

Dynamisches Ressourcen Scheduling auf Grafik-Prozessoren

Mit dem Dissertationspreis der Gesellschaft für Informatik (GI) als beste Informatikdissertation im deutschsprachigen Raum wurde in diesem Jahr erstmals ein Österreicher ausgezeichnet: Markus Steinberger. Er promovierte im Jahr 2013 zum Thema „Dynamic Resource Scheduling on Graphics Processors“ an der Technischen Universität Graz.

Die Dissertation beschäftigt sich mit dem Problem, wie sich eine höhere Rechenleistung erreichen lässt, wenn traditionelle Ausführungsstrategien den Bedürfnissen der Informatikanwendungen nicht nachkommen können, sprich: wenn Anwendung auf Grund ihrer

Komplexität nur Bruchteile der eigentlich zur Verfügung stehenden Rechenleistung (des Chips) nutzen können. Im nachfolgenden Beitrag stellt Markus Steinberger seine Arbeit und sein Forschungsgebiet vor.

Innerhalb der letzten Jahre hat ein Paradigmenwechsel im Bereich der Computertechnologie stattgefunden. Da die Chipproduktion an die physikalischen Grenzen gestoßen ist, ist eine Erhöhung der Prozessorktaten nicht mehr zielführend. Der einzige Ausweg, um die weiterhin steigende Nachfrage nach immer mehr Rechenleistung bedienen zu können, ist Parallelisierung. Während die ersten parallelen Prozessoren ihre Rechenleistung einfach damit ausspielten, mehrere Anwendungen gleichzeitig auszuführen, sind solch einfache Strategien heutzutage nicht mehr ausreichend, um die volle Rechenleistung moderner paralleler Chips auszureizen. Jegliche Algorithmen müssen von Grund auf darauf ausgerichtet sein, gleichzeitig von einer Vielzahl an Prozessoren abgearbeitet zu werden.

Der aktuell leistungsfähigste parallele Chip ist die Graphics Processing Unit (GPU, dt. Grafikprozessor). Auf einer aktuellen GPU finden sich mehr als 3000 individuelle Rechenkerne, deren effiziente Nutzung für eine Vielzahl an Forschungsgebieten essentiell geworden ist. Simulationen in

diversen Gebieten wie Genetik, Welt- raumforschung oder Medizin sind Paradebeispiele für den Einsatz von Grafikprozessoren. Trotz der Wichtigkeit dieser Simulationen gestaltet sich die Ausführung auf Grafikprozessoren in vielen Bereichen als äußerst schwierig. Eine Portierung auf die GPU erreicht meist nur einen Bruchteil der möglichen Leistung. Die Gründe für dieses Problem finden sich im GPU Programmier- und Ausführungsmodell.

Unsere Arbeit beschäftigt sich mit dieser Problematik und zeigt, wie neuartige, dynamische Schedulingstrategien fähig sind, eine unflexible GPU in eine parallele Rechenmaschine zu verwandeln, die sich an hochdynamische Algorithmen anpasst und somit ihre wahre Rechenleistung freilegt. Dazu präsentieren wir ein Ausführungsmodell, das eine effiziente Abarbeitung von inhomogenem, zeitlich veränderlichem Parallelismus ermöglicht. Den Grundstein dieses Modells bildet ein anpassungsfähiger Scheduler, basierend auf hocheffizienten Warteschlangen. Dieser Scheduler kombiniert Arbeitspakete verschiedener Art zum Zwecke schnellerer, kooperativer Ausführung. Naturgemäß erlaubt er die gleichzeitige Nutzung der GPU durch mehrere Algorithmen bei fairer Ressourcenaufteilung. Eine detaillierte Kontrolle der Ausführungsabfolge innerhalb eines Algorithmus ermöglichen wir durch frei definierbare, dynamische Prioritäten. Zur Komplettierung der Ressourcenverwaltung stellen wir einen dynamischen Speicherverwalter vor, welcher zehntausende Anfragen gleichzeitig bedienen kann.

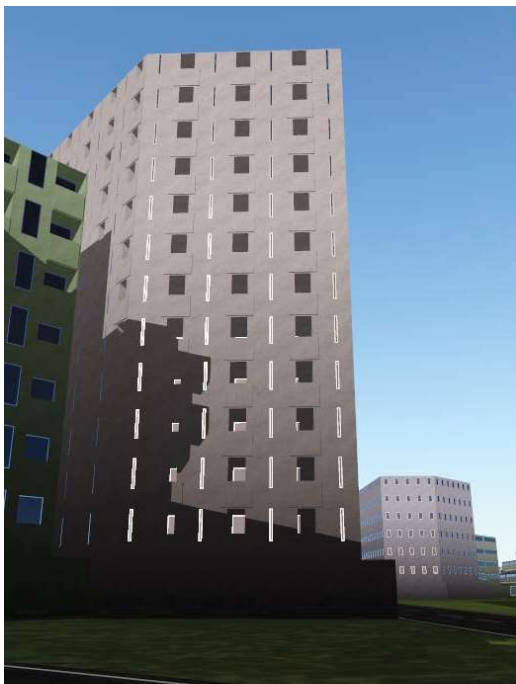

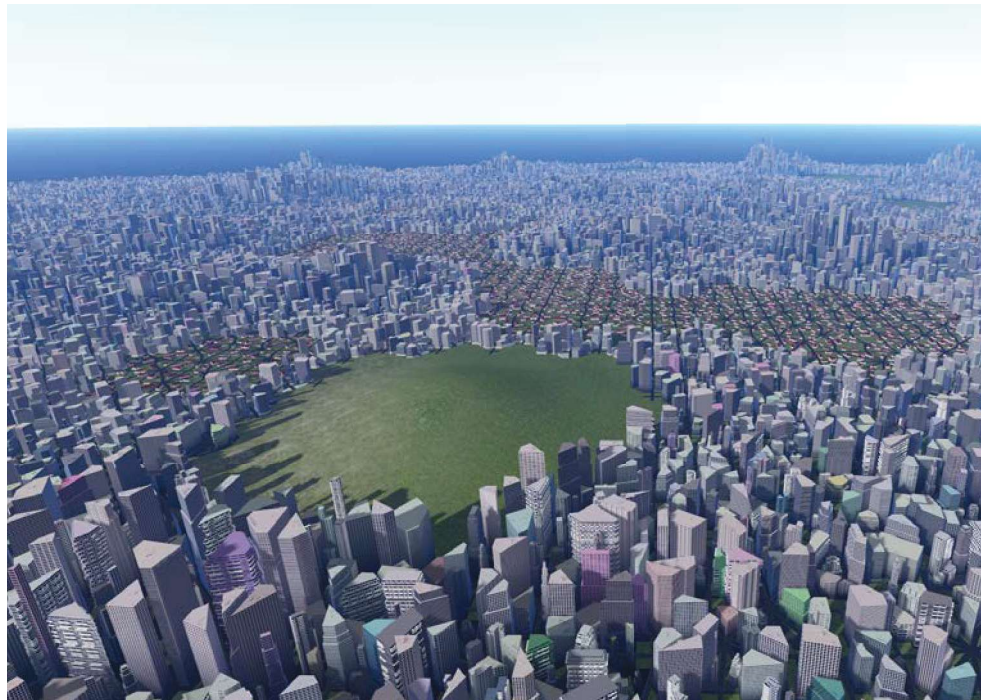


Abbildung 1: Die Darstellung von Städten in Echtzeit wird möglich.
Bild: © Steinberger

Abbildung 2: Die sichtbaren 28km² dieser Stadt beinhalten 47 000 Gebäude und werden von 240 Millionen Regeln generiert. Dank unseres Scheduling ist es möglich, die Geometrie der Stadt 20 Mal in der Sekunde zu generieren.
Bild: © Steinberger

Ergebnis unserer Forschung zur Ressourcenverwaltung auf Grafikprozessoren ist nicht nur der zurzeit schnellste Warteschlangenalgorithmus, der effizienteste Speicherverwalter und der momentan einzige autonome Scheduler mit Unterstützung unterschiedlicher Grade an Parallelismus, sondern auch eine Weiterentwicklung des Standes der Technik in den Bereichen Bildsynthese, Visualisierung und geometrischer Algorithmen. So beschleunigt unser Modell globale Beleuchtungssimulationen durch die Fokussierung der verfügbaren Rechenleistung auf jene Bereiche, die den stärksten Beitrag zur Verbesserung der Bildqualität erwarten lassen. Im Bereich der Visualisierung analysieren wir drei Techniken zur Volumendarstellung und zeigen, wie durch intelligentes Scheduling suboptimale Ausführungskonfigurationen vermieden werden können. Schlussendlich beschreiben wir die erste Grammatik, die komplette Städte in Echtzeit auf der GPU generiert und darstellt (siehe Abbildung 2). Eine Generierung mit traditionellen Methoden würde Stunden benötigen und den verfügbaren Arbeitsspeicher sprengen.

Da der Grad an paralleler Verarbeitung in Zukunft noch weiter ansteigen wird, ist es von essentieller Wichtigkeit, genügend Parallelismus innerhalb von Algorithmen zur Verfügung zu stellen. Grundsätzlich gibt es drei Möglichkeiten zukünftige Generationen von Grafikprozessoren mit ausreichend Parallelismus zu versorgen: Ein größerer Anteil an Parallelismus kann durch algorithmische Anpassungen freigelegt werden; mehrere Algorithmen können gleichzeitig ausgeführt werden; oder ein alternatives Scheduling kann die volle, im Algorithmus vorhandene Parallelität auf die Hardware abbilden. In allen drei Fällen sind unsere Ausführungs- und Programmieretechniken von großem Nutzen. Daher kann unsere Arbeit das Design zukünftiger GPU-Programmiersprachen, -Compiler und -Scheduler positiv beeinflussen. 



Der GI-Dissertationspreis

Mit dem GI-Dissertationspreis würdigen die beteiligten Gesellschaften – die Gesellschaft für Informatik e.V. (GI), das German Chapter of the ACM (GChACM), die Österreichische Computer Gesellschaft (OCG) und die Schweizer Informatik Gesellschaft (SI) – jährlich eine herausragende wissenschaftliche Arbeit, die aufzeigt, wie Algorithmen auf modernen, parallelen Architekturen effizient implementiert werden können.

In diesem Jahr hat das Bundesministerium für Bildung und Forschung (BMBF) die Überreichung des GI-Dissertationspreises übernommen. Die Auszeichnung fand am 24. September 2014 in Anwesenheit des deutschen Staatssekretärs Georg Schütte und des Präsidenten der Österreichischen Computer Gesellschaft, Reinhard Goebel, auf der INFORMATIK 2014 in Stuttgart statt.



Dipl.-Ing. Dr. Markus Steinberger schloss sein Diplomstudium (Tele-
matik) an der Techni-
schen Universität Graz

mit Auszeichnung im Jahre 2010 ab. Danach verfolgte er ein Doktorstudium zum Thema GPU Scheduling, Visualisierung und Geometrische Modellierung, welches er im Jahre 2013 unter der Betreuung von Prof. Dieter Schmalstieg abschloss. Dank seiner ausgezeichneten schulischen und studentischen Leistungen, wurde ihm die Ehre einer Promotio sub auspiciis Praesidentis rei publicae zuteil. Von 2013 bis 2014 arbeitete er in der Computer Vision Forschung Gruppe von Kari Pulli bei NVIDIA in den USA. 2014 erhielt er als erster Österreicher den GI Dissertationspreis. Seine weiten Forschungsinteressen spiegeln sich in der Vielzahl der Preise wider, die seine Publikationen gewonnen haben. Darunter befinden sich ACM CHI, IEEE Infovis, Eurographics, ACM NPAR, EG/ACM HGP Best Paper und Honorable Mention Awards. Seit 2014 leitet er am Institut für maschinelles Sehen und Darstellen an der Technischen Universität Graz die GPU- und Parallelverarbeitungsgruppe.