Aufgabenblatt zum Praktikum Optimierende Compiler

Prof. Dr. Andreas Koch
Jens Huthmann, Julian Oppermann

TECHNISCHE UNIVERSITÄT DARMSTADT

Sommersemester 13
Aufgabenblatt 1

Einleitung

In der Praxis interagieren verschiedene Optimierungsverfahren miteinander. So kann beispielsweise nach der Erkennung konstanter Ausdrücke eine Entfernung von "toten" Bedingungen (immer wahr oder falsch) mit größerem Erfolg vorgenommen werden, als vorher. Um diese Untersuchungen zu erleichtern, soll um den Bantam-Compiler herum eine interaktive Shell entwickelt werden, mit der durch zeilenorientierte Kommandos verschiedene Passes aufgerufen und Zwischenergebnisse dargestellt werden können.

Zur Vereinfachung der weiteren Projektphasen sollen Sie den Compiler dann so modifizieren, dass er nur eine Untermenge der Sprache Bantam als Eingabe akzeptiert und nicht mehr unterstützte Konstrukte als Fehler ausgibt.

Aufgabe 1.1 Interaktive Oberfläche

Implementieren Sie eine interaktive Oberfläche für den Bantam-Compiler in Form einer Kommandozeile. Aus dieser heraus sollen einzelne Passes dann gezielt durch den Benutzer aufrufbar sein, wobei gegebenenfalls auch noch zusätzliche Parameter auf der Kommandozeile angegeben werden können.

Die Oberfläche soll leicht in einer methodischen Form um weitere Kommandos erweiterbar sein (die im Laufe des praktischen Teils der Veranstaltung von Ihnen entwickelt werden).

Für die schon jetzt im Bantam-Compiler **vorhandenen** Passes sollen dabei folgende Kommandos verwendet werden:

`read` *filename* zum Einlesen einer Bantam-Quelldatei (lexikalische und syntaktische Analyse).

`check` zur Durchführung der Kontextanalyse.

`ast2cfg` zur Umwandlung des AST in einen Drei-Adresscode CFG.

`codegen` zur Code-Erzeugung.

`dumpast` zur grafischen Ausgabe des ASTs als DOT (Graphviz) Graph.

`write` *filename* Ausgabe des MIPS-Codes in eine Datei.

`exit` zum Beenden einer Sitzung.

Eine Beispielsitzung könnte also so aussehen:

```
$ bantamc-shell
bc> read HelloWorld.btm
Syntactic Analysis ...
bc> check
Contextual Analysis ...
bc> ast2cfg
CFG generation ...
bc> codegen
Code Generation ...
bc> write HelloWorld.asm
Compilation was successful
bc> exit
$
```

Beachten Sie, dass die Kommandos teilweise aufeinander aufbauen und dies auch überprüfen müssen. So ist zum Beispiel ein `check` nur nach einem `read` sinnvoll, ein `codegen` setzt einen vorhandenen CFG, der mit `ast2cfg` erzeugt wurde, voraus, und ein `write` ein erfolgreiches `codegen`.

Hinweise zur Shell

In erster Linie dient die Shell natürlich dazu, Ihren Compiler interaktiv zu steuern. Je nach Bedarf können einzelne Passes (ggf. mehrfach) aufgerufen oder weggelassen werden. Der von Ihnen entwickelten Compiler wird nach Abgabe von uns getestet, indem zahlreiche, meist kleine Testprogramme kompiliert und ausgeführt werden – jeweils mit und ohne Optimierungen.

Um diese aufwendige Prozedur zu beschleunigen, werden die meisten Tests automatisch von einem speziell dafür entwickelten Skript erledigt. Dieses Skript steuert Ihren Compiler über die Shell und ist daher auf die Einhaltung der geforderten Befehls- und Ausgabeformate angewiesen.

Allgemeine Funktionsweise

Die Shell liest Tastatureingaben und führt die eingegebenen Befehle aus. Statusmeldungen und sonstige Ausgaben werden auf den Bildschirm geschrieben. Das Befehlsformat ist bewusst sehr einfach gehalten:

- Jede eingegebene Zeile entspricht genau einem Befehl.
- Der eigentliche Befehl steht am Anfang der Zeile. Durch je ein Leerzeichen getrennt können ein oder mehrere Parameter folgen.
- Befehle und Parameter bestehen nur aus Buchstaben (a. . . z, keine Umlaute), Zahlen, Bindestrich, Unterstrich und Punkt. Insbesondere enthalten sie keine Leerzeichen. Es sind also keine Maßnahmen erforderlich, um Sonderzeichen zu “maskieren”.

Falls bei der Bearbeitung eines Befehls ein Fehler auftritt, muss dies deutlich gekennzeichnet werden, z. B. indem der Fehlermeldung der Text `ERROR:` vorangestellt und bei Bedarf ein Stack-Trace ausgegeben wird. Schwerwiegende Fehler (z. B. Typfehler bei der Kontextanalyse) können auch dazu führen, dass die Kompilierung ganz abgebrochen und kein Maschinencode erzeugt wird.

Befehlsklassen und Ausgabe

Die zu implementierenden Befehle unterscheiden sich teilweise enorm hinsichtlich ihrer Ausgaben: Während der Befehl `check` sich nur bei Fehlern kurz zu Wort meldet, liefert `dumpast` (siehe unten) Unmengen Text. Um die Auswertung der Ausgaben zu vereinfachen (und die eigentliche Shell-Ausgabe übersichtlich zu halten), müssen besonders “geschwätzige” Befehle ihre Ausgabe in Dateien umleiten können. In Bezug auf die Shell lassen sich die zu implementierenden Befehle somit grob in drei Klassen unterteilen:

a) Shell- und Dateioperation: Dabei handelt es sich um die Befehle `exit` sowie `read` und `write`, mit denen Quelltext aus einer Datei gelesen bzw. Maschinencode in eine Datei geschrieben wird. Diese Befehle sollen nur eine kurze Statusmeldung ausgeben (eine Zeile genügt).

b) Analyse und Transformation: Befehle wie `check`, `codegen` und später `dvnt` führen mehr oder weniger aufwendige Berechnungen bzw. Transformationen mit dem gerade geladenen Programm durch. Dabei können kurze Statusmeldungen während oder nach der Bearbeitung durchaus nützlich sein (z. B. eine Statistik über die von `dvnt` optimierten Teilausdrücke). Die Ausgabe sollte 20 Zeilen jedoch nicht überschreiten.

c) Dump-Ausgabe: Befehle wie `dumpast` führen selbst keine Transformationen durch, sondern geben “nur” die internen Datenstrukturen zur Kontrolle in einen DOT-Graphen aus. Da diese Ausgaben sehr umfangreich sein können, akzeptieren diese Befehle einen zusätzlichen Parameter, über den eine Datei angegeben werden kann, in die die Dump-Ausgabe erfolgt. Ansonsten wird ein fester Dateiname angenommen.

Um das Debuggen zu erleichtern, steht es Ihnen natürlich frei, Ihre Shell geeignet zu erweitern, beispielsweise indem Sie für Ihre eigenen Tests ausführliche Statusmeldungen über einen einen zusätzlichen Parameter (wie z. B. `-v`) einschalten.

Befehlsabfolge

Die Befehle können gewöhnlich nicht in beliebiger Reihenfolge ausgeführt werden. Einige Befehle dürfen nur einmal aufgerufen werden, andere auch mehrfach. Eine robuste Shell muss daher erkennen, ob ein bestimmter Befehl zu einem Zeitpunkt erlaubt ist oder nicht. Ihre Shell muss jedoch mindestens folgende Abläufe beherrschen:

a) Nach dem Start werden `read` und `check` zuerst aufgerufen.

b) Da alle weiteren Schritte auf Basis des CFG erfolgen wird der AST mit `ast2cfg` in einen CFG umgewandelt.

c) Falls das Programm als CFG vorliegt (nach `ast2cfg`), können CFG-Transformationen und -Dumps ausgeführt werden.

d) Dumps können mehrfach durchgeführt werden, z. B. nach jeder Transformation.

e) `codegen` und `write` werden zuletzt aufgerufen.

f) Die Shell kann jederzeit durch `exit` beendet werden.

Je nach Art der Passes können weitere Einschränkungen bzw. Vorbedingungen gelten. Diese werden in den jeweiligen Aufgabenstellungen genauer festgelegt. Sie können vereinfachend davon ausgehen, dass pro Shell-Aufruf nur eine Quelltextdatei bearbeitet wird und dass nach `codegen` keine Transformationen oder Dumps mehr durchgeführt werden. Bei der Durchführung eines Befehls kann sich herausstellen, dass die Kompilierung wegen eines Fehlers abgebrochen werden muss (z. B. wegen Syntaxfehlern im Bantam-Programm, unzulässiger Befehlsabfolge in der Shell oder gar einem internen Fehler im Compiler). Ihre Shell muss diese Situation erkennen und die Ausführung weiterer Befehle ggf. verweigern. Je nach Schwere des Fehlers ist auch denkbar, dass sich Ihre Shell selbst beendet (natürlich mit einer aussagekräftigen Fehlermeldung und bei Bedarf mit Stack-Trace).

Hinweise zur Implementierung

Die Shell liest Befehle vom Eingabestrom stdin (`System.in` in Java) und schreibt Statusmeldungen auf den Ausgabestrom stdout (`System.out`) bzw. stderr (`System.err`). Bei einem "normalen", d. h. interaktiven Aufruf der Shell werden stdin und stdout letzten Endes mit Tastatur bzw. Bildschirm verbunden. Unser Testskript klinkt sich an dieser Stelle ein und sendet selbst Befehle nach stdin bzw. liest und verarbeitet Statusmeldungen von stdout und stderr. Es ersetzt damit Tastatur und Bildschirm. Im Gegensatz zum "echten" interaktiven Aufruf sind einige Besonderheiten zu beachten: Das Testskript "tippt" schneller: Alle auszuführenden Befehle werden auf einmal an die Shell gesendet. Daher kann es passieren, dass ...

- ein Aufruf von `System.in.read()` möglicherweise mehrere Befehle enthält. Beim interaktiven Aufruf ist stdin üblicherweise zeilengepuffert und `read()` liefert immer nur eine Zeile. Tipp: Verwenden Sie `java.io.BufferedReader.readLine()` in Verbindung mit `java.io.InputStreamReader` zum Lesen der Befehle.
- sich noch zusätzliche, jedoch nicht mehr durchführbare Befehle im Eingabepuffer befinden, nachdem die Kompilierung wegen eines Fehlers abgebrochen wurde.
- Der Eingabestrom ist endlich, d. h. irgendwann wird beim Lesen das "Dateiende" (EOF) erreicht. Die Shell muss auch in diesem Fall "sauber" beendet werden und darf z. B. nicht in einer Endlosschleife hängen bleiben.

Abweichungen und Erweiterungen

Die hier beschriebenen Anforderungen sind nur die *Minimalanforderungen*. Ihre Shell darf natürlich schöner, größer, besser sein (was auch zur Notenverbesserung führen kann). Bitte vergessen Sie dabei nicht, alle unterstützten Befehle Ihrer Shell in der Readme-Datei zu dokumentieren (einschließlich der Vorbedingungen) und Abweichungen ggf. mit dem Betreuer abzusprechen.

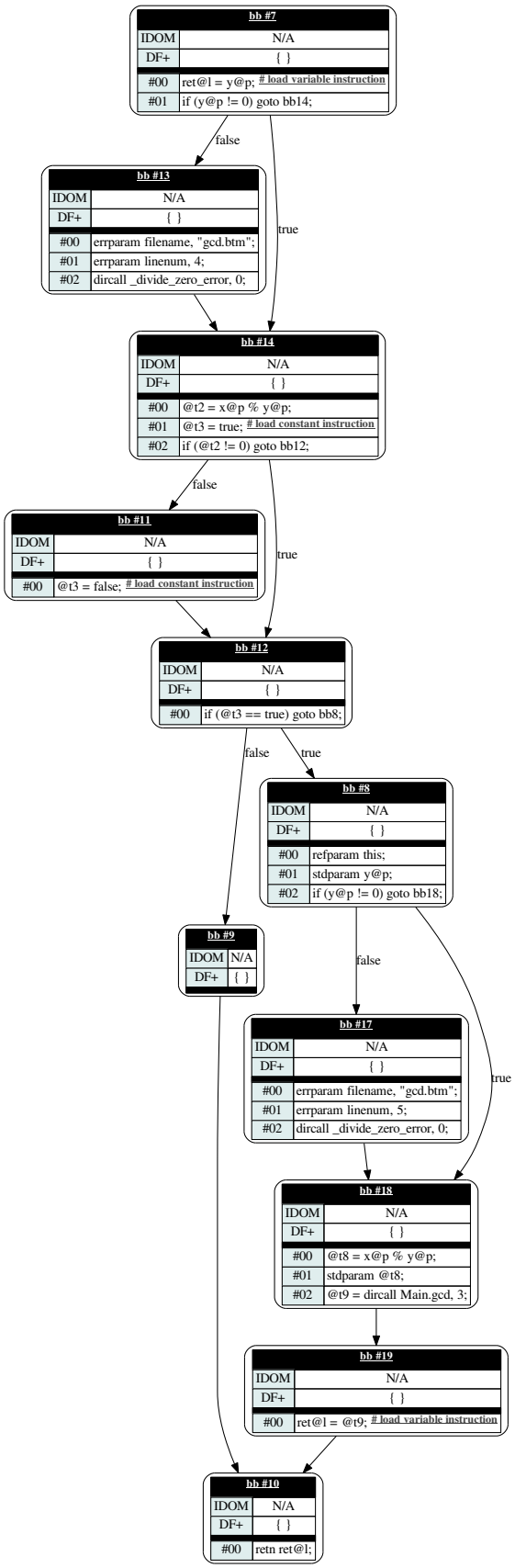
Aufgabe 1.2 Ausgabe des CFG

Der CFG ist *die* entscheidende Darstellung des aktuellen Programms im Bantam-Compiler. Er wird durch verschiedene Optimierungsschritte transformiert. Zum Debugging ist es daher essentiell, sich vor/nach jedem Schritt eine entsprechende Anzeige zu verschaffen.

Hierzu sollen Sie eine Anzeige der internen Struktur des CFGs (sogenannter *dump*) mit allen in der Vorlesung vorgestellten Attributen implementieren. Das Kommando in der Compiler-Shell soll hierzu `dumpcfg` lauten und jeden CFG des aktuellen Programms in einen eigenen DOT-Graphen schreiben. Der DOT-Graph wird anschließend mit GraphViz (`www.graphviz.org`) visualisiert. Ein Beispiel für das Ergebnis sehen Sie in Abbildung 1.

Beachten Sie bei ihrer Implementierung das im Laufe der Vorlesung weitere Attribute des CFGs vorgestellt werden können und Sie ihre Ausgabe entsprechend leicht um diese erweitern können.

Eine detaillierte Darstellung ist bei der Implementierung und der Fehlersuche in den nächsten beiden Aufgabenteilen sicherlich hilfreich.




```
int gcd(int x, int y) {
  int ret = y;
  if (x % y != 0) {
    ret = gcd(y, x % y);
  }
  return ret;
}
```

Abbildung 1: Beispiel füer einen CFG

Aufgabe 1.3 Beispielprogramme in Bantam-Java

Zum Testen ihrer Optimierungen steht Ihnen eine Sammlung an generischen und echten Problemen als Bantam-Java Programme zur Verfügung. Diese Sammlung können Sie von der Webseite zur Veranstaltung herunterladen.

Damit Sie aber auch einen Einblick in die Bantam-Java Sprache bekommen sollen sie noch zusätzliche Testprogramme schreiben. Implementieren Sie eines oder mehrere Testprogramme mit insgesamt mindestens 100 Zeilen Code. Falls Sie knapp (10 Zeilen) drunter sind, bitte nicht mit unsinnigen Formatierungen Zeilen gewinnen. Dann lieber ein etwas kleineres Programm abgeben. Sinn dieser Aufgabe ist es, dass Sie ein Verständnis der Sprache bekommen.

Am Ende der Veranstaltung planen wir einen Wettbewerb, wo die Qualität der Optimierungen, aber auch die Optimierungsschwierigkeit der Testprogramme verglichen werden sollen. Dieser wird aber keine Auswirkung auf die Note haben, sondern es wird einen kleinen Preis von Prof. Koch geben.

Die neuen Programme werden nach der Abgabe dieses ersten Aufgabenpakets von uns zusammengefasst und als Archiv auf der Veranstaltungswebseite veröffentlicht.

Programmierstil

Die von Ihnen erstellten Programme werden in der Endfassung erfahrungsgemäß zwischen 7.000 und 18.000 Zeilen Java umfassen. Um dem Betreuer das Verständnis und Ihnen die Wartung zu erleichtern, sollen Sie von Anfang an einen sauberen und disziplinierten Programmierstil praktizieren.

Bei der Implementierung sind die Konventionen aus *Writing Robust Java Code* weitgehend einzuhalten. Dieses Dokument liegt als PDF auch auf der Web-Seite der Vorlesung. Ergänzend soll folgendes beachtet werden:

- Achten Sie darauf, dass Klassen nicht zu komplex werden (zu viele Instanzvariablen, zu viele Methoden). Bei deutlich mehr als 20 dieser Konstrukte sollten Sie die Klasse aufteilen.
- Analoges gilt für die Komplexität von einzelnen Methoden. Auch hier sollten Sie bei mehr als 100 Programmzeilen Länge die Methode aufteilen.
- Verwenden Sie statt Abfragen von `instanceof` echte objekt-orientierte Konstrukte (z.B. polymorphe Methoden).

Der Test und die Abnahme Ihrer Programme wird vom Betreuer auf Linux mit dem SUN Java Development Kit (JDK) Version 1.7 erfolgen.

Dokumentation

Die Lösungen werden nur durch das unten beschriebene `README` und die in das Java-Programm eingebetteten JavaDoc-Direktiven und Kommentare dokumentiert. Achten Sie daher darauf, dass Sie von diesen beiden Möglichkeiten ausreichend und aussagekräftig Gebrauch machen!

Kommentare sollen am Anfang jeder von Ihnen modifizierten oder neu erstellten Quelldatei, pro Klasse und pro Instanzvariable und Methode verfasst werden. Bei Verwendung relativ kurzer Methoden und aussagekräftiger Bezeichner können sich Kommentare innerhalb von Methoden auf wenige wirklich wichtige Stellen beschränken. Bei komplizierteren Methoden soll der Ablauf aber durch eine größere Anzahl an aussagekräftige Kommentaren im Methodenrumpf verdeutlicht werden.

Der Dateikopfkommentar muss neben einer allgemeinen Beschreibung auch eine Historie von Änderungen enthalten. Jeder Eintrag in dieser Historie beschreibt unter Angabe von Datum/Uhrzeit und Namen des Autors auf 1-2 Textzeilen die Natur der Änderungen. Alternativ kann hier auch das Log Ihres Versionskontrollsystems (z.B. CVS oder besser SVN) verwendet werden.

Diese Angaben sind für den Betreuer wichtig, damit im Kolloquium die für ein Thema passenden Ansprechpartner gefunden werden!

Abgabe

Grundsätzlich schickt jede Gruppe spätestens zum Abgabetermin in einem `.jar`- oder `.zip`-Archiv die Quelldateien Ihrer Version des Bantam-Compilers an

`oc@esa.informatik.tu-darmstadt.de`

mit dem Subject `Abgabe 1 Gruppe` *N*, wobei Ihnen *N* bereits in der Vorlesung mitgeteilt wurde. In dem Archiv sollen nicht nur die eigenen, sondern *alle* (auch unmodifizierten) Quellen des Bantam-Compilers enthalten sein. Ebenso legen

Sie eventuell verwendete zusätzliche externe Bibliotheken in Form ihrer jeweiligen `.jar`-Dateien bei (aber siehe Abschnitt Plagiarismus).

Wichtig: Um unseren Testaufwand überschaubar zu halten, lesen Sie bitte die letzten drei Absätze nocheinmal. Sie sollen abgeben:

- Die *vollständigen* Quellen (also die `.java`-Dateien!)
- Aber *nur diese*, nicht noch irgendwelche Altdaten Ihrer eigenen Testläufe (soweit nicht explizit in der Aufgabenstellung angefordert). Räumen Sie also Ihre Verzeichnisse auf, bevor Sie sie in ein Archiv packen!
- Die eventuell benötigten Fremdbibliotheken
- Alles zusammen in einem `.jar` oder `.zip`-Archiv. Also kein `.7z`, `.tar.gz` oder `.tbz`

Daneben enthält das Abgabearchiv eine Datei `README.txt`, die enthält

- die Namen der Gruppenmitglieder und die jeweils bearbeiteten Themen
- eine Übersicht über die neuen und geänderten Dateien mit jeweils einer kurzen (eine Zeile reicht) Beschreibung ihrer Funktion.
- Hinweise zur Compilierung der Quellen. Geben Sie eine `javac`-Kommandozeile an bzw. verweisen Sie auf mitgelieferte Makefiles oder ANT Build-Dateien. *Nicht* ausreichend ist ein Hinweis auf eine von Ihnen verwendete IDE (wie Eclipse, NetBeans etc.).
- Angaben über weitere Bibliotheken (beispielsweise JSAP, log4j, JUnit etc.), die Sie eventuell verwendet haben. Diese Bibliotheken legen Sie bitte dann auch als `.jar` Dateien in das abgegebene Archiv.

Beurteilung

Die Beurteilung dieser ersten Abgabe erfolgt in jedem Fall via E-Mail. Kolloquien finden nur bei Bedarf statt.

Gruppenarbeit

Bei den Tests von Optimierungen ist es häufig so, dass auf den ersten Blick alles funktioniert. Aber dann in einem Quelltext ein “Sonderfall” auftaucht, der vom bisherigen Optimierungscode noch nicht oder nur fehlerhaft bearbeitet wird. Unterschätzen Sie also den Testaufwand nicht!

Falls in Ihrer Gruppe eine Situation entstehen sollte, in der einzelne Mitglieder deutlich zu wenig (oder zu viel!) der anfallenden Arbeitslast bewältigen, sprechen Sie den Betreuer bitte *frühzeitig* auf die Problematik an. Nur so kann durch geeignete Maßnahmen in Ihrem Interesse gegengesteuert werden. Nach der Abgabe ist es dafür **zu spät** und Sie tragen die Konsequenzen selber (z.B. wenn sich eines Ihrer Team-Mitglieder wegen seiner Verpflichtungen beim Wasser-Polo nur stark eingeschränkt den Mühen der Programmierung widmen konnte, und Sie daher eine unvollständige Lösung abgeben mussten).

Plagiarismus

Der Fachbereich Informatik misst der Einhaltung der Grundregeln der wissenschaftlichen Ethik großen Wert bei. Zu diesen gehört auch die strikte Verfolgung von Plagiarismus. Mit der Abgabe einer Lösung zu den Programmierprojekten bestätigen Sie, dass Ihre Gruppe die alleinigen Autoren des neuen Materials bzw. der Änderungen des zur Verfügung gestellten Codes sind. Im Rahmen dieser Veranstaltung dürfen Sie den Code des Bantam-Compilers von der FG ESA Web-Seite zu Optimierende Compiler sowie Code-Bibliotheken für nebensächliche Programmfunktionen (Beispiele siehe oben) frei verwenden. Mit anderen Gruppen dürfen Sie sich über grundlegenden Fragen zur Aufgabenstellung austauschen. Detaillierte Lösungsideen dürfen dagegen *nicht vor Abgabe*, Artefakte wie Programm-Code oder Dokumentationsteile *überhaupt nicht* ausgetauscht werden. Bei Unklarheiten zu diesem Thema (z.B. der Verwendung weiterer Software-Tools oder Bibliotheken) sprechen Sie bitte Ihren Betreuer gezielt an.

Weitere Infos unter `www.informatik.tu-darmstadt.de/plagiarism`