Karlsruher Institut für Technologie

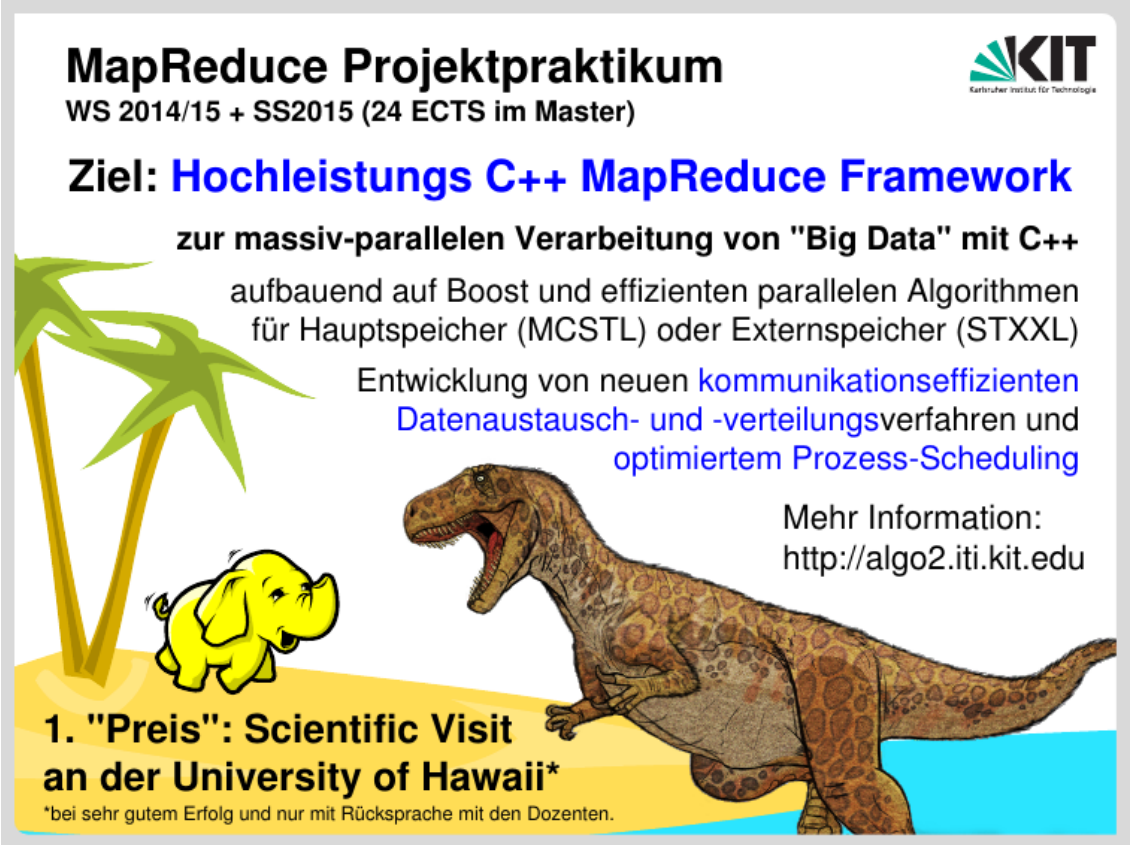
Institut für Theoretische Informatik (ITI)
Algorithmik II
Prof. Dr. Peter Sanders

Verteilte Datenverarbeitung mit MapReduce

Projektgruppe „Praxis der Forschung“
Wintersemester 2014/15

1 Beschreibung

In der Veranstaltung „Praxis der Forschung“[1,2] haben Masterstudenten die Möglichkeit in einer Projektgruppe für zwei Semester an einem gemeinsamen forschungsorientierten Thema zu arbeiten. Dabei lernen die Teilnehmer in einer Gruppe wissenschaftlich zu arbeiten, ein größeres Softwareprojekt zu designen und umzusetzen, und die algorithmisch wichtigen Aufgaben im Projekt effizient zu lösen.



2 Inhalt unseres Projekts

MapReduce ist ein einfaches Programmiermodell zur massiv-parallelen Verarbeitung von sehr großen Datenmengen. Es besteht aus den zwei Funktionen map und reduce, die mehrstufig und zusammen angewandt erstaunlich flexible Datenverarbeitung ermöglichen.

Diese Datenanalyse macht einen Kernbereich des Themenkomplexes „Big Data“ aus, bei dem aus sehr großen Datenmengen durch stichhaltige Analysen wertvolle Datenwerte extrahiert oder zusammengefasst werden sollen. Ein in der Presse bekanntes Beispiel ist Googles Versuch die

[1] http://formal.iti.kit.edu/projektgruppe/
[2] https://sdqweb.ipd.kit.edu/wiki/Praxis_der_Forschung:_Modellgetriebene_Software-Entwicklung_-_Teil_1_SS14

Ausbreitung von Grippe vorherzusagen.[3] Neben solchen Analysen wird MapReduce von mehreren Suchmaschinen eingesetzt um Indexberechnungen und andere Aggregationen auf hunderten Maschinen auszuführen.

Die Stärke von MapReduce ist ein einfaches Programmierinterface, das von der Komplexität des Parallelismus und der verteilten Datenhaltung in einem Rechencluster vollständig abstrahiert. Die darunterliegende Verarbeitung wird von einem MapReduce Framework organisiert. Es gibt nur wenige open-source MapReduce Frameworks, wovon Hadoop[4] das bekannteste ist. Hadoop ist in Java geschrieben und wird von Yahoo und Facebook eingesetzt.

In der Projektgruppe „Praxis der Forschung“ sollen ein oder mehrere open-source MapReduce Frameworks für C++ konzipiert und implementiert werden. Neben der Entwicklung eines tragfähigen softwaretechnischen Designs sollen insbesondere die Kern-Algorithmen des Frameworks untersucht und optimiert werden. Beispielsweise stellen die Datenverteilung, das Scheduling von map/reduce Teilprozessen, der Austausch (shuffle) der Zwischendaten, Fehlertoleranz und ähnliche Teilfacetten bereits beträchtliche Herausforderungen dar.

Im Laufe des Projekts wird die Zielrichtung und Schwerpunkte der Frameworks mit den Teilnehmern festgelegt. Am Anfang steht wahrscheinlich ein einfacher Prototyp, der alle Berechnung lokal und sequentiell durchführt, und anhand dessen die softwaretechnischen Interfaces entwickelt werden. Als erste mögliche Erweiterung könnte die Berechnung dann auf einer Mehrsockel-Maschine mit non-uniform memory access (NUMA) parallelisiert werden. Wiederum andere Aspekte müssen betrachtet werden, wenn die MapReduce-Verarbeitung auf einem verteilten, massiv-parallelen Rechencluster statt finden soll, wobei je nach Cluster lokal nur RAM oder sowohl RAM als auch Festplattenspeicher vorhanden ist.

Das im Projekt entwickelte Framework und dessen Prototypen sollen mit verschiedenen einfachen beispielhaften MapReduce-Anwendungen wie verteilte Wortzählung, parallele Bildverarbeitung und PageRank Berechnung auf einer Matrix getestet werden.

Wir suchen für die Projektgruppe 3-5 engagierte Studenten, die Vorkenntnisse in C++ haben, und Spaß daran haben verteilte Algorithmen zu entwickeln, umzusetzen und zu analysieren.

2.1 Voraussetzungen:

- Grundkenntnisse in C++ und Lernbereitschaft für mehr.
- Gute Bekanntschaft mit Linux.
- Gute Noten in PSE und Algorithmen 1 und 2, oder algorithmisches Talent.
- Fähigkeit in einem Team mit anderen Studierenden zu arbeiten.

2.2 Wir bieten:

- Ein interessantes Forschungsprojekt im top-aktuellen „Big Data“ Themenbereich.
- Unterstützung und Erfahrung bei der Umsetzung des Projekts.

Bei sehr gutem Erfolg winkt ein scientific visit an der University of Hawaii, wo eine Forschungsgruppe an ähnlichen Themen arbeitet.

3 Kontakt / Betreuung

- Timo Bingmann
- Christian Schulz
- Sebastian Schlag
- Michael Axtmann

[3] http://en.wikipedia.org/wiki/Google_Flu_Trends
[4] http://hadoop.apache.org/