Universität Stuttgart
Institut für Parallele und Verteilte Systeme

IPVS

SimTech

Mehrgitter-Verfahren in gemischter Genauigkeit auf GPUs

Daniel Fink

Zwischenvortrag

-

Bachelorarbeit

16. August 2019

Agenda

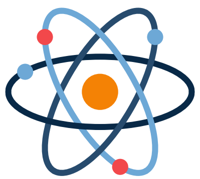
- Motivation
- Zielsetzung
- Mehrgitterverfahren
- Gemischt genaue Mehrgitterverfahren
- Analyse der Verfahren
- Roadmap und Fazit
- Anmerkungen und Fragen

Motivation

- Was bedeutet gemischte Genauigkeit?

→ Verwendung unterschiedlicher Datentypen für geringere(n) { Ausführungszeit, Speicherbedarf, Kommunikationsaufwand }

- Welche Datentypen gibt es?

Double	Single	Half
 64 Bit lang $\epsilon \approx 10^{-16}$	 32 Bit lang $\epsilon \approx 10^{-8}$	 16 Bit lang $\epsilon \approx 10^{-4}$ GPU

- NVIDIA Tesla V100:

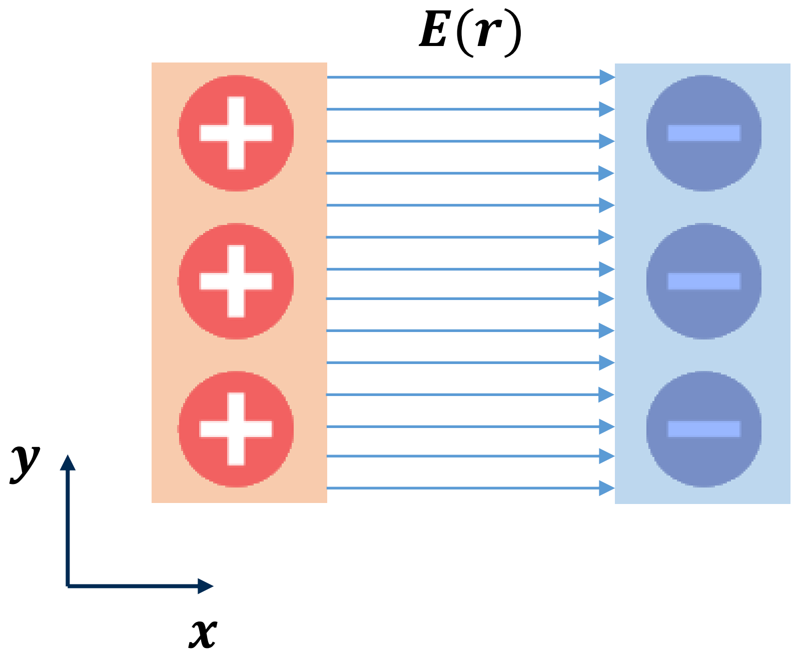
Single = ½ Double
Half = ¼ Double } Sowohl Rechen- als auch Ladezeit → Gesamt: Faktor 2 bzw. 4

Motivation

- Anwendungsbereich: Simulationen (Lösen von PDEs, ODEs, …)
- Hier: Poisson-Gleichung

$$-\triangle u = f \quad \text{in} \quad \Omega = (0,1) \times (0,1)$$
$$u = g \quad \text{auf} \quad \partial\Omega$$

wobei $u \in C^2$ und $f, g \in C$



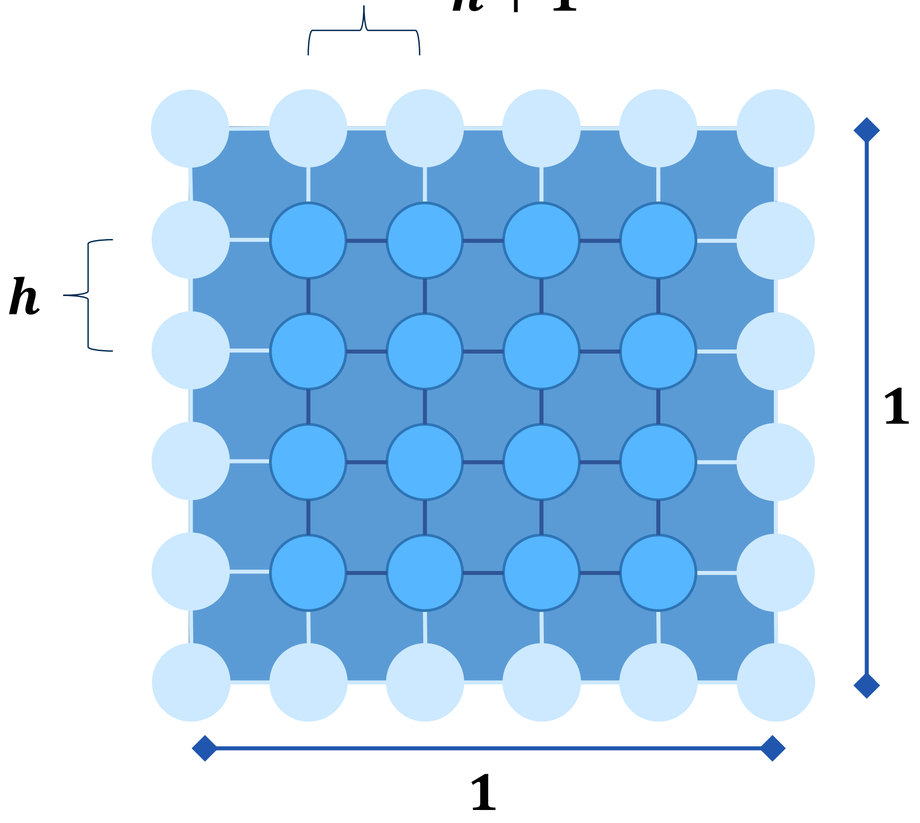
Motivation

- Anwendungsbereich: Simulationen (Lösen von PDEs, ODEs, …)
- Hier: Poisson-Gleichung

  $$-\triangle u = f \quad \text{in} \quad \Omega = (0,1) \times (0,1)$$
  $$u = g \quad \text{auf} \quad \partial\Omega$$

  wobei $u \in C^2$ und $f, g \in C$
- Vorgehen:
  - Diskretisieren (Hier: Finite Differenzen)

$$\boldsymbol{h = \frac{1}{n+1}}$$



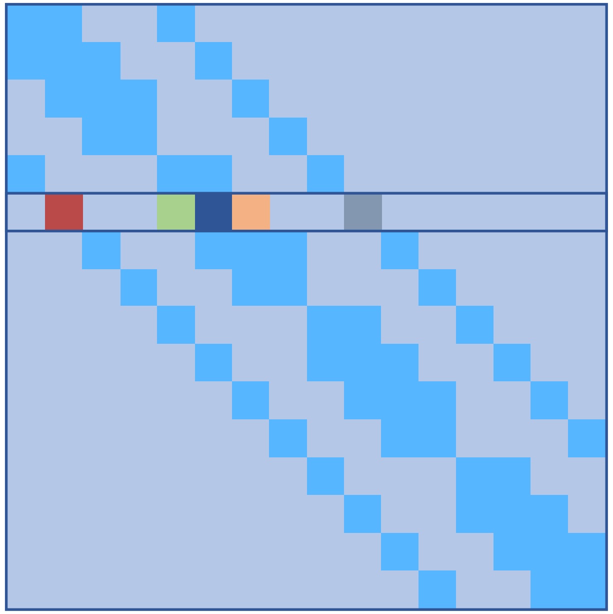
Motivation

- Anwendungsbereich: Simulationen (Lösen von PDEs, ODEs, …)
- Hier: Poisson-Gleichung

  $$-\triangle u = f \quad \text{in} \quad \Omega = (0,1) \times (0,1)$$
  $$u = g \quad \text{auf} \quad \partial\Omega$$

  wobei $u \in C^2$ und $f, g \in C$
- Vorgehen:
  - Diskretisieren (Hier: Finite Differenzen)

  → (dünnbesetztes) Gleichungssystem $\boldsymbol{Au} = \boldsymbol{f}$



Systemmatrix $\boldsymbol{A} \in \mathbb{R}^{\boldsymbol{n}^2 \times \boldsymbol{n}^2}$

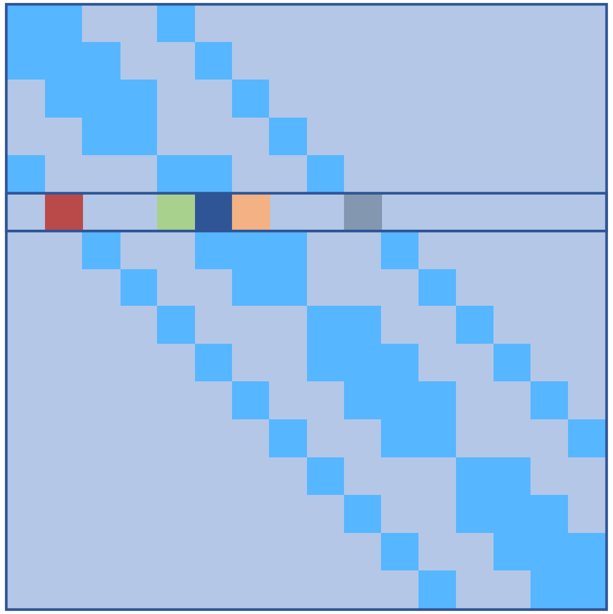
Motivation

- Anwendungsbereich: Simulationen (Lösen von PDEs, ODEs, …)
- Hier: Poisson-Gleichung

$$\begin{aligned} -\triangle u &= f \quad \text{in} \quad \Omega = (0,1) \times (0,1) \\ u &= g \quad \text{auf} \quad \partial\Omega \end{aligned}$$

  wobei $u \in C^2$ und $f, g \in C$
- Vorgehen:
  - Diskretisieren (Hier: Finite Differenzen)

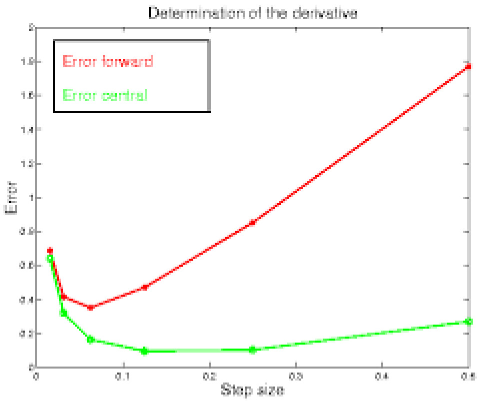
  → (dünnbesetztes) Gleichungssystem $\boldsymbol{Au} = \boldsymbol{f}$
  - Lösen mittels Mehrgitterverfahren
  - Hierarchischer Aufbau → gemischte Genauigkeit



Systemmatrix $\boldsymbol{A} \in \mathbb{R}^{n^2 \times n^2}$

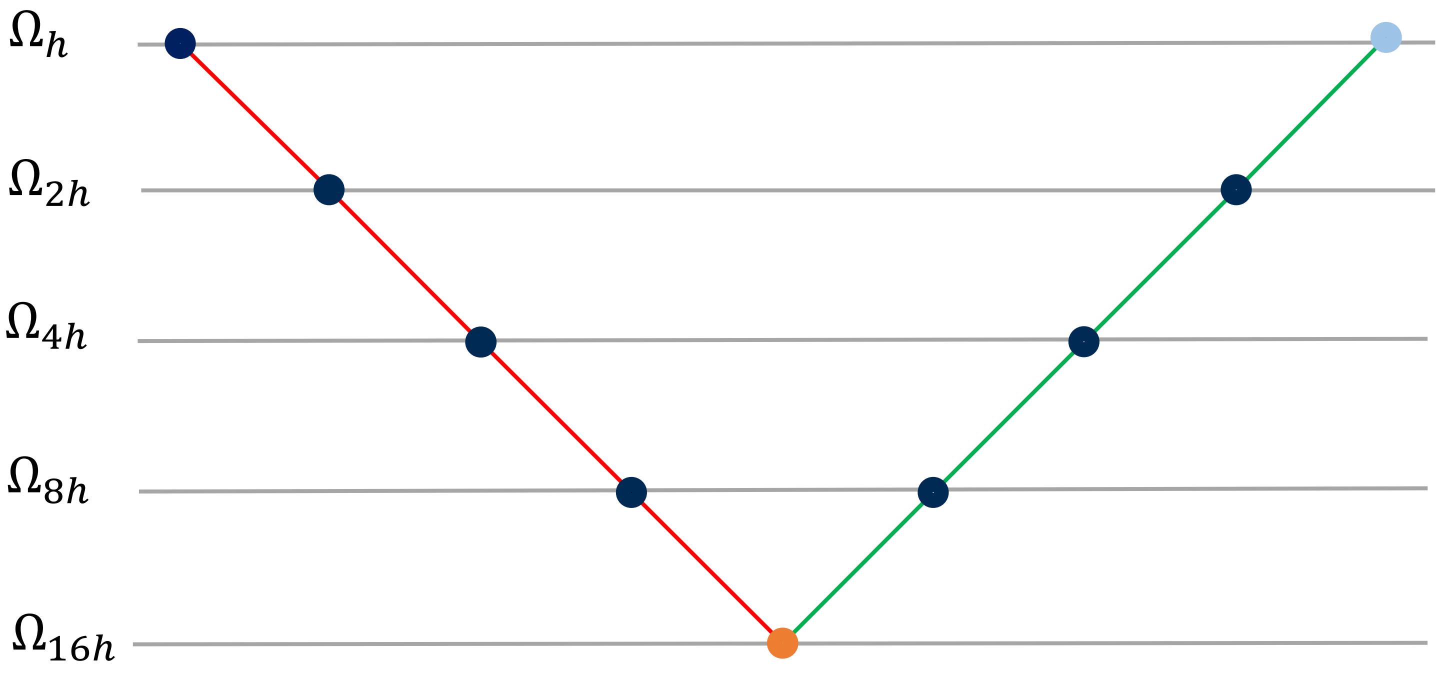
Zielsetzung

- Mehrgitterverfahren mit Double, Single und Half
- Problem wird in doppelter Genauigkeit gestellt
- Lösung soll am Ende in doppelter Genauigkeit vorliegen
- Vergleich mit MG-Verfahren in Double



Mehrgitterverfahren

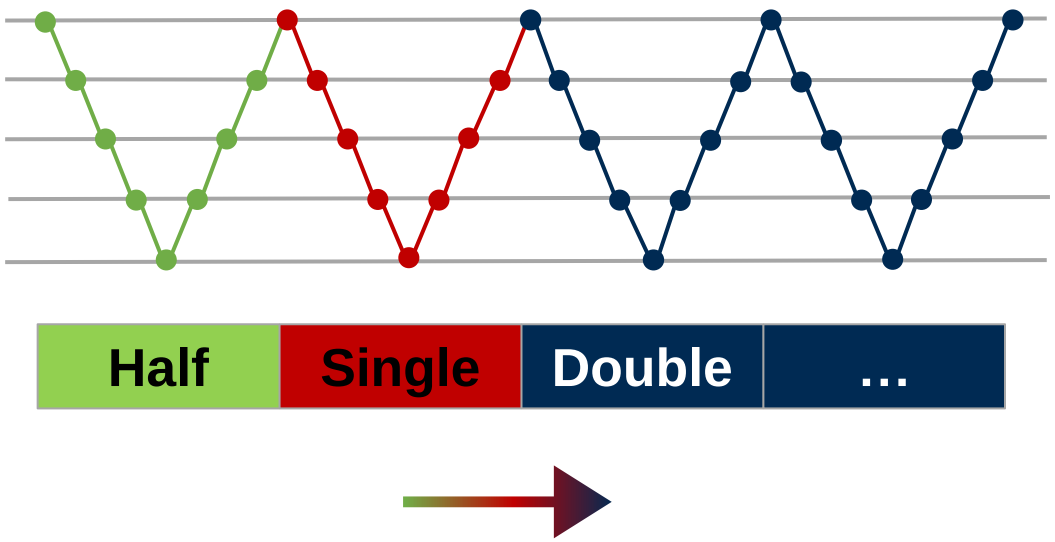
V-Zyklus



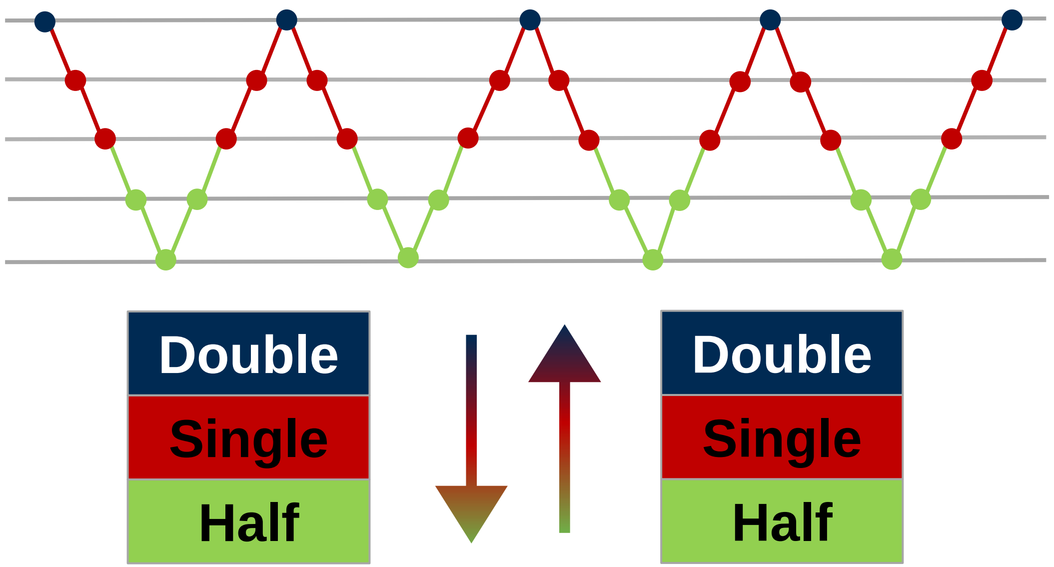
Relaxiere $\boldsymbol{Au} = \boldsymbol{f}$	Jacobi-Verfahren
Restringiere $\boldsymbol{r}$	halb-gewichtet
Relaxiere $\boldsymbol{Ae} = \boldsymbol{r}$	Jacobi-Verfahren
⋮	
Löse $\boldsymbol{Ae} = \boldsymbol{r}$	Jacobi-Verfahren
Prolongiere $\boldsymbol{e}$	Bilineare Interpol.
Relaxiere $\boldsymbol{Ae} = \boldsymbol{r}$	Jacobi-Verfahren
⋮	
Approximation $\boldsymbol{u}^*$	

Gemischte genaue MG-Verfahren

Horizontal-Verfahren



Vertikal-Verfahren



Horizontal-Verfahren

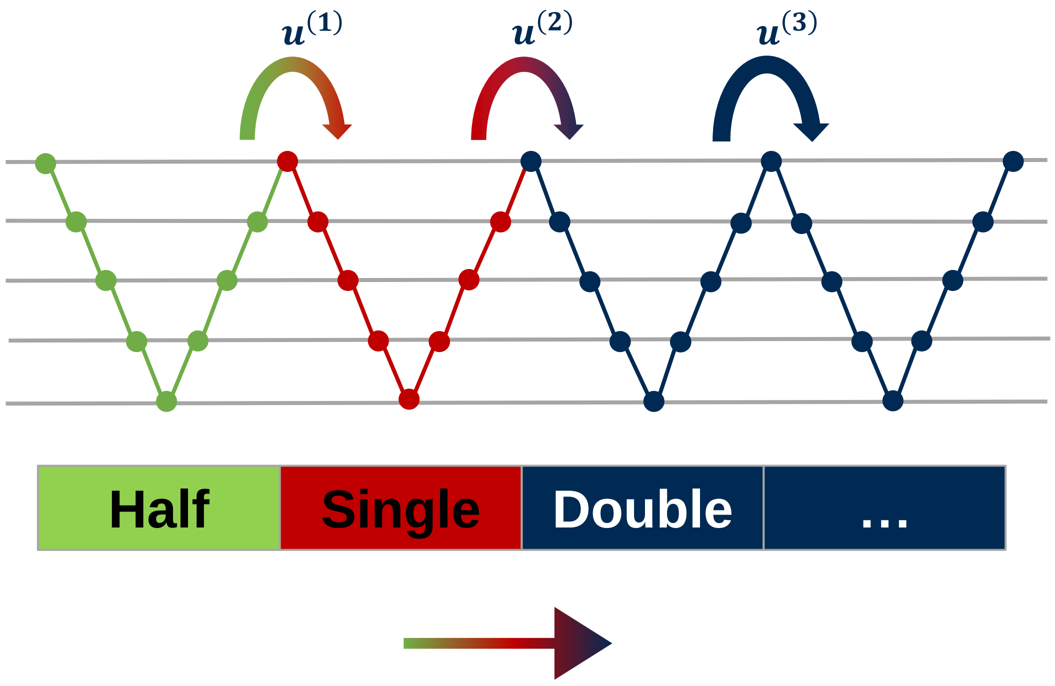
- V-Zyklus → Verbesserte Startlösung
- Ohnehin mehrere Iterationen notwendig

→ Half/Single-MG als Vorkonditionierer

- Wann ändert man den Datentyp?

→ Konvergenzraten oder feste Vorgabe

- Theoretischer Speed-Up:



Angenommen $\left\{ \begin{array}{l} \frac{1}{4} \text{ in Half} \\ \frac{1}{4} \text{ in Single} \\ \frac{1}{2} \text{ in Double} \end{array} \right\}$ Speed-Up = 1,46

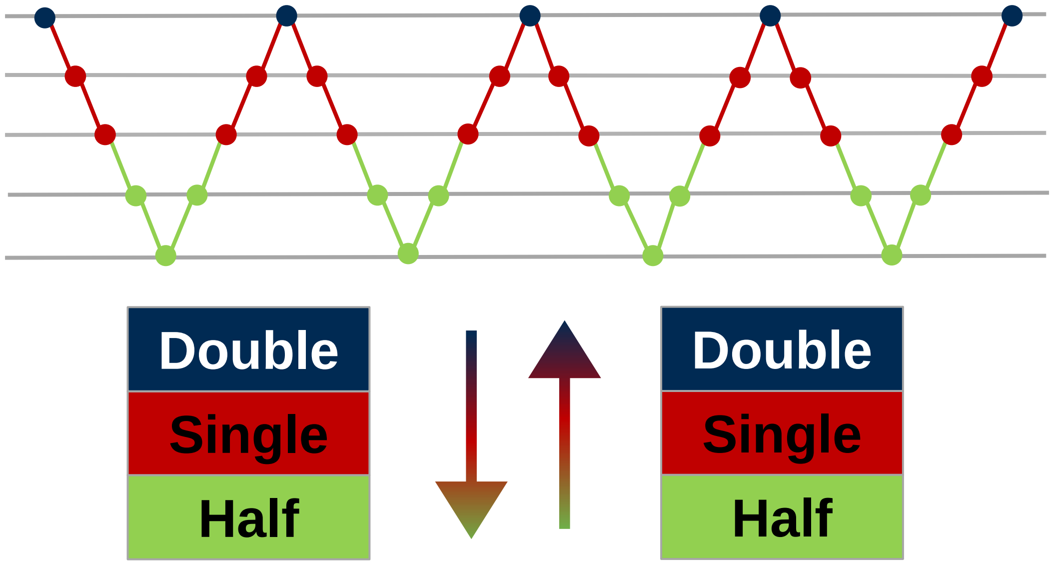
Vertikal-Verfahren

- Größerer Diskretisierungsfehler auf groben Gittern

→ Double viel zu genau

- Wann ändert man den Datentyp?

→ Feste Vorgabe der einzelnen Levels

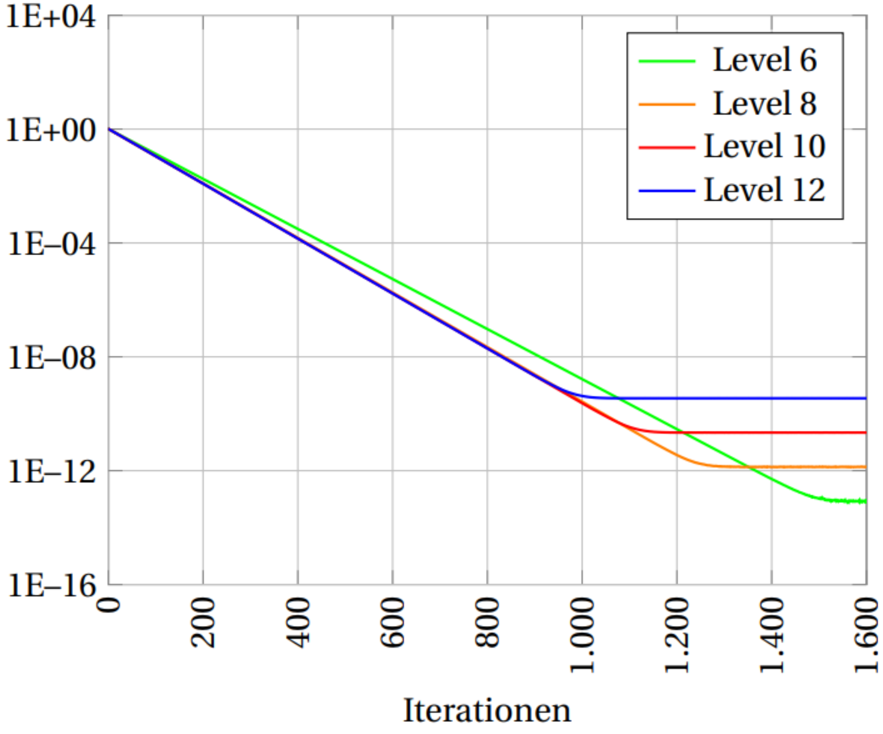


- Theoretischer Speed-Up (bei 12 Gittern):

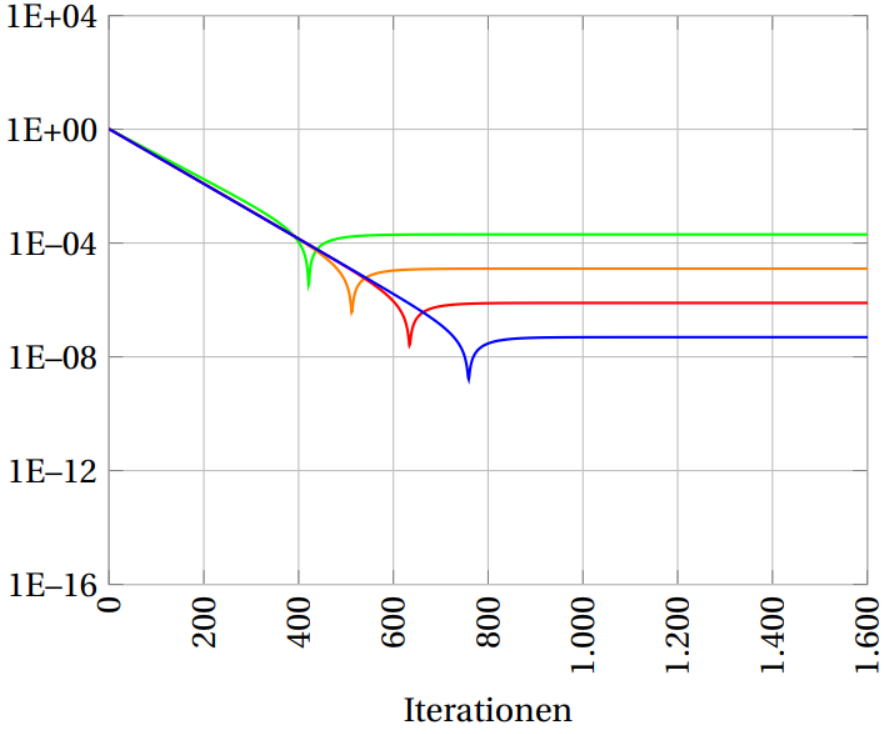
Angenommen
Level 12 in Double
Level 11 – 9 in Single
Level 8 – 5 in Half
Speed-Up = 1,36

MG-Verfahren in Double

Relatives Residuum $\frac{\|f - Au^{(m)}\|}{\|f\|}$



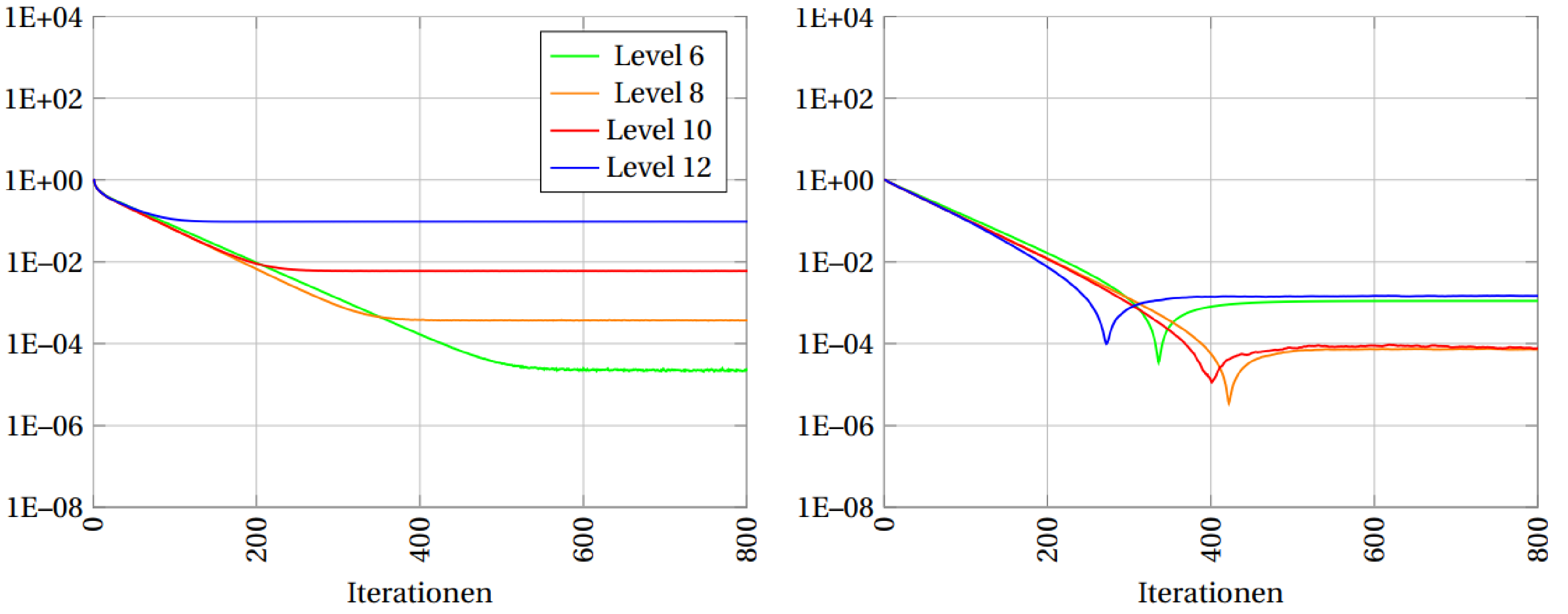
Relativer Fehler $\frac{\|u_{exakt} - u^{(m)}\|}{\|u_{exakt}\|}$



MG-Verfahren in Single

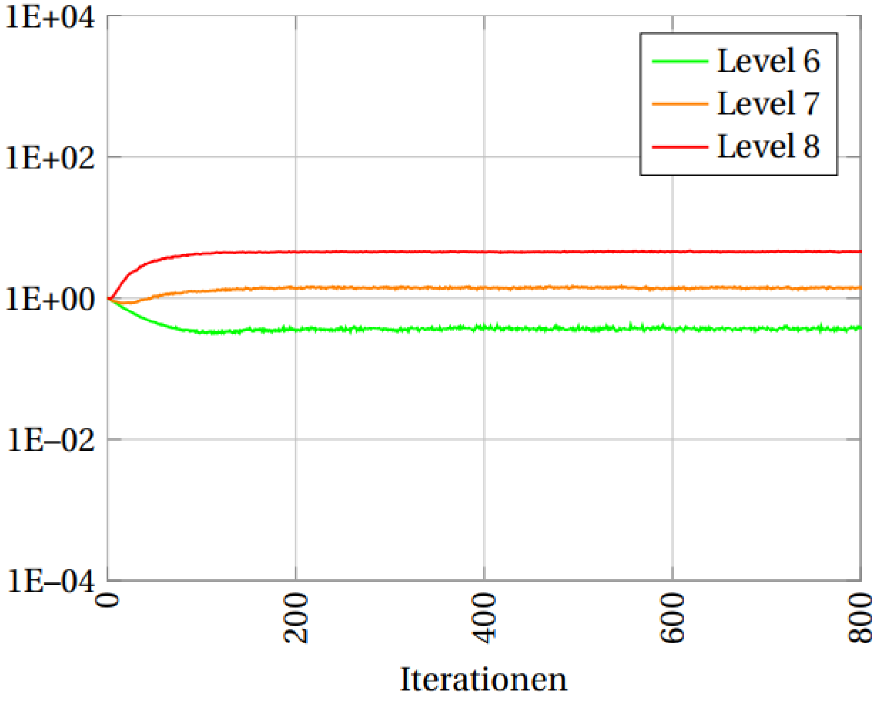
Relatives Residuum $\frac{\|f - Au^{(m)}\|}{\|f\|}$

Relativer Fehler $\frac{\|u_{exakt} - u^{(m)}\|}{\|u_{exakt}\|}$

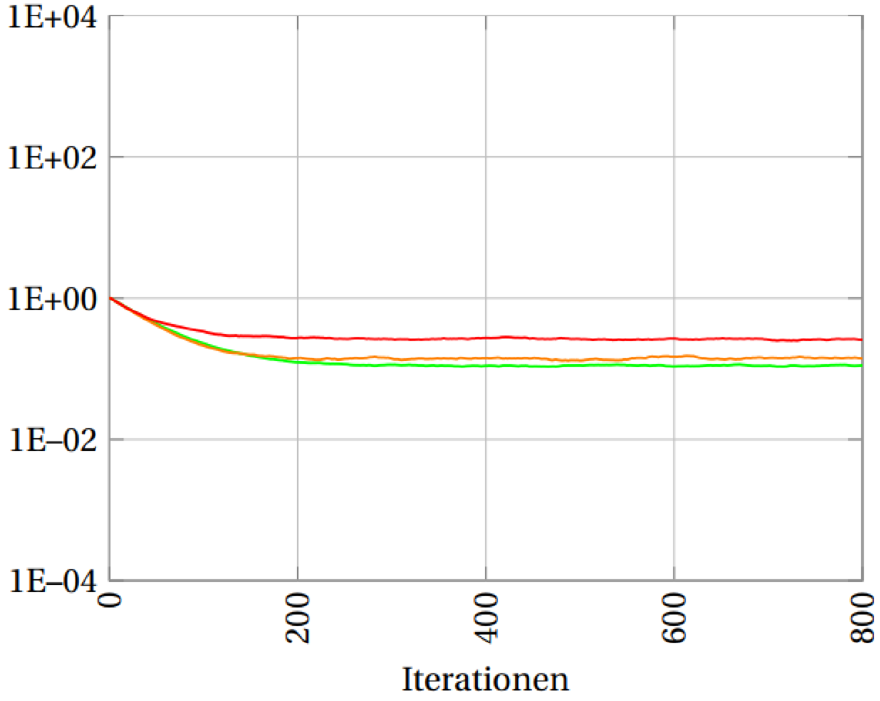


MG-Verfahren in Half

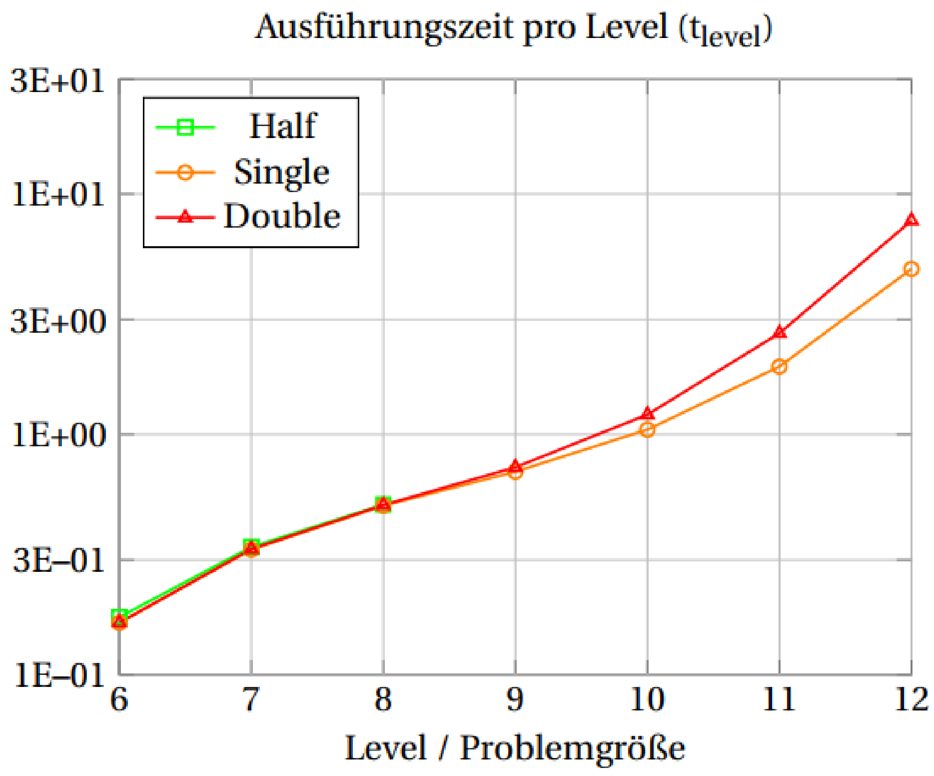
Relatives Residuum $\frac{\|f - A u^{(m)}\|}{\|f\|}$

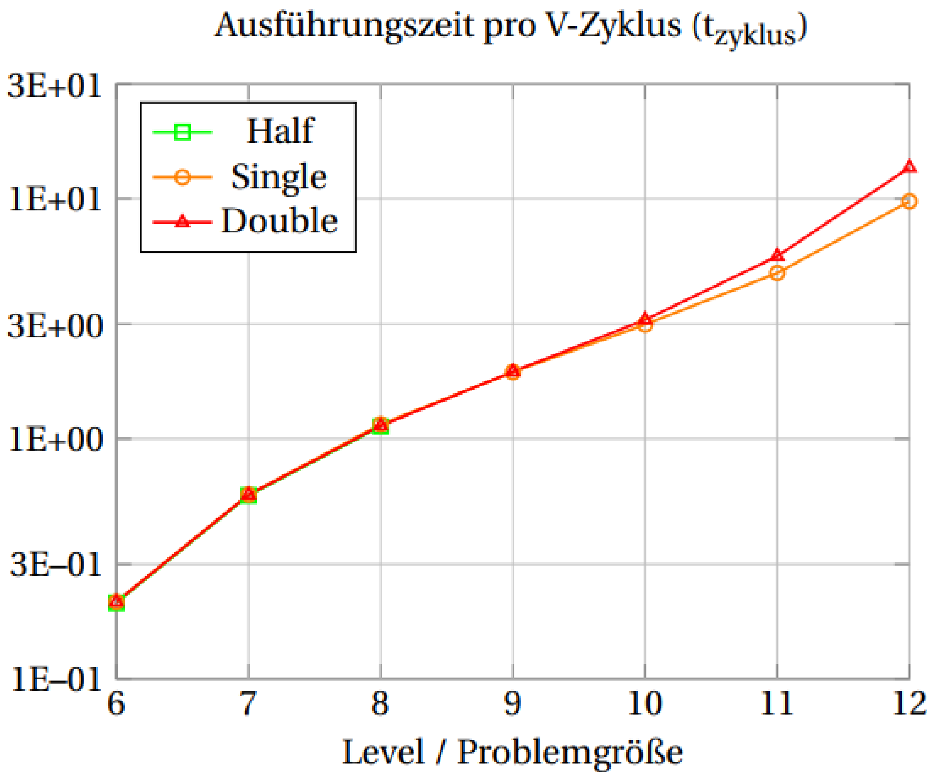


Relativer Fehler $\frac{\|u_{exakt} - u^{(m)}\|}{\|u_{exakt}\|}$



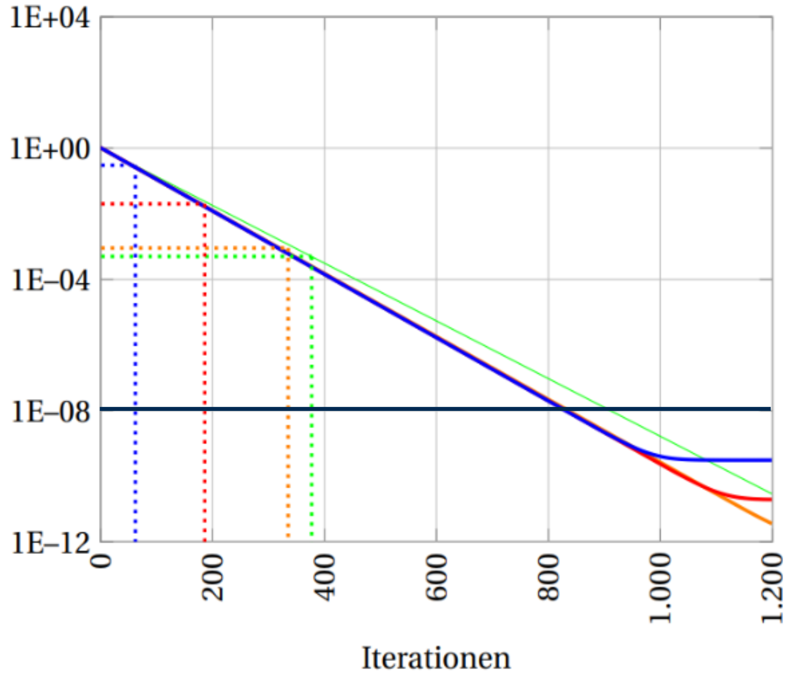
Ausführungszeiten



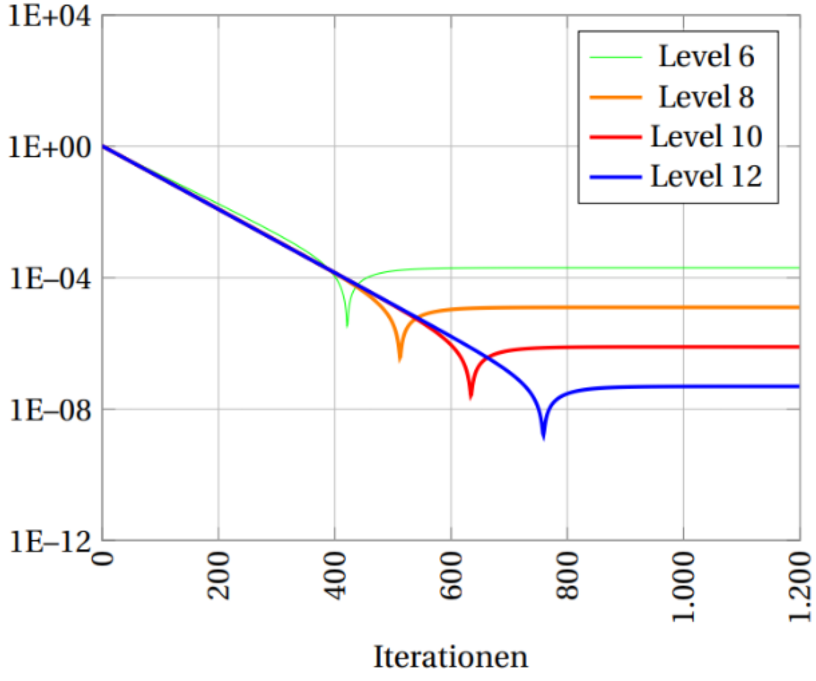


Horizontales Single-Double-MG-Verfahren

Relatives Residuum $\frac{\|f - Au^{(m)}\|}{\|f\|}$

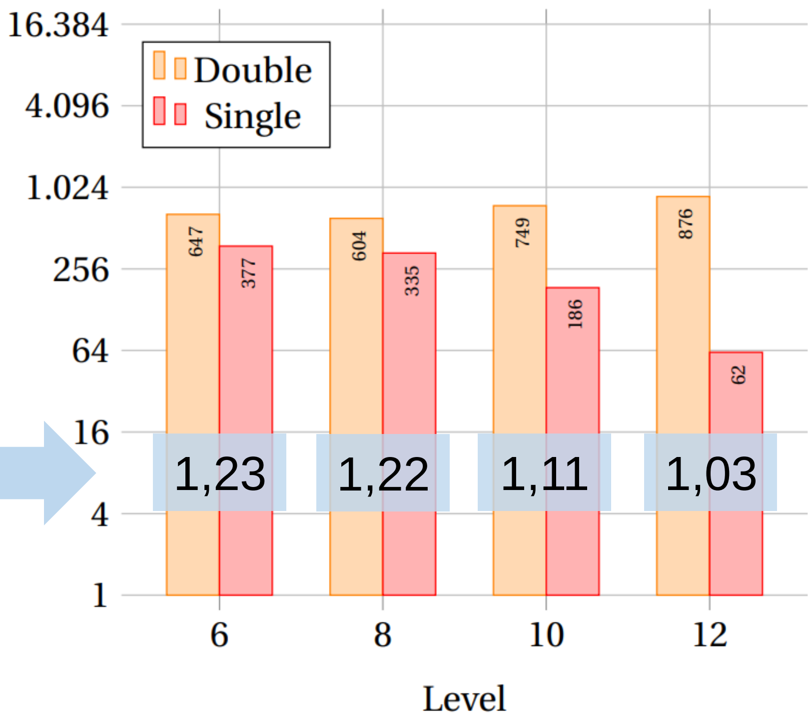


Relativer Fehler $\frac{\|u_{exakt} - u^{(m)}\|}{\|u_{exakt}\|}$



Horizontales Single-Double-MG-Verfahren

Theoretischer Speed-Up



Ausführungszeit			
Level	FP64	FP32/64	Faktor
6	1,33E+00	1,29E+00	1,03
8	1,51E+00	1,46E+00	1,03
10	2,11E+00	2,03E+00	1,04
12	7,64E+00	7,41E+00	1,03

Gemessener Speed-Up

Roadmap

Was wurde schon gemacht

- Einarbeitung in CUDA
- Beschreibung (F)MG-Verfahren
- Literaturrecherche „Gemischte Genauigkeit“

SimTech-Propädeutikum

- Herleitung der gemischt genauen Verfahren
- Theoretische Analyse der Verfahren (Aufwand)
- Implementierung der Kernel mit Textur- und geteiltem Speicher
- Implementierung Horizontal/Vertikal-Verfahren mit Single und Double
- Numerische Analyse des Diskretisierungsfehlers

Was noch kommt

- Verwendung von Half2
- Implementierung Horizontal/Vertikal-Verfahren mit Half2
- Numerische Analyse der Verfahren
- Herleitung einer oberen Schranke für den Speed-Up
- Analyse der Kernel (Rechen- und Datendurchsatz)
- Vergleich mit Low-End-Grafikkarte
- Gemischt genaues Full-MG
- 1D-Struktur in CUDA
- MG mit Tensor-Core-LR-Zerlegung
- …

Ausblick

Fazit

Positiv	Negativ
▪ Kaum Auswirkungen auf die Fehlerentwicklung ▪ Diskretisierungsfehler verhält sich wie erwartet ▪ Lässt sich auf alle (F)-MG-Verfahren übertragen ▪ Sehr viel Potential für weitere Untersuchungen	▪ Feintuning für die einzelnen Datentypen erforderlich ▪ Abbruchkriterien für den Wechsel des Datentyps sind schwer zu definieren ▪ Verwendung von Half nur für kleine Probleme möglich

Fragen und Anmerkungen

Quellen

- Jinn-Liang Liu. Poisson’s Equation in Electrostatics. http://www.nhcue.edu.tw/~jinnliu/proj/Device/3DPoisson.pdf. Abgerufen: Juni 2019.
- IEEE Computer Society. IEEE Standard for Floating-Point Arithmetic. http://www.dsc.ufcg.edu.br/~cnum/modulos/Modulo2/IEEE754_2008.pdf. Abgerufen: Juni 2019.
- ufcg.edu.br/~cnum/modulos/Modulo2/IEEE754_2008.pdf. Abgerufen: Juni 2019. Thomas Jahn. “Implementierung numerischer Algorithmen auf CUDA–Systemen”. Abgerufen: Juni 2019. Diplomarbeit. Universität Bayreuth.
- NVIDIA Corporation. NVIDIA CUDA - Compute Unified Device Architecture - Programming Guide - Vers http://developer.download.nvidia.com/compute/cuda/1.0/NVIDIA_CUDA_Programming_Guide_1.0.pdf. Abgerufen: Mai 2019.
- The Khronos Group Inc. Khronos OpenCL Registry. https://www.khronos.org/registry/OpenCL/. Abgerufen: Juni 2019.
- NVIDIA Corporation. Whitepaper - NVIDIA’s Next Generation CUDA™Compute Architecture: Fermi™. https://www.nvidia.com/content/PDF/fermi_white_papers/NVIDIA_Fermi_Compute_Architecture_Whitepaper.pdf. Abgerufen: Februar 2019.

Quellen

- NVIDIA Corporation. NVIDIA Tesla V100 GPU Architecture. https://images.nvidia.com/content/volta-architecture/pdf/volta-architecture-whitepaper.pdf.
Abgerufen: Februar 2019.
- James H. Wilkinson. Rundungsfehler. pub-Springer:adr-B: pub-Springer, 1969.
- Cleve B. Moler. “Iterative Refinement in Floating Point”. In: J. ACM 14.2 (Apr. 1967), S. 316–321. ISSN: 0004-5411. DOI: 10.1145/321386.321394. URL: http://doi.acm.org/10.1145/321386.321394.
- Alfredo Buttari et al. “Mixed Precision Iterative Refinement Techniques for the Solution of Dense Linear Systems”. In: The International Journal of High Performance Computing Applications 21.4 (2007), S. 457–466. DOI: 10.1177/1094342007084026. URL: https://doi.org/10.1177/1094342007084026.
- Alfredo Buttari et al. “Using Mixed Precision for Sparse Matrix Computations to Enhance the Performance while Achieving 64-bit Accuracy”. In: ACM Transactions on Mathematical Software 34 (Juli 2008). DOI: 10.1145/1377596.1377597.

Quellen

- Dominik Göddeke, Robert Strzodka und Stefan Turek. “Performance and accuracy of hardware-oriented native-, emulated- and mixed-precision solvers in FEM simulations”. In: International Journal of Parallel, Emergent and Distributed Systems 22.4 (2007), S. 221–256. DOI: 10.1080/17445760601122076. URL: https://doi.org/10.1080/17445760601122076.
- Dominik Göddeke. Wissenschaftliches Rechnen. Vorlesungsskript. März 2017.
- Wolfgang Hackbusch. Multi-grid methods and applications. Springer, 1985.
- Dominik Göddeke. “Fast and Accurate Finite-Element Multigrid Solvers for PDE Simulations on GPU Clusters”. Diss. Technische Universität Dortmund, Feb. 2010.
- NVIDIA Corporation. CUBLAS Library - User Guide. https://docs.nvidia.com/cuda/pdf/CUBLAS_Library.pdf. Abgerufen: August 2019.
- NVIDIA Corporation. NVIDIA TESLA V100 GPU ACCELERATOR. https://images.nvidia.com/content/technologies/volta/pdf/tesla-volta-v100-datasheet-letterfnl-web.pdf. Abgerufen: Juni 2019
- NVIDIA Corporation. NVIDIA TESLA V100 GPU ACCELERATOR. https://www.anandtech.com/show/12576/nvidia-bumps-all-tesla-v100-models-to-32gb. Abgerufen: Juni 2019.

Quellen

- T. Washio C.W. Oosterlee. On the Use of Multigrid as a Preconditioner. https://pdfs.semanticscholar.org/dcc1/9ad91450753e7f47157e323fade5c2b4e320.pdf. Abgerufen: Juli 2019.
- Osamu Tatebe. The Multigrid Preconditioned Conjugate Gradient Method. http://www.hpcs.cs.tsukuba.ac.jp/~tatebe/research/paper/CM93-tatebe.pdf. Abgerufen: Juli 2019.
- **Bilder**
  - Scientific Datatype, https://cdn.icon-icons.com/icons2/539/PNG/512/atom_icon-icons.com_53030.png
  - Graphical Datatype, https://icon-library.net/images/games-icon/games-icon-18.jpg
  - Neural Network Datatype, https://icon-library.net/images/icon-artificial-intelligence/icon-artificial-intelligence-6.jpg
  - Time Measurement, https://cdn2.iconfinder.com/data/icons/social-productivity-line-black-1/3/14-512.png
  - Numerical Error, https://encrypted-tbn0.gstatic.com/images?q=tbn:ANd9GcRF2dNX6SpaYge-O3CF-XLMrAI0l1hNNtXzwDAI4txzKA0EpuwH