Re-Allocation of Resources during Releases *

Md Tajmilur Rahman
Concordia University
Montreal, QC H3G 1M8
438-932-2288, +1
mdt_rahm@encs.concordia.ca

Peter C. Rigby
Concordia University
Montreal, QC H3G 1M8
514-848-2424, +1
peter.rigby@concordia.ca

ABSTRACT
This section will be written at the end.

Categories and Subject Descriptors
K.6.3 [**Software Management**]: Software Development; Software Resource Management, Resource Reallocation

General Terms
Experiment, Human Factors, Resource Management, Reallocation

Keywords
Resource Reallocation, Software Releases

1. INTRODUCTION
Software projects are notorious for going over budget and schedule. Rush periods are often get seen before a major release that turn the developers into dinosaurs as Frederick Brooks likens in his benchmark study "The Mythical Man Month" [4]. This "Rush To Release (RTR)" can be prompted either by external forces such as decisions by management to include new features in the release or to release earlier to beat a competitor. Alternatively, the rush may simply be due to inappropriate or unrealistic scheduling. Whatever the reason is it is an obvious. Regardless of the causes, the rush to release stresses developers and often requires developers to work on unusual, high priority or critical areas of the system. In this paper we study how RTR effects project organization and introduces technical debt. The key research questions that we expect to answer with our methodology are as follows:

1. Do developers work on different areas of the system around the time of release?
2. Are there certain areas of the system that receive increased attention (i.e. do developers focus on a smaller set of files around releases)?
3. Do the areas of code that are modified around the time of release have higher defect densities than code that is modified during normal development?

We observe a reallocation of the resources among the software development teams in a large project to identify that an improper reallocation or inappropriate reorganization causes a disruptive event take place in a software development process. We attempt to identify a projectâĂŹs different release times and calculate the difference between two consecutive releases to discover which are the new areas have been worked between two consecutive releases. The commit log data that we are working on for this purpose will help us to extract a lot of information like calculating the developers' working areas and time-frame of each release. This information will help us to identify the criteria of the resources, their roles, file ownership and nativeness (âĹĘÆđ) in the domain. This knowledge of nativeness (âĹĘÆđ) will guide us understanding the method of reallocation. Very few research works have been performed regarding the re-allocation of resources. Robert van Engelen worked for similar kind of a research to understand the resource allocation dynamics across the software projects [8]. He mainly tried to reallocate development resources amongst projects for increasing the satisfactory level of consumer or customer while we are focusing on the impact on code-base like the complexity of script files. Robert proposed a project-entropy metric in his work to understand if there is any limit for a particular reallocation does not lead to user satisfaction. Here entropy is to represent disorder and chaos to understand degradation of software and its inherent complexity. In his work resources may not just be the developers but also can be any other resources necessary for a software project development. We have organized this paper as follows. In Section 2, we describe some background and motivations followed by some summaries of related works in section 3. Section 4 will describe about the ownership of files and ownership of a set of files or a directories. We will try to understand how native a code-base is to a developer or a development team. In Section 5, some analysis to determine reallocation has been performed in a release or where reallocation needs to be performed will be presented. What changes in nativeness (âĹĘÆđ) occurs after the reallocation. Section 6 will give us the result to show how change in âĹĘÆđ puts impact on

*Copyright note

the outcome of a software. Finally section 7 will give us an idea of our future work followed by the conclusion in section 8.

2. BACKGROUND

Sometimes project members with interdependent tasks usually may not communicate effectively; coordination breakdowns occur, which results in integration failures [12]. There may have lower developers productivity [13, 5] which may cause inefficient run in the rush moments in a release period. There is a substantial and important body of literature on risk in software engineering. Boehm identified the most important risks encountered by software project managers and described successful risk management practices [3, 14, 2]. Some of the risks identified are related to disruptive events, such as the introduction of a new technology, but most are macro risks associated with running a project, such as developing the wrong functionality. General risk mitigation strategies can be difficult to apply to specific disruptive events. There may be various kinds of disruptive events for example, as a release approaches; developers take shortcuts that introduce technical debt. If it is not repaired, the long term quality of the system will suffer. Another example can be placed, if a lead developer who owns an important part of the code-base leaves and if steps to train other developers were not taken, it will become a dead area of the system and will be difficult to modify and maintain. Also often management reorganizes the developers on a companyâĂŹs projects, with the result that developers move to code-bases for which they have less experience. The reorganization introduces new perspectives and expertise that can lead to innovation; however, it can also result in a drop in productivity and the unnecessary re-writing of large portions of the system that the new developers do not understand. In this paper, we plan to take the measures on this last example among them mentioned above. We want to study that proper re-organization or efficient re-allocation of resources based on meaningful criteria can bring better outcome of a software development project by identifying quantify outcomes (Example: number of defects found).

3. RELATED WORKS

Many people have worked with pretty relevant ideas but I didn't find many very similar to our motivation. Hindle worked on release pattern discovery via partitioning [1]. In this research they proposed a method of observing, analyzing and summarizing the results of metrics of revisions found near releases. They have characterized a project's behavior around the time of major and minor releases. This is done by partitioning the observed activities like the art-effect checkins around the dates of major and minor releases, then look for reasonable patterns. Hindle divided the revisions in each release in 4 different classes, Source Code, Testing, Building, and Documentations. Actually this paper worked in a reverse way than Cook did [11]. Cook inserted sensors and monitors into the development process but Hindle and Michael analyzed the data to understand what happened in the past. Another research work we would like to mention was done by Damian where they have worked on the role of domain knowledge and cross functional communication among the OOS development teams [6]. Posnett did some dual ecological measures of focus in software development [7]. Posnett's measure was for the more general view that unifies developer focus and artifact ownership. Posnett analyzed i) developer artifact contribution to network to a predator-prey food web ii) drew upon ideas from ehology to produce a novel and iii) conceptually unified view of measuring focus and ownership. Another study was done by F. Rahman about the authorship of the code-bases in OSS development [9].

4. METHODOLOGY AND DATA

This section presents our methodology for discovering information which can give us the idea to get the answers to our research questions. We have collected the development history data of Linux kernel. Actually it is a database containing the all the commit log records by the Linux kernel developers since 2005. We are going to present the steps involved in this process and then we will follow up with an application of our methodology in a case study. Our methodology can be summarized as: Extracting Data for revisions and releases (Section 4.1); Partitioning the version numbers (Section 4.2); Get time-span between each release (Section 4.3); Calculate developer areas (Section 4.4); Finding code area owners (Section 4.5). Finding merging time and development time within a release period that we found in section 4.1 (Section 4.6).

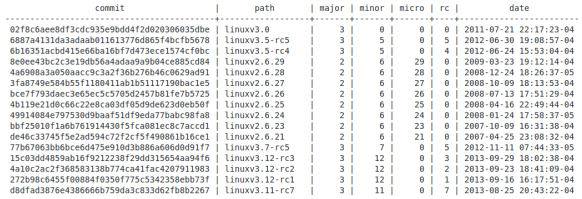
4.1 Extracting Data

We went for the VCS of a target project and either mirror the repository or download every revision and commit log history data. From DVCSs such as Git we extract the revisions and release information. We wrote Perl scripts to extract date that was further processed to obtain details. Manual inspection was used to resolve problems and things like that in cases where all automated techniques failed. We then put them into a database. We have used PSQL database to create tables in, to store our extracted data. These extracted will be analyzed by us later on. Per each revision the information extracted includes the commit id, tree id, author of the revision, date of revision, the name of the revised file, parent and child info for the revision and the detail log information. Once extraction is complete we are ready to partition the version numbers (Section 4.2) and duration of each release (Section 4.3).

4.2 Partitioning Release Numbers

We stored the git commit log extracted data into a table named git commit and git revision where all the basic and log information for a particular commit was mentioned in the first table and second one containing which commit belongs to which version of Linux kernel development and change details like path modified, new path created due to the change, how many addition and how many deletion occurred in a particular commit etc. By joining these tables we can easily get the dates of each version and from the version number which is a combination of different types of releases we can get determine which commit belongs to which version and release, and also what type of release that is as well. Figure 1 shows the picture of the data how it looks like.

```
Find out the release candidates:
update git_refs_tags set rc = cast(
  substring(
    path from
      position( '-rc' in path )
```



commit	path	major	minor	micro	rc	date
02f8c6aee8df3cdc935e9bdd4f2d020306035dbe	linuxv3.0	3	0	0	0	2011-07-21 22:17:23-04
6887a4131da3adaab01161377dd865f4bcfb5678	linuxv3.5-rc5	3	5	0	5	2012-06-30 19:08:57-04
6b16351acbd415e66ba16bf7d473ece1574cf0bc	linuxv3.5-rc4	3	5	0	4	2012-06-24 15:53:04-04
8e0ee43bc2c3e19db56a4adaa9a9b04ce885cd84	linuxv2.6.29	2	6	29	0	2009-03-23 19:12:14-04
4a6908a3a050aacc9c3a2f36b276b46c0629ad91	linuxv2.6.28	2	6	28	0	2008-12-24 18:26:37-05
3fa8749e584b55f1180411ab1b51117190bac1e5	linuxv2.6.27	2	6	27	0	2008-10-09 18:13:53-04
bcef7793deec3e05ec5c5705d2457b81fe7b5725	linuxv2.6.26	2	6	26	0	2008-07-13 17:51:29-04
4b119e21d0c66c22e8ca03df05d9de623d0eb50f	linuxv2.6.25	2	6	25	0	2008-04-16 22:49:44-04
49914084e797530d9baaf51df9eda77babc98fa8	linuxv2.6.24	2	6	24	0	2008-01-24 17:58:37-05
bbf25010f1a6b761914430f5fca081ec8c7accd1	linuxv2.6.23	2	6	23	0	2007-10-09 16:31:38-04
de46c33745f5e2ad594c72f2cf5f490806b1b16ce1	linuxv2.6.21	2	6	21	0	2007-04-25 23:08:32-04
77b67063bb6bce6d475e910d3b886a606d0d91f7	linuxv3.7-rc5	3	7	0	5	2012-11-11 07:44:33-05
15c03dd4859ab16f9212238f29dd315654aa94f6	linuxv3.12-rc3	3	12	0	3	2013-09-29 18:02:38-04
4a10c2ac2f368583138b774ca41fac4207911983	linuxv3.12-rc2	3	12	0	2	2013-09-23 18:41:09-04
272b98c6455f00884f0350f775c5342358ebb73f	linuxv3.12-rc1	3	12	0	1	2013-09-16 16:17:51-04
d8dfad3876e4386666b759da3c833d62fb8b2267	linuxv3.11-rc7	3	11	0	7	2013-08-25 20:43:22-04

Figure 1: Releases extracted from Version

```
        + 3 for 2
  ) as integer
) where path ~ E'-rc';
```

Find out major release version number:

```
update git_refs_tags set major = cast(
  substring(
    path, 'v([0-9]+)\.?'
      ) as integer
);
```

Find out minor release version number:

```
update git_refs_tags set minor = cast(
  substring(
    path, 'v[0-9]+\.([0-9]+)\.?'
      ) as integer
);
```

Find out micro release version number:

```
update git_refs_tags set micro = cast(
  substring(
    path, 'v[0-9]+\.[0-9]+\.([0-9]+)\-?'
      ) as integer
);
```

We are getting the release dates also with every commit record. So we can understand which commit belongs to which version of release, is this a major release or minor or micro. Another information we have captured that is **rc** which means that the particular commit was for a release candidate.

4.2.1 Explanation

A Linux kernel development release version having maximum length of information looks like **linuxvA.B.C-rcP** where A, B, C, P are numeric. If a release versioning looks like "linuxv2.6.13" it tells us that this particular release is the 13th minor release under the 6th major release of kernel version 2. "linuxv2.6.13**-rc1**" says that after the 13th minor release been released development for the next release has been started and rc1 is the first candidate on the way to the next release. It doesn't ensure that next release is also going to be another minor release.

4.3 Get Time-Span Between Release

Another information that we require is what are the durations of the releases of Linux kernel development. To achieve that we joined the table in figure 1 and the table "git_commit". The following psql did the job pretty easily for us:

```
select
grt.path as release, grt.major, grt.minor,
grt.micro, grt.rc, gc.committer_dt
from
git_refs_tags grt, git_commit gc
where
gc.commit = grt.commit
order by
gc.committer_dt
```

This query gives us all the dates of commit for each and every release, so we can easily find out the duration between two consecutive releases (all including release candidates RC). This information is more important to us because we need to see what is the actual development period and what is the merging or other fixing period within a release time-span and also we need to understand how developers are working, what is the impact of their changes made in the code-base during a particular release period. Table 1 shows a small portion of this information. Preior to calculate the areas of the developers in secion 4.4 we require this information.

Table 1: Different Releases

Release	Type	Start Date	End Date
linuxv2.6.12-rc2	rc	-	2005-04-16
linuxv2.6.12-rc3	rc	2005-04-16	2005-04-20
linuxv2.6.12-rc4	rc	2005-04-20	2005-05-07
linuxv2.6.12-rc5	rc	2005-05-07	2005-05-24
linuxv2.6.12-rc6	rc	2005-05-24	2005-06-06
linuxv2.6.12	micro	2005-06-06	2005-06-17
linuxv2.6.13-rc1	rc	2005-06-17	2005-06-29
linuxv2.6.13-rc2	rc	2005-06-29	2005-07-05
linuxv2.6.13-rc3	rc	2005-07-05	2005-07-13
linuxv2.6.13-rc4	rc	2005-07-13	2005-07-28
linuxv2.6.13-rc5	rc	2005-07-28	2005-08-02
linuxv2.6.13-rc6	rc	2005-08-02	2005-08-07
linuxv2.6.13-rc7	rc	2005-08-07	2005-08-23
linuxv2.6.13	micro	2005-08-23	2005-08-28
linuxv2.6.14-rc1	rc	2005-08-28	2005-09-12
linuxv2.6.14-rc2	rc	2005-09-12	2005-09-19
linuxv2.6.14-rc3	rc	2005-09-19	2005-09-30
linuxv2.6.14-rc4	rc	2005-09-30	2005-10-10
linuxv2.6.14-rc5	rc	2005-10-10	2005-10-20

Note that the very first row has no information for the "Start Date" because we don't have any data prior to release "linuxv2.6.12". From the extraction we found complete records for releases "linuxv2.6.13" to "linuxv2.6.39" then "linuxv3.0" to "linuxv3.11". We have last couple of release candidates and the release date for "linuxv2.6.12" and first couple of release candidates for "linuxv3.12" which is not useful for our calculation. So we are having 340 releases including major, minor, micro and rc; 40 complete releases from "linuxv2.6.12" to "linuxv3.11". We have total 400441 distinct commits and for 372943 of them have change information for 77082 different files in our hand where we have 14599 developers' working history among 14621 distinct developers.

4.4 Calculate Developers' Area

Developers are working throughout the release time. Developers' Are (DA) means the files developers work in. It doesn't refer the to file itself but it is important to understand which files are being touched by which developers. We see that authors are committing files almost everyday throughout a release. We have the commit log data in our hand and we know what is the duration for each and every releases (i.e. start and end dates). Now it is possible to find out which commits were made for which files by a developer within a given time range of a release.

4.4.1 DA in a Release Period

We select all the release information from the table where we stored the releases with their corresponding time-span. After that we select author name and other relevant information from the joining of "git_revision" and "git_commits" tables for every release. We insert these information into another table so that we can use the data for our further analysis.

The following pseudo code shows us the way of achieving the desired result in this section:

Algorithm 1 Calculate Developers Area

Require:
commit_date $\geq$ *release_start_date*
commit_date $<$ *release_end_date*
Ensure:
getallreleases
while *prev_release_date* $\leftarrow$ *release_start_date and release* $\leftarrow$ *release_version* **do**
  if *release* **then**
    if *commit_date* $\geq$ *release_start_date* and *commit_date* $<$ *release_end_date* **then**
      selectauthor, path, commits, churn
      insertdataintotabledev_area_rel
    end if
  end if
end while

A relatively straightforward discipline of Linux Kernel Development team is followed with regard to the merging of patches for each release [10]. At the beginning of each development cycle, the "merge window" is said to be opened. At that time, code which is deemed to be sufficiently stable (and which is accepted by the development community) is merged into the mainline kernel. The bulk of changes for a new development cycle (and all of the major changes) will be merged during this time, at a rate approaching 1,000 changes ("patches," or "changesets") per day. Strategically linux starts a new kernel release with the merging and fixing to get their development branch ready to start development for the next release. So there are two main segments in a release period, one is merge window or merge period (MP) and another is development window what we are going to call Release Development Period (RDP).

Similarly we have calculated DA in both MP and RDP as described in the following subsections.

4.4.2 DA in a Merge Period

Merge Perid can be determined from the previous release (any of major, minor or micro releases) date to the first RC date. For example if the date of release for "linuxv2.6.12" is 2005-06-17 and date of pushing the first RC for the next release "linuxv2.6.13-rc1" is 2005-06-29 then this time period of of 12 days is being called the merge window [10] or MP. Table 2 shows some merging periods of different releases.

Table 2: Release Merging Periods

Release	Start Date	RC Date
linuxv2.6.13	2005-06-17	2005-06-29
linuxv2.6.14	2005-08-28	2005-09-12
linuxv2.6.15	2005-10-27	2005-11-11
linuxv2.6.16	2006-01-02	2006-01-17
linuxv2.6.17	2006-03-20	2006-04-02
linuxv2.6.18	2006-06-17	2006-07-06
linuxv2.6.19	2006-09-19	2006-10-04
linuxv2.6.20	2006-11-29	2006-12-13
linuxv2.6.21	2007-02-04	2007-02-20
linuxv2.6.22	2007-04-25	2007-05-12
linuxv2.6.23	2007-07-08	2007-07-22
linuxv2.6.24	2007-10-09	2007-10-23
linuxv2.6.25	2008-01-24	2008-02-10

We understand DA or Developers' Area is the files in the code-base, developers are working from the dataset that we have in our hand. We have the commit logs and every record says about who committed the file when did he/she worked on the file. If we run an SQL query then we can easily find out for every author and for every particular file how many times it's been committed by an author and what changes (here, addition + deletion = churn) he/she has been made. So we are calculating the DA and storing the information into a table. A part of that table is shown in table 3.

Table 3: DA in Merge Period

Author	File Path	Commits	Churn
D. S. Miller	include/.../pci.h	1	8
V. Hanquez	arch/.../cpu.c	1	14
A. Bunk	drivers/.../shmem.c	1	2
J. Juhl	arch/.../generic.c	1	3

In this table we see in a merging period inside a release period D. S. Miller has made change in pci.h file only once wich the churn number 8. In the similar way we calculate DA in RDP and RTR period as well.

4.4.3 DA in a RD Period and RTR

As like MP we can extract the RDP from the. RDP here in Linux Kernel development process is being considered as from the release date of RC1 (i.e. first release candidate opening the gate for the development of the next release) to the next release (major, minor or micro) date. For example if the date of release for "linuxv2.6.13-rc1" after releasing "linuxv2.6.12" is 2005-06-29 and date of publishing the next release "linuxv2.6.13" is 2005-08-28 then this time period of of 60 days is being called the RDP. Table 4 is showing some development periods of different releases.

Table 4: Release Development Periods

Release	RD Start Date	Release Date
linuxv2.6.13	2005-06-29	2005-08-28
linuxv2.6.14	2005-09-12	2005-10-27
linuxv2.6.15	2005-11-11	2006-01-02
linuxv2.6.16	2006-01-17	2006-03-20
linuxv2.6.17	2006-04-02	2006-06-17
linuxv2.6.18	2006-07-06	2006-09-19
linuxv2.6.19	2006-10-04	2006-11-29
linuxv2.6.20	2006-12-13	2007-02-04
linuxv2.6.21	2007-02-20	2007-04-25
linuxv2.6.22	2007-05-12	2007-07-08
linuxv2.6.23	2007-07-22	2007-10-09
linuxv2.6.24	2007-10-23	2008-01-24
linuxv2.6.25	2008-02-10	2008-04-16

So along these RDs of each release period developers get involved in the development for the next release. As we mentioned earlier RTR may happen as the release approaches. Keeping this in our aim list we noticed an important slot in these RDs of releases. We see in between two consecutive releases there is a merging period of time, after that development starts by pushing RCs and once RC releases stops and it takes some days of time for the preperation to publish the release. We are calling this last part as RTR Period because we asume RTR may happen in this period of time before release. Table 5 represents some releases with their corresponding RTR (i.e. from end date to rc and date of release).

Table 5: RTR Periods

Release	RC End Date	Release Date	Days
linuxv2.6.13	2005-06-29	2005-08-28	5
linuxv2.6.14	2005-09-12	2005-10-27	8
linuxv2.6.15	2005-11-11	2006-01-02	9
linuxv2.6.16	2006-01-17	2006-03-20	8
linuxv2.6.17	2006-04-02	2006-06-17	12
linuxv2.6.18	2006-07-06	2006-09-19	7
linuxv2.6.19	2006-10-04	2006-11-29	14
linuxv2.6.20	2006-12-13	2007-02-04	10
linuxv2.6.21	2007-02-20	2007-04-25	7
linuxv2.6.22	2007-05-12	2007-07-08	8
linuxv2.6.23	2007-07-22	2007-10-09	9
linuxv2.6.24	2007-10-23	2008-01-24	5
linuxv2.6.25	2008-02-10	2008-04-16	8

4.5 Finding Code-Area Owners

... ...

4.6 Jaccard Distance

Once we found different releases and the release periods as well we calculate the developers' are to see what changes developers are doing in releases. We also want to know what is the difference being created by making changes between two consecuitive releases. If I explain it more then I can say, developers are working in many different files in a release and may work in a same file again for the next release or releases.

5. CONCLUSIONS

This section will be written at the end.

6. ACKNOWLEDGMENTS

... ...

7. REFERENCES

[1] R. C. H. A. Hindle, M. W. Godfrey. Release pattern discovery via partitioning: Methodology and case study. In *Proceedings of the Fourth International Workshop on Mining Software Repositories at ICSE Workshops MSR '07*, volume 35, page 19, Minneapolis, MN, USA, 2007. MSR.

[2] B. Boehm. Get ready for agile methods, with care. *ACM Computer*, 35(1):64–69, 2002.

[3] B. W. Boehm. Software risk management: Principles and practices. *IEEE Softw.*, 8(1):32–41, January 1991.

[4] F. Brooks. *The Mythical Man-Month*. Addison-Wesley, 1975.

[5] J. S. Daniel Damian, L. Izquierdo. Awareness in the wild: Why communication breakdowns occur. pages 81–90, 2007.

[6] K. I. Daniela Damian, Helms Remko. The role of domain knowledge and cross-functional communication in socio-technical coordination. In *35th International Conference on Software Engineering, ICSE*, pages 442–451, 2013.

[7] P. D. Daryl Posnett, Raissa D'Souza. Dual ecological measures of focus in software development. In *Proceeding ICSE '13 Proceedings of the 2013 International Conference on Software Engineering*, pages 452–461, 2013.

[8] S. Datta. *Project-entropy: A Metric to Understand Resource Allocation Dynamics across Software Projects*. PhD thesis.

[9] P. D. F. Rahman. Ownership, experience and defects: a fine-grained study of authorship. In *Proceeding of the 33rd international conference on Software engineering, ACM*, pages 491–500, 2011.

[10] L. K. O. Inc. The linux kernel archives. https://www.kernel.org/doc/Documentation/development-process/2.Process.

[11] A. L. W. J. E. Cook. Automating process discovery through event-data analysis. In *Proceedings of the 17th international conference on Software engineering in ICSE âĂŹ95*, volume 35, pages 73–82, Seattle, Washington, USA, April 1995. ACM Press.

[12] I. Kwan. Does socio-technical congruence have an effect on software build success? a study of coordination in a software project. *IEEE Transactions on Software Engineering*, 37(3):307–324, May-June 2011.

[13] J. H. M. Cataldo, P. Wagstrom. Identification of coordination requirements: implications for the design of collaboration and awareness tools. *CSCW*, pages 353–362, May-June 2006.

[14] K. L. M. Keil, P. E. Cule. A framework for identifying software project risks. *Commun. ACM*, 41(11):76–83, November 1998.

APPENDIX

A. HEADINGS IN APPENDICES

Apendix section will be written later with the following subsections may be:

A.1 Introduction

...

A.2 Background and Motivation

A.2.1 Related Works

...

A.2.2 Methodology and Data

...

A.3 Conclusions

...

A.4 Acknowledgments

...

A.5 References

... ...