

UNIVERZITET U BEOGRADU  
MATEMATIČKI FAKULTET



Vladimir Vuksanović

UNAPREĐENJE INFRASTRUKTURE LLVM  
ČUVANJEM ORIGINALNE LOKACIJE PRI  
DEBAGOVANJU IZDVOJENOG KODA

master rad

Beograd, 2023.

**Mentor:**

dr Milena VUJOŠEVIĆ JANIČIĆ, vanredni profesor  
Univerzitet u Beogradu, Matematički fakultet

**Članovi komisije:**

dr Filip MARIĆ, redovan profesor  
Univerzitet u Beogradu, Matematički fakultet

dr Mirko SPASIĆ, docent  
Univerzitet u Beogradu, Matematički fakultet

**Datum odbrane:** \_\_\_\_\_



**Naslov master rada:** Unapređenje infrastrukture LLVM čuvanjem originalne lokacije pri debugovanju izdvojenog koda

**Rezime:**

**Ključne reči:** izdvajanje koda, kompajler, debager, informacije za debugovanje, projekat LLVM

# Sadržaj

<b>1</b>	<b>Uvod</b>	<b>1</b>
<b>2</b>	<b>Kompajleri</b>	<b>2</b>
2.1	Struktura modernih kompajlera . . . . .	2
2.2	Kompajlerske optimizacije . . . . .	3
2.3	Kompajler LLVM . . . . .	4
<b>3</b>	<b>Debageri</b>	<b>12</b>
3.1	Informacije za debugovanje . . . . .	12
3.2	Format DWARF . . . . .	12
3.3	Debager LLDB . . . . .	12
<b>4</b>	<b>Optimizacija izdvajanjem koda</b>	<b>13</b>
4.1	Implementacija u kompajleru LLVM . . . . .	15
4.2	Izdvajanje koda i debug lokacije . . . . .	19
<b>5</b>	<b>Implementacija rešenja</b>	<b>22</b>
5.1	Implementacija u kompajleru . . . . .	22
5.2	Implementacija u debageru . . . . .	31
<b>6</b>	<b>Rezultati</b>	<b>39</b>
<b>7</b>	<b>Zaključak</b>	<b>43</b>
	<b>Bibliografija</b>	<b>44</b>

# Glava 1

## Uvod

# Glava 2

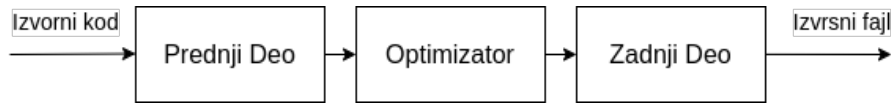
## Kompajleri

### 2.1 Struktura modernih kompajlera

Kompajleri su veoma kompleksni softverski sistemi. Kako postoje već dugo vremena i čine ključne alate za razvoj softvera, kompajleri su tema brojnih istraživanja. Kao rezultat tih istraživanja dobijeno je dosta informacija o dobrim načinima za struktuiranje i implementaciju kompajlera. Moderni kompajleri se obično sastoje iz tri dela [2]:

- Prednji deo – Analizara i prevodi izvorni kôd na neku internu reprezentaciju. Pritom vrši leksičku, sintaksičku i semantičku analizu nad kodom. Ukoliko korisnički kôd nije ispravan, kompilacija se završava ovde.
- Optimizator – Optimizuje internu reprezentaciju kako bi se postigao neki cilj. Može da radi u jednom ili više prolaza.
- Zadnji deo – Prevodi validnu internu reprezentaciju do programa koji može da se izvrši na ciljnoj mašini.

Samo prednji i zadnji deo su potrebni za generisanje mašinskog koda. Optimizator je opcioni deo nastao iz potrebe da dobijeni kôd bude performantan. Svaki dobar kompajler ipak sadrži ovaj deo, i on je možda čak njegov najveći deo. Odnos ovih delova je prikazan na slici 2.1. U svakom trenutku program je enkodiran u neku vrstu interne reprezentacije. Ta reprezentacija postaje sve bliža mašinskom kodu u svakom koraku. Ovakva struktura omogućava lako dodavanje podrške bilo za druge programske jezike i, bitnije, za održavanje novih arhitektura.



Slika 2.1:

## 2.2 Kompajlerske optimizacije

Kompajleri osim prevođenja izvornog koda do mašinskih instrukcija takođe mogu da ga modifikuju za bolje performanse sve dok se optimizovani kôd isto ponaša kao originalni. Performanse mogu da se odnose na vreme izvršavanja programa, memorijski prostor potreban prilikom izvršavanja ili memorijski prostor potreban za skladištenje izvršnog fajla. U zavisnosti od faze kompilacije u kojoj se rade, optimizacije mogu biti nezavisne ili zavisne od ciljne arhitekture. Nezavisne optimizacije rade za sve ciljne arhitekture i oslanjaju se na smanjivanje ukupnog broja operacija koje treba izvršiti. Zavisne optimizacije iskorišćavaju detalje arhitekture kako bi podigli performanse. To se ogleda u korišćenju instrukcija koje rade više operacija u isto vreme ili instrukcija koje rade brže od alternativnih. Čest primer na arhitekturi `x86_64` je zamena instrukcije `mov eax, 0` sa instrukcijom `xor eax, eax`.

Primeri nekih bitnih optimizacija su:

- **Savijanje konstanti (eng. *constant folding*)** – programeri često ostavljaju konstantne matematičke izraze bez promenljivih u kodu zato što je čitljivije nego da izračunaju vrednost. Kako izraz ne sadrži promenljive, nema razlog da se on evaluiira za vreme izvršavanja svaki put već se to radi u vreme kompilacije.
- **Eliminacija mrtvog koda (eng. *dead code elimination*)** – statičkom analizom je moguće u specijalnim slučajevima utvrditi da se neki deo programa nikada neće izvršiti. Brisanjem tog koda štedi se memorija, ubrzava proces kompilacije i može da se ubrza vreme izvršavanja zbog bolje organizacije keš memorije.
- **Globalna numeracija vrednosti (eng. *global value numbering*)** – neki izrazi ili delovi izraza se mogu javiti na više mesta u kratkom rasponu. Ukoliko su izračunati jednom, nije potrebno računati ih ponovo već je moguće obezbediti da prethodni rezultat ostane dostupan u memoriji dok nije ponovo potreban.



- **Razmotavanje petlji (eng. *loop unrolling*)** – izvršavanje petlji sa kratkim telom dovodi do velikog broja skokova u maloj količini vremena. Skokovi su skupa instrukcija i njihovim smanjenjem mogu znatno da se povećaju performanse programa. Razmotavanje se radi tako što se telo petlje ponovi više puta u jednoj iteraciji i broj iteracija se isti broj puta smanji. Ovo je moguće samo u posebnim situacijama kada kompajler ima dovoljno informacija o načinu izvršavanja petlje.

Postoji nekoliko nivoa optimizacije koji se mogu podesiti u prilikom poziva kompajlera. Izbor nivoa određuje koje optimizacije treba da se izvrše a koje ne. Način odabira i nivoi optimizacije su drugačiji na različitim kompajlerima. Na kompajlerima `gcc` i `clang` korisnik može da izabere koji nivo tako što prosledi zastavicu `-O` iza koje se javlja neki od karaktera `0`, `1`, `2`, `3` ili `s`<sup>1</sup>. `O0` je podrazumevani nivo ukoliko nijedan drugi nije specificiran i označava da se kôd uopšte ne optimizuje. Ovaj režim se koristi za pravljenje debug verzija programa. Više reči o tome će biti u poglavlju 3.1. `O1` i `O2` redom uključuju sve veći broj optimizacija podrazumevano da one drastično ne povećavaju memorijsku složenost programa. `O3` je najveći nivo optimizacije vremena izvršavanja. On podrazumeva sve što se koristi u `O2`, ali uključuje i optimizacije koje mogu da znatno povećaju memorijsko zauzeće programa. Isporučene verzije programa se često kompajliraju na ovaj način. Sa druge strane `Os` nivo takođe podrazumeva sve optimizacije iz `O2` ali dodatno i optimizacije koje smanjuju memorijsko zauzeće na uštrb vremena izvršavanja. Ovaj nivo se koristi za uređaje sa ugrađenim računarom (eng. *embedded devices*) koji imaju malu količinu memorije.

## 2.3 Kompajler LLVM

Projekat LLVM započet je 2000. godine na Univerzitetu Illinois od strane Krisa Latnera. Ubrzo se projektu pridružio i njegov mentor Vikram Adve. Cilj projekta je bio proučavanje tehnika kompajliranja u SSA obliku (eng. *Static Single Assignment*) koja podržava statičku i dinamičku kompilaciju proizvoljnih programskih jezika. Inicijalno, naziv LLVM je bio akronim za „virtuelna mašina niskog nivoa” (eng. *Low Level Virtual Machine*). Od tada akronim se više ne koristi, ali je ime ostalo nepromenjeno. Danas, projekat pokriva veliki broj biblioteka i alata za programiranje

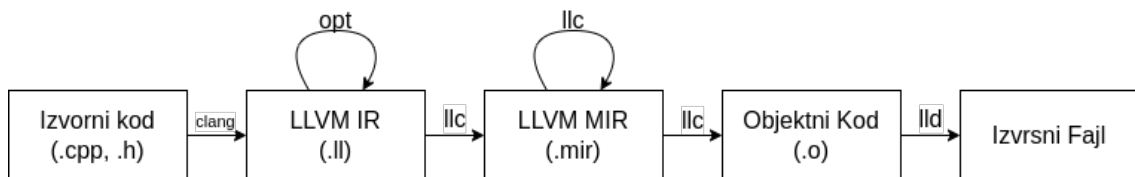
---

<sup>1</sup>Ovo nije potpuna lista. Postoje nivoi koji nisu navedeni.

niskog nivoa [4]. Svaki deo projekta je dizajniran kao biblioteka tako da se može ponovo upotrebiti za implementiranje drugih alata. Celokupan izvorni kôd je javno dostupan na servisu GitHub. Prednost pristupu otvorenog koda je da svako ko želi da poboljša kôd ili ispravi neku grešku može to da učini. Velika zajednica se formirala oko projekta što je znatno doprinelo njegovoj popularnosti. Mnoge firme koriste svoje verzije kompajlera LLVM bilo za podršku neke arhitekture ili kao osnovu za novi programski jezik.

Glavnu komponentu projekta LLVM čini kolekcija biblioteka u okviru istoimenog potprojekta. One implementiraju optimizator i zadnji deo kompajlera i koriste se od strane većine ostalih alata. Drugi bitan potprojekat je clang. On implementira prednji deo kompajlera za jezike C, C++ i Objective-C. LLVM implementira i svoj debager koji se zove LLDB. Više reči o njemu će biti u glavi 3.3. Postoji još veliki broj biblioteka i pomoćnih alata koji nisu pomenuti zato što nisu predmet ovog rada.

Kompajler LLVM je relativno nov u odnosu na druge popularne C++ kompajlere i prati moderniji dizajn. Za razliku od kompajlera GCC, koji je napisan u jeziku C i ima monolitnu strukturu, LLVM uživa u pogodnostima proširenja koje nudi jezik C++ pritom koristeći modularnu arhitekturu. Dok je najveći deo projekta napisan u programskom jeziku C++, postoji interfejs za povezivanje sa jezicima C i Python. Implementacija kompajlera LLVM prati opštu strukturu kompajlera prikazanu u poglavlju 2.1 [1]. Korake kompilacije izvršavaju različiti alati. Odnos različitih reprezentacija koje program ima u toku kompilacije i alata koji konvertuju iz jedne reprezentacije u drugu je prikazan na slici 2.2.



Slika 2.2:

U nastavku će biti opisani redom prednji, srednji i zadnji deo LLVM kompajlera sa fokusom na poslednja dva kako su oni bitniji za ovaj rad.

## Prednji deo

Projekat LLVM sadrži više različitih prednjih delova kompajlera. Druge organizacije i firme takođe proizvode i održavaju prednje delove kompajlera LLVM za svoje jezike. Iz tog razloga, LLVM podržava veliki broj programskih jezika koji uključuju C, C++, Objective-C, Fortran, Haskell, Swift i druge. Prednji deo za programske jezike C i C++ koji su fokus ovog rada naziva se clang. Pre nastavka dalje, potrebno je razjasniti terminologiju. Naime, postoji nekoliko značenja termina clang. Može se smatrati da je to prednji deo kompajlera, kao u skladu sa prethodnim korišćenjem izraza, zatim se može podrazumevati da je to alat koji upravlja celim procesom kompilacije ili može da bude biblioteka koja implementira funkcionalnost prednjeg dela kompajlera. Nadalje se podrazumeva prva interpretacija.

Funkcija ovog dela kompajlera je da izvorni kôd, u ovom slučaju napisan u jezici-ma C ili C++ prevede u LLVM međureprezentaciju. Usput se izvršavaju sve provere ispravnosti izvornog koda i prikazuju se greške i upozorenja korisniku. Program se detaljno analizira iz nekoliko koraka i pritom dolazi do transformacija izvornog koda.

Prvo se prepoznaju osnovne jedinice gramatike programskog jezika, odnosno lekseme. Lekseme se pretvaraju u različite tipove tokena u zavisnosti od toga šta označavaju. Na primer za ključne reči programskog jezika postoje različite vrste tokena, dok svi identifikatori imaju isti tip tokena. Tokeni osim tipa imaju i informacije o opsegu koji zauzimaju u izvornom kodu pomoću kog se može pristupiti njihovom tekstu. Nad nizom tokena se zatim radi sintaksna analiza ili parsiranje u skladu sa pravilima gramatike jezika. Ta pravila se zadaju korišćenjem kontekstno-slobodne gramatike. Primenom pravila gradi se sintakšno stablo gde unutrašnji čvorovi predstavljaju operacije a njihova deca su argumenti. Parsiranje utvrđuje da li program ima ispravnu strukturu ali ne proverava da li taj program ima smisao. To je zadatak semantičke analize. Program može da ima dobru strukturu, ali da nema validno značenje. Prolaskom kroz sintakšno stablo se prikupljaju informacije o funkcijama, promenljivima i drugim objektima i smeštaju se u tabelu simbola. Korišćenjem te tabele proverava se ispravnost sintaksnog stabla. Primer provere koja se vrši na ovom nivou je da li je korišćena promenljiva deklarirana pre toga. Rezultat svih ovih analiza je program u obliku apstraktnog sintaksnog stabla (eng. *Abstract Syntax Tree*, *AST*).

U biblioteci clang parsiranje i semantička analiza su usko povezani, a lekser se poziva po potrebi od strane parsera [14]. Bitan deo leksera čini pretprocesor koji izvršava direktive poput `#include` i `#define`. Prilikom primene pravila gramatike

parser zahteva tokene od leksera i delegira posao kreiranja sintaksnog stabla semantičkom analizatoru. Ukoliko se pronade greška u bilo kojoj analizi, program se ne može kompajlirati i poruka o razlogu se vraća korisniku. U suprotnom, obilaskom apstraktnog sintaksnog stabla generiše se LLVM međukod. Komanda za prevođenje programa do LLVM međukoda prikazana je u listingu 2.3.

```
clang -S -emit-llvm -o output.ll input.c
```

Slika 2.3:

Ukoliko se program uspešno prevede do međukoda, onda je izvorni kôd validan za jezik u kome je napisan. Još uvek je moguće da dođe do greške prilikom povezivanja zbog ne postojanja definicija nekih funkcija, ali će kompilacija do asemblerskog ili objektnog koda biti uspešna. Naravno, činjenica da se program kompajlira ne znači da on radi ono što je programer zamislio niti da će se program uvek uspešno završiti.

## Optimizator

Prednji deo prevodi izvorni kôd u LLVM mašinski nezavisan međukod. U terminologiji se koristi skraćenica IR (eng. *Intermediate Representation*). Izbor reprezentacije međukoda je veoma bitan kada kompajler treba da podrži i više programskih jezika i više ciljnih arhitektura. On treba da je na dovoljno visokom nivou da ne zavisi od arhitekture i u isto vreme na dovoljno niskom nivou da ne zavisi od programskog jezika [13]. LLVM međukod je u SSA obliku. To znači da je svakoj promenljivoj vrednost dodeljena samo jednom i da svakoj upotrebi promenljive prethodi njena definicija. Ovaj oblik je izabran zato što je pogodan za izvršavanje optimizacija.

U listingu 2.4 je dat primer LLVM međukoda za sitan program. Svaki fajl sa LLVM međukodom čuva informacije o jednom modulu. Na početku fajla se nalazi ime fajla u kojem se nalazi izvorni kôd, informacije o načinu zapisa podataka i naziv ciljne arhitekture. U nastavku su globalne promenljive i funkcije. One se prepoznaju po tome što im imena imaju prefiks '@'. Svaka funkcija se sastoji od niza osnovnih blokova, a svaki blok od niza instrukcija. Osnovni blok je jedinica čije instrukcije se uvek linearno izvršavaju. Ako se uđe u osnovni blok on će se uvek izvršiti do kraja. Dakle instrukcije grananja smeju i moraju da se pojave na kraju bloka. Granice osnovnih blokova su označene kao labele u jeziku C. Takođe, početak funkcije započinje prvi osnovni blok. U okviru funkcije lokalne promenljive su zamenjene virtuelnim registrima. Identifikatori virtuelnih registara počinju karakterom '%'. Zbog

SSA oblika, svakom od njih se može dodeliti vrednost tačno jednom. Smatra se da tih registara ima neograničen broj. Instrukcije se sastoje od naziva (`alloca`, `call`, `icmp`, ...), tipa (`i32`, `float`, `label`, ...) i operandada. Ukoliko instrukcija vraća vrednost, on se smešta u novi virtuelni registar. Dodatne informacije o instrukcijama i globalnim objektima se čuvaju u vidu metapodataka. Oni se mogu prepoznati po tome što počinju karakterom '!'. Između ostalog, oni se koriste za informacije za debugovanje. Više reči o primeni metapodataka će biti u glavi 3.1. Kompletan pregled jezika LLVM IR dostupan je na internetu [5].

Optimizacije su implementirane u vidu prolaza (eng. *pass*) koji obrađuju modul, funkciju ili petlju. Prolazi nasleđuju apstraktnu klasu `PassInfoMixin` i implementiraju funkciju `run` sa argumentima za odgovarajuću jedinicu obrade. Prolazi se dele na analize i transformacije. Analize prikupljaju podatke dok transformacije koriste te podatke kako bi izmenili kôd. Transformacija zahteva različite vrste analiza i može da ih poništi ako više ne važe posle promene koda. Optimizacije se primenjuju na program inkrementalno. Program dobijen kao izlaz iz jedne optimizacije postaje ulaz sledećoj optimizaciji. Dakle konačan program zavisi od redosleda izvršavanja optimizacija. Loš redosled može da rezultuje znatno lošijim performansama izvršnog fajla. Svaki nivo optimizacije propisuje koje optimizacije treba da budu primenjene i kojim redosledom.

Za primenu optimizacija na LLVM IR fajlu se koristi alat `opt`. On prima ime na optimizacija koje treba da pokrene i fajl sa ispravnim IR programom koji da optimizuje. Takođe moguće je pokretanje korisnički napisanih prolaza učitanih kao dinamičke biblioteke [11]. Primer korišćenja alata je prikazan u listingu 2.5.

## Zadnji deo

Pre generisanja mašinskog koda, postoji jos jedna, mašinski zavisna, interna reprezentacija. U terminologija projekta LLVM, ona se naziva MIR (eng. *Machine Intermediate Representation*). Ona je dosta bliža ciljnoj arhitekturi i sastoji se od instrukcija te arhitekture. MIR kôd se zapisuje u formatu YAML. Tela funkcija su parsirana kao obična niska. Struktura instrukcija je slična kao za IR kôd. Najveće razlike su u tome što su instrukcije vezane za konkretnu arhitekturu i virtuelni registri su zamenjeni pravim registrima. Kompletan pregled jezika dostupan je u okviru dokumentacije projekta LLVM [6].

Veliki i kompleksan korak u prevođenju koda je spuštanje sa mašinski nezavisne na mašinski zavisnu reprezentaciju. Taj korak se naziva izbor instrukcija (eng. *In-*

```

1  ; ModuleID = 'helloworld.c'
2  source_filename = "helloworld.c"
3  target datalayout =
4  ↪ "e-m:e-p270:32:32-p271:32:32-p272:64:64-i64:64-f80:128-n8:16:32:64-S128"
5  target triple = "x86_64-redhat-linux-gnu"
6
7  @.str = private unnamed_addr constant [3 x i8] c"%d\00", align 1
8  @.str.1 = private unnamed_addr constant [14 x i8] c"hello, world!\00", align 1
9
10 ; Function Attrs: noinline nounwind optnone uwtable
11 define dso_local i32 @main() #0 {
12     %1 = alloca i32, align 4
13     %2 = alloca i32, align 4
14     store i32 0, ptr %1, align 4
15     %3 = call i32 @__isoc99_scanf(ptr noundef @.str, ptr noundef %2)
16     %4 = load i32, ptr %2, align 4
17     %5 = srem i32 %4, 2
18     %6 = icmp eq i32 %5, 0
19     br i1 %6, label %7, label %9
20
21 7:                                     ; preds = %0
22     %8 = call i32 @printf(ptr noundef @.str.1)
23     br label %9
24
25 9:                                     ; preds = %7, %0
26     ret i32 0
27 }
28 declare dso_local i32 @__isoc99_scanf(ptr noundef, ...) #1
29
30 declare dso_local i32 @printf(ptr noundef, ...) #1
31
32 ...
33
34 !llvm.module.flags = !{!0, !1, !2}
35 !llvm.ident = !{!3}
36
37 !0 = !{i32 1, !"wchar_size", i32 4}
38 !1 = !{i32 7, !"uwtable", i32 2}
39 !2 = !{i32 7, !"frame-pointer", i32 2}
40 !3 = !{"clang version 15.0.7 (Fedora 15.0.7-2.fc37)"}

```

Slika 2.4:

```
opt -passes="mem2reg,instrnamer" -o input_opt.ll input.ll
```

Slika 2.5:

*struction Selection*). Postoje tri implementacije izbora instrukcija:

- **SelectionDAG** — Podrazumevana implementacija izbora instrukcija na većini arhitektura.
- **GlobalIsel** — Nova implementacija izbora instrukcija sa modularnim dizajnom, poboljšanim performansama i mogućnosti da optimizuje veći deo kôda. Trenutno nije završena implementacija, ali se planira da u budućnosti zameni **SelectionDAG**.
- **FastIsel** — Pristup izboru instrukcija koji radi veoma brzo, ali generiše neoptimizovan kôd. Ne podržava spuštanje svih ilegalnih instrukcija i u tim situacijama se oslanja na **SelectionDAG**.

Ovaj rad se fokusira na implementaciju **SelectionDAG** zbog svoje stabilnosti i zato što je podrazumevana za arhitekturu **x86\_64**.

**SelectionDAG** je dobio ime po načinu reprezentacije koda za vreme izbora instrukcija. Naime, odgovarajuća klasa se takođe zove **SelectionDAG**. Osnovni blokovi programa su predstavljeni usmerenim acikličkim grafovima (eng. *Directed Acyclic Graphs*). Čvorovi tog grafa su instrukcije koje imaju tip **SDNode**, a grane su različite zavisnosti koje postoje između instrukcija. Postoje dva posebna čvora: ulazni čvor i koren. Ulazni čvor obeležava početak osnovnog bloka i služi samo za postavljanje veza. Sa druge strane, koren označava kraj bloka. Na kraju obrade je vezan za poslednju instrukciju osnovnog bloka. Konstrukcija se radi prolaskom kroz LLVM međukod pomoću klase **SelectionDAGBuilder** koja implementira obrazac „posetilac”. Za svaku instrukciju se kreira čvor na osnovu njenog koda. Konstrukcija čvora se radi direktno u klasi **SelectionDAG**. Veze se kreiraju kada neki čvor koristi izlaznu vrednost drugog čvora. Tip te vrednosti određuje tip veze.

Prethodno konstruisan graf nije pogodan za spuštanje instrukcija i nad njim se vrše dodatne transformacije. On se dodatno optimizuje zamenom grupe čvorova sa jednostavnijim grupama uz pomoć algoritama za uparivanje stabala. Te optimizacije mogu da budu i mašinski nezavisne i zavisne. Graf takođe može da sadrži nepodržane tipove i operacije za ciljnu arhitekturu. Njih je potrebno zameniti odgovarajućim podržanim varijantama. Ove transformacije se zovu legalizacija tipova i operacija. Način legalizacije zavisi od arhitekture i implementiran je u odgovarajućim **TargetLowering** klasama. Prolazi optimizacije i legalizacije se rade nekoliko puta u određenom redosledu i rezultuju grafom spremnim za spuštanje instrukcija.

Proces spuštavanja instrukcija kada su sve potrebne pripreme izvršene je prilično jednostavan. Kao i prilikom optimizacija, algoritam se zasniva na uparivanju čvorova. Čvorovi svake `SelectionDAG` strukture traže svoj pandan za ciljnu arhitekturu. Arhitektura definiše ta uparivanja u svojoj podklasi klase `SelectionDAGISel`. Ona definiše funkciju `Select` koja prima objekat tipa `SDNode` i vraća čvor sa mašinski zavisnim kodom operacije. U ovom koraku se ne zamenjuju baš svi čvorovi već neki prolaze u sledeću fazu.

Poslednji trag LLVM međukoda koji je opstao u programu a da je potrebno da se eliminiše su virtuelni registri. U toku izvršavanja program će imati ograničen broj registara i ne može svaki od njih da se koristi za svaku operaciju. Dodeljivanje registara je korak kada se virtuelni registri zamenjuju konkretnim registrima za ciljnu arhitekturu. Po potrebi dodaju se nove instrukcije za čuvanje i učitavanje vrednosti registara. Dobijeni `SelectionDAG` predstavlja instrukcije potpuno ispravnog MIR programa. Ostalo je još da se one rasporede u linearni redosled uz poštovanje zavisnosti između čvorova. Ovo rezultuje kodom u MIR obliku.

I na ovom nivou postoje optimizacioni prolazi. Optimizacije na ovom nivou su vezane za ciljnu arhitekturu. Svi mašinski zavisni prolazi nasleđuju klasu `MachineFunctionPass` i implementiraju funkciju `runOnMachineFunction` koja se poziva za svaku funkciju u izvornom kodu. Način izvršavanja prolaza je sličan kao za mašinski nezavisan međukod.

Poslednji prolaz na MIR kodu je `AsmPrinter` koji emituje asmeblerski ili objektni kôd za ciljnu arhitekturu. Na ovom mestu se završava zadatak kompajlera. Dalje, ukoliko je emitovan asmeblerski kôd on se prevodi do objektnog koristeći asmebler. Objektni kôd se povezuje sa bibliotekama i drugim objektnim kodom pomoću poveziivača (eng. *linker*) i dobija se izvršni fajl.

Alat koji obavlja posao zadnjeg dela kompajlera je `llc`. Njemu se opciono zadaje vrsta izlaza (asmeblerski ili objektni kôd) i nivo optimizacije. Bitan parametar koji `llc` prihvata je ciljna arhitektura. Ona se zadaje argumentom `-mtriple` i vrednosti u vidu niske koja sadrži arhitekturu, proizvođača, operativni sistem i okruženje. Komanda za prevođenje optimizovanog mašinski nezavisnog međukoda do asmeblerskog fajla je prikazana u listingu 2.6.

```
llc -filetype=asm -mtriple="x86_64-unknown-linux-gnu" -o output.s  
↪ input_opt.ll
```

Slika 2.6:



## Glava 3

# Debageri

3.1 Informacije za debugovanje

3.2 Format DWARF

3.3 Debager LLDB

## Glava 4

# Optimizacija izdvajanjem koda

Izdvajanje koda (eng. *outlining*) je kompajlerska optimizacija koja smanjuje količinu memorijskog prostora koji zauzima program, ali pritom potencijalno povećava njegovo vreme izvršavanja. Ona pronalazi segmente koda (uzastopne nizove instrukcija) koji se ponavljaju u programu, izdvaja ih u zasebnu funkciju i menja pojavljivanja tog segmenta sa pozivom ka novoj funkciji. U nekim slučajevima moguće je da se ubrza i vreme potrebno za kompilaciju programa [7].

Ova optimizacija je posebno korisna za uređaje sa malom količinom memorije, najčešće uređajima sa ugrađenim računarom (eng. *embedded devices*) poput pametnih satova, mp3 plejera ili urađaja zasnovanih na razvojnom sistemu Arduino. Vreme izvršavanja programa može da bude oštećeno ako se izdvoji deo koda koji se bas često izvršava (eng. *hotspot*). To je zato što je poziv funkcije je skupa instrukcija. Takođe, osim samog poziva, stek mora da bude pripremljen za poziv i parametri premešteni na odgovarajuća mesta. Sve to dodaje trošak koji nije prethodno postojao. Izdvajanjem koda koji se retko izvršava, sa druge strane, moguće je da dođe do poboljšanja u vremenu izvršavanja zbog bolje organizacije keš memorije. Nažalost, u vreme kompilacije je teško odrediti koliko puta će se izvršiti segment koda, time ni razlika u vremenu izvršavanja programa nije lako predvidiva.

U listingu 1 je prikazan primer C koda. Redovi 11-14 i 16-19 predstavljaju dobre kandidate za izdvajanje. Vrš se iste operacije samo sa drugačijom promeljivom. Ukoliko se ovaj kôd prevede do asemblerskog koda komandama prikazanim u listingu 2 tako da se pokrene samo optimizacija izdvajanja koda, rezultujući kôd je prikazan u listingu 3. Može se primetiti da je konstruisana nova funkcija `outlined_ir_func_0`. Ona sadrži sve instrukcije iz izdvojenih redova, a ti redovi su zamenjeni pozivom ka toj funkciji.

```
1  int global;
2
3  __attribute__((noinline)) int foo(int a, int b) {
4      return a / b * 2;
5  }
6
7  int main() {
8      int x = 0;
9      int y = 1;
10
11     int c1 = x + y;
12     c1--;
13     global += 2;
14     foo(c1, global);
15
16     int c2 = x + y;
17     c2--;
18     global += 2;
19     foo(c2, global);
20
21     return c1;
22 }
```

Listing 1: Primer C programa koji je dobar kandidat za izdvajanje koda

```
clang -Xclang -disable-llvm-passes -g -O3 -S -emit-llvm test.c -o test.ll
opt -passes="iroutliner" -ir-outlining-no-cost=true -S test.ll -o test_opt.ll
llc -filetype=asm test_opt.ll -o test.s
```

Listing 2: Instrukcije za prevođenje koda sa uključenim izdvajanjem instrukcija

Ova optimizacija predstavlja suprotan proces poznatijoj optimizaciji, umetanju koda (eng. *inlining*). Umetanje koda podrazumeva zamenu poziva ka funkciji sa telom te funkcije. Najčešće se primenjuje na kratke i jednostavne funkcije, koje se pozivaju na više mesta. Time se poništava cena poziva funkcije, što poboljšava vreme izvršavanja, ali pritom pošto su neki delovi koda duplirani dolazi do povećanja memorijskog zauzeća koda.

Iako imaju suprotne efekte, obe optimizacije je moguće koristiti i zajedno, na primer za delimično umetanje [16]. Ova ideja koristi izdvajanje koda kako bi iz funkcije namenjene za umetanje bili izdvojeni „hladni” delovi koda i time se smanjile negativne posledice umetanja. U tom kontekstu izdvajanje koda ne traži sekvence koje se ponavljaju već koristi heuristike da pronađe kôd koji se retko izvršava.

U jezicima C i C++ postoji ključna reč `inline` koja nagoveštava kompajleru da

```
1  main:                                     # @main
2  .Lfunc_begin1:
3      .loc    0 11 0 is_stmt 1              # outline.c:11:0
4      .cfi_startproc
5      ...
6      movq    %rbx, %rdi
7      movq    %r14, %rsi
8      callq   outlined_ir_func_0
9      ...
10     movq    %rbx, %rdi
11     movq    %r14, %rsi
12     callq   outlined_ir_func_0
13     ...
14     .type    outlined_ir_func_0,@function  # -- Begin function outlined_ir_func_0
15 outlined_ir_func_0:                       # @outlined_ir_func_0
16 .Lfunc_begin2:
17     .loc    0 0 0 is_stmt 1              # outline.c:0:0
18     .cfi_startproc
19     ...
```

Listing 3: Isecci iz asemblerskog koda koji sadrži funkciju dobijenu izdvajanjem koda

primeni umetanje koda na označenu funkciju. Kompajler nije primoran da primeni optimizaciju, već je slobodan da izabere da li će funkciju umetnuti ili ostaviti poziv. Za razliku od toga, za izdvajanje koda ne postoji sličan mehanizam. Ukoliko kompajler podržava ovu optimizaciju, na njemu je da izabere koji delovi koda će biti izdvojeni. Ova optimizacija nije podrazumevano uključena, već programer može da je uključi na podržanim kompajlerima najčešće kompilacijom programa sa opcijom `-Os` (ili `-Oz` za clang) koja optimizuje program za memoriju umesto za vreme.

## 4.1 Implementacija u kompajleru LLVM

Kompajler LLVM, kao jedan od najpopularnijih kompajlera za jezik C++ ima podršku za veliki broj optimizacija. Jedna od tih optimizacija je i izdvajanje koda. Postoje dve implementacije izdvajanja koda u okviru kompajlera LLVM:

1. `MachineOutliner` — na mašinski zavisnom međukodu (MIR) [12]
2. `IROutliner` — na mašinski nezavisnom međukodu (IR) [10]

Obe su implementirane kao prolazi (eng. *pass*) koji rade na nivou modula tako da omoguće izdvajanje koda iz različitih funkcija. Ideja je da obe implementacije komplementiraju jedna drugu i daju bolje rezultate nego ako bi se koristila samo

jedna od njih. Oba mesta implementacije imaju svoje prednosti i mane. Na mašinski zavisnom međukodu su dostupne informacije o ciljnoj arhitekturi čime je moguće dosta bolje proceniti cene izdvajanja instrukcija. Bolji model cena instrukcija bitno utiče na određivanje delova koda koji će biti izdvojeni. Problem na ovom nivou je što je u tom trenutku već izvršena alokacija registara. Kandidati za izdvajanje su ograničeni samo na sekvence koje koriste iste registre, iako možda postoji više sekvenci koje vrše isto izračunavanje samo sa drugim registrima. Sa druge strane, na mašinski nezavisnom međukodu se koriste virtuelni registri. Oni ne moraju da se poklapaju između sekvenci tako da je moguće izdvojiti mnogo veći deo koda. Međutim ovde se sada javlja problem procene dobitka memorije zato što se bez konkretne ciljne arhitekture ne može odrediti veličina instrukcija. Zbog korišćenja procena za cene instrukcija u obe implementacije je moguće da rezultat izdvajanja koda ima veću veličinu od početnog koda.

Obe implementacije se zasnivaju na istoj ideji, samo su adaptirane za međukod koji optimizuju. U nastavku će prvo biti objašnjen opšti algoritam, a zatim specifičnosti obe implementacije.

## Opšti algoritam za izdvajanje koda

Opšti algoritam za izdvajanje koda sadrži dva osnovna koraka: odabir kandidata i izdvajanje koda.

Algoritam za odabir kandidata se zasniva na ideji da se program može tretirati kao niz numeričkih vrednosti gde se svakoj instrukciji dodeljuje jedna vrednost. Instrukcije se dele u tri grupe:

- Legalne instrukcije su one koje je dozvoljeno izdvojiti. Tu spada većina instrukcija: aritmetičke, logičke...
- Ilegalne instrukcije su one čije izdvajanje je zabranjeno kao što su pozivi funkcija ili instrukcije vezane za obradu grešaka.
- Nevidljive instrukcije ne utiču na izdvajanje, ovo su primarno instrukcije koje se koriste za debugovanje.

Pojedinačne vrednosti dodeljene instrukcijama se nadovezuju za svaki osnovni blok i svaki od njih se terminira jedinstvenom vrednosti kako bi se sprečilo izdvajanje koda preko granice blokova. Nadovezivanjem tih reprezentacija osnovnih blokova se dobija reprezentacija celog programa.

Nad takvom reprezentacijom programa sada mogu da se primene algoritmi za traženje ponavljajućih podniski (eng. *longest repeated substring*). Implementirani algoritam korsiti sufiksno stablo. Sufiksno stablo je struktura podataka koja omogućava efikasno izvršavanje nekih algoritama nad niskama. Formalno, sufiksno stablo niske  $S$  dužine  $n$  je korensko stablo za koje važi:

- stablo ima tačno  $n$  listova koji su numerisani od 1 do  $n$
- osim korena, svaki unutrašnji čvor ima bar 2 deteta
- svaka grana je označena nepraznom podniskom niske  $S$
- nikoje dve grane koje polaze iz istog čvora nemaju zajednički prefiks
- niska dobijena konkatenacijom oznaka svih grana na putu od korena do lista numerisanog brojem  $i$  je sufiks niske  $S$  koji počinje od  $i$ -tog karaktera

Ovo stablo se sagrađi u linearnoj složenosti u odnosu na broj karaktera niske pomoću Ukonenovog algoritma [15]. Broj čvorova u ovom stablu je linearno zavisn od broja karaktera, što znači da je obilazak ovog stabla takođe linearne složenosti. Traženje svih ponavljajućih podniski se može obaviti jednostavnim obilaskom stabla, takođe linearne složenosti.

Kada se odrede kandidati prelazi se u fazu izdvajanja koda. U petlji se odabira grupa kandidata za izdvajanje sa najvećim izračunatim doprinosom. Način računanja doprinosa zavisi od implementacije. U oba slučaja se dobija kao procena razlike u veličini programa ukoliko se grupa ne izdvoji i ako se izdvoji. Za odabranu grupu se generiše nova funkcija, a svi kandidati se zamenjuju pozivom te funkcije. Pritom se pazi da se održi semantika koda na tom međukodu. Posle svakog odabira kandidata, ta grupa kao i svi drugi kandidati koji se preklapaju sa njom se brišu iz sufiksnog stabla i postupak se ponavlja sve dok postoje grupe koje sa pozitivnim doprinosom. Ukoliko ih više nema, prolaz izdvajanja koda je završen.

### Specifičnosti na mašinski zavisnom međukodu

Izdvajanje koda na mašinski zavisnom međukodu je implementirano u klasi `MachineOutliner`. Ovo je jednostavnija od dve implementacije. U nastavku će biti objašnjeni algoritmi za heširanje, procenu dobitka i način generisanja izdvojenog koda.

Heširanje instrukcija na ovom nivou je jednostavno. Poređenje se vrši po jednakosti. Dakle instrukcija i operandi moraju da budu identični. To predstavlja jedno od najvećih ograničenja na ovom nivou. Ukoliko dve sekvence vrše isto izračunavanje ali njihovim operandima su dodeljeni različiti registri, neće biti moguće izdvojiti taj kôd. Još uvek važi klasifikacija instrukcija, tako da instrukcije klasifikovane kao ilegalne nije moguće izdvojiti. Klasifikaciju instrukcija radi `TargetInstrInfo`.

Procena dobitka memorije se računa kao razlika između veličine programa bez i sa uključenom optimizacijom. Kako se ona radi na nivou jedne grupe kandidata, dovoljno je posmatrati samo instrukcije koje pripadaju toj grupi. Veličina grupe kandidata bez optimizacije ne mora da se proceni, jer je već dostupna kao proizvod veličine jedne grupe i ukupnog broja grupa. Veličina sa primenjenom optimizacijom zahteva više računa. Još uvek postoji jedan kandidat iz grupe koji predstavlja telo izdvojene funkcije. On se sabira sa procenama cena svih poziva izdvojene funkcije i cenama dodatnih instrukcija za kreiranje novog okvira za funkciju. Sve to ukupno čini procenu sa uključenom optimizacijom. Cene instrukcija na ciljnoj arhitekturi kao i procene poziva i se dobijaju kroz interfejs `TargetInstrInfo`.

Grupe se sortiraju po benefitu. Izbacuju se svi kandidati koji se preklapaju sa prethodno izdvojenim kodom. Generiše se nova funkcija sa imenom `OUTLINED_FUNCTION_X` gde je `X` redni broj izdvojene funkcije. Kopiraju se instrukcije proizvoljnog kandidata, generišu se informacije za debugovanje. Svi kandidati se zamenjuju pozivom ka generisanoj funkciji.

### Specifičnosti na mašinski nezavisnom međukodu

Izdvajanje koda na mašinski nezavisnom međukodu radi prolaz koji se zove `IROutliner`. Ona je znatno kompilkovanija nego mašinski nezavisna varijanta, ali pruža više mogućnosti za izdvajanje. U nastavku će biti objašnjeni algoritmi za heširanje, procenu dobitka i način generisanja izdvojenog koda.

Za određivanje kandidata se koristi pomoćna analiza, `IRSimilarityIdentifier`. Umesto poređenja celih instrukcija, poredi se samo tip instrukcije i operanda i dodatni parametri za neke instrukcije (npr. ime pozvane funkcije ili tip predikta za instrukciju poređenja). Operandi ne utiču na heš vrednost instrukcije zato što se oni mogu izvući kao argumenti izdvojene funkcije. Zbog toga je moguće izdvajanje dosta više koda.

Dobitak se računa slično kao na mašinski nezavisnom nivou, kao razlika bez i sa uključenom optimizacijom. Kako je ovaj međukod dalji od krajnjeg koda, procena

postaje teža jer se oslanja na manje precizne podatke. Cena bez optimizacije se ovaj put računa kao suma procena veličina svih kandidata. Cena sa optimizacijom sadrži prosečnu veličinu kandidata i procene veličina instrukcija za smeštanje i učitavanje potrebnih argumenta. Procene cena instrukcija nakon transformacije na ciljnu arhitekturu se dobija kroz interfejs `TargetTransformInfo`.

Usled postojanja argumenata izdvojenoj funkciji, način izdvajanja postaje komplikovaniji. Kandidati se izdvajaju iz dva pokušaja. Prvi put kada se izdvoje određuju se svi argumenti koje će funkcija imati. Svi potrebni argumenti su poznati tek nakon prvog prolaska tako da se drugi put se izdvajanje vrši uz zamenu argumenata. Jedan od izdvojenih kandidata se proglašava za konačnu izdvojenu funkciju i njemu se dodeljuje ime `outlined_ir_func_x` gde je `x` redni broj izdvojene funkcije. Informacije iz svih izdvojenih kandidata se spajaju sa izabranom funkcijom i zatim brišu. Na kraju, svaki kandidat se zamenjuje pozivom ka funkciji prilikom čega se dodaju i instrukcije za čuvanje i učitavanje argumenata.

## 4.2 Izdvajanje koda i debug lokacije

Debug lokacije predstavljaju pozicije u izvornom kodu koje odgovaraju instrukciji u izvršnom fajlu. Sastoje se od fajla, reda i kolone. U idealnom slučaju, svaka instrukcija u izvornom kodu bi trebalo da ima odgovarajuću lokaciju, ali to nije slučaj pogotovu u optimizovanom kodu. Veliki broj optimizacija menja kôd na način koji može da naruši tačnost debug lokacija. Neki od najčešćih slučajeva su brisanje nepotrebnih instrukcija ili pomeranja instrukcija van tela petlje. Kompajleri se trude da održe što veću količinu debug lokacija, ali nekada to ili nije moguće ili nema smisla posle izvršavanja optimizacije.

U nastavku se podrazumeva da se koristi format DWARF za čuvanje informacija za debugovanje. Lokacije za debugovanje se interno čuvaju u tabeli linija u `.debug_line` sekciji izvršnog fajla. Ta tabela preslikava adresu iz tekst (kôd) segmenta izvršnog fajla na odgovarajuću lokaciju (fajl, red, kolona) u izvornom kodu i neke dodatne informacije o toj lokaciji. Format DWARF propisuje način enkodiranja tabele linija u vidu konačnog automata, ali postoje alati koji prikazuju podatke u obliku tabele. Listing 4 prikazuje upotrebu nekoliko alata za dekodiranje tabele linija, a u listingu 5 se nalazi primer jedne tabele.

Debuggeri prikazuju lokaciju iz tabele koja je vezana za prvu adresu koja je manja ili jednaka adresi instrukcije na kojoj je program zaustavljen. Na taj način ukoliko



```
llvm-dwarfdump --debug-line [file]
readelf -wL [file]
objdump -WL [file]
```

Listing 4: Komande za ispis tabele linija koristeći alate llvm-dwarfdump, readelf i objdump

Address	Line	Column	File	ISA	Discriminator	Flags
-----	-----	-----	-----	-----	-----	-----
0x0000000000401110	1	0	1	0	0	is_stmt
0x000000000040111b	2	6	1	0	0	is_stmt prologue_end
0x0000000000401122	3	3	1	0	0	is_stmt
0x000000000040112b	4	2	1	0	0	is_stmt
0x000000000040112f	4	2	1	0	0	is_stmt end_sequence

Listing 5: Tabela linija ispisana pomoću alata llvm-dwarfdump

trenutna instrukcija nema svoju debug lokaciju, prikazaće se lokacija prve prethodne instrukcije koja ima lokaciju. Standard propisuje i specijalne debug lokacije koje se poznaju po tome sto imaju red 0. Ove lokacije znače da je kôd kompajlerski generisan i da ne postoji odgovarajuća lokacija u izvornom kodu. Prilikom izvršavanja programa red po red (koristeći komandu **step** u debageru LLDB) preskaču se sve instrukcije sa istim redom kao prethodna ili sa redom 0. Na instrukciju koja nema dodeljenu lokaciju je moguće stati izvršavanjem programa instrukciju po instrukciju, postavljanjem tačke prekida na adresu te instrukcije ili usled zaustavljanja programa zbog izuzetka koji se desio u toj instrukciji.

Prilikom izdvajanja koda jedna instrukcija ima više mogućih debug lokacija u zavisnosti od mesta poziva izdvojene funkcije. Ovo predstavlja više problema:

- U vreme izvršavanja programa nije moguće odrediti koja od ovih lokacija je korektna bez dodatnih informacija
- Format DWARF ne podržava postojanje više različitih lokacija za istu adresu [3]

Radi održavanja korektnosti, prilikom izdvajanja koda brišu se lokacije svih izdvojenih instrukcija. Kao posledica toga mogu se javiti nepogodnosti prilikom debugovanja:

- Izvršavanje korak po korak pravi velike skokove kada dođe do izdvojenog koda, bez bilo kakvog obaveštenja korisniku

- Postavljanje tačaka prekida na izdvojenu instrukciju precizirajući njen red nije moguće

Primer debugovanja sa izdvojenim kodom je prikazan u listingu 6. Redovi 11-14 i 16-19 su izdvojeni u funkciju. Kada debager koračanjem naredbu po naredbu dođe do nekog od tih regiona, u sledećem koraku ga celog preskoči.

```
$ lldb outline
(lldb) breakpoint set 11
(lldb) run
   8          int x = 0;
   9          int y = 1;
  10
-> 11          int c1 = x + y;
  12          c1--;
  13          global += 2;
  14          foo(c1, global);
(lldb) next
  13          global += 2;
  14          foo(c1, global);
  15
-> 16          int c2 = x + y;
  17          c2--;
  18          global += 2;
  19          foo(c2, global);
(lldb) next
  18          global += 2;
  19          foo(c2, global);
  20
-> 21          return 0;
  22      }
```

Listing 6: Proces debugovanja programa sa izvojenim kodom. Izdvojeni delovi koda su preskočeni prilikom izvršavanja korak po korak.

## Glava 5

# Implementacija rešenja

Implementacija rešenja nadograđuje verziju 16 projekta LLVM, što je najnovija stabilna vezija u vreme pisanja rada. Celokupan izvorni kôd je dostupan na servisu GitHub na adresi <https://github.com/vvuksanovic/llvm-project/tree/outline-debug-info>. Komande za kompilaciju projekta su prikazane u listingu 5.1.

```
$ git clone https://github.com/vvuksanovic/llvm-project.git
$ cd llvm-project
$ git checkout outline-debug-info
$ mkdir build && cd build
$ cmake -G Ninja -DCMAKE_ENABLE_PROJECTS="clang;lldb" -DCMAKE_BUILD_TYPE=Release
$ ninja
```

Slika 5.1: Komande za prevođenje kompajlera LLVM

Rešenje se sastoji iz dva dela. Prvo je potrebno u okviru kompajlera generisati potrebne informacije za debugovanje, a zatim ih iz debagera iskoristiti za poboljšanje procesa debugovanja. U nastavku će redom biti opisana oba dela implementacije.

### 5.1 Implementacija u kompajleru

Zadatak ovog dela je da prilikom primene optimizacije izdvajanja funkcija zapamti originalne lokacije u izvornom kodu, prenese ih kroz sve faze kompilacije i ispiše ih u sekciju za debugovanje izvršnog fajla. Potrebne informacije se generišu u okviru implementacija optimizacije, u prolazima `IROutliner` i `MachineOutliner`.

Kako tabela linija ne podržava duplirane vrednosti adrese, ona nije u optičaju za čuvanje vrednosti. Ideja je da se debug lokacije sačuvaju u `.debug_info` sekciji umesto u `.debug_line`. Verzija 5 standarda DWARF je uvela novu etiketu, `DW_TAG_call_site`, koja bi trebalo da sadrži podatke o mestu poziva funkcije [8]. Trenutno se već koristi za pružanje informacija o vrednostima parametara prilikom poziva funkcije [9] i da li je poziv repni ili repni rekuzivan. Ovo je pogodno mesto za dodavanje informacija o izdvojenim instrukcijama zato što su one baš vezane za mesto poziva izdvojene funkcije.

Implementacija trenutno podržava čuvanje podataka na prilikom izdvajanja koda mašinski nezavisnoj međureprezentaciji. Kako to uključuje spuštanje kroz mašinski zavisnu reprezentaciju, jednostavno je dodati podršku za izdvajanje koda i na toj reprezentaciji. Jedini deo koji nedostaje je generisanje podataka. Za izbor instrukcija se koristi `SelectionDAG` implementacija, tako da će čuvanje podataka raditi za sve podržane arhitekture. Glavni fokus ove implementacije je `x86_64` arhitektura, ali količina koda specifičnog za tu arhitekturu je minimalna.

Dakle, debug lokacije moraju da budu prenesene od `IROutliner` optimizacionog prolaza, kroz izbor instrukcija do mašinski zavisnog međukoda i konačno do asemblerskog fajla u okviru debug sekcije. Svaka od navedenih stavki je detaljnije objašnjena u nastavku.

## Čuvanje debug lokacija prilikom optimizacije

Kao što je objašnjeno u poglavlju 4.2, optimizacioni prolaz `IROutliner` prilikom izdvajanja instrukcija briše sve podatke o njihovim lokacijama u izvornom kodu. Kako izdvajanje funkcija ostavlja kôd sa manje instrukcija nego što je bilo pre toga, ne postoji dovoljno mesta da se smeste potrebne informacije. Kao prvi korak očuvanju ovih podataka ovaj rad uvodi novu unutrašnju (eng. *intrinsic*) funkciju za debugovanje, `llvm.dbg.outlined`, čiji poziv se generiše odmah posle poziva ka izdvojenoj funkciji i dodatno je opisuje odgovarajućom debug lokacijom izdvojene instrukcije.

Unutrašnje funkcije su funkcije koje su ugrađene u sami kompajler takve da on može da ih implementira na optimalan način [13]. One predstavljaju mehanizam proširenja LLVM jezika takav da ne zahteva modifikaciju svih prolaza kada se nešto doda u jezik. Imena svih unutrašnjih funkcija počinju prefiksom „`llvm.`“, on će u nastavku biti izostavljen. Posebnu klasu unutrašnjih funkcija čine funkcije za debugovanje. U ovu grupu spadaju `dbg.declare`, `dbg.value` i `dbg.addr` za praćenje

promenljivih i `dbg.label` za obeležavanje labela u izvornom kodu. One ne utiču na krajnji izvršni kôd nego se koriste za popunjavanje informacija za debugovanje.

Kao pomoć za identifikovanje izdvojenih instrukcija, uveden je i novi metapodatak `outline_id`. Taj metapodatak će se nalaziti uz svaku izdvojenu instrukciju kao i uz instrukciju poziva izdvojene funkcije. On nema nikakve dodatne podatke nego je umesto toga deklarisan kao jedinstven. To znači da će se za svaku upotrebu kreirati nova instanca umesto da se ponovo koristi postojeća kao što je podrazumevano.

Sada kada postoji način za referisanje instrukcija, moguće je definisati deklaraciju `dbg.outlined` funkcije. Deklaracija ima dva parametra, oba su metapodaci koji pokazuju na `outline_id` identifikatore redom za odgovarajuću izdvojenu instrukciju i instrukciju poziva uz koju stoji. Ovi parametri ne mogu da budu zaključeni implicitno na osnovu mesta na kom se nalaze zbog potencijalnih optimizacija koje mogu da se izvrše kao i promene redosleda instrukcija (npr. u izdvojenoj funkciji). Dodatno, instrukcija sadrži i debug lokaciju koju pružima od izdvojene instrukcije koju zamenjuje.

Kreiranje nove unutrašnje funkcije za debugovanje je veoma jednostavno i radi se na deklarativan način. Jedino što je potrebno uraditi je dodati deklaraciju u `Intrinsics.td` fajlu sa nazivom, u parametrima. Za lakše upravljanje tom funkcijom i njenim parametrima napravljen je novi tip instrukcije `DbgOutlinedInst` koja nasleđuje `DbgInfoIntrinsic` klasu. Ta instrukcija služi kao omotač i koristi se kao alias za instrukciju poziva `dbg.outlined` funkcije. Instanciranje ove instrukcije se radi uz pomoć nove funkcije u klasi `DIBuilder`. Metapodatak `outline_id` se registruje u definicionom fajlu `FixedMetadataKinds.def`. Implementacija se nalazi u klasi `DebugInfoMetadata` i nju je potrebno registrovati u `Metadata.def`. Dodatno, uvedena je podrška za čitanje i pisanje LLVM IR fajlova kao i za njihovu proveru ispravnosti.

Za svaku izdvojenu instrukciju koja ima debug lokaciju se generiše odgovarajuća `dbg.outlined` instrukcija i dodaje se posle instrukcije poziva. Isečak LLVM međukoda programa nakon izdvajanja funkcije je prikazan u listingu 5.2.

Funkcije generisane izdvajanjem koda je potrebno na neki način razlikovati od drugih funkcija. Trenutno se već u mašinski nezavisnom međukodu dodaje zastavica (eng. *flag*) `DIFlagArtificial` ali to nije dovoljno da se razlikuje od drugog kompajlerski generisanog koda. Iz tog razloga, dodata je nova vrsta zastavice, `DIFlagOutlined`, koja označava samo izdvojene funkcije. U kasnijoj fazi kompilacije ona će se prevesti u DWARF atribut `DW_AT_LLVM_outlined`. Nova zastavica se

```
define dso_local i32 @main() #2 !dbg !30 {  
    ...  
    call void @outlined_ir_func_0(ptr %x, ptr %y, ptr %c1), !dbg !44,  
    ↪ !outline_id !45  
    call void @llvm.dbg.outlined(metadata !46, metadata !45), !dbg !44  
    call void @llvm.dbg.outlined(metadata !47, metadata !45), !dbg !48  
    call void @llvm.dbg.outlined(metadata !49, metadata !45), !dbg !50  
    call void @llvm.dbg.outlined(metadata !51, metadata !45), !dbg !43  
    call void @llvm.dbg.outlined(metadata !52, metadata !45), !dbg !53  
    call void @llvm.dbg.outlined(metadata !54, metadata !45), !dbg !53  
    call void @llvm.dbg.outlined(metadata !55, metadata !45), !dbg !53  
    call void @llvm.dbg.outlined(metadata !56, metadata !45), !dbg !57  
    call void @llvm.dbg.outlined(metadata !58, metadata !45), !dbg !57  
    call void @llvm.dbg.outlined(metadata !59, metadata !45), !dbg !57  
    call void @llvm.dbg.outlined(metadata !60, metadata !45), !dbg !61  
    call void @llvm.dbg.outlined(metadata !62, metadata !45), !dbg !63  
    call void @llvm.dbg.outlined(metadata !64, metadata !45), !dbg !65  
    ...  
}  
...  
define internal void @outlined_ir_func_0(ptr %0, ptr %1, ptr %2) {  
    ...  
    %3 = load i32, ptr %0, !outline_id !46  
    %4 = load i32, ptr %1, !outline_id !47  
    %add = add nsw i32 %3, %4, !outline_id !49  
    store i32 %add, ptr %2, !outline_id !51  
    %5 = load i32, ptr %2, !outline_id !52  
    %dec = add nsw i32 %5, -1, !outline_id !54  
    store i32 %dec, ptr %2, !outline_id !55  
    %6 = load i32, ptr @global, !outline_id !56  
    %add1 = add nsw i32 %6, 2, !outline_id !58  
    store i32 %add1, ptr @global, !outline_id !59  
    %7 = load i32, ptr %2, !outline_id !60  
    %8 = load i32, ptr @global, !outline_id !62  
    %call = call i32 @foo(i32 noundef %7, i32 noundef %8), !outline_id !64  
    ...  
}
```

Slika 5.2: Primer LLVM međukoda nakon optimizacije izdvajanja koda sa novom instrukcijom i metapodatkom.

dodaje u trenutku generisanja izdvojene funkcije u klasi `IROutliner`, a stara zastavica `DIFlagArtificial` je zadržana radi održanja kompatibilnosti unazad. Pristup ovoj zastavici je omogućen dodavanjem `isOutlined` metode u okviru `DISubprogram` klase koja predstavlja metapodatak sa podacima o funkciji.

## Čuvanje debug lokacija prilikom izbora instrukcija

Na kraju prethodne glave su potrebne informacije za debugovanje postale dostupne posle izvršavanja optimizacija na mašinski nezavisnom međukodu. Naredni korak je sprovesti te informacije do sledećeg nivoa, odnosno mašinski zavisnog međukoda. Proces konverzije iz mašinski nezavisnog do zavisnog međukoda u infrastrukturi LLVM se naziva izbor instrukcija. Postoji više implementacija izbora instrukcija, ali u ovom radu je trenutno podržana samo implementacija `SelectionDAG`.

Od uvedenih podataka za potrebe ovog rada, za `dbg.outlined` i vezu između `DIOutlineId` i instrukcija za koje su vezane potrebno je implementirati dalje prevođenje. Svi metapodaci ostaju u istom obliku u ovoj fazi kompilacije. Na kraju izbora instrukcija sve `dbg.outlined` instrukcije će biti zamenjene ekvivalentnom MIR instrukcijom `DBG_OUTLINED` sa istim parametrima i debug lokacijom. Instrukcija `DBG_OUTLINED` je registrovana u definicionim fajlovima `Target.td` i `TargetOpcodes.def`. Klasa `SelectionDAG` tretira instrukcije za debugovanje drugačije od ostalih instrukcija. One se takođe pretvaraju u čvorove usmerenog acikličnog grafa ali umesto da se ubace u usmeren aciklični graf, one su ostavljene po strani u odvojenim strukturama podataka. Zahvaljujući ovome te instrukcije će ostati neizmenjene prilikom izvršavanja većine faza izbora instrukcija. Po ugledu na druge instrukcije za debugovanje, uveden je tip cvora `SDDbgOutlined` i dodato polje u klasi `SelectionDAG` za njihovo skladištenje. Kako se gradi objekat tipa `SelectionDAG`, sve `dbg.outlined` instrukcije se pretvaraju u `SDDbgOutlined` i čuvaju u novom polju. Na kraju izbora instrukcija, kada je potrebno poredati instrukcije svaki `SDDbgOutlined` sadrži redni broj koji je dobio prilikom konstrukcije i na osnovu njega se `DBG_OUTLINED` emituju na očekivanom mestu u kodu.

Na sličan način kao za čvorove instrukcija za debugovanje, tako se i podaci o `outline_id` metapodatk u instrukcija čuva odvojeno od grafa. Konkretno, čuva se u novom polju u okviru promenljive `SDCallSiteDbgInfo`. To je mapa koja preslikava `SDNode` u objekat koji sadrži dodatne informacije o čvorovima. Prilikom emitovanja mašinskih instrukcija proverava se postojanje ovog podatka za svaku instrukciju i ukoliko postoji dodeljuje se instrukciji.

```
name: main
...
body:
  ...
  CALL64pcrel32 @outlined_ir_func_0, outline-id !45, debug-location !44
  ADJCALLSTACKUP64 0, 0, debug-location !44
  DBG_OUTLINED !46, !45, debug-location !44
  DBG_OUTLINED !47, !45, debug-location !48
  DBG_OUTLINED !49, !45, debug-location !50
  DBG_OUTLINED !51, !45, debug-location !43
  DBG_OUTLINED !52, !45, debug-location !53
  DBG_OUTLINED !54, !45, debug-location !53
  DBG_OUTLINED !55, !45, debug-location !53
  DBG_OUTLINED !56, !45, debug-location !57
  DBG_OUTLINED !58, !45, debug-location !57
  DBG_OUTLINED !59, !45, debug-location !57
  DBG_OUTLINED !60, !45, debug-location !61
  DBG_OUTLINED !62, !45, debug-location !63
  DBG_OUTLINED !64, !45, debug-location !65
  ...
```

Slika 5.3: Isečak LLVM mašinski zavisnog međukoda dobijen izborom instrukcija koda iz prethodne glave.

Najveća količina informacija se izgubi prilikom izbora instrukcija. Identifikatori koji stoje iz instrukcije se mogu izgubiti ako se instrukcija zameni drugom bez kopiranja njenih dodatnih podataka. Na primer, prilikom spuštanja mašinski nezavisne instrukcije poziva funkcije na mašinski zavisan nivo za arhitekturu X86 pravi potpuno novi čvor. Zato što se `outline_id` čuva u mapi ciji je ključ `SDNode`, zamenom nekog čvora će se izgubiti ta veza osim ako se taj podatak eksplicitno prenese. Pomenuti primer je popravljen u implementaciji, ali još uvek postoji veliki broj slučajeva kada će se ovaj problem ispoljiti.

Isečak LLVM mašinski zavisnog međukoda programa nakon izdvajanja funkcije je prikazan u listingu 5.3.

### Ispisivanje debug lokacija prilikom emitovanja asemblerskog koda

Poslednja faza kompilacije koju radi direktno kompajler je emitovanje asemblerskog koda. U okviru nje se mašinske instrukcije dobijene izborom instrukcija



pretvaraju u odgovarajuće asemblerske instrukcije za tu arhitekturu. U isto vreme se generišu i informacije za debugovanje u izabranom formatu. Sledeće faze kompilacije rade assembler i linker koji će bez problema preneti sve informacije redom do objektnog odnosno izvršnog fajla. Dakle ovo je poslednji korak koji je potrebno izmeniti za potrebe ovog rada.

Svi podaci o izdvojenim instrukcijama treba da se pretvore u informacije za debugovanje. Uključivanje informacija za debugovanje ne sme da menja krajnji kôd tako da nijedna instrukcija neće biti dodata u tekst segment programa. Ova implementacija podržava samo ispisivanje u formatu DWARF. Proizvođači kompajlera mogu da dodaju svoja proširenja standarda sve dok ne menjaju vrednosti propisane standardom DWARF. Kako se etikete i atributi mapiraju na celobrojne numeričke vrednosti, proizvođač može da koristi jedan slobodno izabran opseg vrednosti sve dok se ne preklapa sa opsegom nekog od drugih kompajlera. Što se tiče kompajlera LLVM, on uvodi svoja proširenja standarda DWARF koja se mogu prepoznati po tome što njihovi nazivi sadrže podnisku LLVM. Implementacija ovog rada se oslanja na nove vrste etiketa i atributa koji će biti opisani u nastavku.

Kao što je pomenuto na početku poglavlja, novi podaci će biti smešteni u okviru postojeće etikete `DW_TAG_call_site` koja opisuje poziv funkcije. Za svaku `DBG_OUTLINED` instrukciju u odgovarajućoj etiketi mesta poziva funkcije biće dodata novouvedena etiketa `DW_TAG_LLVM_outlined_ref`. Nova etiketa se zove `DW_TAG_LLVM_outlined_ref` i sadrži četiri atributa:

- `DW_AT_decl_file` — Identifikator fajla lokacije u izvornom kodu
- `DW_AT_decl_line` — Broj reda u izvornom kodu
- `DW_AT_decl_column` — Broj kolone u izvornom kodu
- `DW_AT_low_pc` — Adresu instrukcije u tekst segmentu programa. U ovom trenutku predstavljenu labelom ispred te instrukcije.

Informacije za debugovanje se u kompajleru LLVM generišu funkciju po funkciju. Međutim, po prirodi ovog zadatka uvek će postojati izdvojena funkcija koju je potrebno obraditi pre nego što je moguće obraditi funkcije koje je pozivaju. LLVM ne pruža mogućnost zadavanja ovakvih zavisnosti, tako da je rešenje pomeriti većinu poslova da se izvrše na samom kraju obrade modula. Obradom svake funkcije se rade potrebne pripreme i čuvaju informacije koje će se na kraju spojiti u konačnu reprezentaciju. Glavna klasa zadužena za upravljanje informacijama za debugovanje

na ovom nivou je `DebugHandlerBase`. To je bazna klasa koja radi opšte obrađivanja informacija zajedničko za sve formate za debugovanje. Za sve poslove vezane za konkretan format ona poziva virtuelne funkcije koje implementira odgovarajuća podklasa za taj format. Ovaj rad se fokusira na implementaciju `DwarfDebug`.

Prvi korak prema ispisivanju informacija za debugovanje je prikupljanje `DBG_OUTLINED` instrukcija iz prethodne faze kompilacije. Na početku obrade funkcije `DebugHandlerBase` poziva instancu klase `DbgEntityHistoryCalculator` da prođe kroz kôd i zapamti unutrašnje instrukcije za debugovanje. Ta klasa je proširena da isto to uradi i za `DBG_OUTLINED` instrukcije. Radi lakšeg budućeg pristupa, grupiše ih po njihovom drugom argumentu koji predstavlja `outline_id` instrukcije poziva. Svi prikupljeni podaci se smeštaju u polje klase `DebugHandlerBase` tako da je vidljivo i podklasama. Druga stvar koja se radi takođe na početku funkcije je zahtevanje za kreiranje labela ispred instrukcija u funkcijama generisanim izdvajanjem koda koje imaju `outline_id`. Te labele se dodaju tek prilikom obrađivanja instrukcije ispred koje treba da se smeste, tako da sve potrebne labele postoje tek na kraju obrađivanja cele funkcije. Celokupnu dalju obradu informacija radi instanca klase `DwarfDebug`. Na kraju obrade svake od funkcija iz korisničkog koda pozvana je funkcija `endFunctionImpl`. Ona prikuplja i emituje sve informacije za debugovanje koje su nastale prilikom obrade te funkcije. Te informacije uključuju:

- DWARF etikete koje predstavljaju mesto poziva funkcije i `outline_id` metapodatak uz instrukciju poziva za svaku funkciju nastalu izdvajanjem koda.
- Labele postavljene ispred svake izdvojene instrukcije zajedno sa njenim `outline_id` podatkom
- Sve ključne informacije iz `DBG_OUTLINED` instrukcija u novoj reprezentaciji vezanoj za DWARF umesto za `MachineInstr`.

Na kraju obrade modula sigurno su spremni svi neophodni podaci iz svih funkcija korisničkog koda. Spajanjem tih podataka na osnovu zajedničkih `outline_id` vrednosti lako je u postojeće etikete dodati sve potrebne informacije za debugovanje.

Funkcije za kreiranje novih i modifikovanje postojećih DWARF etiketa nalaze se u klasi `DwarfCompileUnit`. Tu su dodate funkcije za generisanje `DW_TAG_LLVM_outlined_ref` etiketa sa ranije opisanim poljima i dodata je pomoćna funkcija za ubacivanje tih etiketa u `DW_TAG_call_site`. Uz to, toj etiketi je dodat atribut `DW_AT_LLVM_outlined` koji označava da je to poziv izdvojene funkcije. Isti atribut se dodaje i etiketi te funk-

cije a ovde je dodat kako bi se kasnije brže odredilo da li se poziva izdvojena funkcija ili ne.

Što se tiče emitovanja instrukcija u tekst segment, u funkciju `emitFunctionBody` klase `AsmPrinter` je dodat slučaj za ispisivanje instrukcije `DBG_OUTLINED` koji ne radi ništa. Promenom ove implementacije da ispiše komentar može se potvrditi da su instrukcije `DBG_OUTLINED` još uvek prisutne i da se nalaze na očekivanom mestu u kodu.

Prethodnim modifikacijama omogućeno je prevođenje do asemblerskog fajla. Asembler nema predstavu o formatu ili informacijama za debugovanje i ne postoje posebne asemblerske direktive za njih. Umesto toga, sekcije za debugovanje se u asemblerskom kodu popunjavaju literalima koji imaju redosled i veličinu propisane formatom. Te vrednosti se direktno prenose u objektni i izvršni fajl. Ti literali imaju smisla samo kada se parsiraju u istom formatu u kom su zapisani. Jedan alat koji to može da uradi je `llvm-dwarfdump`. Listing 5.4 prikazuje deo novih etiketa parsiranih iz objektnog fajla pomoću alata `llvm-dwarfdump`. Kako bi imao podršku za prikazivanje ovih etiketa potrebno je koristiti verziju alata koja je sagrađena uz ovaj projekat.

```
0x00000093: DW_TAG_call_site
    DW_AT_call_origin      (0x000000dc)
    DW_AT_call_return_pc   (0x00000000000001179)
    DW_AT_LLVM_outlined    (true)

0x00000099: DW_TAG_LLVM_outlined_ref
    DW_AT_decl_file        ("test.c")
    DW_AT_decl_line        (11)
    DW_AT_decl_column      (0x0b)
    DW_AT_low_pc           (0x00000000000001194)

0x0000009e: DW_TAG_LLVM_outlined_ref
    DW_AT_decl_file        ("test.c")
    DW_AT_decl_line        (11)
    DW_AT_decl_column      (0x0f)
    DW_AT_low_pc           (0x00000000000001196)
```

Slika 5.4: Isečak DWARF stabla koji prikazuje etiketu poziva funkcije i njenu decu sa podacima o izdvojenim instrukcijama

## 5.2 Implementacija u debageru

Programi prevedeni kompajlerom sa izmenama iz prethodnog poglavlja uz neophodne opcije će sadržati debug informacije o izdvojenom kodu. Pošto implementacija predstavlja novo proširenje standarda DWARF, nijedan postojeći debager nema podršku za interpretiranje tih informacija. U ovom poglavlju će biti predstavljena implementirana podrška za uvedeno proširenje u debageru LLDB. Prvo će biti omogućeno učitavanje novih podataka, a zatim rešeni najveći nedostaci trenutne podrške za debugovanje izdvojenih instrukcija. To uključuje ispisivanje tačnih mesta u izvornom kodu kada se proces zaustavi, podršku za izvršavanje korak po korak i postavljanje tačaka prekida.

### Učitavanje podataka

Za rešavanje opisanih problema potreban je način za mapiranje adrese izvršnog fajla na lokaciju u izvornom kodu i obrnuto. Obično oba ova mapiranja radi tabela linija (eng. *line table*), ali za izdvojene instrukcije ta dva smera su nezavisno implementirana na različite načine zato što su im potrebni različiti podaci a koji nisu svi dostupni u vreme kompilacije.

Interfejs za učitavanje i pristup informacijama za debugovanje u debageru LLDB je ostvaren kroz klasu `SymbolFile`. On pruža funkcije za parsiranje podataka iz različitih sekcija za debugovanje i njihovo dohvaćanje u reprezentaciji nezavisnoj od formata u kome su napisane. Većina podataka se ne učitava odmah po pokretanju debagera već tek kada su ti podaci neophodni. Tako na primer dohvaćanjem podataka za jedinicu kompilacije (eng. *compile unit*) kroz funkciju `GetCompileUnitAtIndex` se izvršava parsiranje te jedinice ukoliko već nije učitana. Za tabelu linija, odgovarajuća funkcija za učitavanje se intuitivno zove `ParseLineTable`, ali joj se pristupa pomoću funkcije `GetLineTable` klase `CompileUnit`. Kao i u ranijim poglavljima, podržani formati su implementirani kao priključci (eng. *plug-ins*). Svaki od njih implementira interfejs `SymbolFile` ili preciznije apstraktnu klasu `SymbolFileCommon`. U slučaju formata DWARF, implementacija se nalazi u klasi `SymbolFileDWARF`.

Podaci koji treba da budu učitani su skoro isti kao za tabelu linija, zbog toga se implementacija za učitavanje ugleda na učitavanje tabele linija. Uvedena je nova tabela linija specijalno za lokacije izdvojenih instrukcija. Za razliku od obične tabele, ova ne koristi specijalnu strukturu podataka već je u pitanju običan niz sortiran po lokaciji u izvornom kodu. Podaci se dobijaju obilaskom čitavog DWARF stabla

i parsiranjem atributa svih `DW_TAG_LLVM_outlined_ref` etiketa. Nakon učitavanja, nova tabela linija se predaje klasi `CompileUnit` preko koje može da joj se pristupi iz skoro celog koda. Bitniji metod pristupa je putem određivanja trenutnog konteksta izvršavanja. U poglavlju 3.3 je rečeno da se informacije o stanju programa dobijaju kroz instancu klase `SymbolContext` koja te podatke dobija iz informacija za debugovanje. Stanje programa podrazumeva jedinicu kompilacije, funkciju i lokaciju u izvornom kodu. Jedan od načina za dobijanje tog konteksta je zadavanjem lokacije u izvornom kodu za koju treba odrediti taj kontekst. Bitno je napomenuti da lokacija dobijena iz konteksta ne mora da se poklapa sa prosleđenom lokacijom. Za kreiranje te instance zadužena je ponovo klasa `CompileUnit`, a funkcija koja to obavlja je `ResolveSymbolContext`. Pored proveravanja najbliže lokacije iz tabele linija, dodata je i provera u izdvojenoj tabeli linija. Ukoliko iz obe tabele postoje takve lokacije, u kontekst se stavlja ona koja je bliža prosleđenoj. Ako samo jedna od tabela ima validan rezultat, koristiće se on. Sada svi koji koriste ovaj način dohvaćanja konteksta mogu da dobiju informacije o izdvojenim instrukcijama. Upotreba ove izmene za rešenje konkretnog problema će biti prikazana kasnije.

Mapiranje u drugom smeru, sa adrese izdvojene instrukcije na lokaciju u izvornom kodu nije jedinstveno. Činjenica da je instrukcija izdvojena znači da se poziva na bar dva različita mesta u izvornom kodu. Zbog toga jednoj adresi će uvek odgovarati bar dve lokacije. Kako bi se dobila tačna lokacija potrebno je uzeti obzir i kontekst izvršavanja programa. Taj kontekst je predstavljen stek okvirom (eng. *stack frame*) funkcije koja je pozvala izdvojenu funkciju. Preciznije, predstavljen je adresom povratka (eng. *return address*) na funkciju kojoj odgovara taj stek okvir. Adresa povratka je adresa sledeće instrukcije koja će biti izvršena nakon povratka iz funkcije. Mapiranje u ovom smeru je smešteno u instancama klase `CallEdge`. Neka su funkcije čvorovi usmerenog grafa, i neka grana između dva čvora postoji za svaki poziv funkcije u smeru od pozivaoca do pozvane funkcije. Može postojati više grana između dve funkcije ukoliko se ista funkcija poziva više puta. Klasom `CallEdge` predstavlja se jedna od tih grana. Ona je dopunjena da sadrži adrese i lokacije svih izdvojenih instrukcija prilikom tog poziva. Prethodno pomenuta klasa `SymbolFile` je zadužena za popunjavanje tih podataka kada su potrebni. Ta implementacija je smeštena u funkciji `CollectCallEdges` i zasniva se na parsiranju etikete poziva funkcije. Svim instancama klase `CallEdge` je moguće pristupiti iz kroz `Function` objekat pozivaoca, ili može da se pristupi konkretnoj instanci na osnovu adrese povratka.

Poslednji neophodan podatak koji je potrebno učitati je atribut `DW_AT_LLVM_outlined`. To je implementirano u okviru postojeće implementacije za učitavanje podataka o funkciji u klasi `DWARFASTParserClang`. Pristup informaciji da li je funkcija dobijena izdvajanjem koda je omogućen kroz metod `IsOutlined` dostupan nad objektima `Function` klase koja opisuje jednu funkciju iz koda koji se debuguje.

### Ispisivanje ispravnih linija izvornog koda

Prilikom debugovanja korisnik može da kontroliše tok izvršavanja programa. On zadaje komande koje govore debageru koje delove koda da izvrši. Svaki put kada se kontrola vrati korisniku, debager ispiše mesto gde se program zaustavio (i možda nekoliko okolnih linija). Na osnovu toga korisnik saznaje informacije o trenutnom stanju programa i bira šta će sledeće da izvrši. U prisustvu dobrih debug lokacija to funkcioniše očekivano. Međutim u slučaju izvojenih instrukcija koje nemaju vrednost u tabeli linija, lokacija prikazana u debageru će biti netačna. Prikazana lokacija će pripadati prvoj ranijoj instrukciji koja ima lokaciju. Kako se izdvojene funkcije se dodaju na adresama iza svih korisničkih funkcija, prva ranija instrukcija može da pripada nekoj drugoj funkciji, ili čak može da joj odgovara lokacija koja označava kompajlerski generisan kôd. Korisnik ne dobija nikakvo obaveštenje o tome šta se desilo već samo može da vidi da je program napravio veliki skok na kraj fajla iako to nije slučaj.

Prvi izazov koji je potrebno rešiti kako bi imalo smisla rešavati bilo koji drugi problem je kako odrediti tačnu lokaciju u izvornom kodu kada se debager zaustavi u izdvojenoj funkciji. Bez toga iako se uspešno implementira bilo koji način zaustavljanja na izdvojene instrukcije, debager bi prikazivao netačne lokacije što samo otežava debugovanje. Trenutno jedini načini da se zastane na izdvojenoj instrukciji je da se program izvršava instrukciju po instrukciju ili da se postavi tačka prekida po adresi. Oba načina se mogu koristiti za proveru ispravnosti implementacije.

Ispisivanje statusne poruke, koja između ostalog uključuje i okolne redove gde je program zaustavljen radi funkcija `GetStatus` klase `StackFrame`. Taj deo koda je izmenjen na sledeći način. Prvo se proverava da li trenutni stek okvir pripada izdvojenoj funkciji. Podatak koji to opisuje je dostupan kroz metod `IsOutlined` klase `Function`, a ona se dobija na osnovu konteksta tog stek okvira. Ukoliko stek okvir ne pripada izdvojenoj funkciji, određivanje linije se svodi na postojeću implemetaciju. U suprotnom koristi se nova implementacija. U glavi 5.2 dodata je podrška za mapiranje adrese na liniju u izvornom kodu. Jedino što je potrebno je dodati

kontekst kako bi se odredila dobra `CallEdge` instanca za instrukciju. Funkcija već ima implementiran mehanizam za pronalaženje te instance na osnovu adrese povratka iz funkcije. Ta adresa se postavlja na vrh steka u trenutku poziva funkcije. Nit programa ima dostupnu listu stek okvira koji postoje u tom trenutku. Povratkom u roditeljski stek okvir lako je pronaći tu adresu. Konačna lokacija je dobijena iz `CallEdge` instance dobijene iz tog roditeljskog stek okvira pomoću metode `Function::GetCallEdgeForReturnAddress` sa pronađenom adresom povratka kao parametrom. Uz lokaciju, u status je dodata i poruka da je proces zaustavljen u izdvojenom kodu.

## Izvršavanje korak po korak

Jedna od najkorišćenijih tehnika debugovanja je izvršavanje korak po korak. Ona omogućava korisniku da interaktivno prati i kontroliše izvršavanje programa. Između dva koraka korisnik može da pogleda vrednosti relevantnih promenljivih i priveri da li se one poklapaju sa očekivanim vrednostima. LLDB nudi nekoliko komandi za koračanje:

- **step-over** — Izvršava sve naredbe do sledećeg reda u izvornom kodu preskačući pozvane funkcije
- **step-into** — Izvršava sve naredbe do sledećeg reda u izvornom kodu pritom ulazeći u pozvane funkcije
- **step-inst** — Izvršava jednu instrukciju izvršnog fajla

U postojećoj implementaciji jedino **step-inst** komandom je moguće stati na izdvojenju instrukciju. Druge komande, kako se oslanjaju na broj reda u izvornom kodu, će ih uvek preskočiti. Cilj ovog poglavlja je prepraviti i njih da rade u skladu sa korisnikovim očekivanjima.

Tokom sesije debugovanja, korsnik zadaje komande debageru kako bi kontrolisao tok izvršavanja programa. Celokupna logika o tome koliko program treba da se izvršava i kada treba da se zaustvi implementirana je pomoću planova niti (eng. *thread plan*). Više različitih planova može da bude aktivno u isto vreme, zbog čega se oni skladište u stek strukturi podataka. Svaka nit ima svoj stek planova i njemu se može pristupiti `GetProcess().GetThreadPlans(GetID())` iz trenutne niti. Kada se program zaustavi zbog nekog događaja svi planovi za tu nit se proveravaju redom od vrha do dna steka. Traži se prvi plan koji može da objasni događaj koji je izazvao

zaustavljanje pozivanjem funkcije `DoPlanExplainsStop` svakog plana. Taj plan se ona upita da li treba da se vrati kontrola korisniku funkcijom `ShouldStop`. Ovde se ažurira logika ukoliko je potrebno da se program zaustavi na nekom sledećem mestu. Posle izvršavanja plana proverava se da li je on završio pozivom funkcije `MischiefManaged`. Ukoliko jeste, on i svi planovi iznad njega u steku su uklonjeni ukoliko prijave da više nisu relevantni. Uvek postoji bar jedan plan na steku. To je instanca `ThreadPlanBase` i on se aktivira kada ni jedan drugi plan ne objašnjava događaj kao na primer za tačke prekida napravljene od strane korisnika.

Postoji više implementacija planova koji imaju različite kriterijume za zaustavljanje. Fokus ovog poglavlja su implementacije `ThreadPlanStepInRange` i `ThreadPlanStepOverRange` koje se koriste redom za `step-into` i `step-over` komande. Oni dele zajedničku nadklasu `ThreadPlanStepRange`. Oba plana rade tako što postavljaju tačke prekida na sledeću instrukciju grananja sve dok program ne izađe iz opsega koji mu je zadat u vreme kreiranja. Taj opseg obuhvata sve instrukcije koje pipadaju istom redu i izvornom kodu. Nije moguće odmah postaviti tačku prekida direktno na konačnu lokaciju zbog mogućih grananja u programu. Razlika između planova je što `step-into` prilikom zaustavljanja na instrukciji poziva funkcije dodaje novi plan za ulazak u funkciju dok `step-over` to ne radi.

Ideja rešenja je da se osim postojeće tačke prekida uvede još jedna koja se uvek postavlja na seldeću izdvojenu instrukciju. Nadklasa `ThreadPlanStepRange` je dopunjena funkcijama za postavljanje te nove tačke prekida koja se dobija na isti način kao u poglavlju 5.2. Posebno u slučaju `step-over` komande sledeća instrukcija grananja se traži uz bezuslovno preskakanje instrukcija poziva. Taj postupak je izmenjen dodavanjem uslova da pozvana funkcija nije nastala izdvajanjem koda. Ukoliko jeste, očekivano ponašanje je da debager ipak korača unutar te funkcije kako to odgovara izvornom kodu programa. U suprotnom poziv instrukcije se ignoriše kao i ranije.

Poslednje što je potrebno učiniti je izmeniti konkretne klase planova da dodavaju i proveravaju tačku prekida na izdvojenoj instrukciji. Razlikuju se dva slučaja u kojima je potrebno dodati ciljnu tačku prekida. Ulaskom u izdvojenu funkciju potrebno je postaviti sledeći cilj da bude prva izdvojena instrukcija sa lokacijom. Ukoliko je program već zautavljen usred izdvojene funkcije onda cilj postaje prva sledeća izdvojena instrukcija sa lokacijom u narednom redu. Prethodni slučajevi se lako razlikuju poređenjem odnosa između trenutnog stek okvira i okvira iz trenutka kreiranja plana. Kako bi se plan uopšte aktivirao prilikom dostizanja cilja, ta



provera je dodata u funkciju `DoPlanExplainsStop`. Obrada samog događaja se radi na početku funkcije `ShouldStop` gde se odmah vraća kontrola korisniku i plan se završava.

## Postavljanje tačaka prekida

Tačka prekida (eng. *breakpoint*) je mesto u programu koje kada treba da se izvrši u debageru izaziva pauziranje izvršavanja i vraćanje kontrole korisniku. LLDB podržava nekoliko načina za postavljanje tačaka prekida:

- pomoću adrese u izvršnom fajlu — Najjednostavniji vid postavljanja tačke prekida. Postavlja se tačno na prosleđenu adresu.
- pomoću imena funkcije — U izvornom kodu se pronalaze sve funkcije sa prosleđenim imenom i za svaku od njih dodaje se nova tačka prekida.
- pomoću imena fajla i reda — Prosleđuje se naziv fajla i red u tom fajlu. Na osnovu informacija za debugovanje se tačka prekida postavlja na prvu instrukciju kojoj odgovara prosleđeni red ili prvi sledeći red ukoliko za traženi red ne postoje informacije.
- pomoću regularnog izraza — Izvorni kôd se pretražuje za poklapanja sa regularnim izrazom i svako poklapanje dodaje novu tačku prekida.

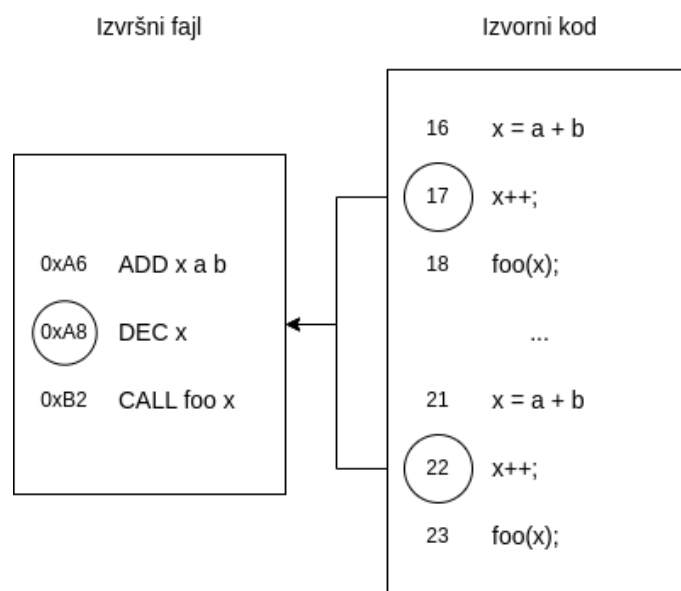
Svi ovi načini se svode na postavljanje preko jedne ili više adresa.

Za potrebe razlikovanja, svakoj napravljenoj tački prekida dodeljuje se jedinstveni identifikator u formatu `X.Y` gde `X` označava identifikator grupe, a `Y` označava identifikator lokacije u toj grupi. Sve tačke prekida nastale istom komandom imaju istu vrednost `X` i različite vrednosti `Y`. Ne može da se desi da dve tačke prekida iz iste grupe imaju istu lokaciju. Ovaj identifikator se koristi za izmenu svojstava tačaka prekida i prikazuje se korisniku kada se program zaustavi usled dostizanja jedne od njih. Uspešnim izvršavanjem komande za postavljanje tačke prekida dobija se poruka nalik na onu u listingu 5.5.

```
* thread #1, name = 'test', stop reason = breakpoint 1.1
```

Slika 5.5: Primer poruke koju prikazuje debager LLDB prilikom postavljanja tačke prekida

Prva dva od prethodno opisanih načina rade očekivano za izdvojene instrukcije dok druga dva ne funkcionišu. Ono što izdvaja druga dva načina je što oni moraju da odrede adresu na osnovu lokacije u izvornom kodu. Tačno je da isti ovaj proces se koristi i za postavljanje tačke prekida po imanu funkcije ali kako izdvojeni kôd ne može da sadrži deklaraciju funkcije, na njega ne utiče dotični problem. Na sreću, potreban mehanizam je već ostvaren dodatnom tabelom linija za izdvojene instrukcije. Međutim i sa tim se javlja drugi problem. Sada prilikom debugovanja dolazi do neočekivanih zastajkivanja na tačke prekida koje korisnik nije postavio. Radi jednostavnosti, nastavak teksta se odnosi na slučaj zadavanja fajla i reda. Sve što će biti rečeno važi i za slučaj koji koristi regularne izraze. Problem je u tome što odgovarajuća adresa pripada izdvojenoj funkciji koja se poziva iz više mesta. Ilustracija ovog problema je prikazana na slici 5.6. Ukoliko korisnik postavi tačku



Slika 5.6: Vizuelni prikaz odnosa izdvojenog koda u izvornom kodu i izvršnom fajlu

prekida bilo na red 17, do prekida će doći i prilikom izvršavanja linije 22 i obrnuto. Korisnik ne treba da bude svestan postojanja izdvojene funkcije jer on debuguje samo na osnovu izvornog koda. Za njega, očekivani ishod treba da bude da se program zautavlja samo na redu koji je on izabrao.

Rešenje problema nije moguće realizovati u vreme postavljanja tačke prekida, već se proverava ispravnosti lokacije radi nakon izazivanja prekida ali pre povratka kontrole korisniku. Dok klasa **Breakpoint** poseduje sve podatke o grupi tačaka prekida koje su nastale istom komandom, informacije o svakoj pojedinačnoj instanci

se nalaze u okviru klase `BreakpointLocation`. Ona je dopunjena poljem koje predstavlja opcionu lokaciju u izvornom kodu koja odgovara toj adresi i poljem logičkog tipa koje obeležava da li je instrukcija izdvojena. Lokacija u izvornom kodu je neophodna samo za podršku izdvojenih instrukcija. Oba dodata polja dobijaju vrednost prilikom kreiranja tačke prekida na osnovu podataka iz tabela linija. Dolaskom na tačku prekida vrši se provera da li treba vratiti kontrolu korisniku kroz funkciju `ShouldStop` klase `BreakpointLocation`. U ovom trenutku je potrebno proveriti da li se lokacija poklapa sa lokacijom koja odgovara trenutnom stanju procesa. Način dobijanja trenutne lokacije je isti kao i u prethodna dva poglavlja, uz pomoć trenutne adrese i konteksta. Nigde u implementaciji se ne nalazi zavisnost na određeni način određivanja tačaka prekida tako da bi trebalo da funkcioniše Posebna pažnja je data da se ova provera radi pre bilo koje druge operacije koja može da izmeni stanje tačke prekida. Na primer pre uvećavanja brojača aktivacije tačke prekida koji korisnik može da korsiti za napredno debugovanje.

## Glava 6

# Rezultati

Modifikovanu verziju kompajlera i debagera je potrebno detaljno testirati u različitim mogućim situacijama kako bi se postarali da se prvenstveno ne smanjuje funkcionalnost u odnosu na nemodifikovanu verziju, a zatim i da se zapravo unapređuje debugovanje izdvojenog koda. Kako je velika količina koda izmenjena, velike su šanse da je došlo do propusta u implementaciji koji nisu detektovani testiranjem na jednostavnim primerima u toku prvobitnog pisanja koda.

Projekat LLVM već uključuje veliki broj regresionih testova koji pokrivaju skoro svaku podržanu funkcionalnost. Ti testovi se pokreću koristeći pomocni alat `lit` (skraćeno za *LLVM Integrated Tester*) i sastoje se od fajlova koji sadrže kôd u nekom formatu i ugrađene direktive u vidu komentara. Direktive mogu da budu ili komande koje se pokreću za izvršavanje testa ili direktive koje zadaju tvrdnje koje test proverava. Testovi se pokreću

Pokrenuti su svi testovi uključeni u LLVM projekat. Većina testova je uspešno izvršena bez problema. To je bilo i očekivano zato što veliki broj testova i ne radi sa uklječenim izdvajanjem koda. Mali deo testova je neuspešno izvršen zato što proverava specifične informacije za debugovanje koje su izmenjene modifikovanim kompajlerom. U tim situacijama, odgovarajući testovi su izmenjeni tako da očekuju pojavljivanje novih etiketa i atributa. Poslednji, poseban slučaj je činio test koji sadrži istovremeno izdvojen i umetnut kôd. Ova situacija nije bila predviđena prilikom pisanja implementacije. Dotični test je prevremeno zaustavljao zbog neispunjene tvrdnje da poziv funkcije koja može da bude umetnuta unutar funkcije koja ima informacije za debugovanje mora da ima lokaciju. Za taj slučaj je dodata posebna podrška. Za pozive koji ispunjavaju opisani uslov se postavlja veštačka debug lokacija.

Ispravkom poslednjeg testa, nije uočeno više problema i svi testovi su se uspešno izvršavali. Sledeće je potrebno isprobati da li implementacija uspešno rešava probleme koje je namenjena da reši. Demonstracija će biti prikazana na istom primeru predstavljenim ranije u listingu 1. Ponašanje debagera na izdvojenom kodu pre modifikacija je opisano u glavi 4.2.

Kompilacijom sa modifikovanim kompajlerom i proverom informacija za debugovanje izvršnog fajla se može se primetiti da su nove etikete generisane na mestima poziva funkcija i da se adrese poklapaju sa adresama instrukcija u izdvojenoj funkciji. Pokretanjem debagera na generisanom fajlu i koraćanjem po programu dobija se ispis prikazan u slici 6.1. Ukoliko se ovaj ispis uporedi sa ispisom pre modifikacija primećuje se da se program zaustavlja na 3 od 4 izdvojene instrukcije što prethodno nije bio slučaj. Takođe, svaki put kad se zastane ispisuje se poruka koja govori da se proces nalazi u izdvojenoj funkciji. Četvrta naredba, `c1--;`, ne izaziva prekid izvršavanja programa. Istraživanjem je otkriveno da se odgovarajuća instrukcija koja je u početku bila instrukcija sabiranja sa -1, u toku izbora instrukcija zamenila sa instrukcijom dekrementiranja za 1. Zbog te zamene izgubljen je `outline_id` koji je ta instrukcija originalno imala, stoga nije bilo moguće pronaći adresu te instrukcije prilikom emitovanja asemblerskog koda. Dakle, još uvek postoji prostor za poboljšanje implementacije.

Još jedna implementirana funkcionalnost koju treba proveriti je postavljanje tačka prekida. U prethodnom primeru je izgledalo kao da se uspesno postavila tačka prekida na liniji 11 koja je izdvojena. Ono što se zapravo desilo je da se tačka postavila na instrukciju poziva izdvojene funkcije koja je dobila lokaciju prve izdvojene instrukcije. Tek sledecim korakom se zapravo doslo do izdvojene instrukcije. Umesto te naredbe na redu 11, tačka će biti postavljena na red 13 kako bi se izbegla prethodna situacija. Ispis debagera je prikazan na slici 6.2. Komanda za postavljanje tačke prekida je uspesno odredila adresu koja odgovara zadatoj liniji i pokretanjem programa, proces je zaustavljen na toj adresi.

Novi regresioni testovi su dodati za kompajler i debager po ugledu na prethodni primer. Kompajleru se testira uspešno generisanje informacija za debugovanje u svakom koraku kompilacije. Debager testovi pokreću debager nad programom sa izdvojenim kodom i proveravaju ispravnost lokacije, da li se program zaustvio na očekivanim mestima i da li je uspešno postavljena tačka prekida.

```
$ lldb outline
(lldb) b 11
(lldb) r
   9          int y = 1;
  10
-> 11          int c1 = x + y;
  12          c1--;
  13          global += 2;
(lldb) n
   9          int y = 1;
  10
-> 11          int c1 = x + y;
  12          c1--;
  13          global += 2;
Note: this function is outlined.
(lldb) n
  11          int c1 = x + y;
  12          c1--;
-> 13          global += 2;
  14          foo(c1, global);
  15
Note: this function is outlined.
(lldb) n
  12          c1--;
  13          global += 2;
-> 14          foo(c1, global);
  15
  16          int c2 = x + y;
Note: this function is outlined.
```

Slika 6.1: Koraćanje po izdvojenom kodu sa izmenjenim debagerom

```
$ lldb outline
(lldb) b 13
Breakpoint 1: where = outline`outlined_ir_func_0 + 11 at outline.c,
↳ address = 0x000055555555551a0
(lldb) r
10
11             int c1 = x + y;
12             c1--;
-> 13             global += 2;
14             foo(c1, global);
15
16             int c2 = x + y;
Note: this function is outlined.
```

Slika 6.2: Postavljanje tačke prekida na izdvojenu naredbu sa izmenjenim debagerom

Glava 7

Zaključak



# Bibliografija

- [1] Amy Brown and Greg Wilson. *The Architecture of Open Source Applications: Elegance, Evolution, and a Few Fearless Hacks*. The Architecture of Open Source Applications. Creative Commons, 2011.
- [2] K.D. Cooper and L. Torczon. *Engineering a Compiler*. Elsevier Science, 2022.
- [3] DWARF Debugging Information Format Committee. *DWARF Debugging Information Format Version 5*.
- [4] The LLVM Foundation. The llvm compiler infrastructure. on-line at: <https://llvm.org/>.
- [5] The LLVM Foundation. Llvm language reference manual. on-line at: <https://llvm.org/docs/LangRef.html>.
- [6] The LLVM Foundation. Machine ir (mir) format reference manual. on-line at: <https://www.llvm.org/docs/MIRLangRef.html>.
- [7] Dick Grune, Kees van Reeuwijk, Henri E. Bal, Criel J.H. Jacobs, and Koen Langendoen. *Modern Compiler Design*. Springer Publishing Company, Incorporated, 2nd edition, 2012.
- [8] Jakub Jelinek. Representation of call sites in the debugging information, 2010.
- [9] Jakub Jelinek. Tracking of values passed as arguments to functions in debug information, 2010.
- [10] Andrew Litteken. Framework for finding and using similarity at the ir level, 2020.
- [11] mayur Pandey and S. Sarda. *LLVM cookbook: over 80 engaging recipes that will help you build a compiler frontend, optimizer, and code generator using LLVM*. PACKT, 2017.

- [12] Jessica Paquette. Interprocedural mir-level outlining pass, 2016.
- [13] S. Sarda and M. Pandey. *LLVM Essentials*. Community experience distilled. Packt Publishing, 2015.
- [14] Anastasia Stulova and Sven van Haastregt. An overview of clang. Technical report, Arm, Cambridge, UK, 2019.
- [15] E. Ukkonen. *On-line Construction of Suffix-trees*. Series of publications / University of Helsinki, Department of Computer Science. A. University of Helsinki, Department of Computer Science, 1993.
- [16] Peng Zhao and José Nelson Amaral. Function outlining and partial inlining. *17th International Symposium on Computer Architecture and High Performance Computing (SBAC-PAD'05)*, pages 101–108, 2005.

## Biografija autora