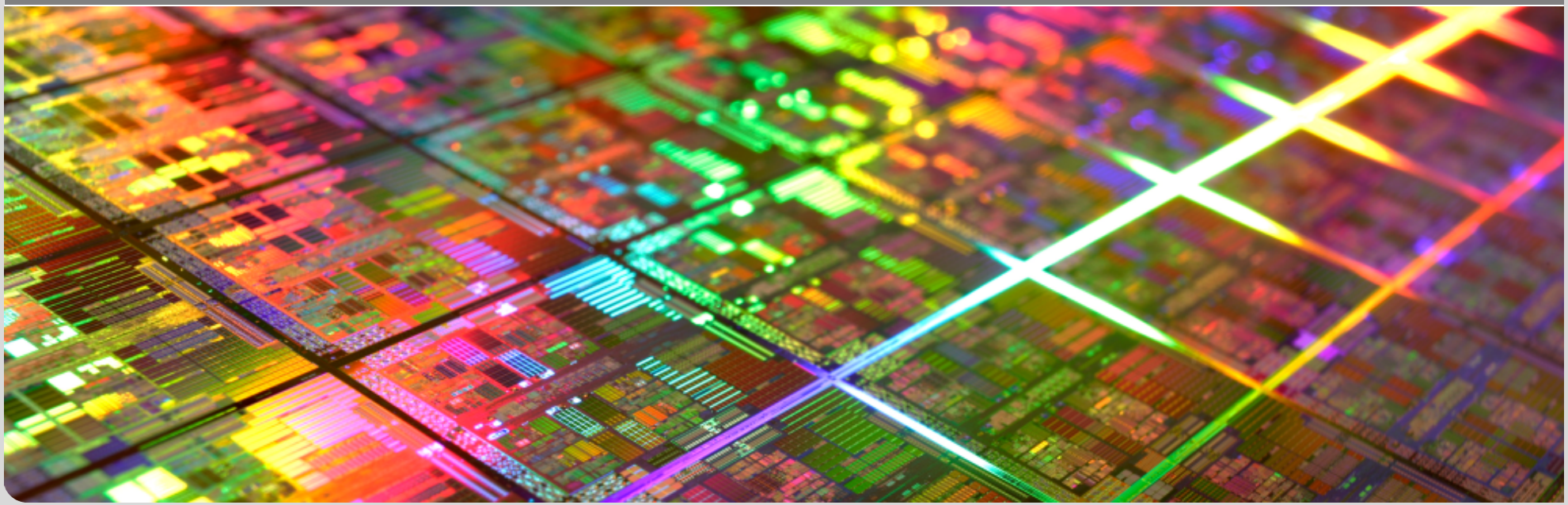
Karlsruher Institut für Technologie

Ausgewählte Kapitel der Rechnerarchitektur

Proseminar/Seminar

Kalmbach, Lehmann | 5. November 2020

PROFESSUR FÜR RECHNERARCHITEKTUR UND PARALLELVERARBEITUNG – PROF. DR. WOLFGANG KARL



Vorstellung

- Institut für Technische Informatik
- Professur: Rechnerarchitektur und Parallelverarbeitung
- Leiter: Prof. Dr. Wolfgang Karl

- Veranstaltungen:
  - Rechnerstrukturen
  - Mikroprozessoren I
  - Heterogene parallele Rechensysteme
  - Seminar/Proseminar
  - Praxis der Forschung
  - Rechnerorganisation

Themenbereiche

- Heterogene parallele Rechensysteme
  - **Reduzierung der Komplexität** für den Nutzer: Programmiersprachliche Konzepte und Laufzeitsysteme
  - **Zuverlässigkeit** und Effizienz für HPC-Systeme
  - **Echtzeitfähigkeit** von heterogenen Systemen

- **Approximate Computing**
  - Reduzierung der Genauigkeit - hinreichend gutes Ergebnis
  - Ansätze auf Anwendungs-, Algorithmen- und Hardware-Ebene

- Fallstudien:
  - Numerik, Analyse von Zeitreihen, sensorbasierte Sortierung, autonomes Fahren, Bildverarbeitung, . . .

Ziele des Seminars

- Erwerb von Kenntnissen aus dem Forschungsgebiet Rechnerarchitektur
- Literatur finden, verstehen, vergleichen, klassifizieren und wiedergeben
- Präsentation der Arbeit

$\rightarrow$ Vorbereitung auf Abschlussarbeit!

Ablauf

1. **10.11.2020 15:30** Vorstellung Themen
2. **11.11.2020** Abgabe Präferenzen
3. **13.11.2020** Themenvergabe
4. Einarbeitung und Gliederung der Ausarbeitung
   - Gliederung muss nicht abgegeben werden
   - **06.12.2020** Anmeldeschluss für die Prüfung im Studierendenportal
5. **21.02.2021** Abgabetermin der Ausarbeitung
6. **ca. 3-4 Wochen nach Abgabe** Präsentation der Ausarbeitung

Einarbeitung und Gliederung

- Muss nicht abgegeben werden!
- Einlesen in gegebene Literatur und
- **Selbständige** Recherche weiterführender Literatur:
  - Bücher, Journals, Konferenzen, Workshops
  - Publikationsverzeichnisse: `ACM Digital Library`, `IEEE Xplore`, `Springer-Verlag`, `CiteSeer.IST`
- Insbesondere “Related Work” beachten, andere Paper der gleichen Konferenz und “Zitiert von”-Relationen
- Erstellung einer Gliederung und grober Textbausteine sowie Literaturliste in Vorlage: `IEEE Transactions Template`
- Best practice: Beim/nach dem Lesen eines Papers gleich Notizen machen und Bibtex-Eintrag speichern

Vollständige Ausarbeitung

- Füllen der in der Gliederung erstellten Punkte
- Herausstellen des:
  - Hintergrundes: Was ist das Szenario?
  - sich ergebenden Problems
  - verwendeten Lösungsansatzes
- Ausführliche Erläuterung, Diskussion, Vergleich der vorgestellten Ansätze
- Bilder zur leicht verständlichen Aufbereitung
- Ausführliches Literaturverzeichnis mit vollständigen Angaben
- 5-6 Seiten doppelspaltig

Abgabe: 21.02.2021

Präsentationserstellung

- Informations- und Wissensweitergabe bzw. -vermittlung
- Anregung zu technischer und wissenschaftlicher Diskussion liefern
- Folien nicht überladen, sparsames Hervorheben
- Vortragsdauer: etwa 30-35 Minuten
- Anschließend 5-10 min Diskussion
- Faustregel: 2 min pro Folie ($\sim$ 15-20 Folien)
- Üben: Vor Spiegel, vor Freunden $\rightarrow$ man merkt ob Inhalt und logische Reihenfolge sinnvoll ist

Abgabe: nach dem Vortrag

Vortrag

- Ziel: **Erfahrung im Halten von Präsentationen gewinnen**
- Inhalte präsentieren: Ergebnisse, Vergleiche, Schlußfolgerungen usw.
- Vortrag vor anderen Seminarteilnehmern und Mitarbeitern des Lehrstuhls
- Üben, üben, üben $\rightarrow$ Sicherheit beim Vortrag

- Vorab: Durchlesen der Ausarbeitungen der übrigen Teilnehmer
- Jeder Teilnehmer überlegt sich mind. 2 Fragen zum Thema
- Lebhafte Diskussion erwünscht

Termin in Abstimmung mit euch
(voraussichtlich Ende März)

Quellen

- Quellen müssen referenziert werden (auch Bilder)!
- Quellen müssen überprüfbar sein
- Kopieren von Texten (auch mit Quellenangabe) ist nicht gestattet.
- Grafiken möglichst selbst erstellen
- Wikipedia nur als Hilfsmittel zur weitergehenden Recherche, nicht als Referenz verwenden!
- BibTeX-Einträge der Quellen sammeln für die Ausarbeitung (z.T. fertig bei ACM zum Download)

Seminar wird benotet

- 40 % Ausarbeitung, 60 % Vortrag

u.a. basierend auf

- Anzahl und Qualität gefundener Artikel
- Wiedergabe und Diskussion der für das Thema wesentlichen Punkte

das bedeutet:

- Man muss sich mit Arbeiten beschäftigen
- Seminar kann auch mit 5,0 bewertet werden

Zur Sicherheit:

- Rücksprache mit dem Betreuer
- **Nicht nach und nicht erst zwei Tage vor Deadline!**

Eigeninitiative

Beispiele für Probleme

- Keine Zeit, Oma ist krank,. . .
- Computer defekt
- Keine Literatur gefunden, Zugang zu IEEE ging nicht,. . .
- Formatvorlage zu kompliziert
- Hab noch nie was mit LaTeX gemacht
- . . .

- Selbständiges Arbeiten wird vorausgesetzt
- Zeitplan einhalten und nicht erst auf Nachfrage tätig werden
- **Bei andauernden Problemen nicht warten bis die Betreuer nachfragen**

Themenüberblick

1. Thema 1: GPU Worst-Case Execution Time Analyse
2. Thema 2: Profiling von Programme hinsichtlich CPU und Speichernutzung
3. Thema 3: Runtime Resource Management for Embedded
4. Thema 4: Performance Vorhersagen auf Heterogenen Systemen mithilfe von Maschinellem Lernen
5. Thema 5: Vergleich neuartiger Grafikprozessoren
6. Thema 6: Pruning-Stategien von neuronalen Netzen zur Verbesserung der Speicherauslastung und Rechenzeit
7. Thema 7: Quantisierungs-Strategien von neuronalen Netzen zur Verbesserung der Speicherauslastung und Rechenzeit
8. Thema 8: Hardware-Beschleuniger speziell für die Anwendung im Bereich der neuronalen Netzen

Thema 1: GPU Worst-Case Execution Time Analyse

- Heterogene Systeme erfordern angepasste Scheduling-Algorithmen
- Nicht nur Scheduling von CPU-Zeit, sondern auch von GPU möglich
- Alle Geräte können dabei unterschiedliche Eigenschaften und stärken haben
- Programme können auch z.B. auf Gerät A (z.B. GPU 1) besser performen als auf Gerät B (GPU 2)
- GPU Scheduling anders als CPU Scheduling
  - Preemptive Scheduling teuer
  - Software kann teil nicht einmal unterbrochen werden ohne bisherige Berechnung zu verlieren

Thema 1: GPU Worst-Case Execution Time Analyse

- Für Scheduling wichtig Laufzeitabschätzung zu haben
- Wie kann man die Laufzeit von GPU Programmen abschätzen?
- Analyse worst-case Execution Time von GPU Anwendungen

Betreuer

- Manuel Kalmbach
  E-Mail: `manuel.kalmbach@kit.edu`

Thema 2: Profiling von Programme hinsichtlich CPU und Speichernutzung

- Auf einigen embedded Plattformen wird häufig kein Scheduling zur Laufzeit gemacht
- Sondern im Voraus einem Programm feste Ressourcen zugewiesen
- Worst-case Ressourcenauslastung nicht optimal
- Wie reelle Ressourcenabschätzung eines Programms erhalten?
- Unterschied der Analyse zur Laufzeit und der statischen Analyse des Programmcodes

Betreuer

- Manuel Kalmbach
  E-Mail: `manuel.kalmbach@kit.edu`

Thema 3: Runtime Resource Management for Embedded

- Im embedded Bereich werden Ressourcen häufig vorab fest einer Software zugewiesen
- Dazu wird eine Abschätzung benötigt, wie viel Ressourcen eine Software braucht
- Abschätzung häufig worst-case
- Führt zu nicht optimalen Ausnutzung der Ressourcen
- Daher Ressource Management zu Laufzeit
- Wie kann das Ressource Management zu Laufzeit gemacht werden und welche Vorteile bringt es?

Betreuer

- Manuel Kalmbach
  E-Mail: `manuel.kalmbach@kit.edu`

Thema 4: Performance Vorhersagen auf Heterogenen Systemen mithilfe von Maschinellem Lernen

- Die Performance Vorhersage ist auf heterogenen Systemen nicht einfach
- Laufzeitverhalten von Programmen kann je nach Hardware variieren
- Laufzeitverhalten kann je nach gleichzeitig ausgeführten Programmen variieren
- Kann maschinelles Lernen die Performance Vorhersage auf heterogenen Systemen verbessern?

Betreuer

- Manuel Kalmbach
  E-Mail: `manuel.kalmbach@kit.edu`

Thema 5: Vergleich neuartiger Grafikprozessoren

- GPUs heutzutage in vielen Bereichen nicht wegzudenken
- Aufgabe: Vergleich der neusten Architekturen der bekannten Hersteller

Betreuer

- Roman Lehmann
  E-Mail: roman.lehmann@kit.edu

Thema 6: Pruning-Stategien von neuronalen Netzen zur Verbesserung der Speicherauslastung und Rechenzeit

- Neuronale Netze finden in vielen Bereichen Anwendung
- Tendenz zu sehr großen Netzen mit Millionen von Parametern
- Pruning ist eine Strategie zum Reduzieren der Parameteranzahl
- Aufgabe: Vergleich und Analyse der gängigen Pruning-Strategien

Betreuer

- Roman Lehmann
  E-Mail: roman.lehmann@kit.edu

Thema 7: Quantisierungs-Strategien von neuronalen Netzen zur Verbesserung der Speicherauslastung und Rechenzeit

- Neuronale Netze finden in vielen Bereichen Anwendung
- Tendenz zu sehr großen Netzen mit Millionen von Parametern führt zu einem sehr hohen Speicherverbrauch und langen Ausführungszeiten
- Quantisierung ist eine Strategie zur Reduzierung des Speicherverbrauchs. Gleichzeitig kann die Berechnung enorm beschleunigt werden.
- Aufgabe: Vergleich und Analyse der gängigen Quantisierungs-Strategien

Betreuer

- Roman Lehmann
  E-Mail: `roman.lehmann@kit.edu`

Thema 8: Hardware-Beschleuniger speziell für die Anwendung im Bereich der neuronalen Netzen

- Neuronale Netze finden in viele Bereichen Anwendung
- Trend zu speziellen Hardware-Beschleuniger für neuronale Netze
- Aufgabe: Analyse und Vergleich spezieller Hardware-Beschleuniger für die Anwendung neuronaler Netze

Betreuer

- Roman Lehmann
  E-Mail: roman.lehmann@kit.edu