TUM NumProg, WiSe 2022/2023

Mitschriften basierend auf der Vorlesung von Dr. Hans-Joachim Bungartz

Zuletzt aktualisiert: 26. November 2022

Introduction

About

Hier sind die wichtigsten Konzepte der NumProg Vorlesung von Dr. Hans-Joachim Bungartz im Wintersemester 2022/2023 zusammengefasst.

Die Mitschriften selbst sind in Markdown geschrieben und werden mithilfe einer GitHub-Action nach jedem Push mithilfe von Pandoc zu einem PDF konvertiert.

Eine stets aktuelle Version der PDFs kann über https://github.com/ManuelLerchner/numprog/releases/download/Release/merge.pdf heruntergeladen werden.

How to Contribute

1. Fork this Repository
2. Commit and push your changes to **your** forked repository
3. Open a Pull Request to this repository
4. Wait until the changes are merged

Contributors

Inhaltsverzeichnis

Floating-Point

Fixed-Point

Representation

Bei Fixed-Point wird die Zahl in eine ganze Zahl und eine Bruchzahl aufgeteilt. Diese werden jeweils "normal" kodiert.

- Nachteile:
  - Schneller Overflow, da nur kleine Zahlen dargestellt werden können
  - Konstanter Abstand zwischen zwei Zahlen. Oft nicht benötigt.

Floating-Point

Representation

Eine Floating-Point Zahl wird in Mantisse und Exponent aufgeteilt. Zusammen mit einem Vorzeichenbit, lässt sich so ein sehr großer Wertebereich darstellen.

- Definition normalisierte, t-stellige Float-Zahl
  - $\mathbb{F}_{B,t} = \{M \cdot B^E \mid M, E \in \mathbb{Z} \land M \text{ ohne führende Nullen bzw: } B^{t-1} \leq M < B^t\}$
  - $\mathbb{F}_{B,t,\alpha,\beta} = \{M \cdot B^E \mid M, E \in \mathbb{Z} \ \& \ M \text{ ohne führende Nullen} \land \alpha \leq E < \beta\}$
- Wobei gilt:
  - B Basis
  - t Anzahl der Bits
  - α kleinster möglicher Exponent
  - β größter möglicher Exponent
- Vorteile:
  - Großer Wertebereich, da variable Abstände zwischen zwei Zahlen

In einem solchen System ist:

- $\sigma = B^{t-1} \cdot B^{\alpha}$ die kleinste positive Zahl, die dargestellt werden kann.
- $\lambda = (B^t - 1) \cdot B^{\beta}$ die größte positive Zahl, die dargestellt werden kann.

Beispiel:

- Mit $B = 10$ und $t = 4$ und $\alpha = -2$ und $\beta = 1$ ergibt sich:
  - $\sigma = 10^{4-1} \cdot 10^{-2} = 10$
  - $\lambda = (10^4 - 1) \cdot 10^1 = 99990$

Formel zur Berechnung des maximalen relativen Abstands zwei Float-Zahlen

Die Resoluition einer Float-Zahl ist der maximale relative Abstand zu einer anderen Float-Zahl. Sie berechnet sich wie folgt:

- $\varrho = \frac{1}{M} \leq B^{1-t}$

Beispiel:

- Mit einer Basis von $B = 2$ und $t = 4$ Stellen ergibt sich; $\varrho \leq 2^{-3} = 0.125$. Damit ist der maximale relative Abstand zwischen zwei Float-Zahlen 0.125.