Energy- and Cost-Efficient Lattice-QCD Computations Using Graphics Processing Units

Matthias Bach

Große internationale Kooperationsprojekte am *Large Hadron Collider* am CERN, sowie zukünftig an der *Facility for Antiproton and Ion Research (FAIR)* am GSI Helmholtzzentrum für Schwerionenforschung GmbH, beschäftigen sich mit dem Verständnis der Quantenchromodynamik (QCD). Diese beschreibt die Wechselwirkung zwischen Gluonen und Quarks, den Bausteinen aller hadronischer Materie. Störungstheoretischen Ansätzen ist die QCD allerdings nur im Bereich hoher Energien zugänglich. *Ab initio* lässt sie sich für niedrigere Energien nur durch die Diskretisierung auf ein euklidisches Gitter in Raum und Zeit rechnen. Dieser Ansatz ist als Gitter-QCD bekannt.

Gitter-QCD-Rechnungen werden aufgrund ihres hohen Rechenbedarfs auf den größten wissenschaftlichen Clustern durchgeführt und haben wiederholt deren Architektur beeinflusst. Seit dem Aufkommen der Nutzung von Grafikprozessoren für nichtgrafische Berechnungen (GPGPU) wurden diese auch für die Berechnung der Gitter-QCD interessant. Anders als traditionell für die Gitter-QCD genutzte Rechner sind sie ein Massenmarktprodukt, was Vorteile in Hinblick auf Preis und Weiterentwicklung verspricht.

Im Rahmen dieser Dissertation wurde CL^{2}QCD entwickelt, eine auf OpenCL basierende Anwendung, welche Gitter-QCD-Rechnungen sowohl auf Grafikprozessoren als auch auf traditionellen Prozessoren ermöglicht. Anders als andere GPGPU-Anwendungen für Gitter-QCD ist CL^{2}QCD nicht auf Grafikprozessoren eines einzelnen Herstellers beschränkt. Mit 120 GFLOPS auf einer AMD Radeon HD 7970 bietet sie den schnellsten $\not{D}$ für doppelt genaue Rechnungen auf einem einzelnen Grafikprozessor. $\not{D}$ ist die rechenintensivste Operation der Gitter-QCD und wird häufig genutzt, um die Leistung verschiedener Systeme zu vergleichen.

Die Rechenleistung von Gitter-QCD-Anwendungen wird vor allem durch die beim Zugriff auf den Speicher verfügbare Datenrate bestimmt. Außerdem unterscheidet sich die Charakteristik von Speicherzugriffen eines Grafikprozessors deutlich von denen eines klassischen Prozessors. So optimiert der Cache eines Grafikprozessors vor allem gleichzeitige Zugriffe unterschiedlicher Threads, während er auf einem klassischen Prozessor aufeinander folgende Zugriffe des gleichen Threads optimiert. Möchte man

größere Datentypen auf hochperformante Datentypen abbilden, indem man das sogenannte *structure of arrays (SoA)*-Pattern nutzt, zeigt sich, dass sich die relative Positionierung der Daten im Speicher stark auf die erreichbaren Datenraten auswirkt. Durch einen im Rahmen dieser Arbeit entwickelten Algorithmus, der den Abstand zwischen den einzelnen Datensegmenten optimiert, lässt sich die bestmögliche Datenrate für Daten beliebiger Größe erzielen. Im $\not{D}$ wird die Rechenleistung durch die Anwendung der untersuchten Optimierungen versechsfacht.

Studien der verschiedenen Kommunikationstechniken zwischen Grafik- und Hauptprozessoren sowie zwischen verschiedenen Grafikprozessoren ermöglichen hochperformante Krylov-Raum-Löser. Die für die Gitter-QCD notwendige Abtastung des Phasenraumes wird durch einen hybriden Monte-Carlo-Algorithmus (HMC) erreicht. In diesem erreicht eine AMD Radeon HD 7970 die vierfache Rechenleistung von zwei AMD Opteron 6220, welche eine optimierte Referenzanwendung nutzen.

Die im Rahmen dieser Arbeit entwickelte Anwendung nutzt die untersuchten Optimierungen, um Grafikkarten bei der Berechnung von Gitter-QCD optimal auszunutzen. CL^{2}QCD ist modular aufgebaut, so dass die Domänenlogik und die Optimierung auf die benutzten Prozessoren entkoppelt sind. Ein Prototyp des Langevin-Algorithmus, einer Alternative zum HMC, wurde innerhalb eines Tages entwickelt und profitierte bereits von allen Optimierungen. Um die in Frankfurt durchgeführten Studien zu ermöglichen, wurde außerdem besonderes Augenmerk auf die erreichte Rechenleistung bei kleinen Gittern gelegt.

Zusätzlich wurde eine für die Domänenlogik transparente Möglichkeit geschaffen, die Ausführung auf mehrere Grafikprozessoren zu verteilen. Dies erlaubt Rechnungen, die nicht in den Speicher eines einzelnen Prozessors passen. Die größte Herausforderung hierbei ist die Kommunikation zwischen den Grafikprozessoren, welche, anders als Prozessoren in traditionellen Gitter-QCD-Systemen, nicht ohne weiteres direkt miteinander kommunizieren können. Bei der Nutzung mehrerer Grafikprozessoren skaliert CL^{2}QCD für hinreichend große Gitter mit einer Effizienz von 85 % linear. Im $\not{D}$ erreichen die vier Prozessoren auf zwei AMD FirePro S10000 bis zu 400 GFLOPS. Im iterativen Gleichungssystemlöser werden 250 GFLOPS erreicht.

CL^{2}QCD bietet nicht nur eine hohe Rechenleistung, sondern übertrifft Systeme ohne Grafikprozessor auch in der Energieeffizienz: bei vierfacher Rechenleistung bietet CL^{2}QCD auch eine vier mal bessere Energieeffizienz. Durch die Nutzung mehrerer Grafikprozessoren in einem einzelnen System lässt sich dieser Vorteil sogar bis zu einem Faktor von 5,25 steigern. Hierbei ist allerdings auf eine hinreichende Kühlung der Grafikprozessoren zu achten.

Bei der Betrachtung der auf die Rechenleistung normalisierten Anschaffungskosten liegen grafikprozessorbasierte Systeme gleichauf mit konventionellen, großen Systemen, die für die Gitter-QCD genutzt werden. Anders als diese können grafikprozessorbasierte Systeme aber von einem einzelnen Grafikprozessor hochskaliert werden. Beispiele für solche grafikprozessorbasierte Systeme sind LOEWE-CSC und SANAM. Auf beiden wurde CL^{2}QCD bereits produktiv genutzt, um Daten für physikalische Studien zu generieren.