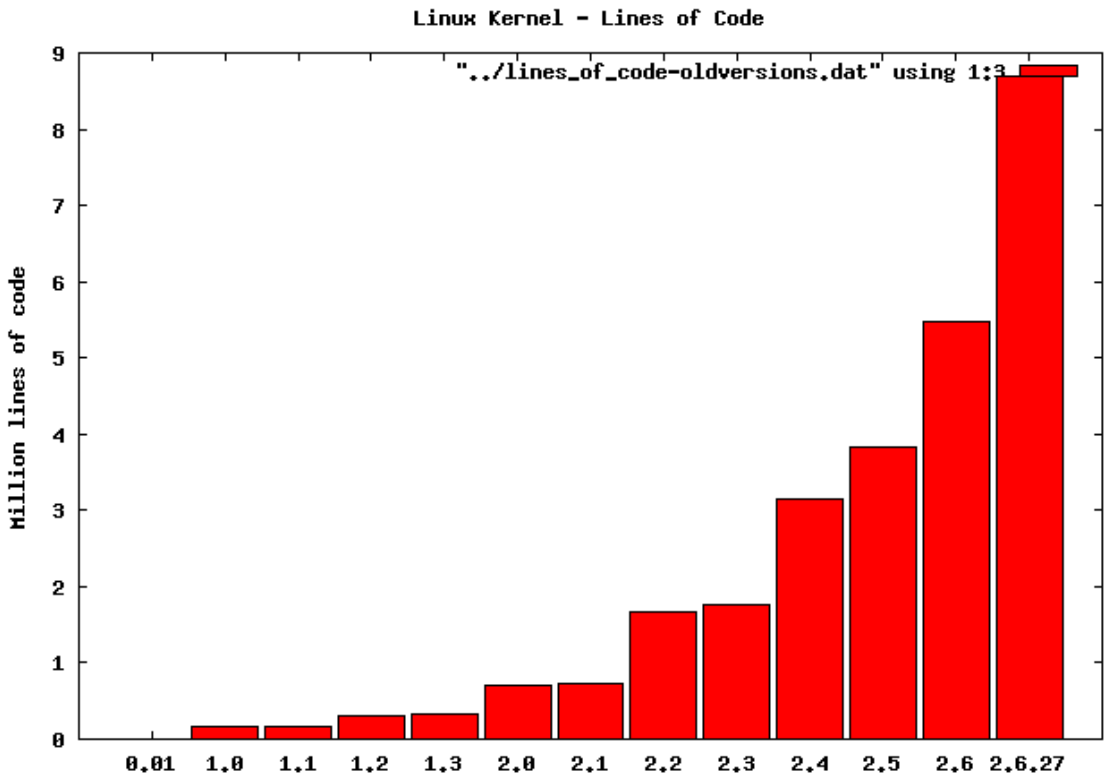
Stable Kernel的社区运作机制

Li Zefan <lizefan@huawei.com>

Linux内核代码规模

- 0.01：91年，~1万
- 1.0：94年，~17万
- 2.0：96年，~71万
- 2.2：99年，180万
- 2.4：01年，337万
- 2.6：03年，593万
- 3.0：11年，1465万
- 3.10：13年，1696万
- 4.0：15年，1930万
- …



Linux内核开发周期（1）

Kernel Release	Version Date	Days of development
3.11	2013-09-02	64
3.12	2013-11-03	62
3.13	2014-01-19	77
3.14	2014-03-30	70
3.15	2014-06-08	70
3.16	2014-08-03	56
3.17	2014-10-05	63
3.18	2014-12-07	63

Kernel Version	Changes (patches)
3.11	10,893
3.12	10,927
3.13	12,127
3.14	12,311
3.15	13,722
3.16	12,804
3.17	12,354
3.18	11,379

Kernel Release	Developers	Companies
3.11	1,266	225
3.12	1,332	244
3.13	1,361	228
3.14	1,446	240
3.15	1,492	237
3.16	1,477	234
3.17	1,433	241
3.18	1,458	239

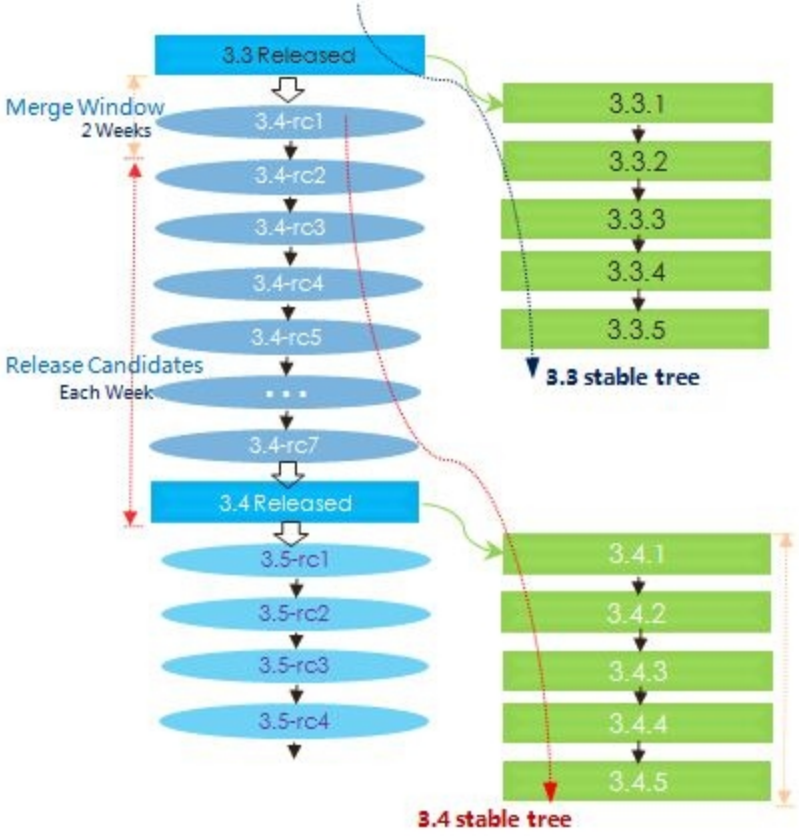
Linux内核开发周期（2）

- Linux内核开发如此活跃，如果保证内核可商用？

什么是stable tree

- Stable tree
  - 基于内核正式版本，只合入bug fix，以及device ID和quirk
  - 除了LTS kernel，其他stable kernel只维护约三个月

- LTS
  - Long-Term Support
  - 由Greg维护两年
  - 两年之后呢？



当前的LTS内核

Version	Maintainer	Released	Projected EOL
4.1	Greg Kroah-Hartman	2015-06-21	Sep, 2017
3.18	Sasha Levin	2014-12-07	Jan, 2017
3.14	Greg Kroah-Hartman	2014-03-30	Aug, 2016
3.12	Jiri Slaby	2013-11-03	2016
3.10	Greg Kroah-Hartman	2013-06-30	Sep, 2015
3.4	Li Zefan	2012-05-20	Sep, 2016
3.2	Ben Hutchings	2012-01-04	2016
2.6.32	Willy Tarreau	2009-12-03	Mid-2015

Fix从何而来？

- 一般来说，bug fix必须来自内核主线

- 具体过程
  - 主动合入：自动（部分手动）分析Linus git tree中的commits
  - 被动合入：用户/开发人员发邮件到stable邮件列表，要求将某个fix合入stable kernel

主动合入bug-fix

1. 主线开发过程中，发现的bug，其中有不少存在于以前的内核版本中

2. 作者/Maintainer在bug fix的changelog中加上Cc stable的标签

3. Stable tree maintainers使用脚本找到所有这类bug fix，合入stable tree

```
author      Masahiro Yamada <yamada.masahiro@socionext.com>   2015-07-15 01:29:00 (GMT)
committer   Greg Kroah-Hartman <gregkh@linuxfoundation.org>   2015-08-05 22:11:48 (GMT)
commit      64526370d11ce8868ca495723d595b61e8697fbf (patch)
tree        6fa4e4c5329cd05b5d8c62b42fe5a540ea779b34
parent      cbfe8fa6cd672011c755c3cd85c9ffd4e2d10a6f (diff)

devres: fix devres_get()

Currently, devres_get() passes devres_free() the pointer to devres,
but devres_free() should be given with the pointer to resource data.

Fixes: 9ac7849e35f7 ("devres: device resource management")
Signed-off-by: Masahiro Yamada <yamada.masahiro@socionext.com>
Acked-by: Tejun Heo <tj@kernel.org>
Cc: stable <stable@vger.kernel.org> # 2.6.21+
Signed-off-by: Greg Kroah-Hartman <gregkh@linuxfoundation.org>

Diffstat
-rw-r--r-- drivers/base/devres.c 4

1 files changed, 2 insertions, 2 deletions

diff --git a/drivers/base/devres.c b/drivers/base/devres.c
index c8a53d1..8754646 100644
--- a/drivers/base/devres.c
+++ b/drivers/base/devres.c
@@ -297,10 +297,10 @@ void * devres_get(struct device *dev, void *new_res,
        if (!dr) {
                add_dr(dev, &new_dr->node);
                dr = new_dr;
-               new_dr = NULL;
+               new_res = NULL;
        }
        spin_unlock_irqrestore(&dev->devres_lock, flags);
-       devres_free(new_dr);
+       devres_free(new_res);

        return dr->data;
 }
```

被动合入bug-fix

发件人 Wang Long

主题 **[request for stable inclusion 3.10 and 3.12] Fix CVE-2014-8173**

收件人 Greg Kroah-Hartman, Jiri Slaby

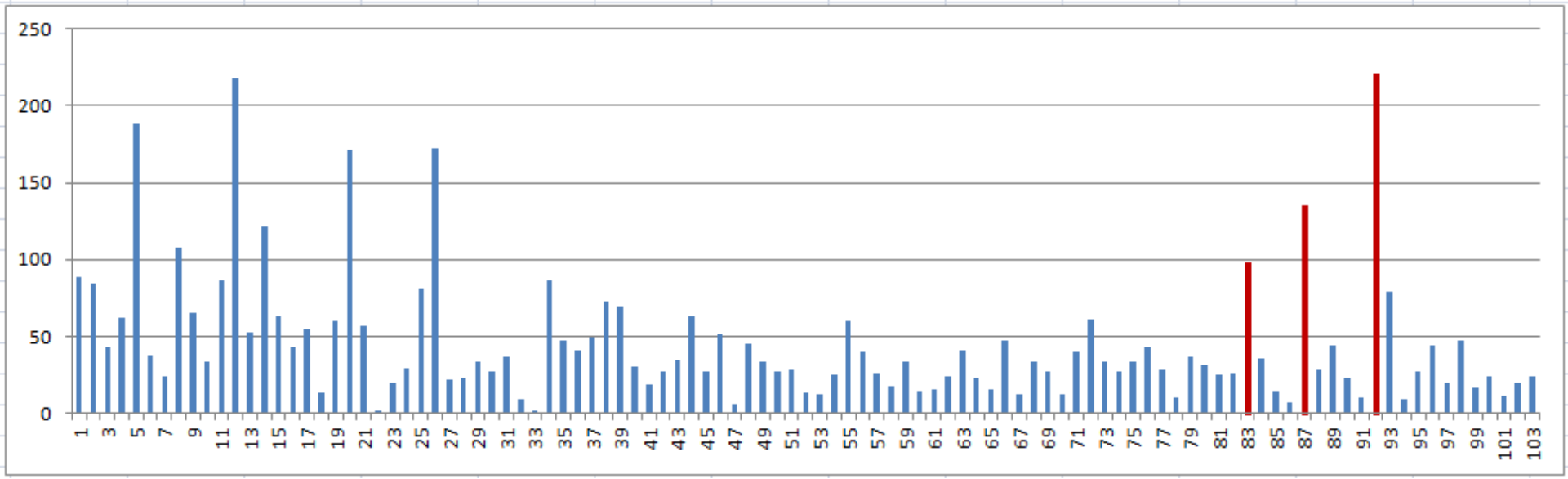
抄送 LKML, stable, Wang Long, peifeiyue 00238447, Sasha Levin, aarcange@redhat.com

```
Hi Greg and Jiri,

The following patch commit ee53664bda169f519ce3c6a22d378f0b946c8178

mm: Fix NULL pointer dereference in madvise(MADV_WILLNEED) support

fix CVE-2014-8173. I wish you could merge this fix into stable 3.10
and 3.12, because the linux kernel before 3.13 on NUMA systems
is affected by it.


Kirill A. Shutemov (1):
  mm: Fix NULL pointer dereference in madvise(MADV_WILLNEED) support

 include/asm-generic/pgtable.h | 5 ++---
 1 file changed, 2 insertions(+), 3 deletions(-)
```

LTS版本发布周期

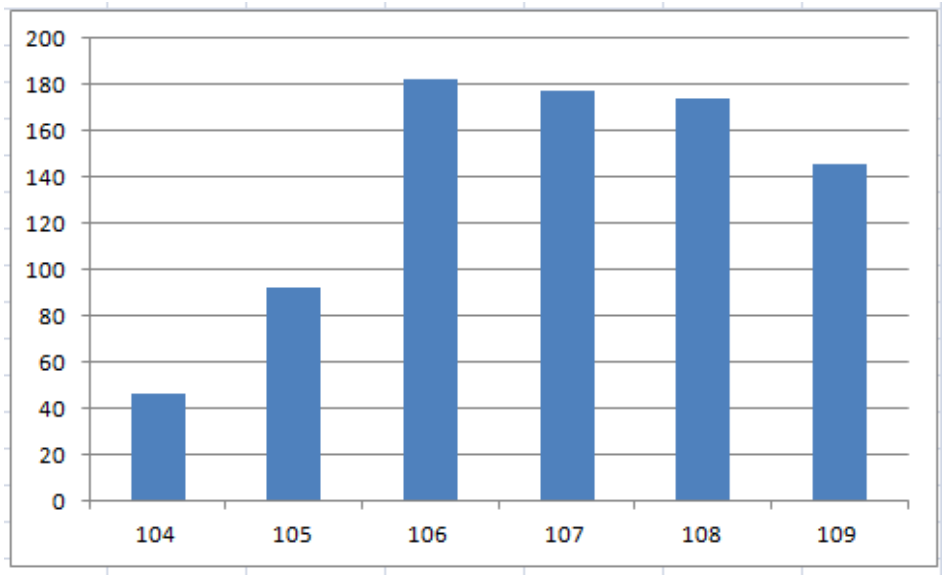
- 一般1~2个星期发布一个新版本

- 以LTS 3.4为例
  - 直到3.4.109，一共合入5000+的patch
  - 合入的patch数，随着时间逐渐减少
  - 为什么会出现红色的spike？



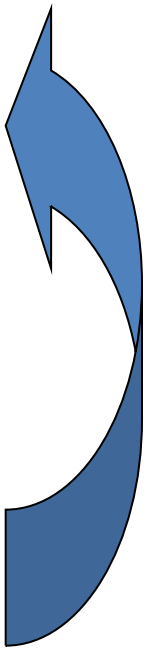
LTS版本发布周期（2）

- 3.4.104开始由我接手，目前2~3个月发布一个新版本



Stable kernel的开发、发布过程

- 1. 持续通过自动化脚本主动合入主线的commits
  - 1. 分析主线git-log的commits，过滤出所有有”cc stable”标签的commits
  - 2. 用”quilt import“命令导入以上的commits
  - 3. 用“quilt push“合入patch，用”quilt delete”删除无法合入的patch
- 2. 持续被动合入主线的commits
- 3. 发布RC版
  - 各个patch抄送相应的作者，由作者review
  - Guenter等人做编译测试、启动测试
- 4. 两天后发布正式版
  - 由吴峰光的LKP做测试



Demo

Stable kernel的不足、问题

- RC版发布后，review不是强制性，很多作者并没有review
- 吴峰光的LKP是在stable kernel版本发布后做测试，而不是在发布前

- 更大的问题：很多bug fix没有被合入到stable tree中。
  - Bug fix的changelog缺少stable tag
  - 很多有stable stag的fix也没有被合入到stable tree

Stable tag的缺失（1）

- 以前我从来不在changelog里加stable tag…
- 如果developer没加tag，maintainer会加上
- 但maintainer也不总是有这个意识…

从 Li Zefan
主题 **[PATCH] tracing: Fix preempt count leak**
至 Steven Rostedt
Cc Frederic Weisbecker, Jiri Olsa, LKML, Hiroyuki KAMEZAWA

```
While running my ftrace stress test, this showed up:

BUG: sleeping function called from invalid context at mm/mmap.c:233
...
note: cat[3293] exited with preempt_count 1

The bug was introduced by commit 91e86e560d0b3
("tracing: Fix recursive user stack trace")

Signed-off-by: Li Zefan <lizf@cn.fujitsu.com>
---
 kernel/trace/trace.c |    6 ++----
 1 files changed, 2 insertions(+), 4 deletions
```

```
commit 1dbd1951f39e13da579ffe879cce19586d0462de
Author: Li Zefan <lizf@cn.fujitsu.com>
Date:   Thu Dec 9 15:47:56 2010 +0800

    tracing: Fix preempt count leak

    While running my ftrace stress test, this showed up:

    BUG: sleeping function called from invalid context at mm/mmap.c:233
    ...
    note: cat[3293] exited with preempt_count 1

    The bug was introduced by commit 91e86e560d0b3ce4c5fc64fd2bbb99f856a
    ("tracing: Fix recursive user stack trace")

    Cc: <stable@kernel.org>
    Signed-off-by: Li Zefan <lizf@cn.fujitsu.com>
    LKML-Reference: <4D0089AC.1020802@cn.fujitsu.com>
    Signed-off-by: Steven Rostedt <rostedt@goodmis.org>
```

Stable tag的缺失（2）

- 有些子系统的maintainer似乎做的不大好…
- git log –no-merges v3.4..v3.10 kernel/sched/rt.c

ce0dbbb sched/rt: Add a tuning knob to allow changing SCHED_RR timeslice
60334ca sched/rt: Further simplify pick_rt_task()
fc79e24 sched/rt: Do not account zero delta_exec in update_curr_rt()
57d2aa0 sched/rt: Avoid updating RT entry timeout twice within one tick period
aa7f673 sched/rt: Use root_domain of rt_rq not current processor
1158ddb sched/rt: Add reschedule check to switched_from_rt()
f3e9478 sched: Remove __ARCH_WANT_INTERRUPTS_ON_CTXSW
a4c96ae sched: Unthrottle rt runqueues in __disable_runtime()
e221d02 sched,rt: fix isolated CPUs leaving root_task_group indefinitely throttled
7f1b439 sched/rt: Fix lockdep annotation within find_lock_lowest_rq()
454c799 sched/rt: Fix SCHED_RR across cgroups
29baa74 sched: Move nr_cpus_allowed out of 'struct sched_rt_entity'
8d3d5ad sched_rt: Avoid unnecessary dequeue and enqueue of pushable tasks in set_cpus_allow

- 12 commits, 4 fixes, 这4个fixes都可以合入到 3.4.x，但都没有stable标签

无法直接合入的patch（1）

- 如果一个bug fix被合入到 3.2.y，那几乎可以肯定这个fix也必须合入到3.4.y。
- 通过脚本分析，发现~450个upstream commits只在3.2.y里，不在3.4.y里.

3.4.80	3.2.54	missing in 3.4.x
3700 patches	4480 patches	~450 patches

无法直接合入的patch（2）

- 为什么3.4会少了几百个commits？
  - 3.4的maintainer是Greg，他使用自动化脚本将主线的commits合入到3.4及其他stable tree上。
  - 如果有一个patch无法直接合到到其中一些stable tree的话怎么办?
    - 如果这个patch只能合入到最新的stable tree…
    - 如果这个patch连最新的stable tree都无法合入…
  - 3.2的maintainer是Ben，他会人工分析每个commit，而不是简单的丢弃合入失败的commit。

无法直接合入的patch（3）

3.4.80	3.2.54	missing in 3.4.x
3700 patches	4480 patches	~450 patches

Total	Needed backporting	Bug fixes	Prerequisites	New device id/quirk	Already backported	Should be dropped
~450	419	325	20	74	14	21

Total	Can’t apply cleanly	No stable tag	Tagged with higher versions	Can apply cleanly
325	238	55	5	27

总结

- 社区通过LTS内核提供适合商用的内核版本
- Bug是不可能完全消灭的
- Stable kernel的维护很成功，但也不是完美的

Q & A