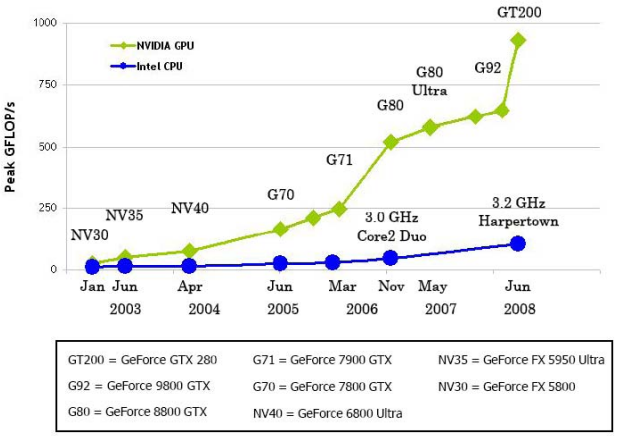
Informationssysteme
Prof. Dr. Felix Naumann
Masterprojekt Sommersemester 2010

Duplikaterkennung auf GPUs

Hintergrund

Duplikaterkennung beschreibt das Auffinden unterschiedlicher Darstellungen gleicher Realwelt-Objekte in einem Datenbestand. Hierzu wird eine Menge von Datensatz-Paaren gebildet und auf diese jeweils eine Ähnlichkeitsfunktion angewendet. Aufgrund der hohen Anzahl von Vergleichen erscheint es sinnvoll, die Ausführung zu parallelisieren.

Die Rechenleistung von GPUs moderner Grafikkarten übersteigt mittlerweile die Rechenleistung von CPUs. So sind beispielsweise in den Top 5 der weltweiten Supercomputer 3 Systeme, die auch GPUs verwenden. Die Anwendungsgebiete für GPUs sind vielfältig und umfassen u.a. Simulationen, Bildbearbeitung und Finanzanwendungen.



http://www.hardwareinsight.com/nvidia-cuda/

GPUs erlauben die parallele Ausführung gleicher Operationen auf unterschiedlichen Daten (Datenparallelität). Nachteilig ist jedoch der Aufwand für das Kopieren von Daten zwischen Hauptspeicher und Grafikkartenspeicher. Weiterhin müssen Algorithmen so optimiert werden, dass eine möglichst hohe Anzahl paralleler Prozesse möglich ist.

Beschreibung

Aufgrund des enormen Rechenpotentials moderner Grafikkarten ist im Rahmen des Masterprojekts zu untersuchen, inwieweit die Verwendung von GPUs zur Duplikaterkennung geeignet ist und die o.g. Nachteile durch die zusätzliche Rechenleistung im Vergleich zur Ausführung auf der CPU wieder ausgeglichen werden. Es ist eine enorme Leistungssteigerung zu erwarten.

Im Rahmen des Masterprojekts sind folgende Aufgaben vorgesehen:

- Implementierung verschiedener Ähnlichkeitsmaße (z.B. Edit-Distance, Soundex) für die GPU und die CPU
- Implementierung verschiedener Algorithmen zur Bildung von Datensatz-Paaren (z.B. naiver Ansatz, Blocking, Windowing) sowie Optimierung für die Ausführung auf der GPU
- Performance-Vergleich der Duplikaterkennung auf der Grafikkarte mit der Ausführung auf einem Mehrprozessor-System (z.B. Quad-Core-Prozessor)
- Einreichung eines Artikels zur Veröffentlichung auf der ICDE 2012.

Das Masterprojekt ist für 3-4 Studenten mit Grundkenntnissen zum Thema Duplikaterkennung (z.B. Vorlesung Informationsintegration, Workshop oder Seminar zum Thema Duplikaterkennung) geeignet. Weiterhin sind Programmiererfahrungen in C bzw. mit CUDA/OpenCL hilfreich.

Kontakt

- Prof. Dr. Felix Naumann naumann@hpi.uni-potsdam.de
- Uwe Draisbach uwe.draisbach@hpi.uni-potsdam.de