

Analyzing the Software Development Life-Cycle using Data-Mining Techniques

OpenTech

Andreas Platschek

<andreas.platschek@opentech.at>

February 3, 2017

- **Generic** qualification approach
- Suitable for up to **SIL2** (IEC 61508 Ed 2)
- Support **multi-core** systems
- **Mainline** Linux kernel + glibc + busybox + tools
- Methods suitable for **pre-existing SW**
- Targeting **SW intensive systems**

Assumptions

- There was a process in place
- The process was followed
- The discrepancies between the actual process and the objectives of IEC 61508 **can** be assessed
- Mitigation of procedural defects is possible

Assurance

- Qualification of involved people
- Structural aspects of organization
- Methods and techniques used
- Judgement of results in a quantitative manner

Route 3_S concept

- Selection and evaluation of divergence
- Assessment of processes used
- Assessment of consistency of results
- Quantification of residual risks

NOTE: This is a very much simplified description but for todays purposes this is good enough - for the full story look at the Route.pdf .

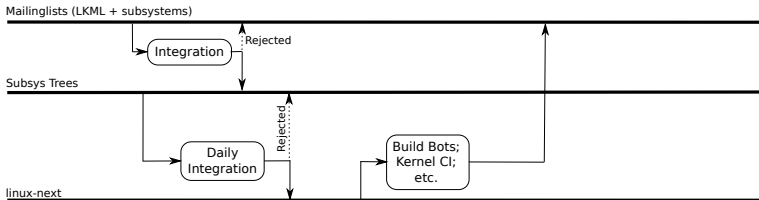
Documented in the git repository in **Documentation/process**.

Examples:

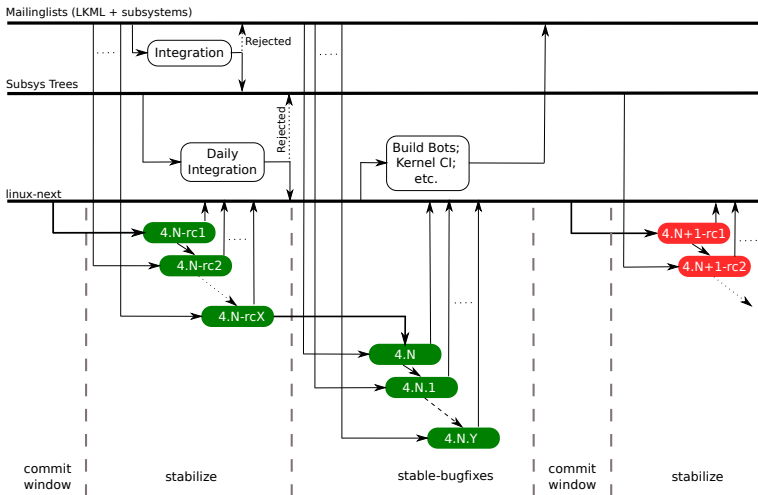
- Formatting of Patches (Subject Line, Body, Sign-Off, etc.).
- Usage of *-by Tags.
- How / What to test before sending in a patch..
- Where to send patches.
- ...

Linux Kernel

Continuous Integration



Linux Kernel Versions



git log – Header

```
commit 87dbf3dc165240f1a3bed1ac7243a6b73c474029
```

```
Author: Tony Lindgren <tony@atomide.com>
```

```
Date: Mon Nov 7 16:50:11 2016 -0700
```

```
ARM: OMAP4+: Fix bad fallthrough for cpuidle
```

```
commit cbf2642872333547b56b8c4d943f5ed04ac9a4ee upstream.
```

```
We don't want to fall through to a bunch of errors for retention  
if PM_OMAP4_CPU_OSWR_DISABLE is not configured for a SoC.
```

```
Fixes: 6099dd37c669 ("ARM: OMAP5 / DRA7: Enable CPU RET on suspend")
```

```
Acked-by: Santosh Shilimkar <ssantosh@kernel.org>
```

```
Signed-off-by: Tony Lindgren <tony@atomide.com>
```

```
Signed-off-by: Greg Kroah-Hartman <gregkh@linuxfoundation.org>
```

git log – Patch

```
diff --git a/arch/arm/mach-omap2/omap-mpuss-lowpower.c b/arch/arm/mach-omap2/om
index 94428b4..7d62ad4 100644
--- a/arch/arm/mach-omap2/omap-mpuss-lowpower.c
+++ b/arch/arm/mach-omap2/omap-mpuss-lowpower.c
@@ -245,10 +245,9 @@ int omap4_enter_lowpower(unsigned int cpu, unsigned int po
     save_state = 1;
     break;
     case PWRDM_POWER_RET:
-    if (IS_PM44XX_ERRATUM(PM_OMAP4_CPU_OSWR_DISABLE)) {
+    if (IS_PM44XX_ERRATUM(PM_OMAP4_CPU_OSWR_DISABLE))
+        save_state = 0;
-    break;
-}
+    break;
     default:
     /*
      * CPUx CSWR is invalid hardware state. Also CPUx OSWR
```

How to Handle Data?

- Distribute to team-members
- Keep data up-to-date
- Support exploratory analysis
- Clean data in one common place / way
- Keep data consistent for all team-members
- Eliminate processing overhead between different analysis scripts

Data Cleaning (1)

Example: Developer Names



Linus Grr.. Torvalds

Linus I'm a moron Torvalds

Linus OCD Torvalds

Linus oopsie Torvalds

Linus sniff Torvalds

Steven Mr. Procrastinator Rostedt

Steven Rostedt (Red Hat)

Steven The King of Nasty Macros! Rostedt

Currently we got **~1250** such cases identified. – Not all of them that informative, mostly differences in lower/upper case, typos like missed/double letters, different version of names that include e-mail address and/or affiliation.

Data Cleaning (2)

Example: (Sub)-Domains

Nr. (Sub)-Domains	Company
26	IBM
19	NEC
13	Linux Foundation
11	Sony
11	davemloft.net
9	linux.org.uk
8	SGI
6	Intel
6	linutronix
6	Samsung

Currently we got **~770** entries in our config file, for everything else we just stick with the domain from the e-mail addresses.

Data Cleaning (3)

Example: Fixes: tags



From: Documentation/process/submitting-patches.rst:

...use the 'Fixes:' tag with the first 12 characters of the SHA-1 ID, and the one line summary. For example:

```
Fixes: e21d2170f366 ("video: remove unnecessary platform_set_drvdata()")
```

Examples found in the wild:

Fixes: Bug 14662 - Dell E5500 kernel panic with KMS

Fixes: NB#106295 - prevent potential kernel crash in the MMC driver

Fixes: IRQ disabled (i915?) when switchig between gnome themes (gnome-theme-man

Fixes: v1.0

Fixes: http://bugzilla.kernel.org/show_bug.cgi?id=14925

Data Cleaning (3)

Example: Fixes: tags

Times Used	Domain
86	tracker.ceph.com
73	bugzilla.kernel.org
35	bugs.freedesktop.org
13	bugzilla.redhat.com
10	forums.grsecurity.net
9	bugs.elinux.org
6	lkml.kernel.org
5	bugzilla.linux-nfs.org
4	bugzilla.netfilter.org
3	lkml.org
3	bugzilla.novell.com
2	sourceforge.net
2	github.com
1	sourceware.org
1	linuxppc.10917.n7.nabble.com
1	git.linaro.org
1	bugs.gentoo.org

Hash Length

Nr. Occurences	Length
68	XXXX
27	XXXXXX
24	XXXXXXX
412	XXXXXXXX
255	XXXXXXXXX
263	XXXXXXXXXX
526	XXXXXXXXXXX
340	XXXXXXXXXXXX
19299	XXXXXXXXXXXXX $\Leftarrow$ 12 - the "proper" value!
1270	XXXXXXXXXXXXXX
215	XXXXXXXXXXXXXXX
163	XXXXXXXXXXXXXXX
252	XXXXXXXXXXXXXXX
46	XXXXXXXXXXXXXXX
13	XXXXXXXXXXXXXXX
26	XXXXXXXXXXXXXXX
13	XXXXXXXXXXXXXXX
5	XXXXXXXXXXXXXXX
10	XXXXXXXXXXXXXXX
11	XXXXXXXXXXXXXXX
2	XXXXXXXXXXXXXXX
3	XXXXXXXXXXXXXXX
2	XXXXXXXXXXXXXXX
5	XXXXXXXXXXXXXXX
2	XXXXXXXXXXXXXXX
752	XXXXXXXXXXXXXXX

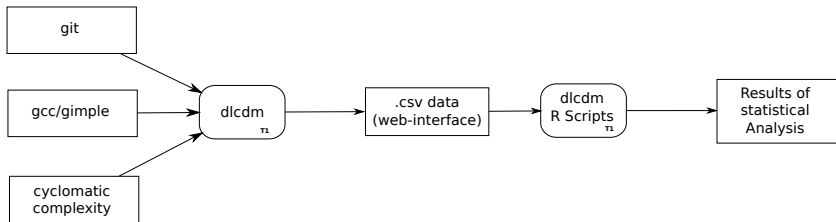
Percentages

Total of Fixes tags: 24349

Fixes tags that could not be resolved: 744 = **3.05%**

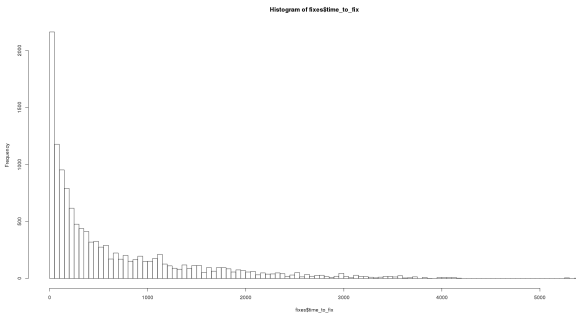
Out of those 254 were URLs and 18 CVEs.

- Distribute to team-members
- Webinterface for browsing data / Overview
- csv download for use in analysis
- Automatically updated
- Extended as needed



Simple R Example

```
> fixes <- read.csv("http://192.168.1.53/commitdm/linux-stable/fixes_data.csv")  
> mean(fixes$time_to_fix)  
[1] 641.1325  
> min(fixes$time_to_fix)  
[1] 0  
> max(fixes$time_to_fix)  
[1] 5383  
> hist(fixes$time_to_fix, 100)
```



Simple Python Example



```
HOST='81.217.60.26'
VERSION_BASE="v4.4"
MAX_DOT = 45

tag_list = []
tag_list.append(VERSION_BASE)
for dot in range(1, MAX_DOT):
    tag_list.append("{0}.{1}".format(VERSION_BASE, dot))

v_data = pd.DataFrame()
v_data['from'] = tag_list[:-1]
v_data['to'] = tag_list[1:]

for i in v_data.index:
    dl_url = "http://{0}/commitdm/linux-stable/{1}/{2}/fixes_data.csv".format(HOST,
                                                                              v_data.ix[i]['from'].replace(".", "_"),
                                                                              v_data.ix[i]['to'].replace(".", "_"))

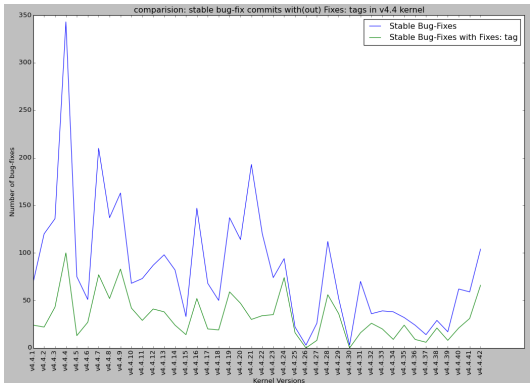
    try:
        read = pd.io.parsers.read_csv(dl_url)
        v_data.set_value(i, 'N_Fixes', len(read.index))
    except:
        v_data.set_value(i, 'N_Fixes', 0)

    dl_url = "http://{0}/commitdm/linux-stable/{1}/{2}/commit_data.csv".format(HOST,
                                                                              v_data.ix[i]['from'].replace(".", "_"),
                                                                              v_data.ix[i]['to'].replace(".", "_"))

    read = pd.io.parsers.read_csv(dl_url)
    v_data.set_value(i, 'N', len(read.index))
```

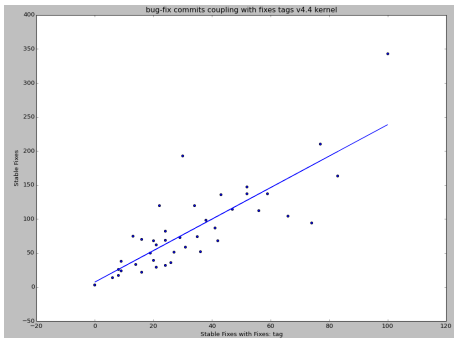
-stable Bug-Fixes Correlation

```
plt.plot(v_data.index, v_data['N'])  
plt.plot(v_data.index, v_data['N_Fixes'])  
plt.xticks(v_data.index, v_data['to'], rotation='vertical')  
plt.title('comparison: stable bug-fix commits with(out) Fixes: tags in {}'.format(VERSION_BASE))  
plt.ylabel("Number of bug-fixes")  
plt.xlabel("Kernel Versions")  
plt.legend(["Stable Bug-Fixes", "Stable Bug-Fixes with Fixes: tag"])  
plt.show()
```

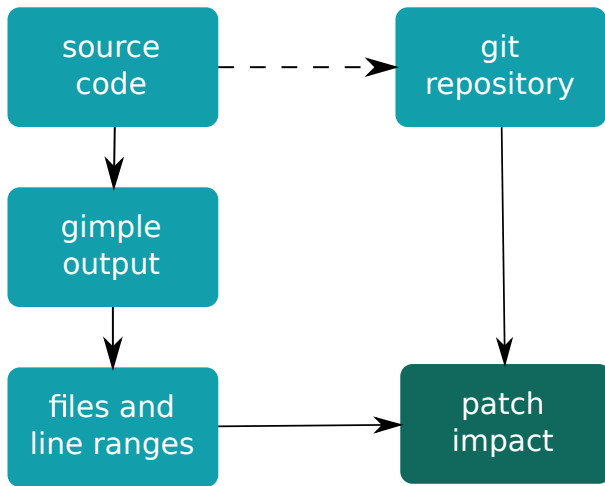


Bug-Fixes: Coupling

```
gradient, intercept, r_value, p_value, std_err = stats.linregress(v_data['N_Fixes'], v_data['N'])  
  
abline_values = [gradient * i + intercept for i in v_data['N_Fixes']]  
  
plt.scatter(v_data['N_Fixes'], v_data['N'])  
plt.plot(v_data['N_Fixes'], abline_values, 'b')  
plt.title('bug-fix commits coupling with fixes tags {} kernel'.format(VERSION_BASE))  
plt.ylabel("Stable Fixes")  
plt.xlabel("Stable Fixes with Fixes: tag")  
plt.show()
```



Patch Impact Analysis



Questions?

Ask now, or e-mail me later!

Andreas Platschek
<*andreas.platschek@opentech.at*>