit einer endlichen Wortbreite, was den darstellbaren Zahlenbereich einschränkt. Welche Auswirkungen kann dies bei der Subtraktion und Addition haben? Können Sie sich vorstellen, wie diese Probleme in der Praxis gelöst werden?

Blatt 3, Aufgabe 5 – Konversion von Gleitkomma-zahlen

design

FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG
TECHNISCHE FAKULTÄT

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

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) $-6{,}25_{10} \cdot 10^{-3} = -0{,}00000\overline{0011}_2$
   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 31	E (8) 30 … 23	M (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:

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

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) $-6{,}25_{10} \cdot 10^{-3} = -0{,}00000\overline{0011}_2$

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 $\mathrm{len}(M)$ 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 $\text{len}(M)$ 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 $\mathrm{len}(M)$ 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?

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

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?

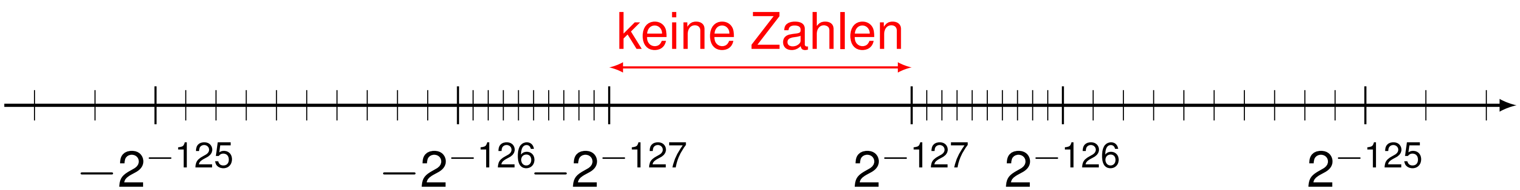
```java
//...
public static void main(String[] args) {
    DecimalIEEE dI = new DecimalIEEE((float) 1e-42); dI.printString();
    DecimalIEEE dI2 = new DecimalIEEE((float) 1e42); dI2.printString();
}
//...
```

Ausgabe

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

IEEE-Standard 754: Normalisierte Darstellung

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

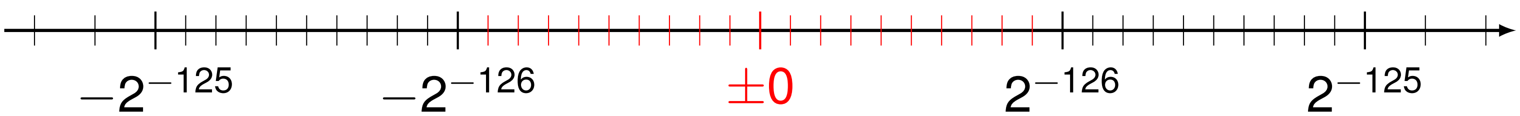


$$Z = (-1)^V \cdot 2^{(E-B)} \cdot (1 + M)$$

IEEE-Standard 754: Denormalisierte Darstellung

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

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



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

Aufgabe 2 – Addition und Subtraktion von Gleitkommazahlen

FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG
TECHNISCHE FAKULTÄT

Aufgabe 2 – Addition und Subtraktion von Gleitkommazahlen

Seien folgende Zahlen im IEEE-Standard 754 einfacher Genauigkeit gegeben:

$$x_1 = 0 \quad 1001\,1001 \quad 0000\,1001\,1001\,0000\,0000\,000$$
$$x_2 = 0 \quad 1001\,1001 \quad 1111\,1111\,0000\,0000\,0000\,000$$
$$x_3 = 1 \quad 1001\,0100 \quad 0011\,0110\,0000\,0000\,0000\,000$$

1. Berechnen Sie $x_4 = x_1 + x_2$.
2. Berechnen Sie $x_5 = x_1 + x_3$.

Aufgabe 2 – Add. und Sub. von GKZ – Verfahrensidee

Wie könnte man zwei Zahlen $x_1 = (-1)^{V_1} \cdot 2^{(E_1 - B)} \cdot (M_1)$ und $x_2 = (-1)^{V_2} \cdot 2^{(E_2 - B)} \cdot (M_2)$ addieren?

Aufgabe 2 – Add. und Sub. von GKZ – Verfahrensidee

Wie könnte man zwei Zahlen $x_1 = (-1)^{V_1} \cdot 2^{(E_1 - B)} \cdot (M_1)$ und $x_2 = (-1)^{V_2} \cdot 2^{(E_2 - B)} \cdot (M_2)$ addieren?

(1) Wir „machen“ die Exponenten „gleich“.
Dazu passe durch Rechtsschieben den Exponenten der kleineren Zahl auf den Exponenten der größeren an.

Aufgabe 2 – Add. und Sub. von GKZ – Verfahrensidee

Wie könnte man zwei Zahlen $x_1 = (-1)^{V_1} \cdot 2^{(E_1-B)} \cdot (M_1)$ und $x_2 = (-1)^{V_2} \cdot 2^{(E_2-B)} \cdot (M_2)$ addieren?

(1) Wir „machen“ die Exponenten „gleich“.
Dazu passe durch Rechtsschieben den Exponenten der kleineren Zahl auf den Exponenten der größeren an.

(2) Wir addieren/subtrahieren die Mantissen ($\rightsquigarrow$ bilde dazu – falls notwendig – das Zweierkomplement).
Ist das Ergebnis < 0, setze das Vorzeichen bit des Ergebnis und bilde das Zweierkomplement.

Aufgabe 2 – Add. und Sub. von GKZ – Verfahrensidee

Wie könnte man zwei Zahlen $x_1 = (-1)^{V_1} \cdot 2^{(E_1 - B)} \cdot (M_1)$ und $x_2 = (-1)^{V_2} \cdot 2^{(E_2 - B)} \cdot (M_2)$ addieren?

(1) Wir „machen“ die Exponenten „gleich“.
Dazu passe durch Rechtsschieben den Exponenten der kleineren Zahl auf den Exponenten der größeren an.

(2) Wir addieren/subtrahieren die Mantissen ($\leadsto$ bilde dazu – falls notwendig – das Zweierkomplement).
Ist das Ergebnis < 0, setze das Vorzeichen bit des Ergebnis und bilde das Zweierkomplement.

(3) Normalisiere das Ergebnis. Dazu:

Aufgabe 2 – Add. und Sub. von GKZ – Verfahrensidee

Wie könnte man zwei Zahlen $x_1 = (-1)^{V_1} \cdot 2^{(E_1 - B)} \cdot (M_1)$ und $x_2 = (-1)^{V_2} \cdot 2^{(E_2 - B)} \cdot (M_2)$ addieren?

(1) Wir „machen“ die Exponenten „gleich“.
Dazu passe durch Rechtsschieben den Exponenten der kleineren Zahl auf den Exponenten der größeren an.

(2) Wir addieren/subtrahieren die Mantissen ($\rightsquigarrow$ bilde dazu – falls notwendig – das Zweierkomplement).
Ist das Ergebnis < 0, setze das Vorzeichen bit des Ergebnis und bilde das Zweierkomplement.

(3) Normalisiere das Ergebnis. Dazu:

   A) Falls das Ergebnis $\geqslant 2$, schiebe das Ergebnis um eins (ggf. mit Rundung) nach rechts und inkrementiere den Exponenten.

Aufgabe 2 – Add. und Sub. von GKZ – Verfahrensidee

Wie könnte man zwei Zahlen $x_1 = (-1)^{V_1} \cdot 2^{(E_1 - B)} \cdot (M_1)$ und $x_2 = (-1)^{V_2} \cdot 2^{(E_2 - B)} \cdot (M_2)$ addieren?

(1) Wir „machen“ die Exponenten „gleich“.
Dazu passe durch Rechtsschieben den Exponenten der kleineren Zahl auf den Exponenten der größeren an.

(2) Wir addieren/subtrahieren die Mantissen ($\rightsquigarrow$ bilde dazu – falls notwendig – das Zweierkomplement).
Ist das Ergebnis < 0, setze das Vorzeichen bit des Ergebnis und bilde das Zweierkomplement.

(3) Normalisiere das Ergebnis. Dazu:
   - A) Falls das Ergebnis $\geqslant 2$, schiebe das Ergebnis um eins (ggf. mit Rundung) nach rechts und inkrementiere den Exponenten.
   - B) Falls das Ergebnis < 1, so schiebe das Ergebnis um ein nach links und dekrementiere den Exponenten.

Aufgabe 2 – Add. und Sub. von GKZ – Verfahrensidee

Wie könnte man zwei Zahlen $x_1 = (-1)^{V_1} \cdot 2^{(E_1-B)} \cdot (M_1)$ und $x_2 = (-1)^{V_2} \cdot 2^{(E_2-B)} \cdot (M_2)$ addieren?

(1) Wir „machen“ die Exponenten „gleich“.
Dazu passe durch Rechtsschieben den Exponenten der kleineren Zahl auf den Exponenten der größeren an.

(2) Wir addieren/subtrahieren die Mantissen ($\rightsquigarrow$ bilde dazu – falls notwendig – das Zweierkomplement).
Ist das Ergebnis < 0, setze das Vorzeichen bit des Ergebnis und bilde das Zweierkomplement.

(3) Normalisiere das Ergebnis. Dazu:

A) Falls das Ergebnis $\geqslant 2$, schiebe das Ergebnis um eins (ggf. mit Rundung) nach rechts und inkrementiere den Exponenten.

B) Falls das Ergebnis < 1, so schiebe das Ergebnis um ein nach links und dekrementiere den Exponenten.

$\Rightarrow$ Wiederhole A) und B) solange bis das Ergebnis entweder = 0 oder $1 \leqslant$ Ergebnis < 2.

Aufgabe 2 – Add. und Sub. von GKZ – Verfahrensidee

Wie könnte man zwei Zahlen $x_1 = (-1)^{V_1} \cdot 2^{(E_1 - B)} \cdot (M_1)$ und $x_2 = (-1)^{V_2} \cdot 2^{(E_2 - B)} \cdot (M_2)$ addieren?

(1) Wir „machen“ die Exponenten „gleich“.
Dazu passe durch Rechtsschieben den Exponenten der kleineren Zahl auf den Exponenten der größeren an.

(2) Wir addieren/subtrahieren die Mantissen ($\leadsto$ bilde dazu – falls notwendig – das Zweierkomplement).
Ist das Ergebnis < 0, setze das Vorzeichen bit des Ergebnis und bilde das Zweierkomplement.

(3) Normalisiere das Ergebnis. Dazu:
  - A) Falls das Ergebnis $\geqslant 2$, schiebe das Ergebnis um eins (ggf. mit Rundung) nach rechts und inkrementiere den Exponenten.
  - B) Falls das Ergebnis < 1, so schiebe das Ergebnis um ein nach links und dekrementiere den Exponenten.
  - $\Rightarrow$ Wiederhole A) und B) solange bis das Ergebnis entweder $= 0$ oder $1 \leqslant$ Ergebnis < 2.

(4) Behandle auftretende Sonderfälle – wie zum Beispiel einen Über-/Unterlauf oder Null.

Aufgabe 3 – Multiplikation von Gleitkommazahlen

FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

TECHNISCHE FAKULTÄT

Aufgabe 3 – Multiplikation von Gleitkommazahlen

Sei folgendes Format für Gleitkommazahlen gegeben, das analog zum IEEE-Standard 754 definiert ist:

V	E (7)	M (8)
15	14 … 8	7 … 0

Multiplizieren Sie folgende Gleitkommazahlen dieses Formats miteinander:

$$x_1 = 0 \;\; 1001\,000 \;\; 1001\,1011$$
$$x_2 = 1 \;\; 1001\,010 \;\; 1110\,1000$$

Aufgabe 3 – Mult./Division von GKZ – Verfahrensidee

Wie könnte man zwei Zahlen $x_1 = (-1)^{V_1} \cdot 2^{(E_1 - B)} \cdot (M_1)$ und $x_2 = (-1)^{V_2} \cdot 2^{(E_2 - B)} \cdot (M_2)$ multiplizieren?

Aufgabe 3 – Mult./Division von GKZ – Verfahrensidee

Wie könnte man zwei Zahlen $x_1 = (-1)^{V_1} \cdot 2^{(E_1 - B)} \cdot (M_1)$ und $x_2 = (-1)^{V_2} \cdot 2^{(E_2 - B)} \cdot (M_2)$ multiplizieren?

(1) Wir multiplizieren/dividieren die Mantissen.
Hier ist aufzupasen, ob normalisierte oder denormalisierte Zahlen vorliegen.

Aufgabe 3 – Mult./Division von GKZ – Verfahrensidee

Wie könnte man zwei Zahlen $x_1 = (-1)^{V_1} \cdot 2^{(E_1 - B)} \cdot (M_1)$ und $x_2 = (-1)^{V_2} \cdot 2^{(E_2 - B)} \cdot (M_2)$ multiplizieren?

(1) Wir multiplizieren/dividieren die Mantissen.
Hier ist aufzupasen, ob normalisierte oder denormalisierte Zahlen vorliegen.

(2) Wir addieren/subtrahieren die „voreingenommenen" Exponenten ($\rightsquigarrow$ bilde dazu – falls notwendig – das Zweierkomplement).
Vorsicht: Der Bias kommt als Hilfsgröße in beiden Faktoren vor, deshalb wird der Bias einmal abgezogen/hinzuaddiert

Aufgabe 3 – Mult./Division von GKZ – Verfahrensidee

Wie könnte man zwei Zahlen $x_1 = (-1)^{V_1} \cdot 2^{(E_1 - B)} \cdot (M_1)$ und $x_2 = (-1)^{V_2} \cdot 2^{(E_2 - B)} \cdot (M_2)$ multiplizieren?

(1) Wir multiplizieren/dividieren die Mantissen.
Hier ist aufzupasen, ob normalisierte oder denormalisierte Zahlen vorliegen.

(2) Wir addieren/subtrahieren die „voreingenommenen" Exponenten ($\rightsquigarrow$ bilde dazu – falls notwendig – das Zweierkomplement).
Vorsicht: Der Bias kommt als Hilfsgröße in beiden Faktoren vor, deshalb wird der Bias einmal abgezogen/hinzuaddiert

(3) Normalisiere die Mantisse.
Binge die Mantisse auf die Form $1, M$ und verschiebe ggf. den Exponenten.

Aufgabe 3 – Mult./Division von GKZ – Verfahrensidee

Wie könnte man zwei Zahlen $x_1 = (-1)^{V_1} \cdot 2^{(E_1 - B)} \cdot (M_1)$ und $x_2 = (-1)^{V_2} \cdot 2^{(E_2 - B)} \cdot (M_2)$ multiplizieren?

(1) Wir multiplizieren/dividieren die Mantissen.
Hier ist aufzupasen, ob normalisierte oder denormalisierte Zahlen vorliegen.

(2) Wir addieren/subtrahieren die „voreingenommenen" Exponenten ($\rightsquigarrow$ bilde dazu – falls notwendig – das Zweierkomplement).
Vorsicht: Der Bias kommt als Hilfsgröße in beiden Faktoren vor, deshalb wird der Bias einmal abgezogen/hinzuaddiert

(3) Normalisiere die Mantisse.
Binge die Mantisse auf die Form $1, M$ und verschiebe ggf. den Exponenten.

(4) Behandle die Vorzeichen getrennt. Dazu wende ein XOR an!
$1 \oplus 0 = 1$ und $0 \oplus 0 = 0$ ...

Aufgabe 3 – Mult./Division von GKZ – Verfahrensidee

Wie könnte man zwei Zahlen $x_1 = (-1)^{V_1} \cdot 2^{(E_1-B)} \cdot (M_1)$ und $x_2 = (-1)^{V_2} \cdot 2^{(E_2-B)} \cdot (M_2)$ multiplizieren?

(1) Wir multiplizieren/dividieren die Mantissen.
Hier ist aufzupasen, ob normalisierte oder denormalisierte Zahlen vorliegen.

(2) Wir addieren/subtrahieren die „voreingenommenen“ Exponenten ($\rightsquigarrow$ bilde dazu – falls notwendig – das Zweierkomplement).
Vorsicht: Der Bias kommt als Hilfsgröße in beiden Faktoren vor, deshalb wird der Bias einmal abgezogen/hinzuaddiert

(3) Normalisiere die Mantisse.
Binge die Mantisse auf die Form $1, M$ und verschiebe ggf. den Exponenten.

(4) Behandle die Vorzeichen getrennt. Dazu wende ein XOR an!
$1 \oplus 0 = 1$ und $0 \oplus 0 = 0$ ...

(5) Behandle auftretende Sonderfälle – wie zum Beispiel einen Über-/Unterlauf, Null oder auch NaN.

Aufgabe 4 – Assoziativität bei Gleitkommadarstellung

Aufgabe 4 – Assoziativität bei Gleitkommadarstellung

Addition und Multiplikation von Operanden in einer Gleitkommadarstellung sind im Allgemeinen **nicht** assoziativ.
Belegen Sie dies, indem Sie mit einem PC-Tabellenkalkulationsprogramm experimentieren. Bestimmen Sie dadurch auch die Anzahl der Bits, die zur Speicherung der Mantisse verwendet wird.
(Beispiel: LibreOffice und Excel liefern bei der Berechnung von:
$10^{20} + 17 - 10 - 10^{20} + 130$ als Ergebnis 130 – dies übrigens ohne irgendeine Warnung.)

Aufgabe 4 – Assoziativität bei Gleitkommadarstellung

Sowohl Excel 2013 als auch LibreOffice 4.2 liefern zum Beispiel folgende Ergebnisse:

	Formel	Ergebnis
A	$10^{20} - 10^{20} + 17 - 10 + 130$	$= 137$
B	$10^{15} - 10^{15} + 17 - 10 + 130$	$= 137$
C	$10^{16} - 10^{16} + 17 - 10 + 130$	$= 137$
D	$10^{20} + 17 - 10 - 10^{20} + 130$	$= 130$
E	$10^{15} + 17 - 10 - 10^{15} + 130$	$= 137$
F	$10^{16} + 17 - 10 - 10^{16} + 130$	$= 136$
G	$10^{20} + 17 + 130 - 10^{20} - 10$	$= -10$
H	$10^{15} + 17 + 130 - 10^{15} - 10$	$= 137$
I	$10^{16} + 17 + 130 - 10^{16} - 10$	$= 136$

Wo liegen die Probleme?

Aufgabe 4 – Assoziativität bei Gleitkommadarstellung

Sowohl Excel 2013 als auch LibreOffice 4.2 liefern zum Beispiel folgende Ergebnisse:

	Formel	Ergebnis
$\vdots$	$\vdots$	$\vdots$
D	$10^{20} + 17 - 10 - 10^{20} + 130$	$= 130$
$\vdots$	$\vdots$	$\vdots$
G	$10^{20} + 17 + 130 - 10^{20} - 10$	$= -10$
$\vdots$	$\vdots$	$\vdots$

Aufgabe 4 – Assoziativität bei Gleitkommadarstellung

Sowohl Excel 2013 als auch LibreOffice 4.2 liefern zum Beispiel folgende Ergebnisse:

	Formel	Ergebnis
$\vdots$	$\vdots$	$\vdots$
D	$10^{20} + 17 - 10 - 10^{20} + 130$	$= 130$
$\vdots$	$\vdots$	$\vdots$
G	$10^{20} + 17 + 130 - 10^{20} - 10$	$= -10$
$\vdots$	$\vdots$	$\vdots$

Problem bei D und G

Aufgabe 4 – Assoziativität bei Gleitkommadarstellung

Sowohl Excel 2013 als auch LibreOffice 4.2 liefern zum Beispiel folgende Ergebnisse:

	Formel	Ergebnis
⋮	⋮	⋮
D	$10^{20} + 17 - 10 - 10^{20} + 130$	$= 130$
⋮	⋮	⋮
G	$10^{20} + 17 + 130 - 10^{20} - 10$	$= -10$
⋮	⋮	⋮

Problem bei D und G

Formel D liefert das falsche Ergebnis, weil $10^{20} + 7$ nicht dargestellt werden kann.

Aufgabe 4 – Assoziativität bei Gleitkommadarstellung

Sowohl Excel 2013 als auch LibreOffice 4.2 liefern zum Beispiel folgende Ergebnisse:

	Formel	Ergebnis
⋮	⋮	⋮
D	$10^{20} + 17 - 10 - 10^{20} + 130$	$= 130$
⋮	⋮	⋮
G	$10^{20} + 17 + 130 - 10^{20} - 10$	$= -10$
⋮	⋮	⋮

Problem bei D und G

Formel D liefert das falsche Ergebnis, weil $10^{20} + 7$ nicht dargestellt werden kann. 10^{20} ist zwar darstellbar, aber aufgrund des hohen Exponenten ist die Genauigkeit für die 7 nicht ausreichend, weshalb auf 10^{20} gerundet wird.

Aufgabe 4 – Assoziativität bei Gleitkommadarstellung

Sowohl Excel 2013 als auch LibreOffice 4.2 liefern zum Beispiel folgende Ergebnisse:

	Formel	Ergebnis
⋮	⋮	⋮
D	$10^{20} + 17 - 10 - 10^{20} + 130$	$= 130$
⋮	⋮	⋮
G	$10^{20} + 17 + 130 - 10^{20} - 10$	$= -10$
⋮	⋮	⋮

Problem bei D und G

Formel D liefert das falsche Ergebnis, weil $10^{20} + 7$ nicht dargestellt werden kann.
10^{20} ist zwar darstellbar, aber aufgrund des hohen Exponenten ist die Genauigkeit für die 7 nicht ausreichend, weshalb auf 10^{20} gerundet wird.
Anschließend wird 10^{20} (= 0) wieder abgezogen und 130 addiert, was das Ergebnis von 130 erklärt.

Aufgabe 4 – Assoziativität bei Gleitkommadarstellung

Sowohl Excel 2013 als auch LibreOffice 4.2 liefern zum Beispiel folgende Ergebnisse:

	Formel	Ergebnis
⋮	⋮	⋮
D	$10^{20} + 17 - 10 - 10^{20} + 130$	$= 130$
⋮	⋮	⋮
G	$10^{20} + 17 + 130 - 10^{20} - 10$	$= -10$
⋮	⋮	⋮

Problem bei D und G

Formel D liefert das falsche Ergebnis, weil $10^{20} + 7$ nicht dargestellt werden kann.
10^{20} ist zwar darstellbar, aber aufgrund des hohen Exponenten ist die Genauigkeit für die 7 nicht ausreichend, weshalb auf 10^{20} gerundet wird.
Anschließend wird 10^{20} (= 0) wieder abgezogen und 130 addiert, was das Ergebnis von 130 erklärt.
Analog bei Formel G.

Aufgabe 4 – Assoziativität bei Gleitkommadarstellung

Sowohl Excel 2013 als auch LibreOffice 4.2 liefern zum Beispiel folgende Ergebnisse:

	Formel	Ergebnis
⋮	⋮	⋮
F	$10^{16} + 17 - 10 - 10^{16} + 130$	$= 136$
⋮	⋮	⋮
I	$10^{16} + 17 + 130 - 10^{16} - 10$	$= 136$

Aufgabe 4 – Assoziativität bei Gleitkommadarstellung

Sowohl Excel 2013 als auch LibreOffice 4.2 liefern zum Beispiel folgende Ergebnisse:

	Formel	Ergebnis
$\vdots$	$\vdots$	$\vdots$
F	$10^{16} + 17 - 10 - 10^{16} + 130$	$= 136$
$\vdots$	$\vdots$	$\vdots$
I	$10^{16} + 17 + 130 - 10^{16} - 10$	$= 136$

Problem bei F und I

Aufgabe 4 – Assoziativität bei Gleitkommadarstellung

Sowohl Excel 2013 als auch LibreOffice 4.2 liefern zum Beispiel folgende Ergebnisse:

	Formel	Ergebnis
⋮	⋮	⋮
F	$10^{16} + 17 - 10 - 10^{16} + 130$	$= 136$
⋮	⋮	⋮
I	$10^{16} + 17 + 130 - 10^{16} - 10$	$= 136$

Problem bei F und I

Formeln F und I liefern das falsche Ergebnis, weil bei der Darstellung von 10^{16} die niederstwertige 1 von $17 = 10001_2$ bei der Addition genau auf die 54. Stelle der Mantisse fallen würde, die Mantisse bei doppelter Genauigkeit (64 Bit) jedoch nur 53 Stellen hat.

Aufgabe 4 – Assoziativität bei Gleitkommadarstellung

Sowohl Excel 2013 als auch LibreOffice 4.2 liefern zum Beispiel folgende Ergebnisse:

	Formel	Ergebnis
⋮	⋮	⋮
F	$10^{16} + 17 - 10 - 10^{16} + 130$	$= 136$
⋮	⋮	⋮
I	$10^{16} + 17 + 130 - 10^{16} - 10$	$= 136$

Problem bei F und I

Formeln F und I liefern das falsche Ergebnis, weil bei der Darstellung von 10^{16} die niederstwertige 1 von $17 = 10001_2$ bei der Addition genau auf die 54. Stelle der Mantisse fallen würde, die Mantisse bei doppelter Genauigkeit (64 Bit) jedoch nur 53 Stellen hat.
Effektiv wird also der Summand 1 abgeschnitten und nur 16 addiert.

Aufgabe 4 – Assoziativität bei Gleitkommadarstellung

Sowohl Excel 2013 als auch LibreOffice 4.2 liefern zum Beispiel folgende Ergebnisse:

	Formel	Ergebnis
⋮	⋮	⋮
F	$10^{16} + 17 - 10 - 10^{16} + 130$	$= 136$
⋮	⋮	⋮
I	$10^{16} + 17 + 130 - 10^{16} - 10$	$= 136$

Problem bei F und I

Formeln F und I liefern das falsche Ergebnis, weil bei der Darstellung von 10^{16} die niederstwertige 1 von $17 = 10001_2$ bei der Addition genau auf die 54. Stelle der Mantisse fallen würde, die Mantisse bei doppelter Genauigkeit (64 Bit) jedoch nur 53 Stellen hat.
Effektiv wird also der Summand 1 abgeschnitten und nur 16 addiert.

Diese Beispiele zeigen auch, dass doppelte Genauigkeit verwendet wird, denn sonst würde auch die 16 verloren gehen.