The Kernel Report

(Plumbers 2010 edition)

Jonathan Corbet
LWN.net
corbet@lwn.net

“Yeah, yeah, maybe you're waiting for flower power and free sex. Good for you. But if you are, don't ask the Linux kernel to wait with you. Ok?”

-- Linus Torvalds
June 28, 2010

Photo: Chris Schlaeger

Last year's kernels

Release	Date	Days	Changes	Developers
2.6.32	Dec 2	84	10,998	1,229
2.6.33	Feb 24	84	10,871	1,152
2.6.34	May 15	80	9,443	1,110
2.6.35	Aug 1	79	9,801	1,145
2.6.36	Oct 20	80	9,501	1,176
Totals:		407	50,614	2,811

Last year's kernels

Release	Date	Days	Changes	Developers
2.6.32	Dec 2	84	10,998	1,229
2.6.33	Feb 24	84	10,871	1,152
2.6.34	May 15	80	9,443	1,110
2.6.35	Aug 1	79	9,801	1,145
2.6.36	Oct 20	80	9,501	1,176
Totals:		407	50,614	2,811

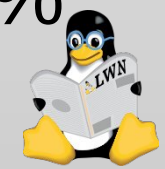
124 changes per day

3560 lines of code added per day

...every day!

Who supports this work

volunteers	18.7%	Fujitsu	1.7%
Red Hat	11.3%	academics	1.6%
Intel	7.6%	Renesas Tech.	1.5%
unknown	5.5%	Atheros	1.4%
Novell	4.8%	Pengutronix	1.2%
IBM	4.3%	Analog Devices	1.1%
Nokia	2.5%	HP	1.0%
consultants	2.3%	Samsung	1.0%
Texas Instruments	1.8%	NTT	0.9%
Oracle	1.8%	New Dream Net	1.0%
AMD	1.7%	Broadcom	0.9%

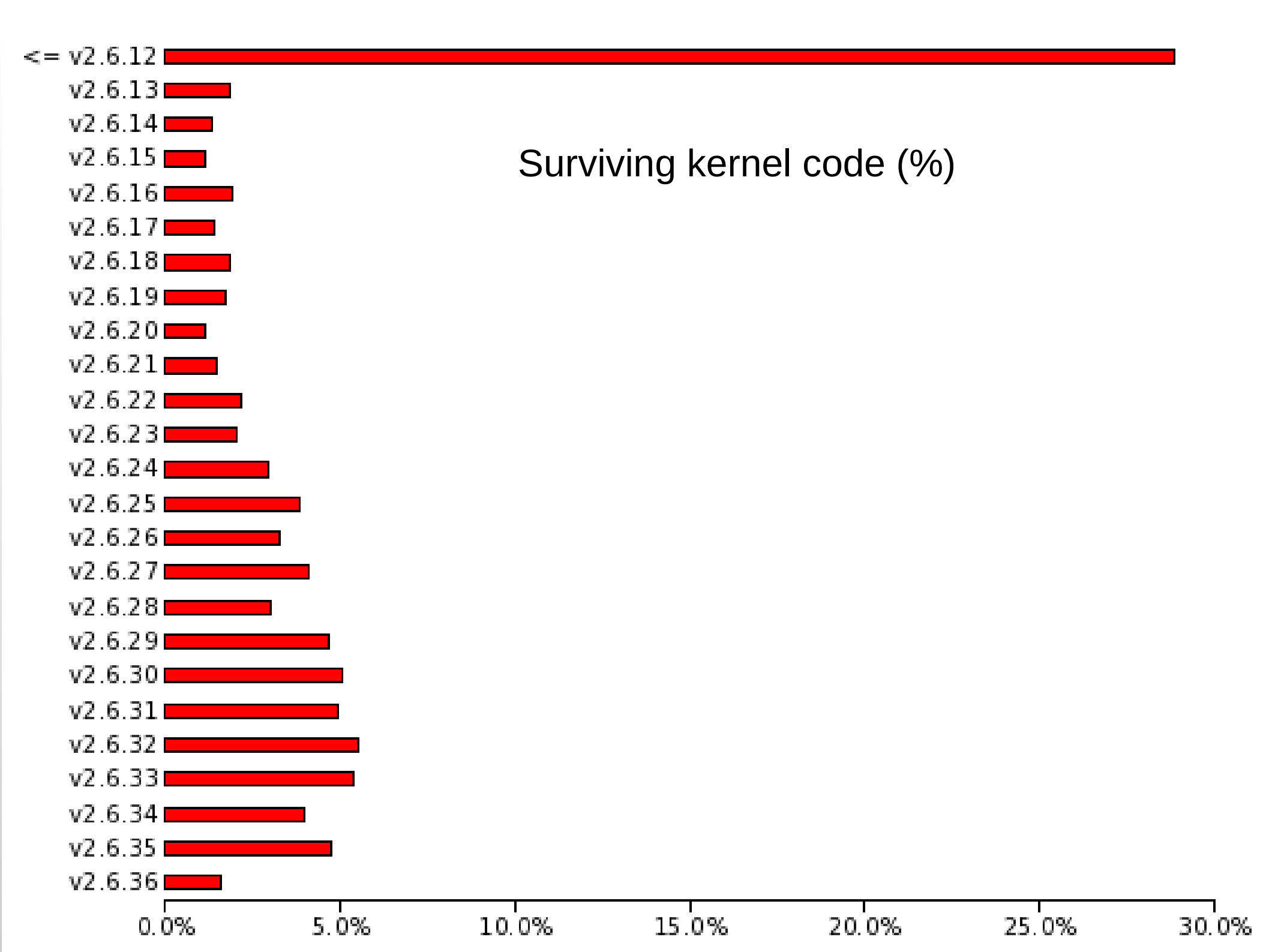


Last year's kernels

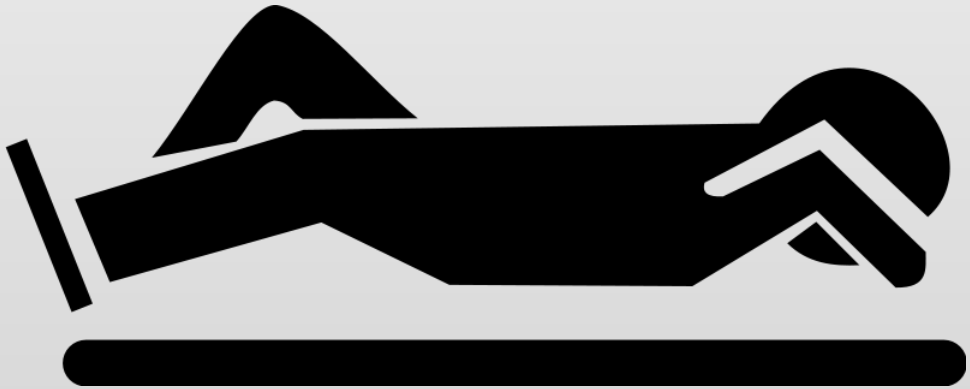
Release	Date	Days	Changes	Developers
2.6.32	Dec 2	84	10,998	1,229
2.6.33	Feb 24	84	10,871	1,152
2.6.34	May 15	80	9,443	1,110
2.6.35	Aug 1	79	9,801	1,145
2.6.36	Oct 20	80	9,501	1,176
Totals:		407	50,614	2,811

Looking back

Last year (2.6.32-36) was one of consolidation and completion.


<= v2.6.12
v2.6.13
v2.6.14
v2.6.15
v2.6.16
v2.6.17
v2.6.18
v2.6.19
v2.6.20
v2.6.21
v2.6.22
v2.6.23
v2.6.24
v2.6.25
v2.6.26
v2.6.27
v2.6.28
v2.6.29
v2.6.30
v2.6.31
v2.6.32
v2.6.33
v2.6.34
v2.6.35
v2.6.36
Surviving kernel code (%)
0.0%
5.0%
10.0%
15.0%
20.0%
25.0%
30.0%

Has kernel development gotten boring?



Next year

...could be a little more exciting...

2.6.37

Due in January

What's coming:

- BKL removal complete! (almost)
- VFS and block scalability work
- Block I/O bandwidth throttler
- Wakeup events, opportunistic suspend
- (Some of) Xen Dom0
- Broadcom wireless driver

Filesystems

2.6.31 Last incompatible Btrfs on-disk change
2.6.33 Reiserfs BKL removal
2.6.34 LogFS merged
Ceph distributed filesystem merged
2.6.35 Direct I/O support in Btrfs
2.6.37 Ext4 scalability work

Filesystem status

Ext4 is ready for production use

Btrfs is getting closer

Btrfs is the default MeeGo filesystem

What's coming: filesystems

Btrfs: completion
- Full RAID support
- Data migration
- Use of btrfs features

Ext4: increased adoption
- *Maybe* snapshots

VFS scalability work

Storage

2.6.31 Storage topology infrastructure
2.6.32 Block scalability improvements
2.6.33 I/O bandwidth controller
Device mapper snapshot merging
2.6.37 Hard barriers removed

What's coming: storage

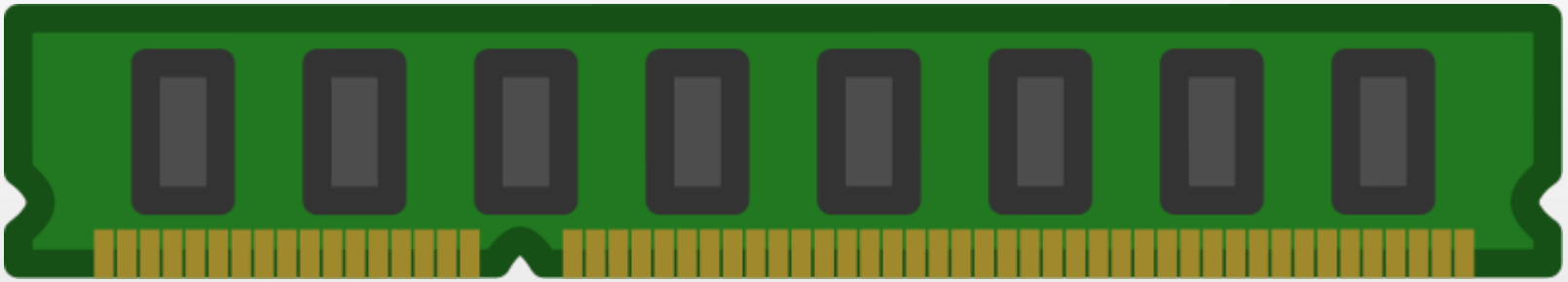
Hardware challenges
- Thin provisioning
- Solid-state devices

RAID unification

Hierarchical storage

I/O controller completion
- 2.6.37

Memory management



2.6.31 Kmemleak
2.6.32 HWPOISON
2.6.32 Kernel same-page merging (KSM)
2.6.35 Memory compaction
2.6.37 Some writeback fixes

What's coming: memory management

Fixing the writeback problem

In-memory swapping/deduplication

- Zcache
- Clearcache/Frontswap/Transcendent memory
- ...

Transparent huge pages

Realtime

2.6.31 perf events
2.6.33 “Raw” spinlock renaming
2.6.37 BKL removal nearing completion

Meanwhile: realtime preemption patch shipped in numerous embedded and enterprise distributions.

What's coming: realtime

Memory management preemptability

Sleeping spinlock merge

- *Maybe* next year...

Numerous scalability issues

- RT tends to find them first

Open problems

- Slab allocator latencies
- Per-CPU data

What's coming: realtime

Deadline scheduling
- What all the cool RT researchers are using

No more priorities
- Instead: WCET, deadline, period

A SCHED_DEADLINE patch exists
- Numerous details to work out yet
- SMP scalability is problematic

Drivers

2.6.31 Radeon memory mgmt and KMS
2.6.33 Nouveau driver merged
2.6.34 VGA switcheroo
2.6.37 brcm80211 driver merged

The graphics problem is almost solved!

Drivers

2.6.31 Radeon memory mgmt and KMS
2.6.33 Nouveau driver merged
2.6.34 VGA switcheroo

The graphics problem is almost solved!

...well...sort of...

How many?

Configuration options in drivers/...

2.6.31	3,890
2.6.32	4,057
2.6.33	4,201
2.6.34	4,300
2.6.35	4,574
2.6.36	4,671

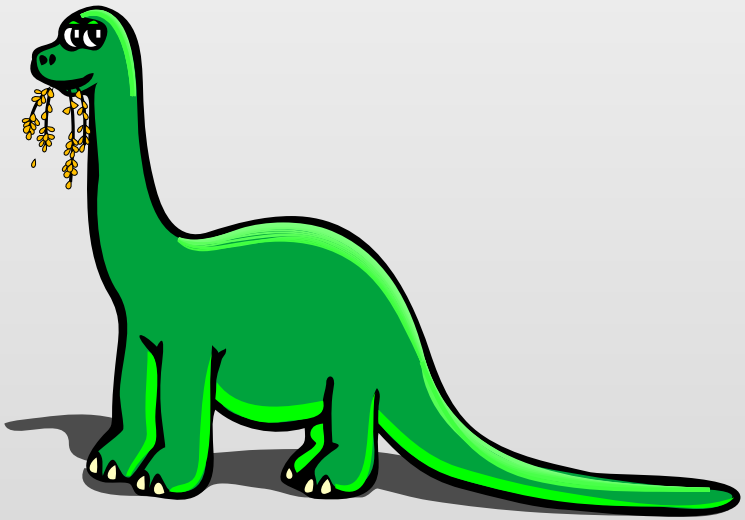
We're adding 156 options each release

What's the problem?

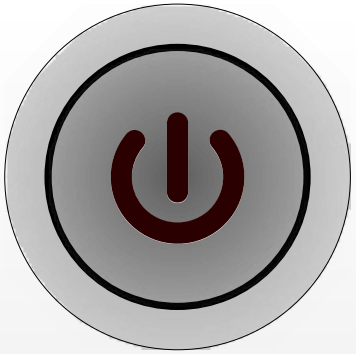
Uncooperative vendors

What's the problem?

Uncooperative vendors



Power management



2.6.32 Framebuffer compression
Framebuffer dynamic clocking
2.6.34 Asynchronous suspend/resume
2.6.35 CPU idle pattern detection
Timer slack
2.6.36 Race-free suspend

How to deal with power-hungry applications?

Approach #1

Fix individual apps
Provide tools
Social pressure
Use QOS control

Approach #2

Suspend system
Stop running apps
Use suspend blockers

The “embedded problem”

Pressures on embedded developers

- Ridiculous deadlines
- Short product cycles
- Lots of secrecy

Results:

- Shipping out-of-tree code
- No community input
- No time to fix things up
- → Code doesn't go upstream

What's coming: power management

Lots of individual hardware fixes

Opportunistic suspend, wakeup sources
    2.6.37

Idle cycle injection

Tracing/measurement

2.6.31 perf events
2.6.32 Scheduler tracepoints
2.6.33 Dynamic ftrace
2.6.35 perf kvm

Perf events

Capture and analysis of hardware events

- Instruction cycles
- Cache misses
- ...

Also software events

- Function calls
- Tracepoints (including dynamic)

What perf can do

Application profiling

- Like prof/gprof - but with kernel profiling too

Who is causing system events?

- Which function causes page allocations?

Statistical analysis

- How variable is a given operation?

See perf.wiki.kernel.org

Ftrace

The native kernel tracing facility

- Function calls
- Latency tracing
- System power states
- MMIO operations
- Stack usage
- Tracepoint hits
- ...

Tracepoints

The key to user-friendly tracing

Linux has them for:

- Interrupts
- Timer events
- Filesystem operations
- Memory management decisions
- Lock operations
- ... >200 in all

Are tracepoints part of the Linux ABI?

SystemTap

Still under development

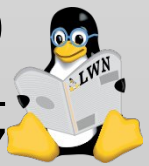
- There are happy users

May never make it into the mainline

- ...the action seems to be elsewhere

Security

CVE-2010-0003	CVE-2010-0006	CVE-2010-0007	CVE-2010-0008
CVE-2010-0148	CVE-2010-0291	CVE-2010-0297	CVE-2010-0298
CVE-2010-0306	CVE-2010-0307	CVE-2010-0309	CVE-2010-0410
CVE-2010-0415	CVE-2010-0437	CVE-2010-0622	CVE-2010-0623
CVE-2010-0727	CVE-2010-0729	CVE-2010-0730	CVE-2010-0741
CVE-2010-1083	CVE-2010-1084	CVE-2010-1085	CVE-2010-1086
CVE-2010-1087	CVE-2010-1088	CVE-2010-1146	CVE-2010-1148
CVE-2010-1162	CVE-2010-1173	CVE-2010-1187	CVE-2010-1188
CVE-2010-1436	CVE-2010-1437	CVE-2010-1446	CVE-2010-1451
CVE-2010-1488	CVE-2010-1636	CVE-2010-1641	CVE-2010-1643
CVE-2010-2066	CVE-2010-2070	CVE-2010-2071	CVE-2010-2226
CVE-2010-2240	CVE-2010-2248	CVE-2010-2478	CVE-2010-2492
CVE-2010-2495	CVE-2010-2521	CVE-2010-2524	CVE-2010-2537
CVE-2010-2538	CVE-2010-2653	CVE-2010-2798	CVE-2010-2803
CVE-2010-2938	CVE-2010-2942	CVE-2010-2943	CVE-2010-2946
CVE-2010-2954	CVE-2010-2955	CVE-2010-2959	CVE-2010-2960
CVE-2010-3015	CVE-2010-3067	CVE-2010-3078	CVE-2010-3079
CVE-2010-3080	CVE-2010-3081	CVE-2010-3084	CVE-2010-3110
CVE-2010-3296	CVE-2010-3297	CVE-2010-3298	CVE-2010-3301
CVE-2010-3310	CVE-2010-3437	CVE-2010-3442	CVE-2010-3477



Questions?