

УНИВЕРЗИТЕТ У БЕОГРАДУ
МАТЕМАТИЧКИ ФАКУЛТЕТ

Ана Д. Митровић

**ПРИМЕНА СКАЛЕ У
ПАРАЛЕЛИЗАЦИЈИ РАСПЛИНУТОГ
ТЕСТИРАЊА**

мастер рад

Београд, 2018.

Ментор:

др Милена Вујошевић Јаничић, доцент
Универзитет у Београду, Математички факултет

Чланови комисије:

др Саша Малков, ванредни професор
Универзитет у Београду, Математички факултет

др Александар Картељ, доцент
Универзитет у Београду, Математички факултет

Датум одбране: 15. јануар 2016.

Наслов мастер рада: Примена Скале у паралелизацији расплинутог тестирања

Резиме:

Кључне речи:

Садржај

1	Увод	1
2	Програмски језик Скала	2
2.1	Опште карактеристике	3
2.2	Функционални део језика	5
2.3	Паралелизам	11
2.3.1	Модел дељења меморије	14
2.3.2	Модел размене порука	16
2.4	Библиотека Akka	18
2.4.1	Структура	19
3	Расплинуто тестирање	24
3.1	Основе тестирања софтвера	24
3.1.1	Нивои тестирања	26
3.1.2	Стратегије тестирања	27
3.2	Основе расплинутог тестирања	29
3.2.1	Историја	31
3.2.2	Ограничења	32
3.3	Врсте програма расплинутог тестирања	33
3.4	Расплинуто тестирање формата датотека	36
4	Структура PDF датотека	41
4.1	Објекти датотека	41
4.2	Организација датотека	43
4.3	Ажурирање датотека	45
5	Имплементација програма за расплинуто тестирање	47
5.1	Улазни параметри	47

САДРЖАЈ

5.2	Организација програма	49
5.3	Actor објекти програма	50
5.4	Креирање PDF датотека	53
5.5	Резултати тестирања	57
6	Закључак	62
	Библиографија	63

Глава 1

Увод

Глава 2

Програмски језик Скала

Скала је језик опште намене настао 2003. године са циљем да превазиђе ограничења програмског језика Јава комбиновањем објектно оријентисане и функционалне програмске парадигме. Мотивација иза овакве идеје је не ограничавати се на једну од ових парадигми и њених предности већ искористити најбоље из оба света. Управо због овакве комбинације парадигми Скала је веома погодна за решавање различитих врста проблема, почевши од малих незахтевних скриптова па све до великих компликованих система. На основу ове прилагодљивости је и добила своје име: реч „скала” означава „скалабилан језик” (енг. *scalable language*), односно језик који ће се прилагођавати и расти заједно са потребама система [29].

Творац Скале је Мартин Одерски (нем. *Martin Odersky*) (1958-), немачки научник и професор на универзитету EPFL (École Polytechnique Fédérale de Lausanne) у Лозани, Швајцарској. Још док је био студент, имао је жељу да напише језик који ће подржавати објектну и функционалну парадигму, говорећи да су ове две парадигме само две стране истог новчића. Желео је да се тај језик преводи у Јава бајт код али и да превазиђе ограничења језика Јава. Први резултат оваквог његовог рада је био језик Funnel, који због свог минималистичког дизајна није заживео. Затим је настала Скала, коју је професор Одерски развијао од 2001. године заједно са својом групом на универзитету EPFL. Познат је и по другим радовима, као што је имплементација GJ (Generic Java) компајлера који је постао основа јавас компајлера [3, 28].

2.1 Опште карактеристике

Ово су неке од најважнијих особина Скале [29]:

Компатибилност са Јавом Скала није сама по себи продужење Јаве али је потпуно компатибилна са њом: њен изворни код се преводи у Јава бајт код који се извршава на Јава виртуелној машини. Из кода писаног у Скали је омогућено коришћење Јава библиотека, класа, интерфејса, метода, поља и типова. Омогућено је и обрнуто, позивање Скала кода из Јаве, мада се оно ређе користи. Такође, Скала омогућава лакшу и лепшу употребу Јава типова уз помоћ имплицитне конверзије омогућавајући употребу својих метода за манипулацију типовима. Компатибилност олакшава програмерима лакши прелазак из Јаве у Скалу јер нису принуђени да се одједном одрекну написаног кода у Јави.

Корен Скале није само језик Јава, иако је Јава имала највећи утицај у њеном стварању. Идеје и концепти из разних језика, како објектно оријентисаних тако и функционално оријентисаних, су инспирисали развој Скале. Међу њима су језици: C# од кога је Скала преузела синтаксне конвенције, Erlang чије идеје су сличне идејама конкурентности базиране на моделу Actors и други: C, C++, Smalltalk, Ruby, Haskell, SML, F# итд.

Објектно оријентисана парадигма Писање програма и смештање података и њихових својстава у класе и објекте тих класа је веома популарно и интуитивно програмерима. Скала је то задржала притом мало изменивши концепт. У многим језицима, укључујући и Јаву, дозвољене су вредности које нису објекти или које нису у склопу објекта. То могу бити примитивне вредности у Јави или статичка поља и методи. Скала то не дозвољава и зато је објектно оријентисан језик у *чистој форми*: „Свака вредност је објекат и свака операција је позив метода” [29]. Елегантан начин коришћења метода налик на операције је један од начина на који се Скала брине да програмерима буде пријатно њено коришћење.

Концизност Скала код има тенденцију да буде краћи од Јава кода, чак се процењује да има барем два пута мање линија кода од Јаве. Постоје и екстремни случајеви где је број линија и десет пута мањи. Ова особина није значајна само због тога што знатно олакшава програмирање, већ олакшава читање кода и откривање грешака којих ионако има мање јер

има мање простора за њих. Ово је нешто што одликује саму синтаксу језика, а веома помажу и разне библиотеке које имају већ имплементиране многе функције које врше послове са којима се често сусрећемо. Лако се и имплементирају и касније употребљавају библиотеке које сами напишемо. Такође, Скала подржава аутоматско закључивање типова (енг. *type inference*) што омогућава изостављање понављања већ наведених типова, што резултује читљивијим кодом.

Висок ниво Како је Скала погодна и за велике и комплексне системе она се труди да се прилагоди њиховим захтевима подижући ниво апстракције у свом коду. Омогућава много једноставнији и краћи начин кодирања разних проблема. Рецимо, уместо да пролазимо кроз ниску карактера карактер по карактер користећи петљу, у Скали се то може урадити у једној линији кода помоћу *предиката* тј. *функцијских литерала* који су детаљније објашњени у поглављу 2.2. Скала код тежи да буде разумљивији и мање комплексан како би олакшао већ комплексан систем који се имплементира.

Статичка типизираност Скала је статички типизиран језик што значи да се типови података који су коришћени знају у време компилације. Неки сматрају ово маном као и да је навођење типова сувишно поред техника тестирања софтвера као што је нпр. тестирање јединица кода (енг. *unit testing*). Ипак, у Скали је статичка типизираност напреднија јер нам дозвољава да изоставимо тип на местима где би он био поновљен и вероватно би нам само сметао. Због оваквих понављања која су неопходна у неким језицима, многи се одрекну предности статички типизираних језика као што су: детектовање разних грешака у време компилације, лакше рефакторисање кода и коришћење типова као вид документације.

Својства Скала уводи појам својстава (енг. *trait*) који деле карактеристике са интерфејсима и наслеђивањем класа у Јави. У својствима као и у интерфејсима можемо декларисати методе, али и дефинисати поља и конкретне методе што није могуће у интерфејсима. Променљиве омогућавају чување стања а методе дефинисање одређеног понашања. То чини својства бога-тијим од интерфејса. Кажемо да се својства „миксују” у класе (енг. *mix in*). У једну класу је могуће миксовати више својстава.

Миксовање се може остварити кључном речју **extends** и у том случају класа наслеђује суперкласу својства. У случају да желимо да класа наследи неку другу класу, онда помоћу **extends** дефинишемо суперкласу а својства миксујемо помоћу **with** као што је илустровано у наредном примеру [29]:

```
class Animal
trait HasLegs
trait Philosophical
/* Superklasa "Animal" i miksovana svojstva "HasLegs" i
   "Philosophical" */
class Frog extends Animal with HasLegs with Philosophical {
  /* ... */
}
```

Највећа разлика између класе и својства односи се на позивање метода надкласе помоћу поља *super*. На пример, позивом метода *super.toString()* у случају класе, тачно се зна који ће метод *toString()* бити позван јер се зна надкласа дате класе. У случају својства, не можемо знати на шта се овај позив односи јер то директно зависи од класе у коју ће дато својство бити миксовано, као и од претходно миксованих својстава. То значи да се *super* позиви одређују *динамички*. Свако миксовано својство може позвати метод претходно миксованог својства. Ово омогућава класама да постигну различито понашање у зависности од редоследа миксованих својстава. Помоћу малог броја дефинисаних својстава може се постићи много различитих циљева комбиновањем редоследа миксовања.

Карактеристика која највише раздваја Скалу од Јаве је њена **функционалност** - Скала у потпуности подржава функционално програмирање. Ова особина је детаљно објашњена у следећем поглављу.

2.2 Функционални део језика

Функционална парадигма се развија од 1960-тих година. У њеној основи леже *ламбда рачун* (енг. *λ -calculus*) и комбинаторна логика. Ламбда рачун представља математичку апстракцију и формализам за описивање функција и њихово израчунавање. Њега је увео Алонзо Черч (енг. *Alonzo Church*) 1930-тих година

а Алан Тјуринг (енг. *Alan Turing*) је 1937. године показао да је експресивност ламбда рачуна еквивалентна експресивности Тјурингових машина. Иако је ламбда рачун првобитно развијен само за потребе математике, данас се он сматра првим функционалним језиком. Други модел рачунања који је утицао на функционалне програмске језике, комбинаторну логику, створили су Мојсеј Шејнфинкел (рус. *Моисей Шейнфинкель*) и Хаскел Кари (енг. *Haskell Curry*) 1924. године са циљем да елиминишу употребу променљивих у математичкој логици [18, 10, 26, 10].

Најстарији виши функционални програмски језик је LISP који је пројектовао Џон Макарти (енг. *John McCarthy*) на институту МИТ (Масачусетски технолошки институт) 1958. године. Након тога се полако развијају и други функционални језици као што су ISWIM, FP, Scheme, ML, Miranda, Erlang, SML, Haskell, OCAML, F#, Elixir итд. Поред функционалних језика постоје и језици који нису функционални али у одређеној мери подржавају функционалне концепте или могу да их остваре на други начин, нпр. Java, C, C++, C#, Python [18, 10].

Разлика између императивне и функционалне парадигме се може описати дефинисањем самог програма [23]:

- Императивна парадигма:

Програм представља формално упутство о томе шта рачунар треба да ради да би урадио неки посао

Програм представља одговор на питање КАКО се нешто РАДИ

- Функционална парадигма:

Програм представља формално објашњење онога што рачунар треба да израчуна

Програм представља одговор на питање ШТА се РАЧУНА

Функционални језици су веома блиски математици због чега се и називају *функционалним* језицима: њихове функције се понашају као функције у математичком смислу. О томе говоре следећа два концепта којима се воде функционални језици [29]:

Први концепт се односи на „**статус**” **функција**. У функционалним језицима функције су вредности *првог реда*, тј. *грађани првог реда*. То значи да се оне могу користити као вредности типова као што су нпр. цели, реални бројеви,

ниске карактера итд. Статус оваквих типова се не разликује од статуса функција, што у многим језицима није случај. Ово омогућава коришћење функција на природан, концизан и читљив начин: као аргументе других функција, повратне вредности функција, њихово чување у променљивим, угњеждавање и слично. Другим речима, функције се креирају и прослеђују без икаквих рестрикција на које смо навикли у императивним језицима.

У Скали можемо користити локалне (угњеждене) функције које су често веома мале и лако се комбинују како би заједно заокружиле један већи посао. Ово одговара функционалном стилу који промовише следећу идеју: програм треба изделити на много мањих функција (целина) од којих свака има јасно дефинисан задатак.

Посебно су корисни и важни *функцијски литерали* који представљају функције које могу бити неименоване (анонимне функције) и прослеђиване као обичне вредности. Назив „функцијски литерал” се користи за функције у изворном коду, а у време извршавања оне се називају „*функцијским вредностима*” (разлика слична разлици између класа и објеката у објектно оријентисаним језицима). Функцијске вредности су објекти које можемо чувати у променљивама али су истовремено и функције које можемо позивати. Пример функцијског литерала који инкрементира цео број изгледа овако [29]:

```
var increase = (x: Int) => x + 1
increase(10)
```

Променљива **increase** је објекат који садржи литерал, али је и функција коју позивамо са аргументом 10.

Други концепт се односи на **непостојање стања** које се иначе представља променљивама у програму. То значи да извршавање метода нема „*бочне ефекте*”, односно да функције које позивамо не мењају податке „у месту” већ враћају нове вредности као резултате свог израчунавања. Избегава се (у чисто функционалним језицима се избацује) кад год је то могуће коришћење променљивих које мењају своје стање, а за промену се користе променљиве код којих то није дозвољено. То су променљиве које се иницијализују само једном. Оне се у Скали дефинишу помоћу кључне речи **val** и покушај промене вредности ове променљиве након што је иницијализована резултирао би грешком. Променљиве које су стандардне у императивним језицима и које можемо мењати по потреби се дефинишу кључном речју **var**.

Непостојање стања има за последицу избацивање итеративних конструкција. Овакав резон може испрва деловати чудно, међутим све што се може остварити кроз итерације се може остварити и *рекурзијом*. Рекурзија је у функционалним језицима потпуно природна и неизбежна. Некада се решења проблема који захтевају стање у оваквим језицима превише закомпликују, али оно се може направити по потреби на друге начине тако да се ово не сматра недостатком. Предност је у томе што не морамо пратити вредности променљивих и њихове промене што олакшава разумевање програма. Један програм у функционалном језику представља низ дефиниција и позива функција а његово извршавање је евалуација тих функција [18].

Као пример метода који нема бочне ефекте може послужити метод **replace** који као аргументе добија ниску карактера и два карактера. Његов задатак је да у датој ниски замени сва појављивања једног карактера другим карактером. Он неће променити ниску коју је добио као аргумент, већ ће вратити нову која више нема појављивања датог карактера који смо заменили новим.

Ова особина метода без бочних ефеката назива се *референтна прозирност* (енг. *referential transparency*): „за било које вредности аргумената позив метода се може заменити својим резултатом без промене семантике програма” [29]. Пример:

```
val p = previous(90)
```

Свако појављивање променљиве **p** можемо заменити изразом **previous(90)**.

Позив метода можемо заменити његовим резултатом јер смо сигурни да ће при сваком позиву са истим вредностима аргумената вратити исти резултат. Вредност резултата не зависи од бочних ефеката и тренутног стања променљивих, већ само од вредности прослеђених аргумената. Због тога, редослед позива таквих метода није важан и може бити произвољан. За овакве методе кажемо да су референтно прозирни или транспарентни. Они су концизнији, читљивији и производе мање грешака јер немају пропратне ефекте. Ипак, у пракси су нам углавном неопходни пропратни ефекти због чега они могу бити присутни на контролисан начин. У том случају, програмски језик више није чисто функционалан [18].

Чисто функционални језици (нпр. Haskell, Miranda) захтевају коришћење имутабилних структура података, референтно прозирних метода и рекурзије. Други језици, као што су Скала, Python, Ruby итд. охрабрују овакво про-

грамирање али не условљавају програмера. Скала омогућава бирање начина за решавање проблема и нуди функционалне алтернативе за све императивне конструкције. На овај начин програмер се полако и без притиска навикава на другачији начин размишљања [18, 29].

Неке од карактеристика које Скала подржава а које су заједничке и другим функционалним језицима су [29]:

Функције вишег реда Функције вишег реда су оне које имају друге функције као своје аргументе. Оне поједностављују код и смањују његово понављање. Ово је пример функције која као аргументе има ниску карактера **query** и другу функцију **matcher**:

```
def filesMatching(query: String,
    matcher: (String, String) => Boolean) = {
  for (file <- filesHere; if matcher(file.getName, query))
    yield file
}
```

Функција **filesMatching** треба да међу датотекама **filesHere** пронађе датотеке које испуњавају неки критеријум. Да се код не би понављао за сваки критеријум појединачно, он се прослеђује функцији као аргумент у виду функције **matcher**. Сам критеријум представља функцију која има две ниске карактера као аргумент, и враћа логички тип - тачно ако је критеријум испуњен и нетачно у супротном.

Следећи пример демонстрира коришћење функције вишег реда из Скала библиотеке:

```
def containsOdd(nums: List[Int]) = nums.exists(_ % 2 == 1)
```

Метод **exists** проверава постојање елемента листе **nums** који задовољава наведени предикат који му је прослеђен као аргумент (дељивост са 2). Постоји доста метода сличних методу **exists**, попут: **find**, **filter**, **foreach**, **forall**, итд.

Каријеве функције Каријеве функције уместо једне листе аргумената узимају аргументе један по један, имајући више листи које садрже по један аргумент. На овај начин омогућују дефинисање функција помоћу већ дефинисаних на следећи начин:

```
def curriedSum(x: Int)(y: Int) = x + y
```

Ово је функција која сабира два цела броја. Када је позовемо на овај начин:

```
curriedSum(1)(2)
```

десиће се два позива функције. Први позив ће узети први параметар **x** и вратити другу функцију која узима параметар **y**, и враћа збир ова два параметра. Прва и друга функција би редом изгледале овако:

```
def first(x: Int) = (y: Int) => x + y
val second = first(1) /* poziv prve funkcije */
second(2) /* poziv druge funkcije */
```

Каријев поступак има пуно примена. Рецимо, можемо дефинисати функцију која додаје број 1 неком целом броју:

```
val onePlus = curriedSum(1)_
onePlus(2) /* --> vraca 3 */
```

Позивамо Каријеву функцију **curriedSum** користећи `_` као знак да на том месту може бити било која вредност (енг. *wildcard*). Суштина је да је функција **onePlus** дефинисана помоћу Каријеве функције која сабира било која два цела броја, а код које је први аргумент везан за број 1.

Упаривање образаца Често је потребно препознати да ли неки израз има одређену форму тј. да ли одговара одређеном обрасцу. То је корисно уколико треба да се имплементира функција која ради различите ствари у зависности од типова аргумената. Уобичајени примери су аритметичке операције или секвенце одређеног типа. Ово је пример са листама:

```
expr match {
  case List(0, _, _) => println("found it")
  case _ =>
}
```

Помоћу кључне речи **match** проверавамо да ли израз **expr** одговара некој листи од три елемента којој је први елемент 0 а друга два било који бројеви. Пошто се израз **expr** пореди редом са случајевима како су наведени, након неуспешног поређења са трочланом листом успешно ће се упарити са `_` који служи као образац који одговара свему (енг. *wildcard pattern*).

Comprehensions Конструкција **for expression** или **for comprehension** се често користи у Скали не само за стандардно итерирање кроз колекције већ и на другачије начине, на пример:

```
val forLineLengths =  
  for {  
    file <- filesHere /* итерирање кроз listu fajlova */  
    if file.getName.endsWith(".scala") /* први uslov */  
    line <- fileLines(file) /* итерирање кроз линије fajla */  
    trimmed = line.trim /* dodela promenljivoj trimmed */  
    if trimmed.matches(".*for.*") /* други uslov */  
  } yield trimmed.length /* повратна vrednost */
```

Овај пример обухвата пуно могућности које пружају **for** изрази. Прво, итерирање кроз листу **filesHere** користећи **<-** метод. Друго, филтрирање тих датотека помоћу услова задатог у **if** изразу. Онда следи једна угњеждена петља где се врши итерација кроз линије тренутног фајла (који је испунио услов, а ако није, прелази се на следећи фајл). Након тога, чување тренутне линије без вишка бланко карактера у променљивој **trimmed** и још један **if** израз. Најзад, помоћу **yield** наредбе враћа се дужина линије чиме се попуњава листа **forLineLengths**. Дакле, можемо користити угњеждене петље, филтрирати резултате, чувати податке у променљивама и попуњавати колекције помоћу **yield** наредбе. Све ове опције нам омогућавају конструкције које подсећају на дефинисање елемената скупова у математици.

Управо функционална својства Скале су кључни разлози због којих је она веома погодна за паралелизацију израчунавања.

2.3 Паралелизам

Потреба за све већом моћи процесора непрекидно расте у складу са захтевнијим апликацијама. „Направи десет пута бржи процесор, и софтвер ће убрзо наћи десет пута више посла” [33]. Другим речима, стандарди расту докле год расту и могућности. Потребно је све брже и ефикасније извршавати захтевна израчунавања. Зато су годинама откривани и развијани разни начини побољшања перформанси процесора као што су повећавање брзине откуцаја часов-

ника, оптимизација тока извршавања и повећавање простора кеш меморије на чипу. Ова побољшања се односе и на секвенцијалне, непаралелне апликације као и на конкурентне апликације. Ипак, ови начини подизања перформанси полако достижу свој максимум. Муров закон каже да се број транзистора на процесорском језгру дуплира на сваких осамнаест месеци, описујући експоненцијалан раст броја транзистора. Као и друге области које експоненцијално напредују, ни ова не може напредовати вечно и у једном тренутку раст више неће бити могућ. Поставља се питање шта ће бити када достигнемо ограничења свих побољшања на која смо навикли? Још увек има простора за раст, али јавила се потреба за новим идејама. Тако је напредак процесора почео све више да се односи на број *централних процесорских јединица* тј. *језгара*. Данаас имамо *multi-core* процесоре који имају у просеку од 2 до 8 процесорских језгара, и *many-core* процесоре са стотинама, па и хиљадама језгара. Сматра се да је конкурентност нова главна револуција у писању софтвера [33, 27].

Како главне улоге у новим побољшањима имају конкурентност и паралелизација, следе њихове дефиниције и појашњења [27, 24, 30]:

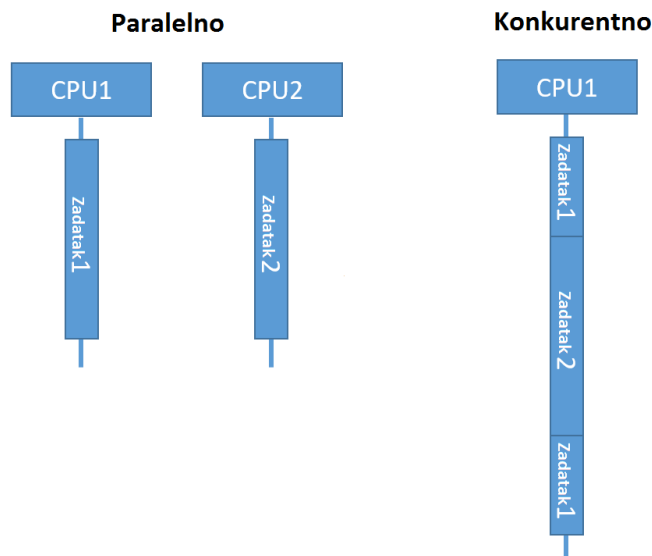
Дефиниција 1. *За два процеса кажемо да се извршавају **конкурентно** ако није унапред позната њихова међусобна временска лоцираност.*

Ово значи да имамо два процеса којима ће распоређивач (енг. *scheduler*), као нпр. Јава виртуелна машина, наизменично да додељује „кришке” времена (енг. *time-slice*). Због овога имамо привид да се ова два процеса извршавају дословно истовремено иако то није случај. Конкурентно извршавање се обично одвија када процеси деле једну процесорску јединицу.

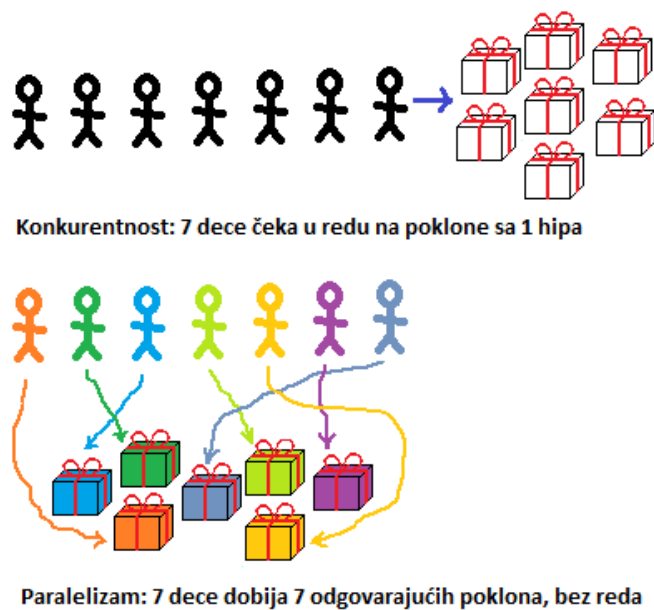
Дефиниција 2. *Два процеса се извршавају **паралелно** ако постоји период времена у коме су оба процеса истовремено активна.*

У случају паралелног израчунавања процеси се извршавају дословно истовремено на различитим процесорским јединицама. Паралелно израчунавање је подскуп конкурентног израчунавања, тј. може се представити као још један од начина на који се постиже конкурентност. Можемо написати конкурентну апликацију која има више просеца или нити, а са додатним језгрима процесора она може постати и паралелна.

Људи често мешају конкурентност и паралелизам и користе их као синониме, иако постоји разлика између њих. Ове појмове не треба тек тако мешати. Њихова различитост је демонстрирана сликама 2.1 и 2.2.



Слика 2.1: Разлика између конкурентног и паралелног израчунавања



Слика 2.2: Разлика између конкурентног и паралелног израчунавања

Оно што је скупо код конкурентног извршавања је промена контекста (енг. *context switch*) а то је промена стања процесора која је неопходна у случају када процесор са извршавања кода једног процеса прелази на извршавање кода другог процеса. Оваква промена се дешава јако често, до неколико стотина хиљада пута у секунди по процесорском језгру, што је незгодно јер узима до-

ста процесорског времена. Такође, подаци о стању процеса заузимају много меморије.

То је довело до појаве **нити** (енг. *thread*). Нит је компонента процеса која се извршава секвенцијално, тако да један процес може имати више нити. Нити омогућавају штедњу што се тиче ресурса ако постоје подаци које те нити деле. Самим тим, добија се и на времену. Нити се, као и процеси, могу извршавати конкурентно и паралелно.

Постоје различити модели тј. стилови паралелног програмирања. Ови стилови говоре о начинима комуникације процесорских језгара, нити, и начинима на који је то остварено. У наставку су објашњена два модела, модел дељења меморије и модел размене порука [27].

2.3.1 Модел дељења меморије

Рад са нитима подржава модел дељења меморије (енг. *shared memory model*). То значи да нити имају део меморије који им је заједнички, тако да могу асинхроно да читају и пишу на те меморијске локације. Ту се јављају разни проблеми. Треба омогућити безбедно дељење ресурса, комуникацију и синхронизацију нити. Случајеви када две или више нити не смеју истовремено имати приступ истим подацима су веома чести и зато се прибегава решењима закључавања као што су мутекси, катанци, монитори, семафори. Сврха ових решења је спречити недозвољене приступе подацима који се деле. Некада је дозвољено истовремено читање података, а некада је строго забрањен приступ свим нитима осим оне која тренутно чита и/или мења податке. Такве одлуке се доносе у зависности од конкретног случаја и потреба [30, 27].

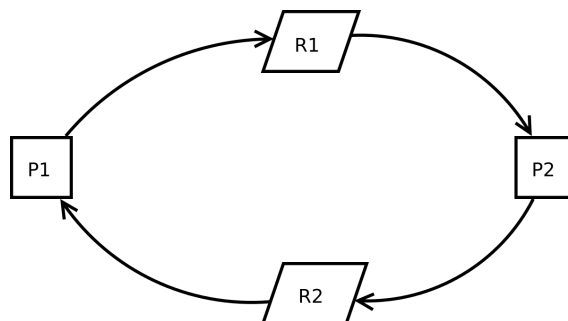
Механизми закључавања решавају проблеме који се тичу приступа подацима, али уводе нове проблеме [30]:

- **Катанци спречавају састављање** Као што је поменуто, функционални језици охрабрују писање мањих целина од којих свака има јасно дефинисан задатак. Ове мање целине на крају „саставе” цео програм. Ово није могуће ако користимо катанце.
- **Превише или премало катанаца** Закључавање и откључавање катанаца представљају скупе операције. Често није јасно колико је катанаца потребно да програм ради како треба а да не утиче на перформансе. Пре-

мало катанаца повећава шансу за недозвољеним приступом а са превише њих долази до пада перформанси и мртвих петљи.

- **Мртва петља и надметање за ресурсима** Ово су неки од проблема карактеристичних за конкурентно програмирање.

Мртва петља (енг. *deadlock*) или *узајамно закључавање* је ситуација када више процеса или нити уђе у стање чекања због тражења ресурса који није доступан. Све ове нити чекају да им нека друга ослободи ресурс, али је то немогуће јер нит коју чекају такође чека неки ресурс. Пример је дат на слици 2.3: Процес P1 тражи ресурс R1 али не успева да му приступи



Слика 2.3: Мртва петља

јер процес P2 држи ресурс R1. Слично, процес P2 тражи ресурс R2 али њега држи процес P1. Два процеса су у застоју јер се међусобно чекају. Проблем мртве петље се може решити тражењем катанаца у унапред дефинисаном редоследу, али и даље представља оптерећење за програмере [2].

Надметање за ресурсима (енг. *race conditions*) представља ситуацију када два или више процеса или нити истовремено читају и мењају исти податак. Један од тих процеса поништи измене другог процеса, па резултат који се добије на крају зависи од редоследа којим процеси мењају дати податак. Редослед је немогуће предвидети тако да се приликом покретања оваквог програма углавном сваки пут добије другачији резултат [2].

- **Тежак опоравак од грешака** Не постоји одређени механизам за опоравак програма са више нити од грешака. Углавном је корисно проћи кроз *.log* датотеку у којој је у различитим тренуцима записано стање стека са циљем праћења позивања функција и промена вредности променљивих.

Писање конкурентних програма је јако тешко. Нити су ниског нивоа, веома блиске хардверу и представљају начин на који сам процесор обавља послове. Постоји мало програмера који су у стању да напишу конкурентан програм без грешака тако да тај посао треба препустити стручњацима у тој области. Један од главних разлога што је ово тежак посао је непредвидивост конкурентних и паралелних програма. Немогуће је утврдити којим редоследом ће се послови обављати, а веома је компликовано доказати да је програм *коректан* тј. да ради управо оно што се од њега очекује. Програмери се ослањају на своје логичко размишљање и надају се најбољим исходима. Друга ствар која је кључни разлог тежине конкурентног програмирања је коришћење променљивих података. Дељени подаци који се временом мењају захтевају синхронизацију и механизме закључавања, отежавајући паралелизацију конкурентног програма. То су разлози због којих је функционална парадигма погоднија од других за паралелизацију. Чим нема мутабилних тј. променљивих структура нема ни проблема које оне носе са собом [30].

2.3.2 Модел размене порука

Модел размене порука (енг. *message passing model*) је један од нових „трендова” који се тичу конкурентности. Овај модел подржава **SN** тј. *Shared Nothing* архитектуру. SN архитектура се односи на дистрибуиране системе и подразумева да се систем састоји из неколико независних чворова. Сваки чвор (тј. независна машина) има своју меморију, дискове и интерфејсе за улаз и излаз. Распоживи подаци се поделе овим чворовима тако да сваки од њих има одговорност искључиво за своје податке. Подаци се међу чворовима никада не деле што значи да сваки чвор има потпуну слободу над својим делом па нема потребе за механизмима закључавања. Како се међу чворовима ништа не дели, отуда назив ове архитектуре [32, 13].

Оваква логика недељивости стоји и иза модела размене порука. Компоненте овог модела међусобно комуницирају искључиво размењујући поруке. Свака компонента овог модела има своје стање (које јесте променљиво) али га никада не дели са другима, а поруке које шаље су непроменљиве. Поруке се могу слати синхронно и асинхронно [29].

Модел размене порука се може имплементирати на више начина, а најуспешнији начин је помоћу такозваног модела **Actors**. Њега је осмислио амерички научник Карл Еди Хјуит (енг. *Carl Eddie Hewitt*) који је 1973. године заједно

са другим ауторима објавио рад који представља увод у модел Actors. Иако се развија годинама, тек је од скоро доказано да је овај модел ефикасан у решавању проблема савремених рачунарских система. Он енкапсулира тежак рад са нитима и олакшава програмирање конкурентности.

Компоненте Actor модела представљају објекти које зовемо *Actors*. Они формирају хијерархијску структуру у којој сваки објекат има свог родитеља који сноси одговорност за њега. Сваки Actor има своје „поштанско сандуче” и његов задатак је да обради сваку поруку коју добије. Поруке се чувају у FIFO (енг. *first-in first-out*) структури тј. *реду* (енг. *queue*) па се поруке обрађују редом којим су стигле. Оно што разликује објекте Actor од других објеката је способност реаговања на поруке одређеном акцијом. Као одговор на поруку Actor може направити још Actor компоненти (своју децу), слати поруке другим компонентама или зауставити рад своје деце или себе. Actor има своје локално стање које је променљиво, али са другима не дели ништа осим порука. Важна особина Actor објеката је да се не блокирају када пошаљу поруку већ настављају са радом, односно слање порука је увек асинхроно.

Предности Actor модела и уопште SN архитектуре се огледају у избегавању свих поменутих проблема конкурентног програмирања чији највећи узрок представљају дељени променљиви подаци. Actor модел је погодан у апликацијама у којима је могуће посао поделити на што више мањих, независних послова. Тада сваки мањи посао представља задатак једне Actor компоненте. Родитељ решава проблеме своје деце и прикупља резултате њихових послова. Дизајн такве апликације личи на организацију послова у компанијама где се послови деле по одељењима, све док ти послови не постану довољно једноставни да их може обавити један радник.

Иако Actor модел у многим случајевима олакшава посао, постоје ситуације када је неопходно да компоненте деле податке између себе. Не треба на силу користити један модел ако природа проблема намеће неки други. Actor модел није универзално решење за све проблеме већ треба препознати случајеве када га није погодно користити. Неке од ситуација када треба прибећи другачијем решењу су следеће [30]:

Дељени променљиви подаци Неки проблеми природно захтевају дељење података. Тада је погодније изабрати рад са нитима, поготово ако се подаци деле само за читање. Рад са базама података и трансакцијама је пример када се програмери обично одлучују за неко друго решење.

Цена асинхроног програмирања Модел дељења порука са собом носи одређену сложеност. Дебаговање и тестирање великих апликација које имплементирају модел дељења порука може бити веома тешко. Наиме, компликовано је праћење асинхронно послатих порука што доводи до тешког проналаска извора проблема. У овом случају је корисна информација о првој послатој поруци. Такође, многим програмерима је тешко да се навикну на нову парадигму што изискује труд и време.

Перформансе Неке апликације захтевају највећу могућу брзину и ефикасност. Иако је Actor модел у већини случајева примењив, нити се налазе на нижем нивоу и самим тим омогућавају боље перформансе.

Разни језици имају библиотеке које имплементирају Actor модел. У следећем поглављу је обрађена одговарајућа библиотека у Скали [30, 16, 9].

2.4 Библиотека Akka

Akka представља имплементацију Actor модела на Јава виртуелној машини. Од верзије 2.10 језика Скала ова библиотека је подразумевана библиотека за коришћење Actor модела. Постојала је и библиотека *Actor*, али је она застарела и од верзије 2.11 се више не користи [15].

Све што се тиче Actor објеката је смештено у својству *Actor* које се уводи *import* наредбом:

```
import akka.actor.Actor
```

Кључне особине које одликују Actor компоненте су слање и примање порука. То се ради на следећи начин [30]:

- Слање порука се врши позивом метода !:

```
a ! msg
```

Објекту **a** се шаље порука **msg**.

- Примање порука се остварује помоћу **receive** блока и упаривања образаца које је објашњено у поглављу 2.2:

```
receive {  
  case pattern1 =>
```

```
...
case patternN =>
}
```

Све компоненте Actor система су хијерархијски распоређене и организоване тако да се свакој приступа на јасно дефинисан начин. Ово је тема следећег поглавља.

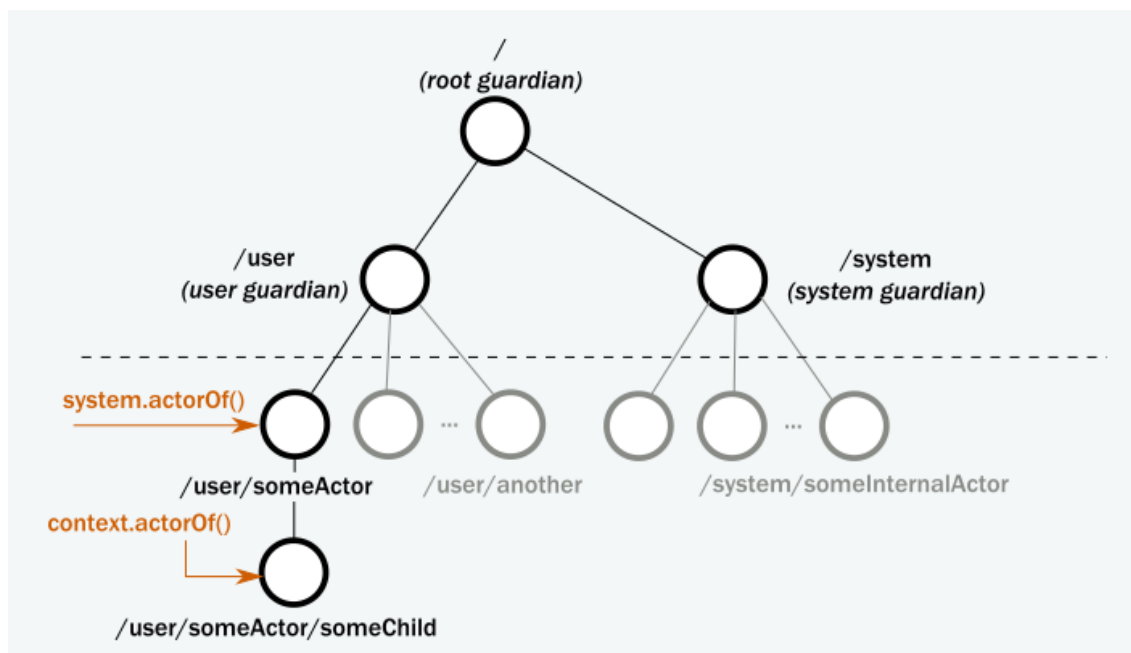
2.4.1 Структура

Све Actor компоненте које припадају једној логичкој компоненти апликације (нпр. једна за управљање базом података а друга за реаговање на захтеве корисника) припадају заједничком систему који се назива **ActorSystem**. То је хијерархијски уређена група компонената које деле исту конфигурацију. На ActorSystem можемо гледати као на менаџера својих компоненти који прави нове и претражује постојеће компоненте. Такође, ActorSystem је задужен за алоцирање $1...N$ нити које ће бити коришћене у апликацији [30, 6].

Све Actor компоненте једног система ActorSystem су смештене у **стабло** које је приказано на слици 2.4. Свака компонента има родитеља коме припада. На самом врху хијерархије налазе се три предефинисане компоненте [6]:

- *root guardian* је родитељ свих других компонената. Чак и он интерно има свог родитеља који се назива „Bubble-Walker”, али је он невидљив корисницима.
- *user guardian* је родитељ свих корисничких компоненти (компоненти које су направили сами корисници) и сва његова деца испред свог назива имају префикс „/user/”.
- *system guardian* је родитељ свих интерних компоненти. Интерна компонента, на пример, може бити компонента коју направи сама конфигурација оног тренутка када се креира систем ActorSystem.

Прву компоненту правимо помоћу метода *actorOf* самог система ActorSystem. Све компоненте направљене на овај начин ће постати деца предефинисане компоненте *user guardian*. Овакве компоненте су на врху хијерархије корисничких компоненти, па за њих кажемо да су на највишем нивоу иако постоје предефинисане компоненте изнад њих. Actor компонента која произведе



Слика 2.4: Стабло Actor компонената

нову компоненту представља њеног родитеља (енг. *parent actor*), а нова компонента је њено дете (енг. *child actor*). Ова радња се постиже методом *actorOf* поља *context* Actor компоненте. Ово поље је типа **ActorContext** који омогућава једној компоненти да има приступ самој себи, свом родитељу и својој деци. Свако дете добија име свог родитеља као префикс свог имена.

Прављењем Actor компоненте или добијањем постојеће претрагом система не добијамо директну референцу на компоненту, већ добијамо показивач на референцу **ActorRef**. Ова референца је задужена за слање порука својој компоненти и штити је од директног приступа корисника. Свака компонента има приступ својој референци преко поља *self*, а током обраде примљене поруке преко поља *sender* има приступ референци компоненте која јој је поруку послала.

Оваква хијерархијска структура је слична структури датотека у систему. У систему ActorSystem свака Actor компонента има путању **ActorPath** која је јединствено идентификује. Пошто се компоненте могу налазити и на више чворова на мрежи тј. извршавати се на различитим Јава виртуелним машинама, први део путање садржи протокол и локацију на којој се налази компонента (идентификацију ActorSystem система у коме се налази). Надаље путању чине надовезани елементи у стаблу раздвојени са „/”, почевши од корена. На пример

[6]:

```
"akka://my-sys/user/service-a/worker1"    // lokalna komponenta
"akka.tcp://my-sys@host.example.com:5678/user/service-b" // udaljena
komponenta
```

Код примера локалне путање протокол је „akka” а ActorSystem је „my-sys”. Међутим, нелокална путања дефинише другачији протокол и уз то име домаћина (енг. *host*) и број порта. У обе путање присутне су поменуте компоненте „root guardian” и „user guardian” што чини део путање „/user”. Након тога следе корисничке компоненте које добијамо проласком кроз стабло све до жељене компоненте.

Још једна сличност ове хијерархије са хијерархијом система датотека је начин на који претражујемо њене елементе. На пример [6]:

```
context.actorSelection("../*") ! msg
```

Метод **actorSelection** враћа референце компоненти које тражимо путем аргумента. Користећи „..” пењемо се један ниво изнад у стаблу тј. добијамо свог родитеља, а како „*” представља знак за било шта (*wildcard*), добијамо сву децу свог родитеља, укључујући и себе. На овај начин лако шаљемо поруку различитим компонентама истовремено.

Сваки ActorSystem има „диспечера” тј. елемент који се назива **MessageDispatcher**. Овај диспечер је задужен за слање порука у одговарајуће поштанско сандуче као и за позивање одговарајућег receive блока Actor компоненте. У позадини диспечера налази се складиште нити која се назива *thread pool*. Складиште садржи одређени број покренутих али неактивних нити које чекају да им се додели задатак. Сваки пут када је потребно извршити нови задатак, проверава се постојање слободне нити у групи. Ако постоји, бира се прва слободна нит. У случају да су све нити тренутно заузете, прва нит која се ослободи ће преузети нови задатак. Оваква идеја се примењује уколико знамо да ће бити пуно нити које ће извршавати кратке задатке, па је ефикасније имати унапред спремне нити него их изнова и изнова правити. На овај начин се такође смањује број нити које се креирају током рада апликације.

Thread pool даје важну гаранцију у конкурентној апликацији, а то је да у сваком тренутку само једна нит може извршавати одређену Actor компоненту. То значи да никада две или више нити не могу истовремено извршавати

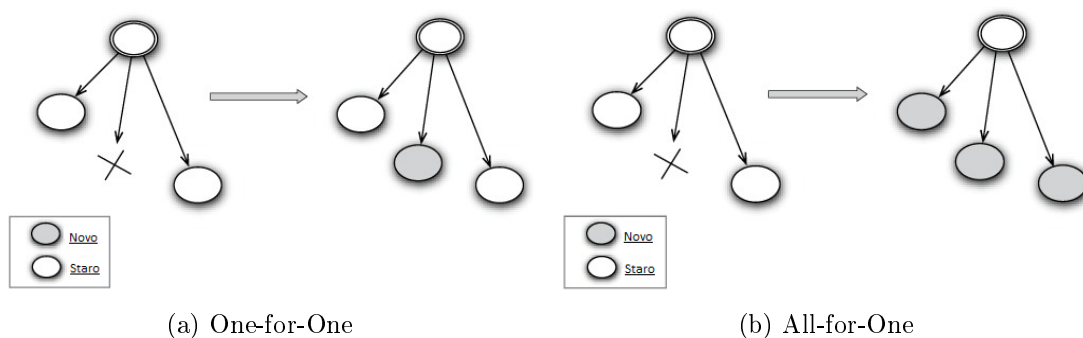
исту компоненту. Једну компоненту могу извршавати различите нити али само у различитим временским интервалима. То нам омогућава да безбедно мењамо податке унутар Actor компоненте докле год те податке не делимо [30, 6].

Компонента која је задужена за примање порука које су послате обустављеној или непостојећој компоненти назива се *dead letter actor*, а њено сандуче *dead letter mailbox*. Након што се одређена компонента угаси, није добра пракса направити нову са истом путањом. Разлог томе је то што не можемо претпоставити редослед догађаја и може се десити да нова компонента прими поруку која је била намењена старој [6].

Један од разлога због којих су компоненте смештене у хијерархијску структуру је управљање њиховим животним циклусима. Компоненте живе до момента када их корисник заустави и тада се прво рекурзивно заустављају сва њена деца. Ниједно дете не може надживети свог родитеља. То је корисно јер се на тај начин спречава цурење ресурса. Такође, препоручује се заустављање компоненте једино из ње саме. Добра је пракса да компонента као одговор на одређену поруку заустави саму себе док се заустављање произвољне компоненте из неке друге избегава.

Други разлог постојања хијерархије је управљање грешкама и изузецима. Акка подржава програмирање у коме се стање грешке посматра као било које друго валидно стање. Немогуће је спречити све грешке, тако да је циљ припремити начине на које ћемо се изборити са њима. У томе помаже структура компонената. Када компонента наиђе на проблем она привремено суспендује сву своју децу и саму себе, и обавештава родитеља о проблему. Њен родитељ одлучује како даље наставити са радом. Другим речима, сваки родитељ је задужен за решавање грешака и изузетака своје деце и тиме представља њиховог *супервизора*. Акција коју родитељ одлучи да предузме назива се *стратегија супервизора* (енг. *supervisor strategy*). Подразумевана стратегија је заустављање и поново покретање (рестартовање) детета и оно подразумева брисање акумулираног стања и враћање на почетак. Ову акцију је могуће предефинисати, тако да поред рестартовања постоје још три опције које родитељ може предузети [30, 6]:

- 1) Настављање са радом са акумулираним стањем



Слика 2.5: Стратегије рестартовања

- 2) Трајно заустављање детета
- 3) Пропагирање грешке на ниво изнад тј. обавештавање свог родитеља о њој и тиме суспендовање себе

Што се тиче самог рестартовања, разликујемо две стратегије: *One-for-One* и *All-for-one*. Ове стратегије говоре о томе шта се дешава са осталом децом супервизора тј. родитеља компоненте која се рестартује и илустроване су на сликама 2.5a и 2.5b [30, 6]:

- **One-for-One** је подразумевана стратегија у којој током рестартовања одређене компоненте остале поменуте компоненте живе и настављају са радом независно од ње. Ова стратегија се примењује када компонента врши посао који не утиче на послове других компоненти.
- **All-for-One** стратегија се примењује када компоненте заједно учествују у неком послу тако да рестартовање једне повлачи рестартовање и осталих компоненти. У овом случају рестартовање једино компоненте у којој се јавила грешка не би било валидно.

Глава 3

Расплинуто тестирање

Грешке у раду било које природе су неминовне, па тако и у програмирању. Развојем софтвера развијају се нове врсте пропуста. „Постоје стотине слабости софтвера које само чекају да буду откривене, и биће откривене када прође довољно времена” [35]. Неке грешке немају велики значај док друге могу оставити озбиљне последице. Постоје програми који користе грешке других програма уметањем вируса, црва, злонамерних скриптова и слично. Циљ напада на софтвер може бити преузимање или пад система, крађа, злоупотреба и мењање постојећих података, производња нових и лажних података итд. Зато је јако важно софтвер тестирати са циљем елиминисања што већег броја грешака које могу угрозити квалитет и безбедност система. [35, 22].

Расплинуто тестирање (енг. *fuzz testing/fuzzing*) је приступ тестирању софтвера који се бави генерисањем великог броја улаза у програм а затим за сваки од улаза посматра његов ток извршавања. Циљ ових улаза је да открију необичне грешке и изузетке у програму. Улази треба да буду неочекивани и генеришу се или потпуно случајно или уз коришћење одређених *хеуристика*. Узимају се у обзир и неисправни улази јер је важно да програм препозна и одбаци овакве улазе без прекида или отежавања даљег рада.

3.1 Основе тестирања софтвера

Тестирање софтвера представља најчешћи вид верификације софтвера. *Верификација* софтвера се бави испитивањем исправности програма тј. провером жељеног понашања програма задатог *спецификацијом*. Постоје *статичка* и *динамичка* верификација. Статичка испитује исправност програма без извр-

шавања кода тј. његовом анализом, док динамичка подразумева испитивање у току извршавања кода [20].

Тестирање се може описати на више начина. Често се описује као процес коме је циљ процена квалитета софтвера. Овакву дефиницију тестирања треба узети са резервом. Тестирање нам свакако даје увид у квалитет софтвера али прилично мали, јер углавном случајеви које тестирамо представљају само кап у океану свих могућих случајева. Мање амбициозна дефиниција представља тестирање као процес који покушава да открије грешке у програму посматрајући његово извршавање у контролисаним условима [25].

Технике тестирања софтвера константно напредују, али треба имати у виду њихова ограничења. Холандски информатичар и добитник Тјурингове награде Едсгер Дајкстра (хол. *Edsger Wybe Dijkstra*) је рекао: „Тестирање софтвера може показати присуство грешака, али никада њихово одсуство” [25]. Тестирање не може доказати исправност софтвера, већ се за то користе друге, математичке технике. Ипак, тестирање има важну улогу у свим фазама развоја софтвера. Помоћу њега смањујемо број промаклих грешака и повећавамо поузданост изграђеног система.

Због ефикасности процеса тестирања, често је пожељна његова аутоматизација. Генерисање тест примера је у већини случајева потребно извршити ручно, мада постоје неке врсте тестирања у којима је могуће аутоматизовати овај процес. Што се тиче извршавања тест примера, аутоматизација је углавном могућа и њу подржавају разни алати за развој софтвера.

Важно је на који начин се приступа тестирању софтвера. Некада је дизајн тестова једнако компликован као дизајн самог програма. Било да се тестирање врши ручно или аутоматизовано, пожељно је имати планове и идеје о тест примерима који су вероватнији да изазову грешку од других. Треба одредити и редослед којим ће се извршавати осмишљени тестови. Редослед је важан јер постоје случајеви када се одређени тест примери не могу извршити пре извршавања неке друге радње. Рецимо, не можемо тестирати брисање податка из базе података ако је она празна тј. ако претходно није успео тест уметања податка у базу. Након извршавања тестова прегледају се добијени резултати и утврђује се спремност тестираног система. Сав посао везан за тестирање може вршити сам програмер који је уједно и тестер. Друга могућност је постојање тестера који поред програмирања разуме технике тестирања, познаје грешке које се често јављају и необичне случајеве који их производе. Он може да ради

заједно са програмером а може бити и раздвојен од њега [12, 19].

„Покривеност кода (енг. *code coverage*) је број неких елемената програма испитаних тестовима у односу на укупан број тих елемената” [19]. Уопштено, покривеност представља неку врсту метрике која говори о броју случајева који су испитани тестовима. Када сматрамо да тестови испитују добар део програма, кажемо да тестови „покривају” велики број случајева тј. да имају велики ниво покривености. У зависности од изабраних елемената програма разликујемо покривеност путања, наредби, грана/одлука, услова, вишеструких услова и функција. Постоје различити алати за рачунање нивоа покривености кода [19].

3.1.1 Нивои тестирања

Тестирање се врши на више нивоа, у зависности од сложености компонента које се тестирају. Почиње се од **тестирања јединица кода** (енг. *unit testing*), тј. независних делова система. Јединице представљају најмање независне целине које обављају неку функцију (нпр. класа или функција). Пре спајања више независних јединица, важно је да оне саме раде правилно, изоловане од других јединица. Други ниво, **компонентно тестирање** (енг. *component testing*), је веома слично првом нивоу. Компоненте су изоловане од других компонента, али су мало веће и јединице које су у њима нису изоловане једне од других. Трећи ниво је **интеграционо тестирање** (енг. *integration testing*). Оно је задужено за тестирање компоненти које заједно чине целину система. То је важно јер се неретко дешава да компоненте које савршено раде самостално не успевају добро да раде заједно и комуницирају једне са другима. Следећи ниво тестира систем као целину и зато се назива **системско тестирање** (енг. *system testing*). На овом нивоу се тестира и хардвер, а тестирање не обухвата само функционалност програма већ и безбедност, капацитет, перформансе, преносивост, итд. На овом нивоу се примењује **истраживачко тестирање** (енг. *exploratory testing*) током кога тестер извршава тест случајеве који нису били у плану, са намером да открије непредвиђене начине коришћења система. Обављају се и **тестови прихватљивости** (енг. *acceptance testing*) које извршавају корисници, проверавајући да ли софтвер испуњава њихова очекивања и потребе. Након измена система врши се **регресионо тестирање** (енг. *regression testing*) које треба да покаже правилан рад неизмењених функција. Такође треба да покаже и да су перформансе новог система макар једнаке перформан-

сама старог система, а пожељно је да су боље. Сви ови нивои се примењују уколико време дозвољава темељно тестирање. У случају мањка времена, тестирање се прилагођава могућностима [12, 19].

3.1.2 Стратегије тестирања

На основу доступних информација о имплементацији софтвера који се тестира разликујемо три основне стратегије тестирања [34, 35, 19]:

Тестирање беле кутије (енг. *white box testing*) или *структурно* (енг. *structural testing*) тестирање подразумева познавање интерне структуре и имплементације програма. Тестерима је доступан целокупан изворни код, тако да је њихов посао да га детаљно изуче како би нашли делове који су подложни грешкама. То је мукотрпан посао јер подразумева пролажење кроз све линије кода којих често има превише за људску обраду. Због тога се углавном користе алати који скенирају код и региструју потенцијалне слабости програма. Након тога их проверава програмер и одлучује да ли су пронађене слабости заиста претња. Алати много помажу али не могу да замене знање и искуство стручњака.

Добра страна ове стратегије је што доступност изворног кода омогућава високу покривеност кода (можемо проћи кроз велики број путања кроз програм, кроз много грана, итд.). Ипак, велики број линија кода повлачи и велику сложеност његовог анализирања. Алати који анализирају код често окарактеришу пуно делова кода као потенцијалне слабости. Тада програмери морају проћи кроз дугачак извештај о њима, а углавном се испостави да је већина записаних слабости лажна узбуна. Такође, ова стратегија је непримењива уколико изворни код није доступан. Треба имати у виду да је доступност кода уобичајена код програма на Linux оперативним системима, што није случај код програма писаних за Windows оперативне системе.

Тестирање црне кутије (енг. *black box testing*) или *функционално* (енг. *functional testing*) не користи никакве информације о интерној структури и имплементацији програма. Ова стратегија се може применити када информације о унутрашњој структури нису доступне, али је корисна и када јесу доступне, поред тестирања беле кутије. Једине информације које се

користе су улаз и излаз из програма. Тестирање се углавном заснива на претпоставкама о тестираном програму, његовим спецификацијама и налажењу прихватљивог броја репрезентативних тест примера.

Једна од предности ове стратегије је примењивост. Без обзира на доступне информације, тестирање црне кутије је увек могуће и једноставније од осталих стратегија јер не изучава специфичности одређеног програма. Погодно је јер га могу примењивати и тестери који нису програмери. Такође, због тога што се тестови не ослањају на унутрашњу структуру софтвера, исти тест можемо употребљавати више пута за различите програме сличне намене. Једноставност овог приступа има и своје мане. Пошто се тестирање врши на основу претпоставки, никада не можемо тачно проценити ефикасност тестирања и ниво покривености кода као код тестирања беле кутије. Такође, ова стратегија није погодна за комплексне нападе где је потребно извршавати тестове одређеним редоследом ради изазивања рањивог стања програма.

Тестирање сиве кутије (енг. *gray box testing*) представља комбинацију претходне две стратегије. Једна могућност је да се тестови базирају на обема техникама беле и црне кутије. Друга могућност је да у некој мери постоје информације о унутрашњој структури софтвера, али не као код тестирања беле кутије. Изворни код није доступан али се увид у структуру софтвера добија преко датотека компајлираног програма тј. његових бинарних извршних датотека. Како су ове датотеке нечитљиве за људе, на њих се примењује процес обрнутог инжењеринга (енг. *reverse engineering*). Обрнути инжењеринг не може да открије изворни код програма, али препознавањем одређених инструкција може да произведе формат који се налази између изворног кода и машинског кода. За разлику од бинарних датотека, овај формат је читљив за људе и омогућава анализу сличну као код структурног тестирања.

Тестирање сиве кутије је хибридно решење које може искористити предности стратегија црне и беле кутије. Уколико изворни код није доступан, често се користи обрнути инжењеринг као допуна чистом тестирању црне кутије. То је могуће јер су бинарне датотеке софтвера увек доступне (осим Веб сервера и апликација). Мана ове стратегије је велика сложеност обрнутог инжењеринга који захтева постојање богатих ресурса.

Свака од поменутих стратегија има своје предности и мане. Ни за једну се не може рећи да је генерално боља од других, јер је свака од њих погодна за откривање различитих врста слабости. Најбоље решење је применити више стратегија ради откривања што више грешака. Различитост стратегија илустрована је сликом 3.1.



Слика 3.1: Разлике између стратегија тестирања

3.2 Основе расплинутог тестирања

Развојем расплинутог тестирања је откривено доста начина генерисања улаза који у великом броју случајева откривају грешке. Најчешће се улази или делови улаза генеришу случајно, праве се улази који су мањи него што би требало (нпр. датотека са мање података), користе се негативне и граничне вредности нумеричких типова података, предугачке ниске карактера, ниске карактера са специјалним карактерима, замењују се суседни битови, итд. Не постоје правила која дефинишу радње које се примењују на подацима, већ само савети који су експериментално откривани годинама.

Савети који помажу при генерисању тест примера представљају *хеуристике*. Хеуристика је приступ решавању проблема који прави изборе на основу искуства тј. изборе који ће највероватније довести до решења проблема. То су избори за које не постоји доказ да доводе до решења, али се у пракси показало да дају добре резултате. Хеуристике се користе у случају да је немогуће или је јако тешко испитати све могуће случајеве неког проблема. Зато се испитају познати случајеви који углавном у кратком временском периоду нађу решење. Решење често не буде најбоље могуће решење, али буде прихватљиво [31].

Кључна предност расплинутог тестирања наспрам других техника тестирања је могућност аутоматизације овог процеса. Ручно креирање тест примера и покретање програма је веома напорно, споро, неефикасно и захтева да га обавља стручњак у области тестирања. Зато се прибегава аутоматизацији што је могуће већег дела посла које тестирање захтева. Тестирање може покренути особа без великог знања о томе. Такође, многе апликације за тестирање су примењиве на велики број различитих програма [34, 35].

Расплинуто тестирање је један од приступа тестирања црне кутије. Програм за тестирање не зна ништа о структури програма који тестира. Проблем који се јавља оваквим тестирањем је тај што формирани улази често откривају грешке које је лако открити јер се тај део кода често извршава, док семантичке грешке које се налазе дубоко у програму остају неоткривене. Зато су развијени приступи алтернативног расплинутог тестирања који одговарају тестирању црне, беле и сиве кутије [35, 8, 14]:

Расплинуто тестирање црне кутије (енг. *black box fuzzing*) је стандардни приступ на који се мисли под појмом „расплинуто тестирање”. Он је најлакши за имплементацију и омогућава тестирање великог броја тест примера за кратко време. Показује се да је често ефикаснији од других приступа уколико њима треба превише времена за генерисање тест примера.

Расплинуто тестирање беле кутије (енг. *white box fuzzing*) пре самог тестирања користи технике које идентификују делове програма који могу да имају грешке. На тај начин ова врста тестирања повећава покривеност кода, открива грешке до којих се теже долази него другим врстама тестирања, али је и компликованија од њих.

Расплинуто тестирање сиве кутије (енг. *gray box fuzzing*) је приступ који има предности претходне две технике а труди се да избегне њихове мане. Тежи се да ефикасност буде што ближа ефикасности приступа црне кутије, али ова техника не генерише улазе „на слепо” већ користи неки паметнији начин, као приступ беле кутије. Уместо анализе кода, користе се технике инструментације да се добије увид у структуру програма.

3.2.1 Историја

Примена тестирања налик на расплинуто је почело још 1980-их година са алатом званим „мајмун” (енг. „*The monkey*”). Он је развијен за тестирање програма на Macintosh рачунарима, који су имали проблема због мањка меморије. Мајмун је симулирао правог мајмуна који неартикулисано удара у тастатуру и миш рачунара и на тај начин тестира рад покренутог програма. Ипак, до 1990-их су интересовања и потребе за тестирањем биле много мање него данас, највише због недостатка интернета и идеја напада на други софтвер. Истраживање безбедности софтвера је почело крајем 1980-их година, када су почели да се развијају напади који се односе на прекорачење бафера. Један од познатих напада је настао 1988. године и назива се „Morris Internet Worm” [34, 35].

Први пројекат развијен за тестирање је написан 1989. године. Идеју за овај пројекат је добио професор Бартон Милер (енг. *Barton Miller*), за кога неки сматрају да је „отац” расплинутог тестирања. Заједно са својим ученицима са курса о оперативним системима тестирао је квалитет и поузданост UNIX апликација. Програм је случајно генерисао ниске карактера и прослеђивао их апликацијама. Иако је тестирање имало веома једноставну идеју, у то време је то био велики помак јер се о расплинутом тестирању знало јакo мало. На Универзитету у Оулу у Финској је 1999. године почело развијање пакета тестова под називом **PROTOS**. Ови тестови су проверавали безбедност разних протокола. Компанија Microsoft је 2002. године новчано помогла овој иницијативи тако да је тим који је развијао PROTOS 2003. године засновао компанију за производњу пакета тестова, под именом **Codenomicon**. Након тога, објављен је програм за тестирање отвореног кода имена **SPIKE**. Он је био напреднији од програма који је развио професор Милер. Брзо је почело и развијање других пројеката попут sharefuzz, mangleme, Hamachi, CSSDIE, FileFuzz, SPIKEfile, notSPIKEfile, итд. Компанија Mu Security је 2005. године почела да развија хардверски уређај коме је циљ мењање података који се шаљу протоколима путем мреже. ActiveX контроле су такође постале мета програма COMRaider и AxMap 2006. године [34, 35]. Временом је потреба за сигурношћу софтвера постајала све већа, тако да су се развијали алати за најразличитије врсте апликација. Методе тестирања, као и апликације које врше тестирање, се и данас развијају и побољшавају. Доступан је велики број алата који примењују најновије технике тестирања и врше разне оптимизације ради веће ефикасности. Оптимизације омогућавају краће време и цену тестирања, једноставност коришћења, мини-

малну људску интеракцију, итд. Неки од алата имају специфичну примену, док су други употребљиви у свакој врсти софтвера. Неки од познатијих алата су AFL, Project Springfield, OSS-Fuzz, libFuzz, BFuzz, Fuddly, Honggfuzz, Peach, Radamsa, итд [1, 34, 35].

3.2.2 Ограничења

Различите врсте тестирања откривају различите врсте слабости. Не постоји универзална техника за откривање свих грешака, већ у зависности од потреба бирамо одређену технику. Неке од врста проблема које расплинато тестирање обично не открива су [34]:

Права приступа Неки програми разликују обичне од привилегованих корисника који имају више права од обичних (рецимо поред читања података могу да их бришу или мењају). Програм за расплинато тестирање не зна ништа о логици програма који тестира, због чега не разликује типове корисника. Извршавање недозвољених функција од стране обичног корисника ће проћи неопажено код програма за расплинато тестирање. Могуће је уметнути логику програма у програм за тестирање, али је то веома сложен процес.

Дизајн програма Неки програми имају мане које се односе само на јако лоше осмишљен дизајн. Пример је апликација која дозвољава свим корисницима приступ одређеним подацима. Велика је грешка не узимање у обзир злонамерне кориснике и претпостављање да ће једино корисници који имају дозволу да приступе подацима то и учинити. Проблем је непостојање аутентикације и ауторизације. Програм за расплинато тестирање не може да препозна овакав безбедносни проблем.

Корупција меморије Једна од врста напада на софтвер може бити приступање недозвољеној меморији. Читањем и писањем на разне меморијске локације нападач може имати одређену контролу над програмом. Недозвољен приступ меморији се препозна на нивоу инструкције машинског кода, након чега се програму пошаље сигнал *SIGSEGV*. Проблем је у томе што неретко програми сами решавају овај проблем тако што „ухвате” сигнал и реше да ли да процес покрену поново или само наставе са радом, што може бити опасно. Програм за расплинато тестирање често не при-

мети овакву врсту слабости јер је она решена (правилно или неправилно) од стране самог програма.

Сложени напади Расплинуто тестирање је корисно за детекцију индивидуалних слабости. Међутим, није се добро показало у случају сложених напада који искоришћавају више слабости истовремено.

3.3 Врсте програма расплинутог тестирања

Програми расплинутог тестирања се разликују по разним карактеристикама. Две основне поделе ових програма се односе на начин на који генеришу тест примере и на тип самог програма који тестирају.

Начин генерисања тест примера има два основна критеријума. Први критеријум представља постојање почетних тест примера од којих се генеришу нови тест примери [34, 35]:

- **Тестери засновани на мутацијама** (енг. *mutation-based fuzzers*) користе постојеће тест примере за генерисање нових, примењујући измене тј. мутације на подацима. Квалитет тестирања зависи од квалитета постојећих тест примера. Ови тестери су обично тестери опште намене.
- **Тестери засновани на генерацијама** (енг. *generation-based fuzzers*) генеришу тест примере „од нуле” тј. независно од постојећих. Једине информације које користе су информације о формату који тест примери треба да задовоље (нпр. формат датотека или одређени протокол). Обично су овакви тестери специфични за одређени проблем.

Програми расплинутог тестирања, независно од тога да ли су засновани на мутацијама или генерацијама, у некој мери користе случајне податке за генерисање тест примера. Могуће је све податке генерисати случајно, што доводи до тежег увиђања тачног низа догађаја који је довео до грешке. Друга могућност је да се на основу спецификације програма креирају тест примери који ће тестирати граничне вредности података, које су често критичне. Најбоље резултате даје генерисање тест примера комбинацијом ове две могућности.

Мутацијски и генерацијски тестери су основни типови тестера. Поред њих постоји још један тип тестера који се заснива на напредној техници која се назива **еволюционо расплинуто тестирање** (енг. *evolutionary fuzzing*). Ова

техника се заснива на „учењу”. На основу претходних корака, тестер закључује које тест примере би требало даље да конструише. Рецимо, за сваки тест пример тестер може да измери покривеност кода, и закључи на који начин да измени тест пример тако да покривеност буде већа. Еволуционо тестирање се ослања на друге технике, попут техника генетских алгоритама [35, 17].

Други критеријум који дели програме за тестирање на основу генерисања тест примера је присуство или одсуство знања о томе како треба да изгледају улази у програм који се тестира [34, 35]:

- **Интелигентни тестери** (енг. *intelligent fuzzers*) познају структуру улазних података програма који се тестира. На тај начин се избегне генерисање података које би програм одбацио пре било каквог рада над њима. Имплементација интелигентног тестера је компликована јер захтева проучавање структуре података који представљају улазе у програм.
- **Неинтелигентни тестери** (енг. *non-intelligent/dumb fuzzers*) конструишу улазне податке у програм „на слепо” тј. не знајући ништа о структури тих података. На тај начин тестер може генерисати велики број улаза који ће брзо бити одбачени од стране програма. Њихова имплементација је једноставнија од имплементације интелигентних тестера.

Ако улаз у програм представља датотека одређеног формата, неинтелигентни тестер генерише тест примере тако што инвертује произвољне битове те датотеке. Са друге стране, интелигентни тестер намерно оставља неке делове датотеке неизмењеним а друге делове мења. Рецимо, ако структуру чине називи поља и њихове вредности, називи ће остати исти док ће се вредности мењати.

Интелигентни тестери углавном имају већу покривеност кода и укажу на већи број грешака од неинтелигентних, али то није правило. Најбоље је креирати тестер који не иде ни у једну крајност, јер то води до лоших резултата. Тестер који не зна ништа о формату улаза уме да креира бескорисне тест примере, али уме и да пронађе неочекивану слабост система да се носи са неисправним улазима. Са друге стране, интелигентни тестер избегава губљење времена непотребним тест примерима, али води до претпостављања правилног рада неких функција система. Неретко се деси да овакав тестер не ради добро јер прави исте претпоставке које је правио и програмер који је писао програм, због којих и долази до одређених грешака. Такође, интелигентни тестер може бити

превише скуп због своје сложености. Балансиран приступ је најбоље решење: направити тестер који је довољно интелигентан да смањи време извршавања програма, али да не буде превише сложен и скуп, као и да не ослаби способност тестирања.

Неки аутори поистовећују тестере засноване на мутацијама са неинтелигентним тестерима, а тестере засноване на генерацијама са интелигентним тестерима. Такво поистовећивање није исправно. Тестери засновани на мутацијама најчешће заиста јесу неинтелигентни јер обично не узимају структуру улаза у обзир већ мењају битове постојећих улаза. Ипак, то није правило. Постоје и интелигентни тестери засновани на мутацијама који парсирају постојеће улазе да не би креирали оне које би програм одмах одбацио. Слично, тестери засновани на генерацијама могу бити неинтелигентни, иако је то веома редак случај. Рецимо, тестер који генерише улазе случајним подацима, без постојећих улаза, технички спада у тестере засноване на генерацијама.

У односу на програм који тестирамо, програм за расплинато тестирање спада у неку од следећих категорија [34, 35]:

Локални тестери (енг. *local fuzzers*) Ово је веома широка категорија која обухвата тестирање разних локалних програма. Класичан пример програма који су погодни за расплинато тестирање су програми који омогућавају обичном кориснику да привремено (за време његовог извршавања) добије одређене привилегије. На Unix системима има неколико таквих подразумеваних апликација (енг. *setuid applications*) а то се може подесити и за друге. Слабости у оваквим програмима представљају велику опасност јер би грешке могле да омогуће кориснику да трајно добије привилегије и извршава произвољан код. Расплинато тестирање проналази слабости ових апликација тако што их покреће са трансформисаним аргументима. Аргументи се могу слати програму у виду аргумената командне линије (енг. *command-line fuzzers*), у виду променљивих окружења (енг. *environment variable fuzzers*), или у виду датотека одређеног формата (енг. *file format fuzzers*).

Тестери удаљених програма (енг. *remote fuzzers*) Ови тестери се баве програмима који су доступни путем мреже, што је посебно значајно због широке употребе интернета. Међу њима постоје две врсте тестера.

Прву врсту чине тестери мрежних протокола (енг. *network protocol*

fuzzers). Тестирају се једноставни и сложени протоколи. Једноставни су они који имају слабу аутентикацију или је уопште немају, користе ASCII карактере уместо бинарних података, итд. Сложени протоколи се састоје углавном из бинарних података међу којима се налази понека ASCII ниска карактера. Аутентикација је сложенија и користи шифровање. Такође, често се врше провере помоћу дужине или збира послатих података.

Другу врсту чине тестери Веб апликација (енг. *Web application fuzzers*). Они комуницирају помоћу HTTP протокола. Веома су важни јер данас имамо јако много апликација на Вебу с којима су се развиле и слабости јединствене за њих попут уметања SQL упита (енг. *SQL injection*), уметање скриптова (енг. *cross site scripting*), онемогућавања услуге (енг. *denial of service*) и многих других.

Меморијски тестери (енг. *in-memory fuzzers*) Ови тестери су корисни када су неке друге врсте неефикасне. Њихове идеје су једноставне и заснивају се на измени аргумената функција програма директно у меморији, при чему се мењају и аргументи који иначе нису доступни кориснику. Имплементације оваквих тестера су веома компликоване. Постоји више приступа оваквом тестирању. Један приступ подразумева замрзавање процеса и чувања његовог тренутног стања које се касније поново користи тј. изазове се без поновног покретања процеса. Након тога, у свакој итерацији се у процес у таквом стању уметну неисправни улазни подаци, посматра се резултат и враћа се на сачувано стање. На овај начин се добија на брзини и ефикасности јер се делови програма који су неважни за тестирање прескачу (замрзавањем стања нас не занима шта се и како се извршило пре тога), и одмах се долази до критичног дела. Ипак, ово је веома сложен процес. Могуће је у програм уметнути улазне податке који не би били могући при нормалном покретању програма. Такође, ако се дошло до грешке оваквим тестирањем, потребно је ту грешку изазвати и нормалним покретањем програма, што може одузети доста времена.

3.4 Расплинуто тестирање формата датотека

Расплинуто тестирање формата датотека је метода тестирања апликација које прихватају специфичне улазе. Улази су одређеног формата, тако да је

циљ тестирања откривање грешака при парсирању улазних датотека. Примери су програми за манипулацију сликама, документима, музиком, итд. Програми су углавном на клијентској страни, али могу бити и на страни сервера (нпр. сервери електронске поште).

Особина овакве врсте тестирања која је чини једноставнијом од тестирања неких других типова апликација је та што се углавном извршава целокупно на једном систему. Рецимо, у случају тестирања мрежних протокола или Веб апликација постоје барем два система: један на ком се налази програм који се тестира, и други на ком се покреће тестирање. Са друге стране, тестер формата датотека теже препознаје да је циљана апликација изазвала грешку или изузетак. При тестирању програма на серверу је најчешће очигледно када је дошло до грешке јер сервер постане недоступан. То није случај код клијентских апликација јер програм за тестирање изнова поново покреће апликацију са новим тест примерима. Без адекватног решења може да се деси да тестеру промакне важна слабост програма. Због тога је важно да тестер формата датотека надгледа програм који тестира помоћу постојећег дебагера или дебагера који је уметнут у сам програм за тестирање. То повећава сложеност тестирања формата датотека [34, 35].

Тестирање се може описати помоћу следећих пет корака [34]:

1. Припремити тест пример методом заснованом на мутацијама или методом заснованом на генерацијама
2. Покренути циљну апликацију и учитати припремљен тест пример
3. Надгледати циљну апликацију (најчешће коришћењем дебагера)
4. У случају откривања грешке, забележити све потребне информације о њој. Ако након одређеног времена ниједна грешка није пронађена, ручно угасити циљну апликацију.
5. Поновити дате кораке.

Расплинуто тестирање формата датотека може бити интелигентно и неинтелигентно, као што може користити и обе методе засноване на мутацијама и генерацијама. У наставку су описане најчешће коришћене две технике и њихове добре и лоше стране [34]:

Неинтелигентно расплинато тестирање засновано на мутацијама

представља веома једноставан приступ чији квалитет зависи од квалитета почетних улаза. Наредни улази се добијају мењањем постојећих, па је важно имати међусобно различите почетне улазе да би што више случајева било покривено. Мењање улаза може бити у виду мењања појединачних бајтова, скупа бајтова, додавањем нових података, итд. Проблем код овог приступа је то што програми пре рада са улазним подацима често проверавају њихову исправност методама као што је провера контролне суме (енг. *checksum*). Улазни податак који је промењен не пролази проверу и бива брзо одбачен од стране програма. Због оваквих провера, тестер успева да испита јако мали део кода. Ово се може решити искључивањем провера што је тежак посао. Такође, овај приступ није ефикасан јер је број свих могућих начина на које можемо променити улазни податак превелик. Могуће је тестирати мали број могућности. Добра страна овог приступа је веома мала количина труда потребна пре покретања тестера, а она се односи на избор почетних тест примера.

Интелигентно расплинато тестирање засновано на генерацијама је при-

ступ који генерише тест примере на основу шаблона формата датотеке који се користи у тестирању. Шаблон многих формата датотека је познат и може се лако наћи, док у случају непостојања шаблона програмер мора сам проучити дати формат датотека. Квалитет тестирања зависи од тога колико добро програмер познаје дати формат датотека, због чега се много времена улаже у његово проучавање. Проучавање одузима доста времена, али је овај приступ ефикаснији јер тежи покривању већег дела кода.

При парсирању неисправних датотека се може наићи на разне врсте грешака. Неке од најчешћих врста грешака које покушавамо да откријемо расплинутим тестирањем су [34, 35]:

Онемогућавање услуге (енг. *denial of service*) Ово је напад на систем у коме се онемогућава његов рад. Нападач преоптерећује систем тако што му шаље јако велики број захтева. Системи подложни овом нападу су апликације на серверу.

Проблеми у раду са целим бројевима Цели бројеви су подложни грешкама које се односе на прекорачење меморије. Разлог за то неретко буде

знак целих бројева. Јављају се грешке при алокацији меморије, радом са индексима низова, поређењем означених и неозначених целих бројева, итд.

Прекорачење стека и хипа До прекорачења меморије било на стеку или на хипу долази из много разлога. Типичан пример подразумева копирање податка превелике величине у бафер фиксиране, мање величине. На тај начин може доћи до корупције меморије и извршавања произвољног кода.

Форматиране ниске карактера Једна од познатих слабости програма је слабост форматираних ниски карактера (енг. *format string vulnerability*). Опасност представљају функције за исписивање података на излаз које као аргументе имају форматирани ниски карактера као што су функције језика C *printf* и *fprintf*. За испис се користе параметри попут *%x*, *%s*, *%d* који говоре о типу конверзије која треба да се примени на аргументе функције. Проблем настаје када функција за испис очекује више аргумената него што јој је послато, тако да чита или пише на меморијске локације које јој нису дозвољене. Нападач успева да изазове овакво понашање тако што функцији за испис даје као аргумент ниску карактера које садржи параметре *%s*, *%x*, итд. На овај начин се на нападнутом систему могу извршити произвољне команде, прочитати недозвољени подаци, могу се изазвати пад система и друге последице.

Надметање за ресурсима Ово је ситуација када понашање програма зависи од догађаја којима не можемо предвидети време извршавања. Вишеничне апликације су склоне овом проблему (описано у поглављу [2.3.1](#)).

Важна особина програма за расплинуто тестирање је начин на који препознаје да је дошло до грешке или изузетка у програму. Повратна вредност апликације се користи као један начин препознавања и интерпретирања грешке. Програмери се ређе одлучују за ту опцију јер постоје други начини који дају више информација о изузетку. Најчешће се у ову сврху користи дебагер, али он може бити незгодан јер се већина дебагера употребљава интерактивно. То може смањити ефикасност аутоматизације расплинутог тестирања, јер захтева ручно поново покретање програма који је изазвао грешку. Други начин је имплементирање неких особина дебагера у сам тестер. То може одузети више

времена и свакако захтева више труда и истраживања, али на овај начин се имплементира дебагер који је прилагођен потребама.

Глава 4

Структура PDF датотека

Један од формата који се користи за приказивање и размену разних врста садржаја попут текста, дугмића, аудио и видео записа је формат *PDF* (енг. *Portable Document Format*). Приказ датотека у PDF формату (скраћено, PDF датотека) не зависи од софтвера, хардвера и оперативног система. Компанија Adobe је изумела овај формат, а данас га одржава Међународна организација за стандардизацију (енг. *International Organization for Standardization*). PDF датотеке имају много могућности, нпр. могу извршавати Javascript код, могу бити заштићене шифром, итд [5, 4].

4.1 Објекти датотека

Свака PDF датотека је сачињена из великог броја *објеката*. На пример, свака страница датотеке је описана *објектом странице* (енг. *page object*). Објекат странице реферише на објекте који описују садржај те странице (текст, слике, итд.).

PDF датотека се састоји од објеката следећих типова [7, 11, 21]:

Истинитосне вредности Представљају логичке вредности и дефинишу се кључним речима *true* (тачно) и *false* (нетачно).

Бројеви Користе се два типа објеката: цели и реални бројеви. Цели бројеви се представљају једном или више цифара којима може претходити знак броја ('+' или '-'). Реални бројеви се представљају исто као цели бројеви, с тим што у себи могу садржати децималну тачку.

Имена Представљају симболе дефинисане секвенцом ASCII карактера која следи одмах иза косе црте. Специјални карактери (нпр. бланко, заграде, итд.) унутар имена се представљају својим хексадецималним записом иза карактера '#'. Имена се користе за описивање садржаја објеката. Сваки објекат има садржај који је значајан за приказ датотеке, а природа самог садржаја се описује помоћу имена „/Type”. На пример, ако желимо да дефинишемо фонт којим је написана датотека, правимо објекат који садржи информације о том фонту. У таквом објекту се након имена „/Type” наводи име „/Font”, као индикатор да објекат садржи информације о фонту. Конкретан фонт дефинишемо именом који представља тај фонт (нпр. фонт „Times-Roman” бисмо дефинисали помоћу имена „/Times-Roman”).

Ниске карактера Представљају низове бајтова. Могу се представити произвољним бројем карактера између '(' и ')’ заграда, или низом хексадецималних записа карактера између '<' и '>' заграда.

Низови Представљају уређене секвенце 0 или више објеката који не морају бити истог типа. Објекти се наводе између заграда '[' и ']’.

Речници Представљају уређену листу парова кључ/вредност. Кључ се представља именом, а вредност било којим типом објекта. Парови се налазе између заграда '<<' и '>>'.

Токови Представљају секвенце бајтова неограничене дужине, због чега се велики блокови података попут слика углавном смештају у њих. Токови се означавају речником који описује податке који се налазе у њему, након кога се блок података наводи између кључних речи **stream** и **endstream**. Речник тока садржи разне информације, попут информација о броју бајтова који подаци заузимају, типу објекта који ток описује, филтерима који се примењују на податке, броју објеката у подацима итд.

Објекат „Null” Представља објекат чији тип и вредност нису једнаки типу и вредности било ког другог објекта. Референца на непостојећи објекат се третира као „Null” објекат. Он се означава кључном речју **null**.

Индиректни објекти Представљају објекте PDF датотеке који су нумерисани бројевима који представљају њихове јединствене идентификаторе.

Идентификатори омогућавају другим објектима да реферишу на индиректне објекте. Индиректни објекти се означавају својим идентификатором након кога се између кључних речи **obj** и **endobj** налази сам објекат. Било који објекат у PDF датотеци може бити индиректан објекат. Објекат странице спада у ову категорију.

4.2 Организација датотека

Због могућности приказа разноликог садржаја, структура PDF датотека је веома комплексна. Као што је приказано на слици 4.1, структура се састоји из четири дела [7, 11, 21]:



Слика 4.1: Структура PDF датотека

Заглавље Прва линија датотеке чини њено заглавље и садржи број PDF верзије која је коришћена у датотеци. У случају да се у датотеци налазе и бинарни подаци, потребно је у заглавље додати и другу линију која садржи најмање четири бинарна карактера (ASCII вредност им је већа или једнака од 128). Друга линија служи као индикатор да се у датотеци налазе бинарни подаци, што је корисно апликацијама које раде са PDF датотекама.

Тело Садржи све индиректне објекте који се налазе у датотеци. Сви објекти су смештени у стабло, чији је корен објекат који се назива каталог објекат (енг. *document catalog*). Каталог објекат је речник који садржи информације о PDF датотеци и реферише на друге објекте који дефинишу податке датотеке.

Подстабло (стабла свих објеката) које садржи све објекте страница се назива „стабло страница” (енг. *page tree*).

Табела „xref” Табела „xref” или „*cross reference*” табела садржи податке о свим објектима у датотеци и омогућава брз приступ било ком објекту датотеке. Уместо претраге целе датотеке у потрази за одређеним објектом, локација објекта се прочита из табеле након чега му се може директно приступити.

Сваки објекат описан је једним редом у табели дужине 20 бајтова. Првих 10 бајтова имају различиту намену у зависности од врсте објекта. Постоје две врсте објеката: *валидни* и *ослобођени* објекти. Ове две врсте објеката, као и само ослобађање су објашњени у поглављу [4.3](#).

У случају валидног објекта, првих 10 бајтова садржи локацију објекта у датотеци. Локација објекта је одређена помоћу удаљености (енг. *offset*) почетка датотеке до почетка дефиниције објекта. У случају ослобођеног објекта, првих 10 бајтова садржи број следећег ослобођеног објекта (референцу на следећи ослобођени објекат у повезаној листи свих ослобођених објеката).

Након првих 10 бајтова, у случају обе врсте објекта следи *генерацијски број* (енг. *generation/revision number*), који се увећава сваки пут када се објекат ослободи. Након генерацијског броја налази се индикатор који говори о врсти објекта: 'f' за ослобођен објекат или 'n' за валидан објекат.

„Trailer” Ова секција датотеке се дефинише кључном речју **trailer**. Овде се налази речник који садржи информације о датотеци као нпр. број редова у табели „xref”, референца на каталог објекат, итд. Најважнија информација је информација о локацији „xref” табеле тј. о удаљености почетка датотеке до почетка дефиниције „xref” табеле. Број бајтова који представља удаљеност се наводи након кључне речи **startxref**. Локација „xref” табеле је важна јер омогућава брз приступ самој табели која даље својом структуром омогућава брз приступ објектима датотеке. Последња линија „trailer” секције садржи индикатор за крај датотеке „*%%EOF*”.

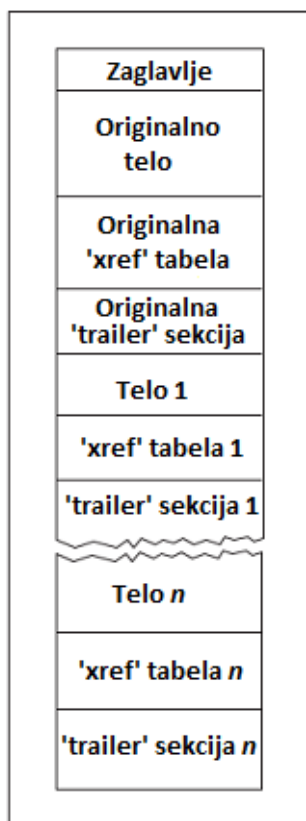
4.3 Ажурирање датотека

PDF датотеке се могу мењати, чиме се мења њихова структура тј. објекти у њима. Ажурирање датотеке не подразумева брисање или мењање постојећих података, већ додавање нових података на сам крај датотеке. Рецимо, заменом слике из датотеке новом сликом, објекат који представља стару слику неће бити обрисан или промењен тако да му садржај представља нову слику. Уместо тога, објекат старе слике биће означен као **ослобођен**, док ће објекат нове слике бити додат на сам крај датотеке. Тачније, ажурирањем датотеке се додаје ново тело, нова „xref” табела и нова „trailer” секција на крај датотеке. Ажурирање се може обављати неограничен број пута, при чему свака итерација (свако ажурирање) додаје три нове секције на крај датотеке. PDF датотека која је ажурирана n пута је илустрована на слици 4.2.

У новом телу се налазе нови или измењени објекти датотеке. Нова „xref” табела садржи искључиво локације нових, измењених и ослобођених објеката. Локације старих објеката се и даље читају из старе „xref” табеле („xref” табеле претходне итерације). До старих објеката се лако долази, јер нова „trailer” секција поред локације нове „xref” табеле садржи и локацију претходне „xref” табеле.

Сваки објекат који се обрише или измени постаје ослобођен, док се они који нису ослобођени називају **валидни** објекти. Сви ослобођени објекти се налазе у једној повезаној листи. Они се третирају као обрисани и не могу бити у садржају датотеке, а референце на њих су једнаке „Null” објекту. Ипак, објекат који је ослобођен може поново постати валидан неким наредним ажурирањем. *Генерациски број* објекта нам говори о укупном броју ослобађања једног објекта [7, 11, 21].

Ажурирање датотеке на овај начин омогућава да се мале измене у великој датотеци брзо сачувају. Такође, некада је важно сачувати оригиналан садржај датотеке [7].



Слика 4.2: Структура ажуриране PDF датотеке

Глава 5

Имплементација програма за расплинуто тестирање

Као демонстрација могућности Скале у паралелизацији задатака развијен је програм за расплинуто тестирање читача PDF датотека (скраћено PDF читача). Паралелизован је процес генерисања PDF датотека, као и њихово отварање у одређеном PDF читачу. Програм користи библиотеку Акка за имплементацију модела размене порука помоћу модела Actors. Што се тиче самог тестирања, имплементиран је интелигентан тестер заснован на мутацијама. Мутације се примењују на садржај датотека из корпуса датотека, који је описан у наставку. Коришћена је верзија 2.12.5 језика Скала, а програм је развијен у окружењу IntelliJ IDEA Community Edition 2017.3.5.

Корпус датотека је директоријум који садржи различите PDF датотеке. Корисник може додавати и брисати датотеке из корпуса. Иницијално је у корпус смештено двадесет шест датотека. Све датотеке садрже типове објеката описане у поглављу [4.1](#), а оно што их разликује је садржај тих објеката и величина датотеке. Обухваћени су типови садржаја као што су текст, слике, аудио и видео записи, линкови, интерактивни дугмићи и тродимензионални објекти.

5.1 Улазни параметри

Покретањем програма се пре почетка тестирања од корисника тражи да унесе:

1. **PDF читач** који жели да тестира. Могуће је изабрати један од понуђених читача или унети путању до извршне датотеке произвољног читача.
2. **Број итерација** који представља број датотека који ће отворити сваки од објеката који су задужени за тестирање. Ти објекти су Actor објекти који су детаљно објашњени у наредном поглављу.
3. **Број Worker објеката** који представљају Actor објекте који конкурентно (паралелно, уколико је могуће) мењају садржај PDF датотека и отварају их у одабраном читачу (детаљније објашњење налази се у наредном поглављу). Минималан број Worker објеката који корисник може да унесе је 1, а максималан зависи од рачунара на коме се извршава програм. Програм рачуна максималан број на основу два критеријума: броја језгара процесора и количине слободне меморије.

Уколико је број језгара процесора x , максималан број Worker објеката се најпре поставља на $3 * x$. Идеја је не преоптеретити рачунар непотребним бројем објеката који се не могу паралелно извршавати јер нема довољно језгара. Превелики број Actor објеката може смањити перформансе због дељења много мањег броја процесора.

Сваки Worker објекат учитава целу PDF датотеку у меморију. Због тога максималан број Worker објеката може бити и мањи од $3 * x$: број се смањује уколико слободне меморије нема довољно да сваки од Worker објеката прочита датотеку. Програм је подешен тако да дозвољава да датотеке заузимају половину слободне меморије.

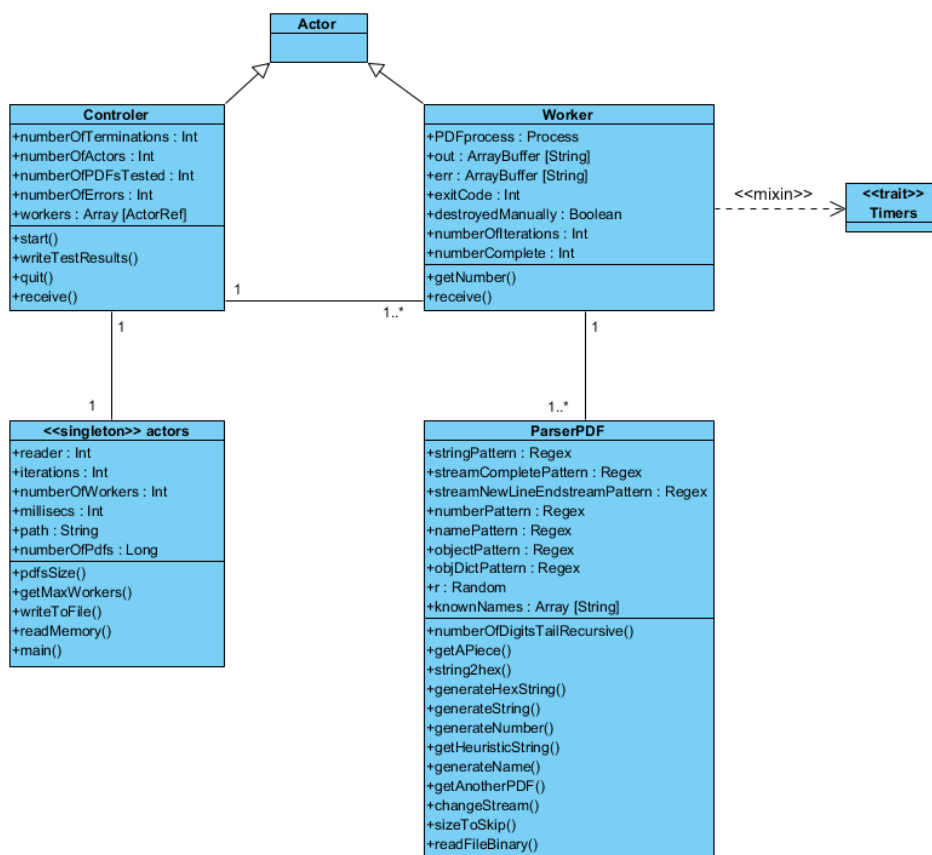
4. **Број милисекунди** који представља време које се додељује PDF читачу да отвори одређену датотеку. Неким читачима као што су Adobe, Foxit и Sumatra довољно је 1000ms да отворе датотеку, док су неки други читачи као што је Slim читач много спорији и потребно им је барем 3000ms. Уколико се читачу не додели довољан број секунди за отварање датотеке, процес отварања ће се угасити пре него што је датотека отворена и програм неће успети да детектује грешку у читачу (уколико она постоји). Због тога је најбоље пре самог тестирања читача проценити отприлике који број милисекунди му је довољан за отварање датотека.

Пре тестирања одређеног PDF читача потребно је подесити га тако да сваку датотеку отвара у новом прозору. На тај начин ће отварање сваке датотеке

представљати независан процес, што је неопходно за правилан рад програма.

5.2 Организација програма

Програм је организован у четири класе које су приказане дијаграмом на слици 5.1. Класе *Controler* и *Worker* представљају Actor објекте задужене за паралелизацију тестирања. Својство *Timers* је миксовано у класу *Worker*, а класа *ParserPDF* је задужена за мутацију садржаја датотека. Главни програм налази се у објекту *actors* који представља једини објекат истоимене класе. Због тога се он означава као *singleton* објекат. У наредним поглављима је детаљније приказан начин рада објеката ових класа.



Слика 5.1: Дијаграм класа

5.3 Actor објекти програма

Сви Actor објекти програма припадају једном систему ActorSystem. На врху хијерархије корисничких компоненти налази се један Actor објекат класе **Controler** који је задужен за покретање и праћење рада свих осталих Actor објеката. Систем ActorSystem и Controler објекат се креирају у функцији *main* која се налази у објекту actors. Њихово креирање и слање поруке Controler објекту је приказано у следећем примеру:

```
val system = ActorSystem("fuzzerSystem");
val a = system.actorOf(Props[Controler],
    name="fuzzer-controller-actor")
a ! "Pocni!";
```

Controler објекат креће са радом када прими поруку „Pocni!”. Тада позива метод *start()*, који је приказан у следећем сегменту кода:

```
def start() = {
    println("Pokrećem Worker-e");
    val workers: Array[ActorRef] = new Array[ActorRef](numberOfActors);
    for(i <- 0 until numberOfActors) {
        /* pravimo Actor-a i saljemo mu poruku za pokretanje */
        workers(i) = context.actorOf(Props(new Worker(i,
            numberOfActors)), name = "myWorker-" + i);
        context.watch(workers(i));
        workers(i) ! "Pokreni";
    }
}
```

Controlor објекат генерише друге Actor објекте који су објекти класе **Worker**. Референце на њих (типа ActorRef) чува у низу, и сваком шаље поруку за покретање (порука „Pokreni”). Метод *watch* омогућава Controlor објекту да „посматра” креиране Worker објекте, што значи да ће бити обавештен сваки пут када неки од њих заврши са радом.

Actor објекти класе Worker су објекти који су задужени за примену расплинутог тестирања. Сваки од њих има свој редни број на основу кога учитава једну од PDF датотека из корпуса датотека. Објекти класе Worker садрже објекат класе **PDFparser**, који примењује мутације над подацима те датотеке,

ГЛАВА 5. ИМПЛЕМЕНТАЦИЈА ПРОГРАМА ЗА РАСПЛИНУТО ТЕСТИРАЊЕ

како је описано у поглављу 5.4. На тај начин Worker објекат добија тест пример тј. нову PDF датотеку коју треба да отвори у PDF читачу који се тестира.

Отварање измењене PDF датотеке се догађа у спољашњем процесу чије се извршавање може надзирати. То се постиже помоћу библиотеке *scala.sys.process*. Ова библиотека омогућава извршавање, посматрање спољашњег процеса и руковање његовим улазом и излазом. Основу ове библиотеке чине класе *Process* и *ProcessBuilder* језика Java. Скала има класе истих имена и оне су коришћене у имплементацији.

Претпоставка је да сваки PDF читач треба да препозна лоше конструисане и злонамерне датотеке, да пријави грешку и нормално настави са радом. У случају да се PDF читач сам угаси након отварања датотеке, то је индикатор да је дошло до грешке коју није умео да обради. Због тога се спољашњем процесу који покреће PDF читач даје одређено време (улазни параметар *број милисекунди* наведен у поглављу 5.1) да отвори датотеку, након чега се анализира његово стање. Уколико је процес и даље активан, то значи да је PDF датотека и даље отворена у читачу или да је читач успешно обрадио грешку насталу при отварању. Тада се овај процес прекида од стране објекта Worker који га је креирао. Са друге стране, уколико процес није активан, то значи да се сам угасио и да је пронађена слабост PDF читача. У том случају се документују стандардни излаз, стандардни излаз за грешке и повратна вредност процеса. То су информације о извршавању процеса. Ове информације се чувају у текстуалној датотеци у директоријуму „results”. Назив текстуалне датотеке садржи префикс „CRASH_” након кога следи редни број Worker објекта. Такође, поред префикса „CRASH_” могу следити и префикси „ERR_” и „NULL_”. Префикс „ERR_” се појављује уколико стандардни излаз за грешке није празан, а „NULL_” уколико је нека од информација о процесу NULL. Чува се и PDF датотека која је узроковала грешку, у директоријуму „fuzzedPDFError”.

Команда која отвара PDF датотеку која се налази у директоријуму „fuzzedPDFcorpus” у читачу *SlimPDFReader* и покретање спољашњег процеса приказани су у следећем примеру:

```
val command = ".\\PDF_citaci\\slim\\SlimPDFReader.exe
               .\\fuzzedPDFcorpus\\" + numberComplete + "_fuzzed.pdf";
/* за nekoliko sekundi saljemo poruku "stop" sami sebi kako bismo
   analizirali stanje procesa */
timers.startSingleTimer("key", "stop", millisecs milliseconds);
```


ГЛАВА 5. ИМПЛЕМЕНТАЦИЈА ПРОГРАМА ЗА РАСПЛИНУТО ТЕСТИРАЊЕ

```
/* pokretanje spoljasnjeg procesa */  
val qb: ProcessBuilder = Process(command)  
PDFprocess = qb run (ProcessLogger((s) => { out ::= s},  
                                (s) => { err ::= s}));
```

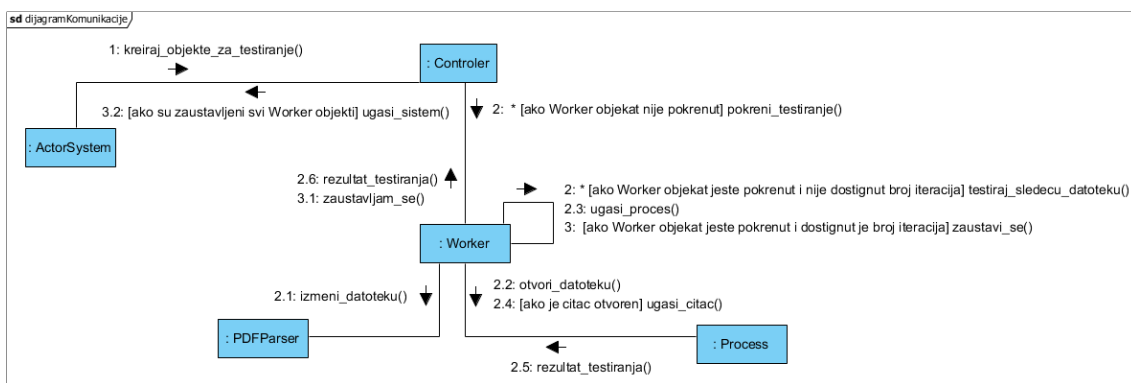
Команда која отвара датотеку је дата као ниска карактера облика „`|odabranCitac.exe .|nazivDatoteke.pdf`”. Чекање од одређеног броја милисекунди пре анализирања статуса процеса се остварује својством *akka.actor.Timers* који садржи објекат *timers* апстрактне класе *TimerScheduler*. Метод *startSingleTimer* овог објекта омогућава *Worker* објекту да сам себи пошаље поруку у одређено време тј. након дефинисаног броја секунди. Као што се види у примеру, објекат ће након *millisecs* милисекунди себи послати поруку „stop”. Када *Worker* објекат прими поруку „stop”, то значи да је истекао период рада спољашњег процеса и да треба да анализира његов статус и да га заустави уколико је још увек активан. Класе *Process* и *ProcessBuilder* се користе за креирање процеса на основу дате команде. Метод *run* извршава дати процес, а њему се прослеђује објекат типа *ProcessLogger* који рукује стандардним излазом и стандардним излазом за грешке. Стандардни излаз се уписује у листу ниски карактера *out*, а стандардни излаз за грешке у листу ниски карактера *err*.

Worker објекти се извршавају паралелно и изнова креирају тест примере и нове спољашње процесе који их отварају у читачу. Број датотека које ће креирати и отворити сваки од њих зависи од *броја итерација* који је унео корисник пре почетка тестирања. Једна итерација подразумева да сваки *Worker* објекат тестира PDF читач отварањем једне датотеке. На пример, ако желимо да сваки *Worker* објекат отвори *x* датотека, треба да унесемо *x* као број итерација. Укупан број датотека које ће програм отворити у читачу је онда $n * x$, где је *n* број *Worker* објеката. Друга могућност је да се програм извршава све док корисник сам не прекине извршавање програма. У том случају, као број итерација треба унети број -1 .

Уколико је дефинисан одређени број итерација (ако је различит од -1), сваки *Worker* објекат завршава са радом када анализира последњи процес који је покренуо. Тада зауставља сам себе, а порука о томе се шаље његовом супервизору, *Controler* објекту. Када *Controler* објекат добије поруку о заустављању од сваког детета, он зауставља цео систем *ActorSystem* чиме се завршава извр-

шавање програма.

Интеракција објеката приказана је дијаграмом комуникације на слици 5.2. Како број Worker објеката зависи од уноса корисника и рачунара на ком се извршава програм, на дијаграму је приказан само један Worker објекат и његова комуникација са другим објектима. Такође, приказ више Worker објеката на дијаграму је незгодан због паралелизма и конкурентности у програму. Не можемо предвидети када ће одређени Worker објекат добити процесорско време, због чега је немогуће приказати тачан редослед корака и порука које објекти размењују. Оно што се може одредити и што је приказано на дијаграму јесу кораци кроз које пролазе сви Worker објекти. Важно је напоменути и да Worker објекти не комуницирају између себе.




Слика 5.2: Дијаграм комуникације

5.4 Креирање PDF датотека

Објекти класе **ParserPDF** су објекти који примењују измене над садржајем датотеке коју добију као аргумент. Ова датотека представља улаз тј. основу за конструисање нове PDF датотеке. Применом мутација над неким подацима постојеће датотеке добија се нова датотека која ће бити коришћена као тест пример.

Због присуства бинарних података у PDF датотекама, садржај целе датотеке се најпре читава као низ бајтова. Низ бајтова се потом конвертује у ниску карактера, јер је потребно претражити садржај датотеке у облику разумљивом људима. Ниска карактера се претражује у потрази за објектима одређених типова, ради њихове измене. Када се измени жељени садржај, потребно је ниску карактера конвертовати натраг у низ бајтова који ће бити уписан у нову

ГЛАВА 5. ИМПЛЕМЕНТАЦИЈА ПРОГРАМА ЗА РАСПЛИНУТО ТЕСТИРАЊЕ

датотеку. При обе конверзије је важно користити једнобајтну кодну шему *ISO-8859-1*. Ова кодна шема је погодна јер чува оригиналан садржај бинарних података. Наиме, у вишебајтним кодним шемама није свака секвенца бајтова валидна, због чега се при декодирању секвенце она може пресликати у неки заменски карактер (нпр. ? или ). То се може догодити током конверзије низа бајтова у ниску карактера. На основу ових заменских карактера је немогуће добити натраг полазну секвенцу бајтова (при конверзији ниске карактера натраг у низ бајтова). Такође, неке једнобајтне кодне шеме не кодирају свих 256 различитих карактера. Због тога је у овом случају потребно користити једнобајтно кодирање које кодира 256 различитих карактера. На тај начин се бајтови јединствено пресликавају у карактере, а касније се ти карактери јединствено пресликавају у бајтове.

Када је датотека претворена у ниску карактера, тада се део њеног садржаја мутира тј. мења. Мутације се примењују над различитим типовима објеката PDF датотеке: бројевима, нискама карактера, именима, индиректним објектима, речницима и токовима. Због тога је потребно препознати појављивања ових објеката у датотеци. Треба напоменути да се унутар токова игнорише појављивање било ког другог објекта јер токови садрже бинарне податке које не треба претраживати као ниске карактера. Препознавање објеката се врши помоћу *регуларних израза*. Скала библиотека која омогућава коришћење регуларних израза је библиотека *scala.util.matching*. Регуларни изрази су дефинисани на следећи начин:

```
/* sabloni koje pretrazujemo u datotekama */
val stringPattern: Regex = ""\"([\\s\\S]+?)|<[0-9a-fA-F]*?>""\".r
val streamCompletePattern: Regex = ""\"(stream)([\\s\\S]+?)(endstream)""\".r
val streamNewLineEndstreamPattern: Regex = ""\"(stream)\\n(endstream)""\".r
val numberPattern: Regex =
    ""\"[+-]?((\\d+\\.?.?\\d+)| (\\d*\\.?.?\\d+)| (\\d+\\.?.?\\d*))""\".r
val namePattern: Regex = ""\"/[a-zA-Z0-9]+""\".r
val objectPattern: Regex = ""\"(\\d+) (\\d+) (obj)([\\s\\S]+?)(endobj)""\".r
val objDictPattern = ""\"(obj)([\\s\\S]*)(<)([\\s\\S]+?)(>)""\".r
```

Применом метода *r* на ниске карактера које представљају регуларне изразе добијамо обрасце који су типа *scala.util.matching.Regex*.

Следећи пример илуструје читање датотеке као низ бајтова, конверзију низа бајтова у ниску карактера, и проналажење свих токова података у ниски карак-

ГЛАВА 5. ИМПЛЕМЕНТАЦИЈА ПРОГРАМА ЗА РАСПЛИНУТО ТЕСТИРАЊЕ

тера помоћу обрасца *streamCompletePattern* (који је дефинисан у претходном примеру):

```
val byteArray: Array[Byte] = Files.readAllBytes(Paths.get(fileName));  
/* pravimo string koji sadrzi podatke iz datoteke */  
var contentString = new String(byteArray,  
    Charset.forName("ISO-8859-1"))  
val matchStreams =  
    streamCompletePattern.findAllMatchIn(contentString).toList
```

Подаци се читају из датотеке чија је путања наведена у ниски карактера *fileName*. Коришћена је библиотека *java.nio.file* и њена класа *Files* која садржи различите статичке методе за манипулацију датотекама и директоријумима. Метод *readAllBytes* чита садржај датотеке као низ бајтова. Од добијеног низа бајтова се конструише ниска карактера *contentString*. Метод *findAllMatchIn* проналази сва појављивања регуларног израза у ниски карактера, а метод *toList* сва појављивања смешта у листу.

Радње које се примењују над објектима које треба изменити су брисање, додавање и измена постојећих података, при чему се над једним објектом не морају применити све радње. На пример, неке објекте можемо заменити новим објектом истог типа, док неке друге можемо добити брисањем једног дела садржаја објекта. Нови објекти могу бити исте или различите величине од објекта који мењају. Могу се добити псеудо случајно или помоћу хеуристика.

Генерисање псеудо случајних података се врши методама класе *scala.util.Random*. На пример, псеудо случајан цео број се може добити позивом метода *Random.nextInt()*. Слично, постоје и методе *Random.nextString()*, *Random.nextDouble()*, итд.

Што се тиче хеуристика, коришћен је приступ који у своје тест примере умеће вредности објеката које често доводе до грешака у раду програма. Типови објеката који су обухваћени су ниске карактера, цели и реални бројеви. Вредности ових типова су прикупљене са два репозиторијума на сајту *github*^{1,2}. Ови репозиторијуми садрже податке који имају велику вероватноћу да изазову разне врсте грешака када се користе као улазни подаци програма. Ту се могу наћи адекватни улазни подаци у зависности од типа улазних података, природе програма који се тестира и типичних врста напада на програм, оперативног

¹<https://github.com/minimaxir/big-list-of-naughty-strings/tree/master/naughtystrings>

²<https://github.com/fuzzdb-project/fuzzdb/tree/master/attack>

ГЛАВА 5. ИМПЛЕМЕНТАЦИЈА ПРОГРАМА ЗА РАСПЛИНУТО ТЕСТИРАЊЕ

система на коме се извршава програм, итд. Нова вредност објекта који се мења хеуристички ће бити случајно одабрана из свих прикупљених вредности тог типа.

Прва измена која се примењује над датотекама је брисање неких индиректних објеката. Број објеката и објекти који ће бити обрисани се одређују случајно, након чега је датотека спремна за претраживање осталих типова објеката. У наставку су објашњени могући начини њихове измене:

Бројеви При измени броја случајно се одређује начин на који ће се генерисати нови број који ће га заменити: псеудослучајно или помоћу хеуристика. Прикупљених бројева има мало и то су граничне вредности (целих и реалних бројева), вредности блиске нули, бројеви 0, -1 , итд. Ови бројеви често изазивају грешке прекорачења меморије. У малом броју случајева, број се мења хеуристички изабраном ниском карактера.

Поред токова података и имена, објекти који се највише појављују у PDF датотекама су бројеви. Они су распрострањени свуда, а највише као идентификатори других објеката. Мењањем референци на друге објекте можемо постићи реферисање на непостојеће, невалидне или објекте погрешног типа.

Ниске карактера При измени ниске карактера случајно се одређује један од три начина измене. Прва два начина су једнака као и код измене бројева: псеудослучајно или хеуристички. Дужине псеудослучајних ниски карактера се одређују случајно тако да некад буду једнаке као ниске карактера које мењају, а некад им се дужина одреди случајно. Што се тиче хеуристика, прикупљених ниски карактера има јако много, а неки примери су приказани у табели 5.1. Најчешће ниске које се користе за тестирање су форматиране ниске карактера, ниске које представљају претрагу по систему датотека, ниске које садрже специјалне карактере, позиве функција, html тагове, итд. Трећи начин мења ниску карактера ниском која представља целу PDF датотеку. Циљ је уметнути другу PDF датотеку у датотеку чији садржај мењамо. PDF датотека која се умеће може бити валидна датотека или датотека коју је раније „покварио” тј. генерисао програм за тестирање.

Имена При измени имена случајно се одређује један од три начина измене. Први начин генерише име псеудослучајно тј. ниску карактера која почиње

знаком `'/'` за којим следи ниска карактера произвољне дужине. Други начин уместо имена умеће хеуристички одабрану ниску карактера, док трећи начин представља замену имена једним од имена који се често користе у PDF датотекама. Често коришћена имена су дефинисана у самом програму и прикупљена су на основу корпуса датотека.

Токови података Објекти овог типа се не могу мењати као обичне ниске карактера већ се конвертују у низ бајтова. Како су токови података већином веома дугачки, није циљ променити цео ток већ променити само његове „парчиће”. „Парчићи” су његови поднизови који најчешће буду промењени псеудо случајним низовима бајтова добијеним методом *Random.nextByte()*. У малом броју случајева, подниз буде замењен низом бајтова који представља другу PDF датотеку (уметање друге датотеке је могуће и код измене ниске карактера). Број поднизова за измену, њихова величина и локација у низу се одређују случајно. Након свих промена у низу бајтова, он се конвертује натраг у ниску карактера. При свакој конверзији користи се кодна шема *ISO-8859-1*.

Речници Речници који се могу променити у датотеци су речници који следе након дефиниције индиректних објеката, тј. након кључне речи *obj*. Они се одређују случајно и мењају се на само један начин: сваки речник одабран за промену копира садржај претходно одабраног речника (у случају првог одабраног речника, његов садржај остаје непромењен).

5.5 Резултати тестирања

У току тестирања читача долазило је до проблема са недостатком меморије. Због тога су одређени делови кода имплементирани на два начина, због чега разликујемо прву и другу верзију програма. У наставку су описане обе верзије, њихово поређење и грешка која се јављала.

Приликом тестирања прве верзије програма дешавало се да програм избаци изузетак: *java.lang.OutOfMemoryError: Java heap space*. До овог изузетка долази када Јава виртуелна машина нема довољно слободне хип меморије. Оно што је било збуњујуће при појављивању ове грешке је што се она манифестовала веома непредвиђено: некад након три сата рада програма, некад након десет минута, а некада се није ни манифестовала. Честим исписивањем количине

ГЛАВА 5. ИМПЛЕМЕНТАЦИЈА ПРОГРАМА ЗА РАСПЛИНУТО ТЕСТИРАЊЕ

System("\\ls -al /\")
<?xml version="\1.0\" encoding="\ISO-8859-1\"?><!DOCTYPE foo [<!ELEMENT foo ANY ><!ENTITY xxe SYSTEM \"file:///etc/passwd\" >]><foo>&xxe;</foo>
\$ENV{'HOME'}
../../../../../../../../etc/passwd%00
Kernel.exec("\\ls -al /\")
eval(\"puts 'hello world'\")
1'; DROP TABLE users- 1
</textarea><script>alert(123)</script>
<IMG SRC=\"javascript:alert('XSS')\"
/..%c1%9c..%c1%9c..%c1%9c..%c1%9c..%c1%9c{FILE}
/../../../../../../../../{FILE}
DCC SEND STARTKEYLOGGER 0 0 0
%d %d %d%d%d%d%d%d%d

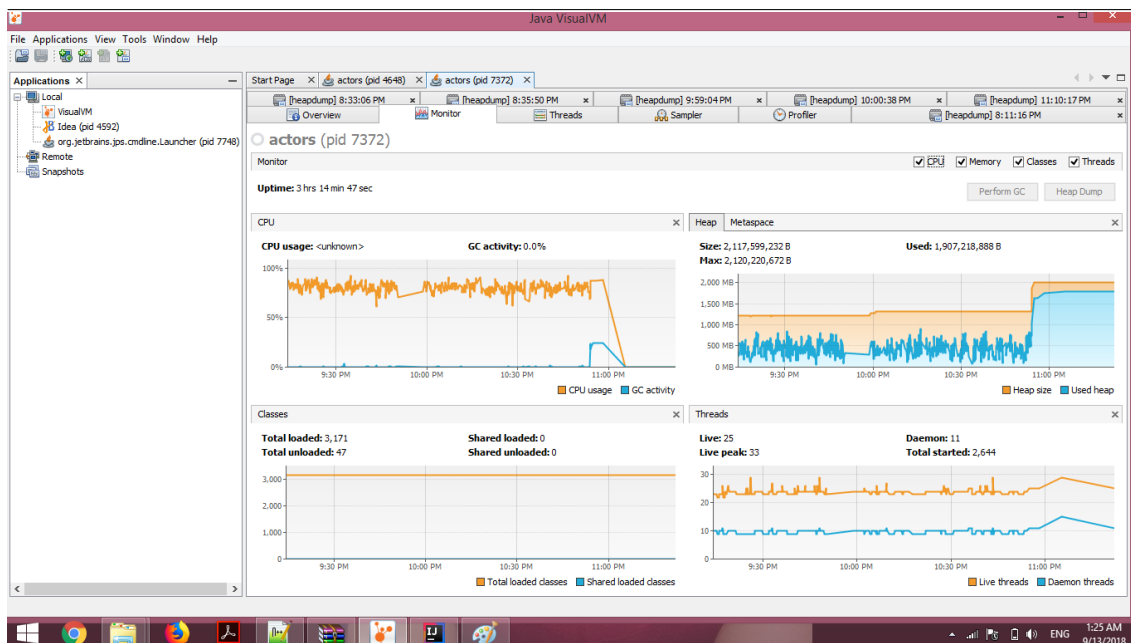
Табела 5.1: Примери ниски карактера

слободне хип меморије заиста се видело да се у неким случајевима меморија редовно ослобађа, док у другима након одређеног времена количина слободне меморије пада све док се не деси избацивање изузетка. Овакво понашање упућивало је на цурење меморије које се не манифестује увек.

Експлицитно постављање референци објеката на `null`, и иницирање сакупљача отпадака из програма није помогло, тако да је наредни корак био покретање програма заједно са профајлером *VisualVM*. Профајлер омогућава лакше и јасније праћење рада програма и меморије, нарочито исписивањем алоцираних објеката у меморији када се деси изузетак.

На сликама 5.3 и 5.4 приказан је пример рада програма у профајлеру. За улазне параметре постављени су читач *Sumatra*, број `-1` као број итерација, број `10` као број *Worker* објеката и број `1000` као број милисекунди. Прва слика приказује целокупан рад програма. Од посебне важности је секција *Heap* коју чини график који приказује количину слободне хип меморије у зависности од времена. Може се видети да је програм радио преко три сата, редовно ослобађајући меморију све док у једном тренутку количина меморије није нагло порасла до границе коју дозвољава Јава виртуелна машина. Када је избачен изузетак, забележени су објекти алоцирани на хипу, што је приказано на другој слици. Највећи део меморије заузимају инстанце

ГЛАВА 5. ИМПЛЕМЕНТАЦИЈА ПРОГРАМА ЗА РАСПЛИНУТО ТЕСТИРАЊЕ



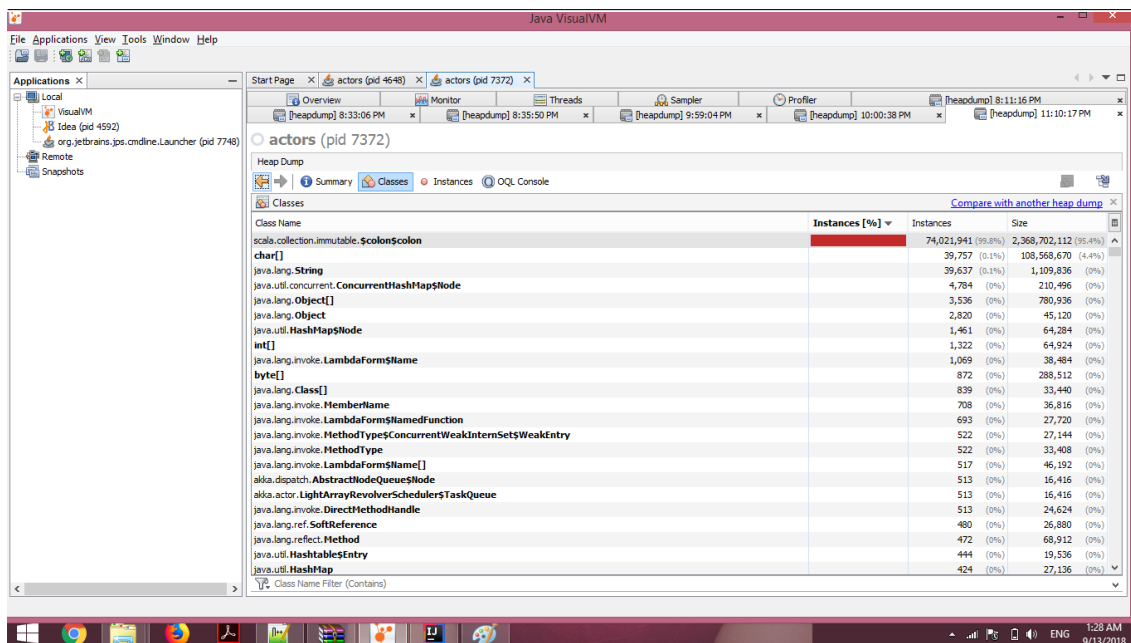
Слика 5.3: Приказ рада програма помоћу профайлера VisualVM

класе `scala.collection.immutable.$colon$colon`, која представља имутабилне листе у Скали. Програм је у једном тренутку заузео меморију великом количином листи чији је тип елемената био `Regex.Match` (што је такође прочитано из профайлера). Објекти овог типа представљају појављивања одређених регуларних израза. Регуларни изрази омогућавају претраживање различитих објеката у датотеци (како је објашњено у поглављу 5.4).

Начин на који је имплементирано претраживање и мењање објеката у првој верзији програма је могао бити разлог преоптерећења меморије листама. Тај начин је подразумевао да се сваки тип објекта у датотеци претражује функцијом `scala.util.matching.Regex.findAllMatchIn`, која проналази **сва** појављивања одговарајућег регуларног израза. Сва појављивања су смештана у листу, након чега се на основу броја појављивања бирао јако мали број објеката који ће бити променени. Када је завршено са свим променама објеката, листа која садржи појављивања је постављена на `null`. Разлог недостатка меморије је могао бити тај да су листе биле превелике да би стале у меморију уколико је број појављивања одређених објеката био превелик. Такође, због конкурентности у програму, могло се десити да више нити у исто време захтева чување великих листи у меморији.

Друга верзија програма је имплементирана тако да избегне конструкцију

ГЛАВА 5. ИМПЛЕМЕНТАЦИЈА ПРОГРАМА ЗА РАСПЛИНУТО ТЕСТИРАЊЕ



Слика 5.4: Приказ алоцираних објеката помоћу профилера VisualVM

и чување листи. Уместо претраживања свих појављивања објеката одједном, појављивања су тражена и чувана једно по једно. Коришћена је функција *scala.util.matching.Regex.findFirstMatchIn* која проналази само **прво** појављивање регуларног израза тј. објекта. Када се пронађени објекат промени, случајно одређени део PDF датотеке се прескаче, након чега се одатле поново тражи прво појављивање објекта одређеног типа. Тај поступак се понавља све до краја датотеке. На тај начин свака нит у меморији чува само по један, тренутно пронађени објекат.

Ипак, друга верзија програма није исправила грешку недостатка меморије. Грешка о недостатку хипа је наставила непредвиђено да се манифестује, појављујући се након различитог периода рада програма. Највећи део заузете хип меморије су и даље биле листе типа *Regex.Match*, иако у имплементацији не постоји део кода за који се може претпоставити да генерише овакве листе. Због тога се претпоставља да је проблем у структурама података које се користе за функционални део Скале.

Због грешке недостатка меморије тешко је упоредити прву и другу верзију програма. Тестирањем друге верзије није се видело побољшање, али се јасно поређење не може дати са сигурношћу због непредвиђеног понашања обе верзије. Друга верзија захтева више времена за обраду једне датотеке, што није

ГЛАВА 5. ИМПЛЕМЕНТАЦИЈА ПРОГРАМА ЗА РАСПЛИНУТО ТЕСТИРАЊЕ

значајно у случају мањих датотека, док је у случају већих датотека разлика приметна.

Иако проблем није отклоњен обе верзије програма су неко време тестирале PDF читаче и постигле одређене резултате. Тестирана су три читача: Foxit, Sumatra и Slim читач. Ниједна верзија програма није успела да пронађе грешку у читачима Foxit и Sumatra. Оба читача су успешно препознала сваку невалидну датотеку пријављујући грешку при отварању. Са друге стране, Slim читач се показао као веома слаб. Обе верзије програма су често изазивале грешке у његовом раду. Један начин на који је Slim читач показао слабост је гашење одмах при отварању неисправне датотеке. Обе верзије програма су детектовале овакво понашање и сачувале датотеке које су изазвале гашење читача. Други начин на који је Slim читач показао неправилно понашање је престанак реаговања након отварања датотеке. У том случају оперативни систем издаје обавештење да је читач престао да реагује, што програм не успева да детектује јер је у том случају спољашњи процес који је задужен за отварање датотеке и даље активан.

PDF датотеке су веома комплексне и њихово дубље проучавање представља највећи простор за побољшање програма. Бољим парсирањем датотека и мењањем њихових објеката на више начина постоје веће шансе за проналазак грешака у читачима који су данас веома јаки и отпорни на нападе. Програм се може побољшати и тако да омогући: детектовање престанка реаговања читача, покретање читача уз дибагер, мењање датотека без учитавања целог њиховог садржаја одједном, итд.

Глава 6

Закључак

Библиографија

- [1] Best Fuzzing Tools 2017. <http://opaida.com/2017/08/12/best-fuzzing-tools-2017/>.
- [2] Description of race conditions and deadlocks. <https://support.microsoft.com/en-us/help/317723/description-of-race-conditions-and-deadlocks>.
- [3] Martin Odersky: Biography and current work. <https://people.epfl.ch/martin.odersky/bio?lang=en&cvlang=en>.
- [4] PDF Reference And Adobe Extensions To The PDF Specification. https://www.adobe.com/devnet/pdf/pdf_reference.html.
- [5] What is PDF? <https://acrobat.adobe.com/be/en/acrobat/about-adobe-pdf.html>.
- [6] Akka Documentation: Version 2.5.13. <https://doc.akka.io/docs/akka/current/index.html>, 2011-2018.
- [7] Adobe Systems Incorporated. Document management - Portable Document Format - Part 1: PDF 1.7. https://wwwimages2.adobe.com/content/dam/acom/en/devnet/pdf/PDF32000_2008.pdf, 2008.
- [8] Marcel Böhme, Van-Thuan, and Pham Abhik Roychoudhury. Coverage-based Greybox Fuzzing as Markov Chain. *ACM CCS 2016*, 2016.
- [9] Paul N. Butcher. *Seven concurrency models in seven weeks: when threads unravel*. The Pragmatic Bookshelf, Dallas, Tex., 2014.
- [10] Felice Cardone and J. Roger Hindley. History of lambda-calculus and combinatory logic. 2006.

- [11] GNUpdf contributors. Introduction to PDF. https://web.archive.org/web/20141010035745/http://gnupdf.org/Introduction_to_PDF, May 2010.
- [12] Lee Copeland. *A Practitioner's Guide to Software Test Design*. Artech House, Norwood, MA, U.S.A., 2004.
- [13] David J. DeWitt, Samuel Madden, and Michael Stonebraker. How to Build a High-Performance Data Warehouse. http://db.csail.mit.edu/madden/high_perf.pdf.
- [14] Vijay Ganesh, Tim Leek, and Martin Rinard. Taint-based Directed Whitebox Fuzzing. *Software Engineering, 2009. ICSE 2009. IEEE 31st International Conference on. 2009*, 2009.
- [15] Marius Herring. Concurrency made easy with Scala and Akka. <http://www.deadcoderising.com/concurrency-made-easy-with-scala-and-akka/>, March 2015.
- [16] Carl Hewitt, Peter Bishop, and Richard Steiger. A universal modular ACTOR formalism for artificial intelligence. 1973.
- [17] Matt Hillman. 15 minute guide to fuzzing. <https://www.mwrinfosecurity.com/our-thinking/15-minute-guide-to-fuzzing/>, August 2013.
- [18] Milena Vujošević Janičić. Programske paradigme: Funkcionalna paradigma. http://www.programskijezici.matf.bg.ac.rs/ppR/2018/predavanja/fp/funkcionalno_programiranje.pdf.
- [19] Milena Vujošević Janičić. Verifikacija softvera: Dinamička analiza softvera. http://www.programskijezici.matf.bg.ac.rs/vs/predavanja/02_testiranje/02_testiranje.pdf.
- [20] Milena Vujošević Janičić. *Automatsko generisanje i proveravanje uslova ispravnosti programa*. PhD thesis, Matematički fakultet, Univerzitet u Beogradu, 2003.
- [21] Dejan Lukan. PDF File Format: Basic Structure. <https://resources.infosecinstitute.com/pdf-file-format-basic-structure/>, May 2018.
- [22] Saša Malkov. Informacioni sistemi: Bezbednost.

- [23] Saša Malkov. Pojam funkcionalnih programskih jezika. <http://poincare.matf.bg.ac.rs/~smalkov/files/fp.r344.2017/public/predavanja/FP.cas.2017.01.Uvod.p2.pdf>.
- [24] Saša Malkov. Razvoj softvera: Konkurentnost.
- [25] Bertrand Meyer. Seven Principles of Software Testing. *Computer*, 41:99–101, 8 2008.
- [26] John C. Mitchell and Krzysztof Apt. *Concepts in Programming Languages*, chapter 4.2, pages 57–67. Cambridge University Press, 2003.
- [27] Cristóbal A. Navarro, Nancy Hitschfeld-Kahler, and Luis Mateu. A survey on parallel computing and its applications in data-parallel problems using gpu architectures. *Communications in Computational Physics*, page 285:329, February 2004.
- [28] Martin Odersky. Scala’s Prehistory. <https://www.scala-lang.org/old/node/239.html>, 2008.
- [29] Martin Odersky, Lex Spoon, and Bill Venners. *Programming in Scala*. Artima Press, Walnut Creek, California, 2 edition, 2011.
- [30] Martin Odersky, Lex Spoon, and Bill Venners. Concurrency programming in Scala. In *Programming in Scala*, chapter 9. Artima Press, Walnut Creek, California, 2016.
- [31] Marc HJ Romanycia and Francis Jeffrey Pelletier. What is a heuristic? *Computational Intelligence*, 1(1):47–58, 1985.
- [32] Ben Stopford. Shared Nothing v.s. Shared Disk Architectures: An Independent View. <http://www.benstopford.com/2009/11/24/understanding-the-shared-nothing-architecture/>, 11 2009.
- [33] Herb Sutter. The Free Lunch Is Over: A Fundamental Turn Toward Concurrency in Software. *Dr. Dobbs’s Journal, C/C++ Users Journal*, December 2004.
- [34] Michael Sutton, Adam Greene, and Pedram Amini. *Fuzzing: Brute Force Vulnerability Discovery*. Pearson Education, One Lake Street, Upper Saddle River, NJ 07458, 2007.

- [35] Ari Takanen, Jared DeMott, and Charlie Miller. *Fuzzing for Software Security Testing and Quality Assurance*. Artech House, 685 Canton Street Norwood, MA 02062, 2008.

Биографија аутора

Вук Стефановић Караџић (*Тришћ, 26. октобар/6. новембар 1787. — Беч, 7. фебруар 1864.*) био је српски филолог, реформатор српског језика, сакупљач народних умотворина и писац првог речника српског језика. Вук је најзначајнија личност српске књижевности прве половине XIX века. Стекао је и неколико почасних доктората. Учествовао је у Првом српском устанку као писар и чиновник у Неготинској крајини, а након слома устанка преселио се у Беч, 1813. године. Ту је упознао Јернеја Копитара, цензора словенских књига, на чији је подстицај кренуо у прикупљање српских народних песама, реформу ћирилице и борбу за увођење народног језика у српску књижевност. Вуковим реформама у српски језик је уведен фонетски правопис, а српски језик је потиснуо славеносрпски језик који је у то време био језик образованих људи. Тако се као најважније године Вукове реформе истичу 1818., 1836., 1839., 1847. и 1852.