What’s new in the Linux kernel

and what’s missing in Debian

Ben Hutchings · MiniDebConf Toulouse 2024

Ben Hutchings

- Working on Linux kernel and related code for Debian and in paid jobs for about 15 years
- Debian kernel and LTS team member, doing various kernel packaging and backporting work

mainline:	6.12-rc7	2024-11-10
stable:	6.11.7	2024-11-08
longterm:	6.6.60	2024-11-08
longterm:	6.1.116	2024-11-08
longterm:	5.15.171	2024-11-08
longterm:	5.10.229	2024-11-08
longterm:	5.4.285	2024-11-08
longterm:	4.19.323	2024-11-08
linux-next:	next-20241112	2024-11-12

Linux keeps changing

- Linus makes a release with new features every 9-10 weeks
- Larger features may take multiple releases to become useful

- Some features need changes elsewhere to enable them:
  - New user-space management tool
  - New version of existing user-space tool
  - Applications and libraries using new API
  - Packaging or infrastructure changes
- I'll talk about new features in Linux 6.6 to 6.12 inclusive

Recap of previous years' features

- **RISC-V:** Now supports performance counters, suspend-to-RAM, membarrier(), memory hotplug, kASLR, CFI (Clang only), drivers in Rust
- **io_uring:** New operations including `waitid`, `{get,set}sockopt`, `bind`, `listen`, `ftruncate`
- **ID-mapped mounts:** Now supported on FUSE, hugetlbfs, zonefs
- **ublk:** Enabled, but need a **ublksrv** package (ITP: #1051678)
- **HID_BPF:** Still not enabled; some packaging work needed
- **Rust:** Support for Ethernet PHY drivers and block drivers in Rust, but progress is slow
- **kTLS:** Now supported for NVMe-TCP

Filesystems (1) — bcachefs

- Modern filesystem competing with btrfs and ZFS:
  - File data is shared copy-on-write, allowing fast snapshots and file copies
  - Full checksumming
  - Optional data compression
  - Optional full disk encryption (AEAD) — but not fscrypt
  - Can use a pool of multiple devices
- Partly based on the existing **bcache** disk caching subsystem
  - Can make effective use of a mix of faster and slower devices
- Still marked *experimental*, but is enabled in Debian kernel
- Some ongoing interpersonal issues with upstream maintenance

Filesystems (2)

- New system calls `listmount()`, `statmount()`:
  - Provide structured information about mounted filesystems
  - Can replace text-based `/proc/self/mount{s,info}`
- New generic ioctls:
  - `FS_IOC_GETUUID`: Get volume UUID
  - `FS_IOC_GETFSSYFSPATH`: Get path to volume metadata in `/sys`
- **tmpfs** supports user extended attributes and quotas
- **XFS** supports online repair of some filesystem errors
- VFS and **XFS** support for filesystem block size > page size

Filesystems (3)

- **ext2** driver is deprecated and won't be updated for Y2038
  - Debian uses the ext4 driver for ext2 volumes, which is not affected
  - Old ext{2,3,4} volumes with 128-byte inodes will need to be replaced by 2038
- **ntfs** driver removed in favour of **ntfs3**
  - Neither is enabled in Debian, though ntfs3 might be ready now

PREEMPT_RT (1)

- **Real-time** means that the *latency* for response to high-priority input can be bounded, with tasks strictly prioritised or guaranteed CPU time
  - Essential for many safety-critical systems: vehicles, industrial control
  - Good for live audio and video production
- Does *not* mean fast; throughput and average latency tend to be worse
- Out-of-tree patch set added config option `CONFIG_PREEMPT_RT`:
  - Changes interrupt and softirq handlers into schedulable tasks
  - Changes most kernel locking to keep preemption enabled
  - Remaining non-preemptible sections are short and should have bounded latency
  - Supports both prioritised and deadline scheduling

PREEMPT_RT (2)

- Many of the changes for real-time were generally useful and are now used in all configurations
- Config option added upstream in Linux 5.3, but dependent on architecture support which was still out-of-tree
- In Linux 6.12:
  - All necessary changes are now upstream for arm64 and x86
  - Out-of-tree patch set still needed for arm and powerpc, and to fix i915 driver
- Debian provides alternate kernel packages with `PREEMPT_RT` for several architectures, starting in version 7 “wheezy”
  - Reminder: Debian comes with no warranty; don’t use these in safety-critical systems!

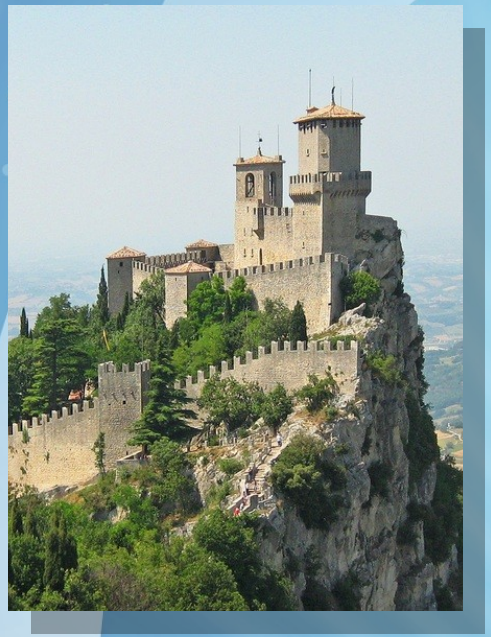
Linux Security Modules (1) — Landlock

- Allows user processes to restrict (sandbox) themselves
- Complementary to system policy applied by AppArmor or SELinux
- Complementary to system call filtering with `seccomp()`
- Now able to:
  - [6.7] Restrict use of TCP ports as client or server
  - [6.10] Disable most ioctls on char and block devices
  - [6.12] Restrict access to “abstract” (non-filesystem) Unix domain sockets

Linux Security Modules (2)

- **AppArmor** can restrict use of io_uring and user namespaces, to reduce kernel attack surface
- New system calls `lsm_list_modules()`, `lsm_get_self_attr()`, `lsm_set_self_attr()`:
  - Generic API for configuration of LSMs
  - Can expose attributes from multiple LSMs, unlike `/proc/self/attr`
  - Preparation for “stacked” major LSMs — e.g. Debian with AppArmor policy in a container on top of Fedora with SELinux policy

Security hardening



- Two mitigations of type confusion attacks against heap use-after-free vulnerabilities:
  - Randomised `kmalloc()` caches: global probabilistic mitigation (compile-time option)
  - Dedicated slab bucket allocator: full mitigation, but new kernel API
- Compile/boot-time option to disable writes to read-only user memory through `/proc/`*`pid`*`/mem`
  - Not done by default due to compatibility concerns
- `mseal()` system call allows user-space to prevent future changes to a memory mapping
- [x86] Intel shadow stack support protects against ROP attacks in user-space

Packaging changes (1)

- Architecture and flavour updates:
  - In: arm64-16k, loong64, mips64r6el, ppc64(el) 4k/64k
  - Out: ~~armel/marvell~~, ~~i386~~, ~~ia64~~
- New binary packages:
  - **linux-bpf-dev** contains header file for BPF CO-RE builds
  - [x86] **intel-sdsi** contains tool for Intel On Demand license activation
- **linux-libc-dev** became Architecture: all (for now)
- Debian-specific ABI numbers replaced by upstream version
- All arch-dependent packages cross-buildable

Packaging changes (2)

Configuration changes to add or improve support for:

[amd64] **Intel** "Meteor Lake" SoCs, IPU3 and IPU6; **System76** systems

[amd64,arm64] Chromebooks and ChromeOS tablet devices from various vendors

[armhf] **NXP** i.MX7 SoCs; **Terasic** DE10-nano

[arm64] **Banana Pi** BPI-R3; **Lenovo** Miix 630, Thinkpad X13s, Yoga C630; **MediaTek** MT8173, MT8183, and other SoCs; **NXP** i.MX8 SoCs; **Pine64** PineTab 2; **Qualcomm** SDA845, RB-series, X Elite, and other SoCs; **Renesas** RZ-series SoCs; **Rockchip** RK3328, RK3399, RK356x, and RK3588 SoCs; **SolidRun** Honeycomb Workstation and HummingBoard-T; **TI** K3-AM642 SoC

[ppc64el] **Raptor** Talos II

[riscv64] **Microchip** Polarfire SoC; **Sophgo** SoCs; **StarFive** JH7110 SoC and VisionFive 2; **T-Head** SoCs

Packaging changes (3)

- Signed modules and optional Lockdown on all architectures
- Mostly reproducible again, but module signing currently conflicts with this
- Modules installed compressed

- CI includes quick build for arm64
- Many changes to internal configuration and build system

Questions?

Credits & License

- Content by Ben Hutchings
  www.decadent.org.uk/ben/talks/
  License: GPL-2+
- Original OpenOffice.org template by Raphaël Hertzog
  raphaelhertzog.com/go/ooo-template
  License: GPL-2+
- Background based on “Serenity” theme by Edward Padilla
  wiki.debian.org/DebianArt/Themes/serenity
  License: GPL-2