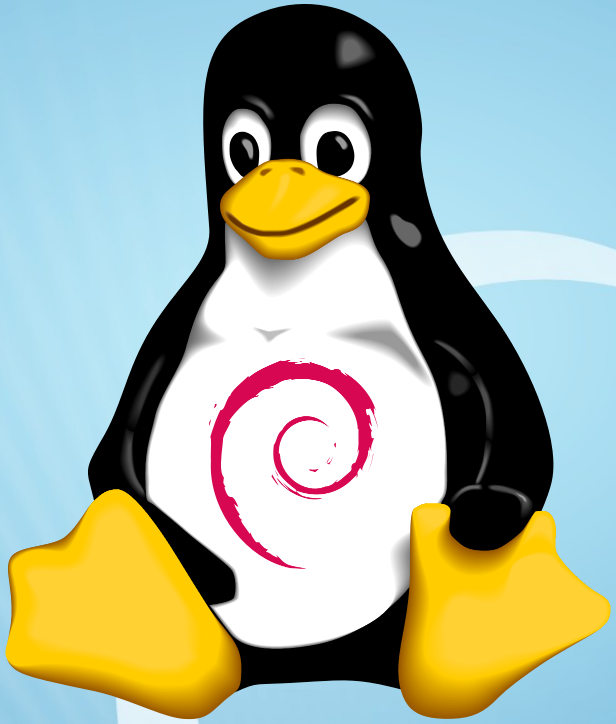
What's new in the Linux kernel

and what's missing in Debian



Ben Hutchings · DebConf 20

Ben Hutchings

- Working on Linux kernel and related code for Debian and in paid jobs for over 10 years
- Debian kernel and LTS team member, doing a lot of the kernel packaging and backporting work
- Formerly maintained Linux long-term stable branches needed by Debian

Linux releases early and often

- Linux has feature releases about 5 times a year, plus stable updates every week or two
- Some features aren't really ready in the first kernel release
- Some will need changes elsewhere to be useful:
  - New user-space tool to configure it
  - 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 5.3 to 5.8 inclusive

Recap of previous years' features

- schedutil: default cpufreq governor on arm64/armhf; usable on x86 after CPU load tracking changes in 5.7
- Zoned recording: dm-zoned-tools *still* needs a sponsor (#882640)
- RISC-V: gained support for EFI boot, huge pages, seccomp, CPU hotplug; still limited SoC and board support
- Y2038: kernel side is done; Debian still needs to decide whether and how to migrate 32-bit arches to new ABI

Better AIO [ongoing] (1)

Linux AIO (asynchronous input/output) introduced in 2.5:

- Restricted to *direct* (uncached) file I/O, so mostly used by database managers that manage their own cache
- Not very efficient—requires *more* system calls than synchronous
- Small set of file operations supported
- Not fully asynchronous: `io_submit` sometimes blocks

POSIX AIO implemented in glibc:

- Uses thread pool calling synchronous system calls
- Not at all efficient

Better AIO [ongoing] (2)

`io_uring` introduced in 5.1 and extended repeatedly since then:

- Supports direct and buffered file I/O, socket I/O, etc.
- Uses submission and completion rings in shared memory
  - Same pattern as used for high-performance I/O devices
  - Allows queueing and de-queueing multiple operations with few system calls
- Application can register a “buffer group” to be used as needed, so e.g. it may `recvmsg()` on many sockets without dedicating a buffer to each
- Application can submit a chain of dependent operations, e.g. `write()`→`fsync()`→`close()` that will stop if any operation fails
- Many file and socket operations supported

User-space library for it (`liburing`) is packaged, but so far only used by QEMU...

BPF everywhere [ongoing] (1)

- Kernel allows user-space to install programs that run in kernel context on certain events
- Started as “Berkeley Packet Filter” in BSD; used to filter packets before copying them to user-space
- Programs run in 32-bit VM with few regs, read access to context (e.g. packet header), no other memory, and no looping
  - Originally interpreted, but can be JIT-compiled to native code
- Linux “Extended BPF” added 64-bit regs and ops, stack, bounded pointer arithmetic, data structures shared with user-space
  - Allows compiling from C to BPF (`clang -target bpf`, `bpf-gcc`)
  - Requires complex verifier in kernel to detect unsafe operations
  - Many features only available to privileged users

BPF everywhere [ongoing] (2)

- Bounded loops allowed [5.3]
- Per-cgroup hooks for filtering `{get,set}sockopt()` [5.3]
- Per-socket hook to monitor TCP round-trip time [5.4]
- BTF (BPF Type Format): compact debug info allows user-space to “relocate” structure access for running kernel [5.4]
- “Dynamic program extensions” like shared libraries [5.6]
- BPF program as TCP congestion control module [5.6]
- BPF became compatible with `PREEMPT_RT` [5.7]
- Linux Security Module using BPF for all hooks [5.7]

Lots of tracing programs available in `bpfcc-tools` package, but not much else is packaged yet

Security hardening [ongoing] (1)

- Heap clear-on-alloc/clear-on-free options [5.3]
- ROP mitigation for x86: functions used to update CR0 and CR4 will never clear security-critical bits [5.3]
- `refcount_t` has full over/underflow checking on all architectures [5.5]
- `openat2()` allows user-space to restrict path lookup [5.6]:
  - Flags to disable symlink or only “magic” link resolution, disable crossing mountpoints, disable looking above starting dir, or treat starting dir as root
  - Useful for programs that inspect untrusted filesystems, e.g. container managers, user-space file servers
  - Allows replacing complex and often buggy “safe” path lookup in user-space

Security hardening [ongoing] (2)

- ROP mitigations for arm64 [5.7, 5.8]:
  - Shadow call stacks (Clang only)
  - Pointer authentication for user-space and kernel return addresses (Arm v8.3)
  - Branch Target Identification for user-space and kernel (Arm v8.5)
- Module loader rejects modules with W+X sections [5.8]

Speculation leak mitigation [ongoing]

- TAA: Some Intel CPUs support “restricted transactional memory”, which reduces need for locking when updating shared data. Conflicts cause transaction to be aborted, but don't stop speculative execution. Similar possibilities for leaks as with earlier “MDS” vulnerabilities.
  - For MDS vulnerable CPUs, existing mitigation covered this
  - For newer CPUs, mitigated by disabling TSX, which required microcode update
- SRBDS: Intel CPUs use “special registers” as buffer between shared RNG/crypto block and CPU cores. Speculative execution could use stale data from a special register, leaking random numbers—often used for cryptographic purposes—between different security contexts.
  - Mitigated by microcode update, which makes RDRAND/RDSEED slower
  - Kernel makes less use of RDRAND, and added option to *disable* mitigation

Packaging changes

- Lockdown is mostly upstream, though we still carry a few patches
  - Dropped support for disabling it through SysRq, because it's possible to synthesise that
  - We now set it to "integrity" level when Secure Boot is enabled, so tracing etc. are allowed
- Dropped support for some NAS devices using Marvell SoCs, as the kernel grew too big for the partition they load it from
- Merged linux-latest source package into linux, so meta-packages stay in sync
- Added shared library and dev packages for libtraceevent
- Added "cloud-arm64" flavour, allowing Arm VMs to use less disk space
- Moved libbpf packages out to their own source package
- Changed debhelper compatibility level from 9 to 12
- Added bpftool package

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