Gleitkommazahlen für Anfänger

10. Dezember 2015
AfRA

Motivation

- Reelle Zahlen annähern
- Zahlen können sehr groß oder klein werden
- Rechenergebnisse sollen genau sein

Definition

- Darstellung einer reellen Zahl x als
  - Basis b
  - normalisierte n-stellige Mantisse $1 \leq m < b$
  - Exponent $e \in E$ und
  - Vorzeichen $\sigma \in \pm 1$
  - sodass $x \approx \sigma \cdot m \cdot b^e$
- Spezielle Darstellung für 0, ∞, NaN
- Beispiel: $-123{,}456 = -1 \cdot 1{,}23456 \cdot 10^2$ ($n = 6$)
- Gleitkommazahl-Typ ist Tripel (b, n, E)

Definition

- $\mathrm{rnd}_{(b,\, n,\, E)}(x)$ rundet reelle Zahl x zu Gleitkommazahl
  - verschiedene Rundungsmodi
- Jede Operation implementiert, als ob exakt berechnet und dann gerundet
- In der Praxis wird 2 oder 10 als Basis verwendet
  - Basis 10 für kaufmännische Rechnungen

Geschichtlicher Überblick

- im Babylonisches Reich, 1800 v. Chr. mit $b = 60$
- Wiederentdeckung im Mittelalter (~1600)
- Formale Beschreibung durch Leonardo Torres y Quevedo (1914)
- Implementierung durch Konrad Zuse (1937)

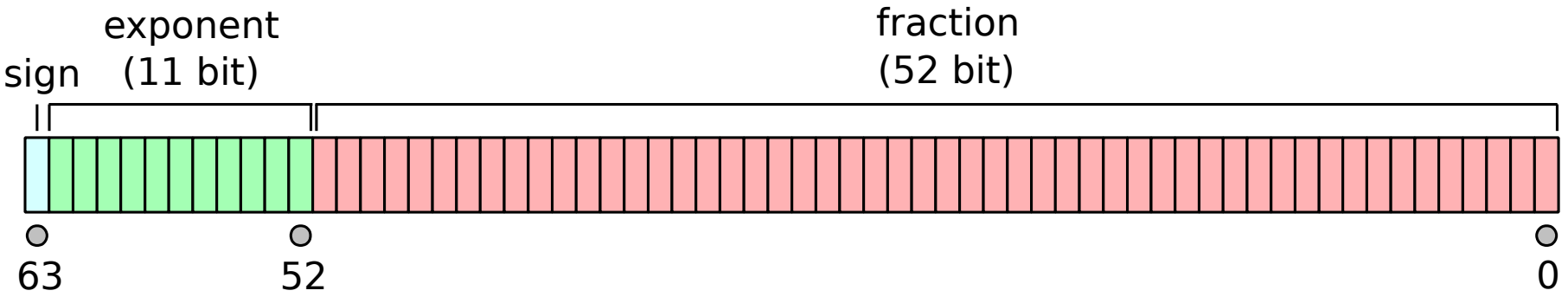
Nachbau des ersten Gleitkommarechenwerks

IEEE 754

- Maßgeblich von W. Kahan für Intel entwickelt $\longrightarrow$ 8087
- Viele Neuerungen: NaNs, subnormale Gleitkommazahlen
- Heutiger Industriestandard
- IEEE 754-2008 mit mehr Features (z. B. $b = 10$)

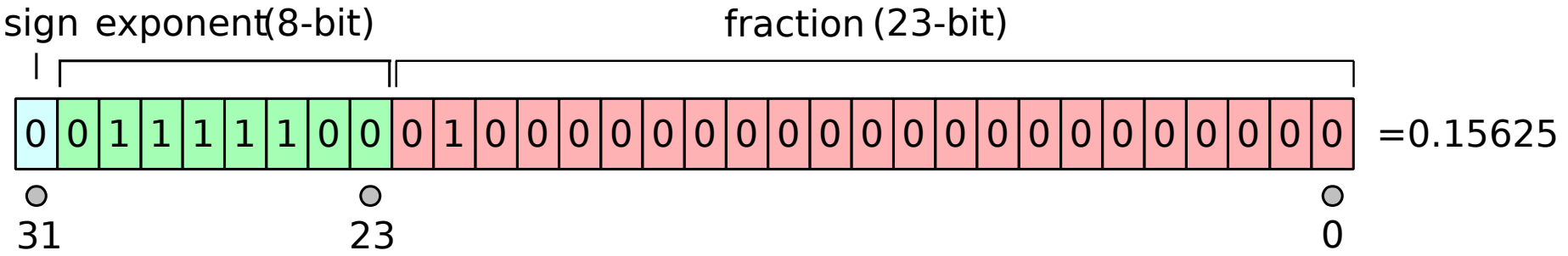
William Kahan

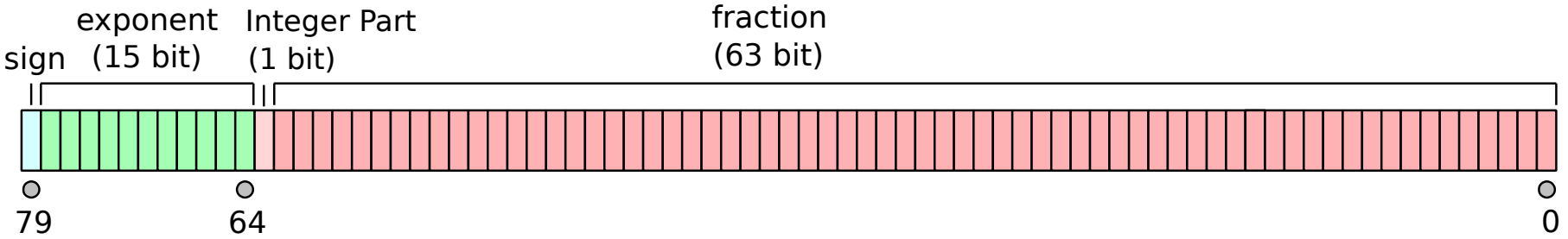
IEEE 754



- Exponent mit Bias (hier 1023) gespeichert
- oberstes Bit der Mantisse wird weggelassen
- Exponent -1023 für 0 und subnormale Zahlen
- Exponent 1024 für ∞ und NaN

IEEE 754


sign
exponent (8-bit)
fraction (23-bit)
0 0 1 1 1 1 1 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
=0.15625
31
23
0


sign
exponent (15 bit)
Integer Part (1 bit)
fraction (63 bit)
79
64
0

Was ist ein NaN?

- Ergebnis undefinierter Operation, z. B. $0/\infty$
- In IEEE 754: Exponent maximal und $m \neq 0$
- *signalling NaN* und *quiet NaN*
  - angezeigt durch MSB der Mantisse (uneinheitlich)
- NaN propagiert durch Berechnungen
- signalling NaN löst Trap aus
  - bspw. für Implementierung zusätzlicher Features

Genauigkeit der Ergebnisse

- Numerische Stabilität
  - Umformung von Formeln häufig nötig
- Auslöschung: $x + y$ ist für $x \approx -y$ ungenau
  - Vermeidung durch Umformung der Rechnung
  - Bsp. Für $x^2 + px + q = 0$ nehme

    $x_1 = -p/2 - \sqrt{p^2/4 - q} \cdot \operatorname{sgn} p, \quad x_2 = q/x_1$

Gleitkommazahlen auf x86

- x87, MMX, SSE, AVX
- x87/MMX, SSE/AVX teilen sich den Registersatz
- MMX, SSE, AVX: Vektorfähigkeiten