



# State of the kernel

Linux Symposium 2010

Jon Masters <jcm@redhat.com>

# About Jon Masters

- Playing with Linux since 1995
- One of the first commercial Linux-on-FPGA projects
- Author of Professional Linux Programming, lead on Building Embedded Linux Systems 2<sup>nd</sup> edition, currently writing “Porting Linux” for Pearson.
- Red Hat Enterprise stuff (kABI, Real Time, etc.)
- module-init-tools, <http://www.kernelpodcast.org/>



# Overview

- Year in review
- Current status
- Future predictions
- Suggestions
- Podcasting
- Questions



# Year in Review



# July 2009 – 2.6.31-rc2 to 2.6.31-rc5 | 2.4.37.34

“It's getting cold in here” -- Greg K-H on Hyper-V

- Ceph filesystem appears in staging
- KVM/AlacrityVM “fork” begins
- KSM (Kernel Samepage Memory) proposed
- Microsoft Hyper-V drivers (54-part patch set)
- VFAT patches from Andrew Trigell
- Kmemleak (Catalin Marinas)



## July 2009 – 2.6.31-rc2 to 2.6.31-rc5 | 2.4.37.34

- IO bandwidth controller (dm-ioband, blkio-cgroup)
- Checkpoint and restart (Oren Laadan)  
clone\_with\_pids(), <http://ckpt.wiki.kernel.org/>
- KVM AMD64 SMP support (Michale Tokarev)
- New MMC maintainer needed (new mailing list)
- Ftrace ringbuffer discussions start
- Git users survey
- Fanotify proposed for merging



# August 2009 – 2.6.31-rc6 to 2.6.31-rc8 | 2.4.37.5

"The recently posted Moorsetown patches finally indicate the arrival of the embedded nightmare to arch/x86" -- Thomas Gleixner

- ARM mailing lists are moved to infradead.org
- TuxOnIce - Nigel Cunningham makes suggestions on the way forward. He and Rafael to work together.
- Thomas Gleixner posts support for Moorsetown
- Kernel-level virtIO server removes up to 4 syscalls and argument continues over virtio vs. vbus
- `socket_operations sock_sendpage` doesn't validate pointer, systems without `vm.mmap_min_addr` hit



## August 2009 – 2.6.31-rc6 to 2.6.31-rc8 | 2.4.37.5

- Runtime power management/selective bus shutdown proposed by Matthew Garrett and Rafael J. Wysocki
- Asynchronous suspend and resume (20% speedup)
- Compcache allows compressed RAM swap device. Requires a swap slot freeing callback be added.
- TRIM support found lacking (Mark Lord)
- localmodconfig and localyesconfig (Steven Rostedt)
- Lazy workqueues (Jens Axboe) to reduce 531 threads
- Linux Foundation hardware distribution suggestions



# September 2009 – 2.6.31 | 2.6.32-rc1

- “BFS is slower than mainline in virtually every measurement” -- Ingo Molnar on Con Kolivas' BFS
- BFS scheduler proposed by Con Kolivas. Frans Pop finds it falls down in tests but seems to offer a smooth and interactive desktop experience nonetheless.
- Threaded interrupt discussion on benefits of tasklets.
- RCU scalability improvements from Paul McKenney
- Timechart (Arjan van de Ven). Later adds suspend and resume time charting for individual devices.
- Mudflap for free-after-use tracking (Yanboe Ye)
- UNREACHABLE macro (Roland McGrath)



## October 2009 – 2.6.32-rc3 (no rc2) to 2.6.32-rc5

- Module symbol resolution speedups (Alan Jenkins) with sorted tables during build. Carmelo Amoroso proposes the same thing during his CELF talk.
- Discussion on eventual removal of md. Suggestion of unified md/dm RAID code in 3-5 years perhaps.
- Robert P. J. Day wiki of unused CONFIG variables
- IO bandwidth throttling discussed at the LPC
- Val Aurora posts the latest version of union mounts
- Mathieu Desnoyers (LTTng) posts uru version 0.2



## November 2009 – 2.6.32-rc6 to 2.6.32-rc8

- Ftrace ported to MIPS and to Microblaze
- Perf gets “perf bench” utility (Mitake Hitoshi)
- Thread sibling renaming support (John Stultz)
- MadWiFi is deemed officially dead (Luis R. Rodriguez)
- Sparse 0.4.2 released, Coverity scans missing
- Dynamic cpumask allows > 4096 CPUs (Rusty)
- Asynchronous page fault support (Gleb Natapov)
- LIRC debates – what to do about going in-tree?



## November 2009 – 2.6.32-rc6 to 2.6.32-rc8

- How to report bugs on “open” firmware? (ar9170)
- PATA gets lots of updates with the intention to kill off the IDE layer (Bartłomiej Zolnierkiewicz)
- Intel folks realize that optimizing for speed can now be faster than optimizing for size (8% thanks to prefetch)
- “Simon and Garfunkel” fallout from VFAT work



## December 2009 – 2.6.32 | 2.6.33-rc1 to 2.6.33-rc2

- “The two-week merge window is not supposed to be 'one day merge window after thirteen days of silence'” -- Linus threatens to shorten the merge window in future cycles.
- ELF support for “Extended Numbering” to handle very large core file dumps (Hatayama Daisuke)
- Google power capping work with idle cycle injection to temporarily reduce power use (Salman Qazi)
- 80 character limits considered harmful. VT100 is dead.



## January 2010 – 2.6.33-rc3 to 2.6.33-rc6

- RT kernel 2.6.31.12-rt20 is released
- Suggestion to enable devtmpfs by default is rejected
- Suggestion to use GFP\_NOIO on suspend rejected
- Linux Power Management Mini-Summit August 9<sup>th</sup>
- Chinang Ma posts analysis showing performance regression from 2.6.18 to 2.6.33-rc4, contrasting with Imbench runs from 2.6.18 to 2.6.27 by Ajay Patel.
- Mathieu Desnoyers PhD is published on his various tracing work and he releases LTTng and urcu updates. He then proposes sys\_membarrier for urcu support.



## February 2010 – 2.6.33-rc7 to 2.6.33-rc8 | 2.6.33

- OOM killer discussion restarted, VmSize vs. RSS. David Rientjes posts completely new OOM killer.
- Bootmem discussions based on Ingo Molnar pointing out that there are 5 different allocators on x86 systems.
- TSC variant/invariant discussions, virt TSC/kvm-clock
- OF Device Tree patches from Grant Likely
- Another update to 2.4 is posted for an e1000 fix
- Kernel.org discusses migration to xz compression
- DecNET orphaned due to “lack of time, space, motivation, hardware and probably experience”



## March 2010 – 2.6.34-rc1 to 2.6.34-rc3

- “It may say 'staging', but that doesn't change the fact that it's in production use by huge distributions. Flag days aren't acceptable” -- Linus Torvalds on nouveau.
- “drm request 3” (Dave Airlie) starts a rant from Linus
- Linus rejects SCSI patches that miss merge window
- Split function and data sections patches updates
- Tejun Heo laments state of 4 KiB sector disk support
- Alex Chiang posts patch implementing PCI slot to device directory information in sysfs



## March 2010 – 2.6.34-rc1 to 2.6.34-rc3

- LMB v. e820 issues raised by e820 update. Logical Memory Blocks (LMB) seen as the future now.
- Multitouch driver discussions surface again
- VMWare post a virtio extension for their balloon driver Avi Kivity favors a separate VMWare balloon.
- Tejun Heo posts support for cpuhogs (stop\_machine)
- kernel.org gets various SSL certificate services thanks to John Hawley and a donation from Thawte.



## April 2010 – 2.6.34-rc4 to 2.6.34-rc6

- Arnd Bergmann continues his work on BKL removal with various TTY patches having previously done fs
- recvmmsg() waits for all packets but not just one
- CONFIG\_PROVE\_RCU incompatible with non-GPL kernel modules on the system
- Darren Hart discusses adaptive spinning for futexes
- Subtle bug found in anon\_vma\_chains (avc) code
- Version 10 of “use lmb with x86” patches
- Steven Rostedt announces version 1.0 of trace-cmd



## May 2010 – 2.6.34-rc7 | 2.6.34 | 2.6.35-rc1

- Discussion is revived on the future of an error reporting subsystem after a mini-summit during Collab Summit
- TSC reliability is discussed (again), especially whether to expose to userspace at all and NTP need for detail on accuracy of the oscillator that it is using
- Steven Rostedt discusses the trials and tribulations of unifying the various (3) ringbuffer implementations
- Transcendent Memory is now called “Frontswap”
- vger.kernel.org has a major power outage



## June 2010 – 2.6.35-rc2 to 2.6.35-rc3

- Lengthy discussion of Google's suspend blockers and the way forward. Alan Stern proposes a compromise.
- The floppy driver (lacking much love) has oopsing issues and Linus offers to pay \$7.99 to “anyone twisted enough to really want to fix that code”.
- Suggestion made to remove support for module removal from modprobe
- Robert P. J. Day announces his “online beginners kernel programming course”, <http://crashcourse.ca/>



## July 2010 - 2.6.35-rc4

- Lengthy discussion of the future of defconfigs
- Btrfs review from Edward Shiskin criticizes variable record size meta-data implementation, utilization.
- Mobile graphics drivers debate heats up.



# Kernel Releases



## 2.6.31 – September 2009

- CUSE (Character devices in USErspace)
- Performance Counters (perf)
- Kmemleak
- USB3 – xHCI (but no hardware)



## 2.6.32 – December 2009

- KSM (Kernel Samepage Merging)
- HWPOISON
- Intel Moorsetown Support
- Per-backing-devie writeback (replaces pdflush)
- Runtime power management
- S+Core architecture (uses Bergmann's asm-generic)
- Several Enterprise kernels



## 2.6.33 – February 2010

- Compcache (compressed RAM-based swap)
- DRBD (Distributed Replicated Block Device)
- Nouveau
- Performance “Events” (prof, probe, etc.)
- Android drivers removed from staging



## 2.6.34 – May 2010

- Asynchronous suspend and resume
- Ceph (distributed filesystem)
- LogFS (transactioned flash filesystem)
- Vhost net (improved KVM networking)
- Merge window intentionally closed early
- LWN reports 9,100 patches and 1,110 developers



## 2.6.35 – August 2010

- `cpu_stop` (`cpuhog`)
- Memory Compaction
- Performance Events for KVM
- Unified Lockup Detector (NMI perf events)



# Current Status



# General

- Overall looking very strong
- 2.6 still going, jokes about 3.0 but unlikely
- Linus increasingly clamping down on merge window
- Bugs, kerneloops, and regressions



# Embedded

- Android suspend blockers
- Compcache
- Ftrace for Microblaze, MIPS
- Intel Moorsetown
- Real Time
- Trace-cmd
- USB3



# Desktop

- Asynchronous suspend and resume
- Module linking speedups
- Runtime power management
- Timechart



# Enterprise/Server

- Asynchronous page fault, Frontswap, etc.
- Ceph, DRBD
- Dynamic cpumask (4096+ CPUs)
- KVM, KSM, venet
- Scalability of RCU, etc.



# Future Predictions

- 2.6.35 next month, development slows slightly
- Fanotify, transparent huge pages will go in
- QoS suspend will go in eventually (suspend blockers)
- RT won't be merged (but getting smaller)
- Hyper-V stays in staging or is removed
- More IO bandwidth controller work
- An Error Reporting subsystem
- Embedded fragmentation



# Recommendations

- Send status emails (like Microblaze and XFS)
- Respond to more questions (many unanswered)
- More civility on the LKML
- Documentation (wiki, etc.)



# Podcast

- <http://www.kernelpodcast.org/>
- Started in May 2009. Over 164,000 downloads and 10,000 unique visitors to the website.
- More Kernel Traffic inspired than Linux Weekly News
- Initially Daily, then weekly...trying very hard(!)
- Lots of work – volunteers welcome
- Pronunciation



# Podcast

- Lots of things happen on an average week
- Pick the most serious and a few fun ones
- Read the entire thread, research the topic
- Write a summary for the week
- Record using audacity/lame
- Distribution and website



# Questions

- The views and opinions expressed here are my own.

